The International Journal of Advanced Manufacturing Technology (2019) 104:285–299
https://doi.org/10.1007/s00170-019-03901-0

ORIGINAL ARTICLE

A pose estimation system based on deep neural network and ICP registration for robotic spray painting application

Zhe Wang[1,2] · Junfeng Fan[1,2] · Fengshui Jing[1,2] · Zhaoyang Liu[1,2] · Min Tan[1,2]

Received: 3 January 2019 / Accepted: 20 May 2019 / Published online: 30 May 2019

Abstract

Nowadays, off-line robot trajectory generation methods based on pre-scanned target model are highly desirable for robotic spray painting application. For actual implementation of the generated trajectory, the relative pose between the actual target and the model needs to be calibrated in the first place. However, obtaining this relative pose remains a challenge, especially from a safe distance in industrial setting. In this paper, a pose estimation system that is able to meet the robotic spray painting requirements is proposed to estimate the pose accurately. The system captures the image of the target using RGB-D vision sensor. The image is then segmented using a modified U-SegNet segmentation network and the resulting segmentation is registered with the pre-scanned model candidates using iterative closest point (ICP) registration to obtain the estimated pose. To strengthen the robustness, a deep convolutional neural network is proposed to determine the rough orientation of the target and guide the selection of model candidates accordingly thus preventing misalignment during registration. The experimental results are compared with relevant researches and validate the accuracy and effectiveness of the proposed system.

Keywords Pose estimation · Spray painting · RGB-D sensor · Deep neural network · ICP registration

1 Introduction

Spray painting is an indispensable procedure in the manufacturing of a wide variety of products, such as furniture, aircrafts, automobiles, and steel structures. In recent years, in order to improve painting quality, promote the manufactural efficiency, and protect the health of workers, robotic spray painting becomes more and more prevalent in manufactural applications. At present, there are numerous researches in robotic spray painting concerning robot dynamic control [1], base position optimization [2], trajectory planning [3], etc.

In robotic spray painting applications, the painting quality is largely dependent on the accuracy and effectiveness of robot end effector (spray gun) path planning and the corresponding robot trajectory planning. Currently, one of the most common trajectory planning methods in this industry is traditional teaching method, which requires massive amount of time to manually configure the specific trajectory in advance, thus largely reducing the efficiency and universal applicability. Another commonly applied trajectory planning method is off-line programming method which utilizes the pre-scanned point cloud model or CAD model of the target as guidance in generating the trajectory [4–7]. This method is able to be applied on a larger variety of targets and requires less manual intervention. Nevertheless, for model-based trajectory generation, the virtual model and environment cannot always be precisely mapped onto the real world. Therefore robot coordinates computed off-line are often inaccurate in actual practice [8]. Hence, a calibration of the relative pose between the actual target

✉ Fengshui Jing
fengshui.jing@ia.ac.cn

Zhe Wang
wangzhe2016@ia.ac.cn

Junfeng Fan
fanjunfeng2014@ia.ac.cn

Zhaoyang Liu
liuzhaoyang2017@ia.ac.cn

Min Tan
tan@compsys.ia.ac.cn

1 The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

2 University of Chinese Academy of Sciences, Beijing 100049, China

and model-based trajectory is a prerequisite for the robot to precisely accomplish the spray painting task according to the generated trajectory. Traditionally, this calibration is accomplished manually by fixing the target to a specifically appointed pose, making the procedure time-consuming and not practical for the manufacture while unavoidably introducing some deviation. Thus, an automatic calibration system that accurately estimates the relative pose is necessary. For precise spray painting application, it is required that the deviation of estimated pose given by the calibration system is controlled within the spray radius of the spray gun.

