Object Pose Estimation Based on Template-matching Using Attention Module and Residual Block

Gaeun Noh[1][0000-0002-6125-8289] and Jong-Il Park[1][0000-0003-1000-4067]

[1] Department of Computer Science, Hanyang University, Seoul, Republic of Korea
{shqmffl486, jipark}@hanyang.ac.kr

Abstract. This paper proposes a method to create synthetic datasets using RGB-D camera and to train a template matching-based object pose estimation network to improve the accuracy of object pose estimation for unseen objects. Previous studies were limited by a lack of adequate training data and a relatively simple network model, which resulted in overfitting. In order to enhance performance, the proposed method incorporates synthetic data of objects with symmetrical shapes or limited textures into the existing datasets. By applying Convolutional Block Attention Module to a ResNet50 based model, the intermediate features of objects were more effectively emphasized and suppressed to improve performance. Through comparative experiments with existing methods, it was confirmed that the proposed method provides higher accuracy for unseen objects compared to the existing methods.

Keywords: Object Pose Estimation · Template-Matching · Deep Learning

1 Introduction

Recently, with the development of deep learning technology, the accuracy of image recognition has increased. In the field of Augmented Reality (AR), the fundamental objective is to utilize technology to recognize objects in images and facilitate interaction between virtual and real environments. Object pose estimation is a crucial technology in Augmented Reality that determines the 3D position and orientation of an object captured in an image by analyzing its shape. Among various techniques, object pose estimation is a key technology in the field of Augmented Reality. There are difficulties in performing object pose estimation, such as occlusion, scale variations, changes in lighting, and separating objects from the background, and various methods have been proposed to overcome these difficulties. [1] proposes a 6D object pose estimation model that estimates the 3D transformation and rotation of an object from an image by determining the center of the object and predicting the distance from the camera, with the orientation being regressed as a quaternion representation. However, it is difficult to accurately estimate the pose when occlusion occurs in the camera image. On the other hand, [2] uses deep networks to extract landmarks of the object in the image and estimate the 6DOF pose through 2D-3D correspondences. This is achieved through the precise multi-precision supervision architecture proposed by predicting landmarks, and it robustly estimates the pose even in the presence of occlusion. It showed higher overall performance than [1], but the accuracy was lower for unseen objects. To address these limitations, a template matching-based object pose estimation research [3] was proposed. This object pose estimation approach that matches the template of an object rendered from multiple views in a CAD model with the highest similarity to the real input image (query image), showing generalization and robustness to occlusion for unseen objects. However, the existing datasets lacks objects with symmetrical shapes or limited textures, making it difficult to estimate poses for complex objects.

In this paper, we suggest to achieve high accuracy in estimating the pose of unseen objects, including objects with symmetrical shapes and with limited textures, in template matching-based object pose estimation. The proposed research uses a mask to remove the background of the captured image before computing the global shape, which can significantly slow down the matching speed. Therefore, to estimate a faster and more accurate pose, the paper proposes to learn a local feature that can be used to match the captured image and synthesized template. In this paper, we create a 3D object synthetic datasets using RGB-D camera to achieve accurate pose estimation even in the case of symmetrical objects and objects with complex shapes and limited textures. Furthermore, we propose a robust model based on ResNet50 with the Convolutional Block Attention Module (CBAM) [6] applied to overcome the limitations of the previous deep learning network models and achieve even more robust object pose estimation.

2 Creating synthetic datasets using RGB-D camera

The process of creating the synthetic dataset for template-matching based object pose estimation is depicted in Fig. 1. The existing datasets used in previous studies on object pose estimation [1, 2, 3, 7] such as LINEMOD, LINEMOD-Occluded, TLESS and YCB-Video, have the limitation of object diversity. To address this, we add the data for objects with symmetrical or complex shape and limited textures. This was done by acquiring the mesh of real objects using an RGB-D camera and incorporating it into the existing CAD model to create the synthetic datasets. The objects were placed on a board printed with markers and captured at various angles using the RGB-D camera to create the synthetic datasets. Each data is composed of RGB and depth images and the extrinsic of the camera calculated through the markers are stored for each image. The point cloud is generated from the RGB and depth images, then converted into a mesh. The converted mesh is used to create a mask image and the 3D bounding box of the object.

