Research on 3D Sampling and Monitoring of Power Supplies Based on Augmented Reality (AR) Technology

Chao Wang, Xi Chen, Ying Wang
State Grid Info & Telecom Group, Beiing Guo Dian Tong NetworkTechnology CO.,LTD, Beijing, China

Abstract

In this paper, three-dimensional object detection and tracking based on neural network is studied. The application of Augmented Reality (AR) in power material sampling monitoring service in China Power Grid is presented, relying on the AR technology of 5G Internet of things to realize real-time capture of the information from power network infrastructure project to goods sampling inspection and monitoring, and to solve the difficult business management problem that many parties need to participate on the spot but can not guarantee the timeliness, the AR technology is used to mark, early-warning and trace the source of the problem materials, so as to ensure the authenticity and traceability of the whole process.

Keywords: *Augmented reality, 3D object detection and tracking, convolutional neural network, real time location and mapping, grid sampling*

I. Introduction

In the real world, physical 3D detection and tracking is one of the core technologies of augmented reality (AR) technology, which mainly applies computer vision model to locate and track specified 3D objects in a continuous image sequence[1]. However, in practical application, when there is more interference in the situation, such as blocking, lighting will affect the detection and tracking of three-dimensional objects, how to adopt a fast, accurate and stable three-dimensional object detection and tracking technology, AR technology in the field of power material sampling and monitoring research is the key issue.

II. 3D MODEL SELECTION

The detection of three-dimensional objects can be divided into two steps: detection and identification, the first step is to obtain the area of the possible three-dimensional objects in the image, and confirm the category of objects; The second step is to accurately obtain the position of three-dimensional objects in space based on categories and regions.

At present, the mainstream methods of 3D object detection are feature-based methods, template-based methods, end-to-end neural network-based methods and Hof forest-based methods. The feature-based method is suitable for texture-rich objects and application, which is essentially the same as camera calibration. The template-based method is generally used to deal with textureless type objects that lack features, the Hof forest-based method is to establish a mapping of three-dimensional object image blocks in place, and the neural network-based method has different principles and usage scenarios according to different network designs.

In recent years, thanks to the powerful expression and classification ability of convolutional neural network, many end-to-end 3D object detection methods based on neural network have emerged. The representative are POSE-CNN, SSD-6D, BB8, DeeP-6d-Pose and other neural network-based methods. Compared to standard sliding window detectors, CNN-based reel networks can quickly locate from a global image to the local area where an object may be located and detect it in that local area.Taking SSD-6D as an e*ample, it is based on SSD object detection algorithm, first of all, in the global image, using the characteristics of SSD network multi-stage scale simultaneous prediction,

ISSN: 0010-8189

quickly identify the two-dimensional bounding box where there may be objects, and then use non-maximum suppression to determine the candidate box; Finally, the 3D bounding box of the object is estimated by the size of the two-dimensional bounding box[2].

III. 3D OBJECT TRACKING AND POSITIONING

Three-dimensional object tracking is to obtain the label of an object and the information of space-time continuity after the initial position through the three-dimensional object detection algorithm, and continuously and steadily obtain the position information of the three-dimensional object in three-dimensional space. The current mainstream approach to 3D tracking can be divided into four categories: model-based approach, feature-based approach, neural network-based approach, Simultaneous Localization and Mapping SLAM method.

Traditional 3D tracking methods perform differently in different scenes. Feature-based methods are more sensitive to illumination changes; model-based methods are more sensitive to occlusion and dynamic blur; neural network-based methods are difficult to maintain stable tracking in large scenes, and SLAM is competent for all kinds of environments. When we register three-dimensional objects in the coordinate system of SLAM system, we can track them immediately and dynamically by simply keeping them updated at a certain frequency[3].