In recent years, researchers carry out some researches in determining the relative pose between the target and model or model-based trajectory. In order to acquire 3D representation of the target, structured light sensor or RGB-D vision sensor is applied in these researches. Chen et al. [9] presented a visual servo approach based on fringe pattern projection for spray path planning. This method is able to adjust the initial model-based path according to the actual target to make the end effector perpendicular to the target surface. However, the large sized visual system is installed on the end effector of the robot and works at relatively a short distance from the target, which would inevitably hinder the painting operation of the spray gun and make it hard to be integrated with existing model-based trajectory generation method. Xu et al. [10] proposed a real-time position and posture measurement device which is small sized and lightweight, making it compact for the integration into spray painting robots. Nevertheless, this system can only detect one small patch of the target at a time and has limited measuring range; thus, it cannot provide consistent adjusting strategy for the whole trajectory. In addition, it calculates the pose of target surface using the projection of three spot lasers, which is inapplicable for complex structures. Lin et al. [11] carried out a spray painting system based on the point cloud model collected by Kinect RGB-D sensor. In this system, the sensor is separated from the spraying robot end effector. Yet it determines the relative pose between the model and the actual object merely from the type and orientation of the target conveying system, so it overlooks the difference every time the target is put on the conveyer thus introduces error. Lin et al. [12] proposed a pose estimation pipeline for auto part painting robots using a combination of iterative closest point (ICP) [13] and genetic algorithm. However, this pipeline only uses the depth information of the RGB-D sensor and adopts RANSAC planar segmentation to segment the target point cloud from the background. Thus, this pipeline is designed under the assumption that the background mainly consists of floor and wall, and its accuracy would decrease in the cluttered industrial environment. Therefore, in robotic spray painting application, it is required to develop an accurate automatic pose estimation system which is able to be easily integrated with the above mentioned model-based trajectory generation methods and can meet the application requirements.

Currently, there are a considerable number of object pose estimation methods proposed by researchers [14–17], although most of these methods have not yet been applied to robotic spray painting applications. Among them, the pose estimation methods based on model registration algorithms such as iterative closest point (ICP) registration [13] are widely accepted. The model registration method is well suitable for spray painting pose estimation system, since the pre-scanned target model initially prepared for the trajectory generation could be used directly as the referenced model in registration. Recently, the model registration is integrated with deep neural network and achieves meritorious performance in pose estimation. Zeng et al. [18] applied fully convolutional network (FCN) [19] for segmenting target from a scene and then they aligned the segmented point cloud with the CAD model using ICP registration to estimate the pose of the target. Similarly, Wong et al. [20] integrated SegNet network [21] and multi-hypothesis registration to improve the estimation accuracy and efficiency. Lin et al. [22] used semantic segmentation network to produce segmented target, and the target point cloud features were aligned with the model features using RANSAC and ICP registration for pose estimation. Yang et al. [23] applied integrated object detection network and ICP registration method on life support robot in actual daily scenarios and obtained increased pose estimation performance. Whereas, in order to obtain favorable segmentation performance, the deep neural networks of these methods generally require large amount of precisely labelled training data, which is inconvenient and inefficient to acquire in actual spray painting application. The quantity of the data is directly related to the preparation time of an applicable pose estimation system and thus determines the productivity. The data augmentation or self-annotation method based on background subtraction is not well suited under dim and unstable light conditions and in cluttered industrial environment. Moreover, for ICP registration in these methods, when two point clouds are relatively far apart and out of range points (outliers) are included, the algorithm might fall into a local minimum and the alignment might become incorrect. In addition, these methods are mainly designed for relatively short perception distance in the laboratory setting. As a result, a pose estimation system that requires relatively small amount of training data while maintaining high accuracy from a relatively large safe distance in the industrial setting is needed in this scenario.

In this paper, a pose estimation system based on deep neural network and ICP registration is presented. This

system possesses three unique properties that guarantee its effectiveness in spray painting applications. Firstly, the sensor in the proposed system is separated from the robot end effector and is placed at a safe distance from the target, thus preventing the interference during the spray painting operation and also protecting the sensor lens from contamination of the paint. Secondly, an orientation determination network based on VGG [25] structure is introduced as well. This network determines the rough orientation of the target to aid the selection of model candidates in ICP registration, thus reducing the misalignment caused by local minima. Furthermore, this selected orientation could assist the spray painting system to decide which generated trajectory to apply on the target in the current pose, thus further improving the compatibility of the proposed system in model-based trajectory generation painting systems. Thirdly, the semantic segmentation network is designed based on U-SegNet [24] structure, which requires a relatively small amount of training data and shows desirable segmenting performance in the poor lighting condition of an industrial setting. This structure is able to further lower the introduction of the out of range points (outliers) in the model registration procedure and improves the pose estimation accuracy.