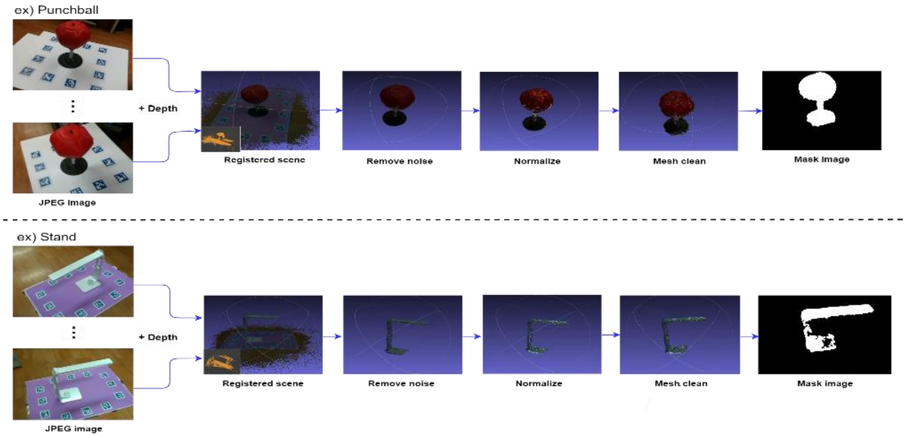

Fig. 1. Creating and processing point cloud scenes using RGB-D camera.
[3D object Synthetic datasets, ex) Punchball(symmetrical) / Stand(complex shape)]

3 Template-matching method

3.1 Framework

The method proposed in this paper is a deep learning-based learning method for estimating an object pose by matching an object template and an real image, based on the template-matching [3] framework. It is configured as shown in Fig. 2. Template matching-based object pose estimation methods generate templates rendered from multiple views and mask images inside the object, not only for seen objects but also for unseen objects and estimate poses by matching them with the most similar templates. Template generation uses BlenderProc [5] to sample synthetic templates for realistically rendered images according to the protocol in [4], and generates 3,084 templates for each object.

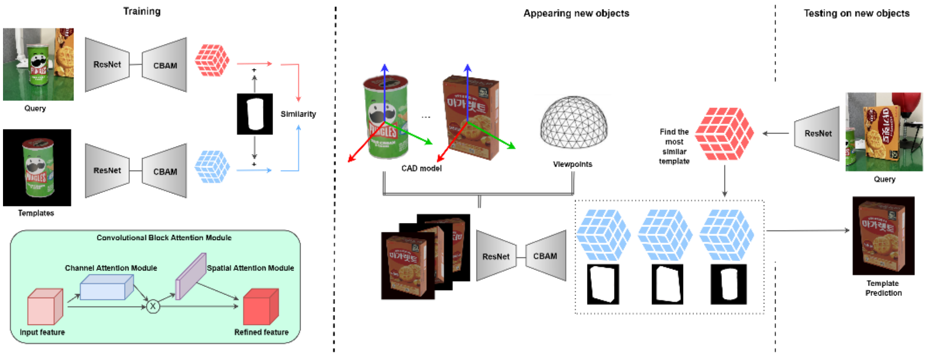

Fig. 2. Framework for Object Pose Estimation Based on Template-Matching.

Considering the template and mask, it can be robust to occlusions arising from real images because only some regions can be compared by deleting the background of real images. And it was confirmed through experiments that the unseen object is more generalized. Therefore, we compare the similarity between the template generated based on the local feature of the image and the real input image based on the local feature similarities and estimate the pose by finding and matching the template with the highest similarity in the template set. It can be partially obscured, and the pose can be stably estimated even though the background of the object is not uniform.

3.2 Network structure

Previous studies have used a CNN algorithm consisting of two convolutional layers and two fully connected layers with subsequential 2x2 max pooling layers as the "base" backbone. However, there was a problem of increasing computation due to large datasets. To solve this problem, the paper applied skip connection by applying ResNet50 to solve the vanishing gradient problem and increased performance by learning a deeper network than the CNN algorithm. In addition, all pooling, FC layers were removed and replaced with two 1x1 convolution layers that output local features, allowing the number of channels to be adjusted and the model to be configured deeper to reduce the computation.

In addition, the CNN algorithm has a problem of learning even the noise contained in the train data due to large computations and overfitting. The suggested method was to select a target object and learn the channel attention to improve the performance of the unseen object. In addition, 7x7 convolutional layer is applied to the channel compressed in the spatial attention to create an attention map. The large receptive field encodes which part of the pixel to focus on while finding spatially important regions, confirming that the performance is improved compared to the previous network with a small amount of computation.

And to reduce the loss, two loss functions were used to experiment. When InfoNCE loss was applied, the result confirmed a lower loss value than the Triplet Loss. It was judged that InfoNCE loss showed higher accuracy because it measures similarity based on mutual information when it was a large dataset (Ablation study).