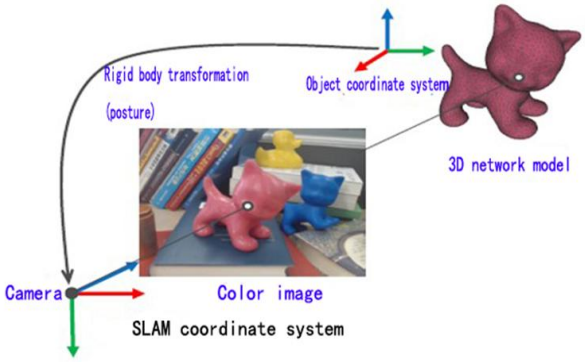

Fig. 1 three-dimensional object tracking and positioning method

IV. 3D Object Detection Based on Neural Networks

The focus of this study is on how to obtain a robust and efficient model that adapts to a variety of environments. Therefore, we compressed and accelerated the detection of neural network-based 3D objects on the new MOBILE NET network architecture, which greatly improved the network processing speed. Compared to traditional methods, our model architecture has three advantages:

1) The ability to detect multiple duplicate 3D objects in the image;
2) The increase in the number of three-dimensional objects does not enhance the time complexity of detection;
3) The complexity of three-dimensional objects does not affect the efficiency of the algorithm[4].

4.1 Model architecture

Our network architecture is based on the modification of YOLOV3 (you only look once). YOLOV3 is based on CNN's 2D detection network model. YOLOV3 uses the residual model as its feature e*tractor, which consists of 53 convolution layers and is named Darknet-53. and predicts the category and location of objects once for each cell. In order to detect objects of different scales, YOLOV 3 uses three scale feature maps for three times detection, YOLOV3 divides the image into s*s cells, which enhances the precision of prediction. The three scales are (13*13),

ISSN: 0010-8189

(26*26), and (52* 52). YOLOV3 outputs feature diagrams at different levels of Darknet-53 and divides them at different scales. Each cell predicts a 2D bounding box for an object, and each cell predicts 3 bounding boxes. The bounding bo* is e*pressed as (tx, ty, tw, th), (tx, ty) is the corner point in the upper left corner of the bounding box, and (tw th) is the length and width of the bounding box. Cells also predict the score of each object's category, and the confidence level (conf) of the box itself. There are a total of (52 * 52) plus (26 * 26) plus (13 * 13) cells, each cell corresponds to 3 outputs.

4.2 Network compression and acceleration

Mobile net is an efficient model proposed by Google for mobile and embedded devices. Based on streamlined architecture, Mobilenet uses depth wise separate convolutions to construct lightweight deep neural network.The main method is to decompose the standard convolution into depth wise convolution and point wise convolution. The advantage of this method is that the amount of parameters and calculation can be greatly reduced.

When the input feature map f is (DF, DF, m), the standard convolution K is (DK, DK, m, n), and the output feature map size is (DG, DG, n), the calculation formula of standard convolution is as follows:

$$G_{k,l,n} = \sum_{i,j}^{n} K_{i,j,m} F_{k+i-1,l+j-1,m} \quad (1)$$

The standard calculation quantity = DK * DK * m * n * DF * DF, Correspondingly, Mobilenet V1 first convolutes the input feature map through DK * DK * 1 * m depth convolution feature map, and the output feature map size is (DG, DG, m). Then, it passes through the point by point convolution layer with the size of (1,1, m, n). The final output is (DG, DG, n). Among them, depth convolution is responsible for filtering, and point by point convolution is responsible for conversion channel. In fact, the Mobilenet model decomposes the input feature graph into m DK * DK * 1, and then performs ordinary convolution calculation on M DK * DK * 1 feature graphs.

The convolution formula of standard convolution is as follows:

$$\hat{G}_{k,l,n} = \sum_{i,j}^{n} \hat{K}_{i,j,m} F_{k+i-1,l+j-1,m} \quad (2)$$

Among them, $\hat{K}$ is the depth of the reel, the co product core is (Dk, Dk, 1, M), where the mth cosmumulation is applied to the mth channel in F, resulting in the mthth channel output on G, the new calculation is: (Dk*Dk*M*DF) plus (M*N*DF*DF).

The ratio of standard and decomposed depth to point-by-point co product calculations is:

$$\frac{D_K \square D_K \square M \square D_F \square D_F + M \square N \square D_F \square D_F}{D_K \square D_K \square M \square N \square D_F \square D_F} = \frac{1}{N^2} + \frac{1}{D_K^2} \quad (3)$$

4.3 Network training

When we want to detect three-dimensional objects, we first select the appropriate network architecture, then input the image according to RGB format, and then divide the image into s*s cells to obtain the control point graph of three-dimensional objects, and generate S*S*D data output. In order to obtain the parameters of the rotation matrix in three-dimensional space, the related operations of Lie groups and Lie algebras are introduced.