In the remainder of this paper, the overall system description and configuration are presented in Section 2. The pose estimation framework and the detailed approaches are presented in Section 3. Then, Section 4 presents the dataset and experiment results. Finally, the paper is concluded in Section 5.

2 System description

In a spray painting application, the pose of the unpainted target, including its orientation and position, varies due to the uncertainty of placement on the operation platform. Hence, the proposed vision system captures the RGB-D data of the target and outputs its pose as a calibration for the robotic spray painting trajectory.

The hardware of the proposed system includes an RGB-D sensor and a computer. As illustrated in Fig. 1, the sensor is placed on a tripod by the side of the target at a certain safe distance (1.5 m) in order to provide suitable interspace for perception without interfering the spray painting end effector. The RGB-D sensor is PERCIPIO FM810 camera, which is an industrial-grade active stereo camera with two IR (Infrared Radiation) cameras, an IR laser projector and an RGB camera. The active stereo technology enables the camera to collect depth data accurately even under the poor lighting condition in an industrial setting. The depth range of the sensor is 0.5 to 6 m, enough to cover the required sensing distance.

The sensor is connected to the computer with a USB cable. The RGB-D data collected by the sensor are saved onto the computer as aligned RGB and depth images. The images are then fed to the proposed pose estimation program that runs on the computer and the estimated target pose is output accordingly.

This system mainly deals with the rotationally asymmetric target, upon which the robotic spray painting operation is conducted according to the target orientation. For rotationally symmetric target, this system could only estimate its actual position, because the definition of the orientation of a rotationally symmetric profile tends to be ambiguous.

For further integration with robotic spray painting system, the estimated pose could be transformed to the pose in the robot base frame using the coordinate transformation between camera and robot base frame obtained from eye-to-hand calibration. Thus, the deviation of the generated trajectory could be rectified according to the pose.

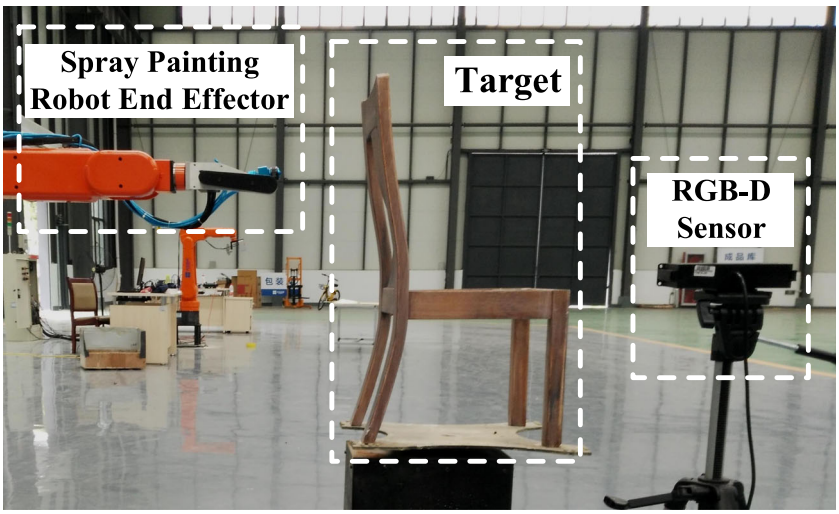

Fig. 1 The relative position of RGB-D sensor, robot end effector, and the target

3 Pose estimation framework and approaches

3.1 Overview of the proposed framework

The proposed pose estimation framework is organized as follows.

Firstly, a convolutional neural network is adopted to determine the rough orientation of the target using the RGB image as input. For the target which could horizontally rotate on the platform, four basic orientations of the target are defined as front, back, left, and right orientations. If the azimuth of the target is in between two basic orientations, the closest and second closest basic orientations are designated as its primary and secondary orientations. The primary and secondary orientations are determined from the largest and the second largest probabilistic outputs of the network, respectively.

Secondly, pixel-wise semantic segmentation based on deep neural network is conducted on the RGB image. The probabilistic mask of the target obtained from the segmentation is binarized and then crops the aligned depth image. Therefore, the segmented 3D point cloud of the target could be generated from the cropped depth image and the intrinsic matrix of the camera.