4 Experiment and result

The experiment in this paper, we used RGB-D camera as well as existing datasets to generate and experiment with various synthetic datasets for additionally symmetric objects(punchball), complex shape objects(stand), and objects with limited textures (mirrors, etc.). The experimental setting was conducted at ubuntu 18.04, with epoch=100, learning rate=1e-4, batch size=8, and Adam as the optimizer. The experiment takes about 12 hours to learn from GeForce RTX 3080.

For the comparative experiment, the total dataset was classified into train:valid:test = 6:2:2 to generate a total of 88,668 data. The class of commonly set data sets was divided into seen object (0, 1, 2, ⋯⋅, 8) and unseen object (9, 10, 11, ⋅⋅, 17). The experiment is divided into seen object and unseen object as a comparison between the proposed network and the previous network. Also, we experiment on existing datasets and synthetic datasets, respectively. The experimental results are shown in Table 1 and Table 2. Table 1 compares the performance of previous and proposed networks for seen object and unseen object using only the existing dataset. As a result of the comparison, it was confirmed that the accuracy of the seen object and the unseen object improved the performance of the proposed network compared to the previous network. Table 2 experiments on seen objects and unseen objects using only synthetic datasets containing complex objects. As a result of the comparison, the accuracy of the proposed network increased by approximately 1.2% in seen object compared to the previous network. In the unseen object, the accuracy increased by approximately 1.4%. The results of the method proposed in this paper prove higher performance not only for existing datasets, but also for synthetic datasets with symmetrical, complex shapes, and limited textures. However, we find that there is a limit to estimate poses of objects that are not in the training datasets because the unseen object has lower accuracy than the seen object.

Table 1. Comparison results using existing datasets of size 224x224

	Our method	Previous method
Seen object	0.9779	0.8386
Unseen object	0.4536	0.4418

Table 2. Comparison results using synthetic datasets of size 224x224

	Our method	Previous method
Seen object	0.9788	0.8569
Unseen object	0.5457	0.4018

5 Ablation Study

This section uses synthetic dataset for the proposed network to reduce training loss, and experiments are conducted by applying InfoNCE loss and Triplet loss, respectively. The result of applying InfoNCE loss reduced the loss by approximately 0.26 compared to the Triplet loss. This was judged to have shown higher accuracy because InfoNCE loss measures similarity based on mutual information when learning large datasets as in experiments.

Table 3. Comparison results of loss function for seen object

	InfoNCE loss	Triplet loss
Our method	0.9743	0.7152

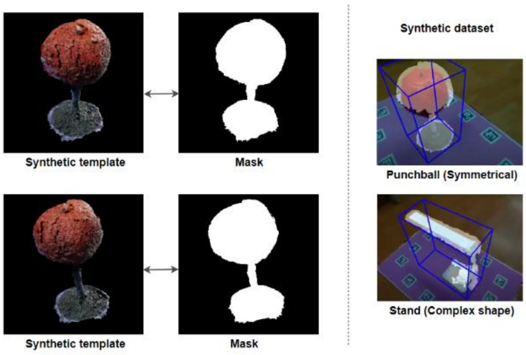

Fig. 3. Synthetic template, mask and special case (synthetic dataset).

6 Conclusion

This paper proposes a method to increase the accuracy of template matching-based object pose estimation not only for seen objects but also for unseen objects by creating a synthetic datasets of 3D objects and applying the Convolutional Block Attention Module to the proposed ResNet50 network. Specifically, the scene of the point cloud was generated and processed as a synthetic dataset and experimented with synthetic datasets of 3D objects and an existing LINEMOD datasets, respectively. As a result, the networks proposed in this paper achieve higher accuracy for complex objects for seen objects and unseen objects than previous networks. In addition, higher accuracy was confirmed when InfoNCE loss was applied than Triplet loss in ablation study. In future research, we will make it possible in real-time by focusing on higher performance object pose estimation for unseen objects.

References

1. Xiang, Yu, et al. "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes." *arXiv preprint arXiv:1711.00199* (2017).
2. Chen, Bo, Tat-Jun Chin, and Marius Klimavicius. "Occlusion-Robust Object Pose Estimation with Holistic Representation." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022.
3. Hu, Yinlin, et al. "Templates for 3D Object Pose Estimation Revisited: Generalization to New Objects and Robustness to Occlusions." *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022.
4. Paul Wohlhart and Vincent Lepetit. Learning Descriptors for Object Recognition and 3D Pose Estimation. In Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
5. Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. Blenderproc. arXiv preprint arXiv:1911.01911, 2019.
6. Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
7. Chen, Bo, Tat-Jun Chin, and Marius Klimavicius. "Occlusion-robust object pose estimation with holistic representation." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.