All 3-by-3 rotational matrices make up the Special Orthogonal Group (so(3)), so(3) is a three-dimensional smooth flow with an expression of:

$$so(3) = \left\{ R \in \mathbf{R}^3 : RR^T = I, \det(R) = 1 \right\} \quad (4)$$

The two-dimensional coordinates, depth values, so(3) of the object, class C probability confidence score and 1 control point confidence score are obtained from the final layer output of 9 control points. The so(3) and control point depth values obtained here are mainly used for calculating loss functions and are generally not exported directly as output, because the accuracy of the three-dimensional information obtained by the prediction, such as

so(3) and the depth value of the control point, is lower than that of the two-dimensional information such as the two-dimensional coordinates of the control point.

$$\begin{aligned} g_x &= f(x) + c_x \\ g_y &= f(x) + c_y \end{aligned} \quad f(x) = \begin{cases} \dfrac{1}{1+e^{-x}} & (Centroid) \\ x & (angular\ point) \end{cases} \tag{5}$$

In the above formula, the predicted 9 control points are 8 corner points and 1 centrity of a three-dimensional object. In the actual design, the neural network does not directly predict the global coordinates (x,y) of the control point, but rather predicts the bias of the control point relative to the upper-left point (cx, cy) of the cell, and then calculates its coordinates in the global. Among the nine control points, (gx, gy) is the global coordinates of the control point, (cx, cYx) is the global coordinates of the corner of the cell where the control point is located, and (x, y) is the result predicted by the neural network. When the control point is 8 corners of the 3D bounding box, f (x) is a unit linear equation. When the control point is the center of mass, f (x) is a 1D sigmoid function. This is to compress the relative displacement of the center of mass relative to the cell to the interval (0-1), and effectively ensure that the center is in the grid cell to perform the prediction. By constraining the range of centroid points, the model can enhance the learning ability of the centroid and get more stability. It helps the model get priority of the cell of the centroid of the object. On this basis, the coordinates of the remaining eight control points are refined[5].

V. Augmented Reality Based on 3D Object Detection and Tracking Technology

For a more robust tracking effect, we registered detected 3D objects in 3D into the SLAM system, tracked them, and obtained an augmented reality application combined with SLAM. We also used VINS-SLAM to track and register 3D objects into SLAM using two-frame initialization, using the control point detected by the two frames and the camera position of the two-frame SLAM. VINS-SLAM is a VIU based on key frame optimization, which can obtain scale information after initialization, which can help us know the scale of the actual three-dimensional object, and track the three-dimensional object through SLAM, which can obtain strong anti occlusion and anti dynamic blur ability[6].

5.1 Positioning and navigation initialization

VINS-SLAM is a SLAM system implemented through VIO, and we get the result of the rotation and position of the IMU, rather than directly obtaining the position of the camera. So we need to align the IMU coordinate system to the camera coordinate system by initializing it. This step can also help us get the actual scale of three-dimensional objects.

VINS-SLAM uses a loosely coupled sensor fusion method to obtain the initial value. At initialization, the results of the IMU pre-integration and the visual SfM in the sliding window are obtained, and by aligning the IMU pre-integration with the visual SfM results, the scale, gravity, velocity and even deviation can be roughly restored[7].

5.2 Registration of 3D objects

When we get the corresponding control point coordinates X1 and X2 of the three-dimensional object in the two images through the three-dimensional object detection, let $\bar{x}1$ = (U, V, l), then we get the unit coordinates of the control point in the camera coordinate system:

$$p_1 = k^{-1}\bar{x}_1 = K^{-1}\begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix}, p_2 = k^{-1}\bar{x}_2 = K^{-1}\begin{bmatrix} u_2 \\ v_2 \\ 1 \end{bmatrix} \tag{6}$$

Let Z1 and Z2 be the depth values of the control point, T1 and T2 represent the pose of the two frames provided by SLAM system, and P represent the coordinates of the control point in the world coordinate system. The results are as follows:

$$z_1 \square p_1 = T_1 \square P \ , \quad z_2 \square p_2 = T_2 \square P \ (7)$$

The two sides of the equation are multiplied by the outer product (P1)x, (P2)x, of P1,P2, that is:

$$\mathbf{O} = (p_1)_x \square T_1 \square P$$
$$\mathbf{O} = (p_2)_x \square T_2 \square P$$

$$\text{Let } A = \begin{bmatrix} (p_1)_x \square T_1 \\ (p_2)_x \square T_2 \end{bmatrix}, \text{then: } AP = \begin{bmatrix} (p_1)_x \square T_1 \\ (p_2)_x \square T_2 \end{bmatrix}.\text{p=0 (8)}$$

The above is a system of linear equations solved by the ordinary least squares. The SVD decomposition of matrix A is calculated, and the singular vector with the smallest singular value is taken as the solution of the three-dimensional coordinate P. the point P is the coordinate of the control point of the three-dimensional object in the world coordinate system of SLAM..