Finally, the results from the first and second steps are integrated for the pose estimation based on ICP registration algorithm. The pre-scanned point cloud model or CAD model of the target is cropped and retrieved according to the primary and secondary orientations derived from the first step. Then, ICP is used to register the segmented point cloud with the two cropped model candidates of primary and secondary orientations, and the ICP fitness scores are computed. The orientation with the optimal fitness score is adopted as the final orientation and the corresponding pose is output as the final estimated pose.

In the off-line preparation process, the abovementioned deep neural networks are trained in advance with the target dataset, which is discussed in detail in Section 4. The complete proposed pose estimation framework is illustrated in Fig. 2.

3.2 Determination of orientation

The accurate determination of rough orientation is fundamental in the proposed system, because the initially determined orientation is used to select the basic orientation of pre-scanned model used in model registration. For ICP registration, if the orientation of target point cloud is largely distinct from the model and large amount of out of range

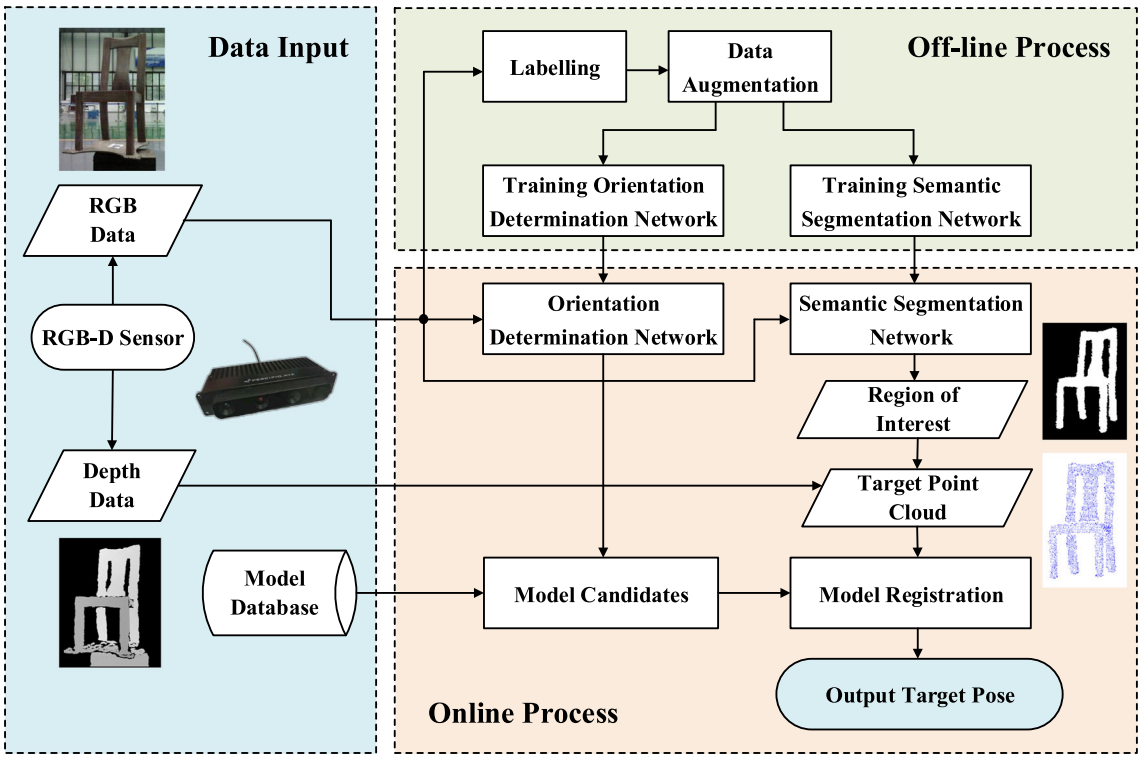

Fig. 2 The proposed pose estimation framework. The off-line process refers to the preparation procedures before the actual application. The online process refers to image processing and model registration procedures during application

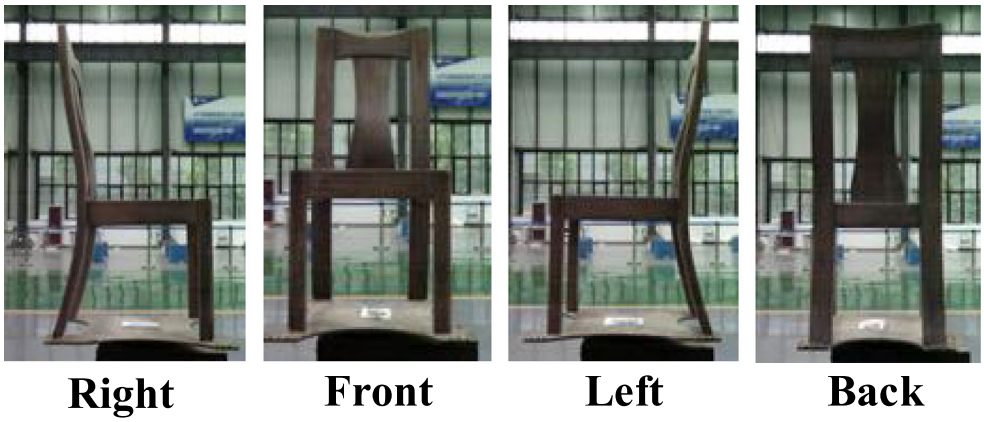

Fig. 3 Four basic orientations of the target

points (outliers) are included, the algorithm is likely to fall into a local minimum and gives inaccurate pose estimation.

As is shown in Fig. 3, the four basic orientations of the target can be defined as right, front, left, and back orientations, which correspond to four basic orientation angles 0° (360°), 90°, 180°, and 270°. The angles within plus minus 45° range of each basic angle are included in corresponding category, as demonstrated in the formula below. If the azimuth of the target is closest to one of these basic orientation angles, this basic orientation is designated as its primary orientation. If the target is placed in the middle of two basic orientations, that is, the azimuth angle is close to the lower or upper bound of each category, the closest and second closest basic orientations are designated as its primary orientation and secondary orientation, respectively.

$$\alpha_{\text{right}} \in (315°, 360°) \cup [0°, 45°],\ \alpha_{\text{front}} \in (45°, 135°],$$
$$\alpha_{\text{left}} \in (135°, 225°],\ \alpha_{\text{back}} \in (225°, 315°] \quad (1)$$

To determine the basic orientation category of the target, an orientation determination network based on VGG [25] classification network structure is proposed. The excellent classification performance achieved by VGG structure derives from the larger depth of the network and the use of convolution filters with a very small receptive field. According to these characteristics, the orientation determination network is devised based on VGG structure with a modification in network depth and filter size to further improve the classification performance. The network structure consists of 15 convolutional layers and 3 fully connected layers as shown in Fig. 4. The input of the network is 224 × 224 RGB image resized from the original image and the output is a probability distribution of four orientation categories.

The primary orientation corresponds to the largest probabilistic output of the network. The secondary orientation is determined by the second largest probability if it is over a certain threshold. Otherwise, there is no secondary orientation. Because the softmax layer enlarges the difference tremendously between the largest and second largest probabilities, in order to attenuate the suppression of the non-maximum value in the comparison, the output of the last fully connected layer in front of the softmax layer is adopt to calculate the largest and second largest probabilities. These probabilities are determined as follows:

$$p = \arg\max_i(v_i)\ i \in \{1, 2, 3, 4\} \quad (2)$$
$$s = \arg\max_i(v_i)\ i \in \{1, 2, 3, 4\} \backslash \{p\}\ \ if\ v_s \geq 0.5 v_p \quad (3)$$