5.3 Effect display of augmented reality

In addition to the advantages of multi view, fast moving and stable tracking in the case of light occlusion, the proposed method can also interact with other virtual objects in SLAM world coordinate system. When the coordinates of 3D objects and virtual objects in the world coordinate system are obtained, they are projected into the coordinate system of AR eyeglass camera. The system can effectively deal with the occlusion between virtual objects and 3D objects, and generate a more realistic augmented reality experience[8]. The figure below shows 3D objects and other virtual objects in slam system.

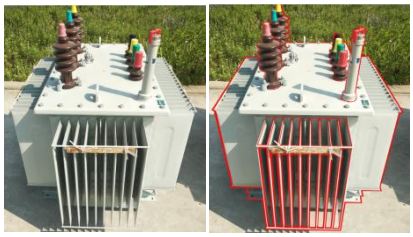

Fig. 2 three-dimensional tracking location of power transformer based on AR

VI. Future Application Prospects and Technological Improvements

In the real detection stage of 3D objects in power grid equipment, in view of the changes in the warehouse environment and the diversity of models and other factors, we put forward a new neural network-based 3D object detection method in combination with YOLOV3 and MobileNet model, realized efficient and robust 3D object detection, combined with real-world rendering technology[9], enhanced the generalization ability of the algorithm, and further ensured the detection results through edge-based optimization technology. Finally, compared with other methods on the LineMod dataset, the feasibility of the proposed algorithm is verified.

Due to the complex environment of power grid material warehouse and the sampling of power equipment in a changing environment, there are still some problems to be solved in the future, such as:

a) Since SLAM system is responsible for the tracking of inspected materials, when the tracking of SLAM is unstable, the tracking effect of inspected equipment will be affected.

b) At present, the sampling material can be tracked under the camera movement and environmental changes, but in the case of continuous movement, it is necessary to repeat the sampling material to the SLAM system registration,

which affects the imaging speed.

c) So far, the tracking of the materials being tested has not used historical position information. It is a continuous optimization process to update the attitude of objects with the historical position information of the inspected materials.

The breakthrough of the above-mentioned technology can realize the more accurate and realistic sampling and monitoring of the inspection materials of the power grid, and improve the quality and efficiency of the materials and equipment of the power grid.

References

[1] HAYASHI M, BACHELDER S. NAKAJIMA M, et al, A new virtual museum equipment with automatic video content generator[C]// International Conference on Cyberworlds(CW). Santander, Spain: IEEE, 2014[377-383]

[2] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, Dieter Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes[C]. In Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018.

[3] Shichao Yang, Sebastian Scherer. CubeSLAM: Monocular 3D Object Detection and SLAM without Prior Models [J]. CoRR, 2018, abs/1806.00557.

[4] Alberto Crivellaro, Mahdi Rad, Yannick Verdie, Kwang Moo Yi, Pascal Ira, Vincent Lepetit. Robust 3D Object Tracking from Monocular Images Using Stable Parts [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2018, 40(6):1465-1479

[5] LarsEgevad,PeterStröm,Kimmo Kartasalo,Henrik Olsson,Hemamali Samaratunga,Brett Delahunt,Martin Eklund. The utility of artificial intelligence in the assessment of prostate pathology[J]. Histopathology,2020,76(6).

[6] Rudy van Belkom. The Impact of Artificial Intelligence on the Activities of a Futurist[J]. World Futures Review,2020,12(2).

[7] Reza Hafezi. How Artificial Intelligence Can Improve Understanding in Challenging Chaotic Environments[J]. World Futures Review,2020,12(2).

[8] Alejandro Díaz-Domínguez. How Futures Studies and Foresight Could Address Ethical Dilemmas of Machine Learning and Artificial Intelligence[J]. World Futures Review,2020,12(2).

[9] Russell T. Warne,Jared Z. Burton. Beliefs About Human Intelligence in a Sample of Teachers and Nonteachers[J]. Journal for the Education of the Gifted,2020,43(2).