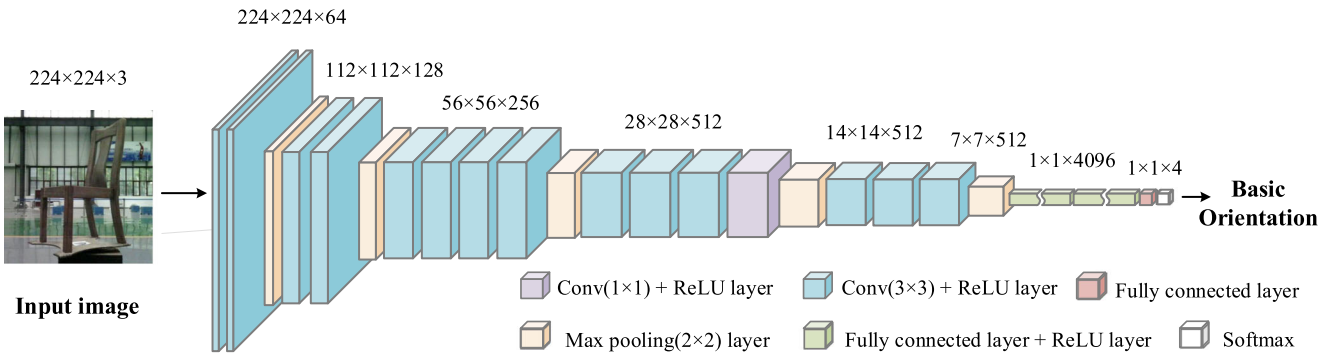

Fig. 4 The network structure of orientation determination network. The height, width and channel of the layers are denoted for each group of convolution layers and fully connected layers

where p and s are the largest and second largest probability indexes and v_i is the output value of the last fully connected layer.

This proposed network can be trained directly using the small quantity of data from industrial setting to provide an accurate determination of rough orientation without the need of any pre-trained model parameters, which is validated in the experiment.

3.3 Semantic segmentation

In order to obtain an appropriate mask for the segmentation of depth data and produce a point cloud representation of the target with few out of range points (outliers), a modified U-SegNet semantic segmentation network is adopted to segment the image. This network requires a relatively small amount of training data and shows desirable segmenting performance in poor lighting conditions.

U-SegNet [24] network is an incorporation of two widely applied segmentation architectures SegNet [21] and U-Net [26]. SegNet is an encoder-decoder architecture that transfers pooling indices from the encoder to the decoder thus achieving more accurate segmentation on the edge of the target while reducing training parameters. U-Net is a fully convolutional network that adopts skip connection between the contracting path and expansive path; thus, it captures multilevel information and can be trained using small quantity of data. The U-SegNet network combines the advantages of the both networks by applying the skip connection from U-Net to the SegNet architectures. This integrated structure reduces the reliance on massive amount of training data and converges faster during training, while achieving a fairly good segmentation performance. This network is first adopted in medical image segmentation domain where the labelled image data are costly to obtain.

In this paper, as is shown in Fig. 5, the U-SegNet structure is modified for the application to the proposed system. In order to improve the perception ability of multilevel information, a skip connection is added between each pair of encoder and decoder layers that shares the pooling indices, adding up to three skip connections in total, while in the original structure, only one skip connection is introduced at the uppermost layer. The skip connections in the proposed modified structure concatenate three encoder layers to the decoder layers respectively thus exploiting the image features of three different scales. Additionally, the input of the network is a 256 × 256 RGB image resized from the original image and the output is a grayscale probabilistic mask of the same size, identifying the target region and the environment. In the spray painting application, the proposed system is designed to estimate the pose of one kind of target at a time. Therefore, the softmax layer designed for multi-class output in the original structure is replaced by a sigmoid function which is designed for binary output. In this paper, we explore several different segmentation networks, including FCN, SegNet, U-net, original U-SegNet, Mask R-CNN [27], and our proposed modified U-SegNet, and find that our proposed U-SegNet is the best selection for RGB image segmentation in this study under the industrial setting.

From the segmentation network, a grayscale probabilistic mask of the target is generated. The gray value of each pixel is in proportion to its probability of belonging to the target region. Hence, the threshold that determines

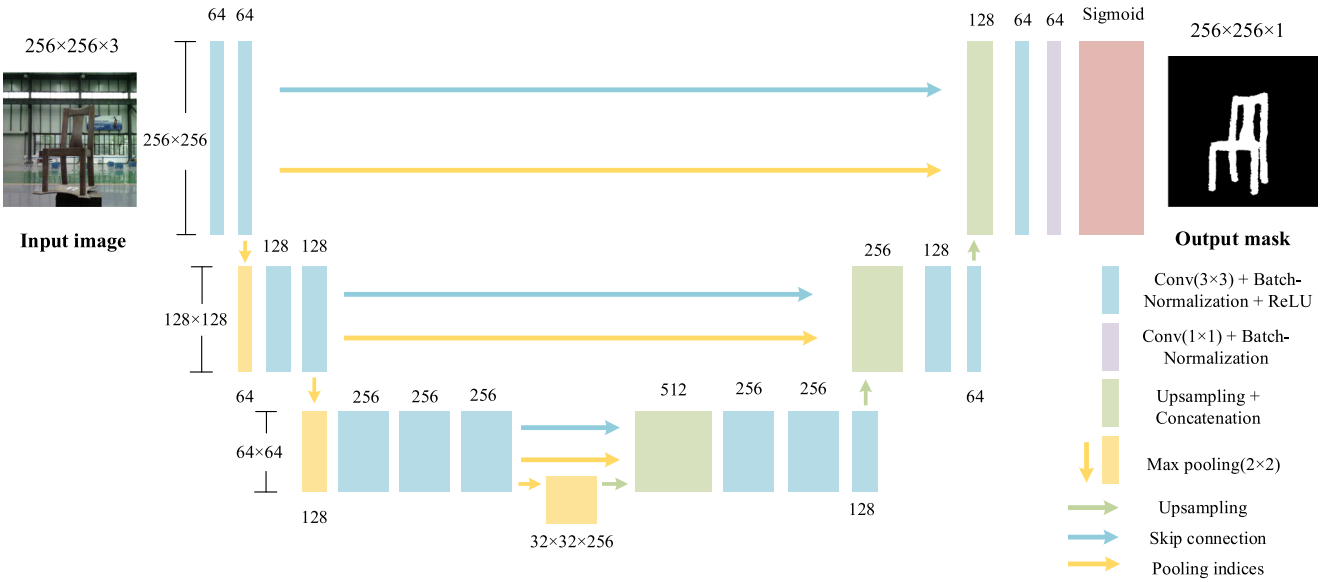

Fig. 5 The modified network structure of U-SegNet network in the proposed system. The major modification is that skip connection is added between each pair of encoder and decoder layers that shares the pooling indices

whether the pixel belongs to the region of interest is also important for the accurate segmentation. Generally, as the threshold increases, the precision rate of segmentation becomes higher and the recall rate becomes lower.

In the proposed system, the RGB image segmentation mask is used to crop the aligned depth image and produce 3D point cloud of the target, given the intrinsic matrix of the camera. If the mask is larger than the target (larger recall), the points from the surrounding objects such as operation platform and robot end effector might be included in the point cloud, thus causing incorrect model registration. On the contrary, if the mask is smaller than the target (larger precision), though a small part of the target is not represented by the point cloud, there are few outliers in the point cloud and the model is fitting with actual part of the target. Therefore, the registration accuracy is not affected if the mask is slightly smaller than the target region. So a slightly higher threshold that favors the increase of precision is adopted in the binarization procedure of the probabilistic mask.

3.4 Model registration

The results from orientation determination and semantic segmentation are integrated for the model registration based on ICP algorithm to provide the pose estimation of the target.

Before the model registration, the pre-scanned model needs to be preprocessed. The model of the target is cropped from four basic orientations defined above and only the front face of the model in corresponding orientation is saved. Hence, four model candidates are provided. This operation largely diminishes the model points that are imperceptible for the sensor in certain orientation, thus reducing the redundant points in the model and preventing ICP registration from converging into a local minimum. In addition, if the model is a CAD model, it should be sampled and converted into a point cloud model for the registration with the segmented target point cloud. The initial model position is set at a reference position of sensor coordinate.

Next, the model is retrieved and two final model candidates corresponding to primary and secondary orientations given by the orientation determination network are selected. The model candidates and segmented target point cloud are downsampled to reduce computational cost. Then ICP is used to register the segmented point cloud with the two model candidates of different orientations, and the ICP fitness score S is a squared error of Euclidean distance computed as follows:

$$S(R,t) = \frac{1}{N}\sum_{i=1}^{N} \|m_i - Rc_i - t\|^2 \tag{4}$$

where m_i and c_i is the model points and target point cloud points, R is the relative rotation and t is the relative translation between model and target. N denotes the total number of successfully aligned points.

From the fitness score, it can be observed that if a large amount of points in the model and the target are not precisely aligned, the score would be fairly high, as is shown in Fig. 6. Hence, the orientation with the smallest fitness score is adopted as the closet orientation and its corresponding pose matrix of rotation and translation is output as the final estimated pose. The introduction of model candidates of primary and secondary orientation gives an alternative model candidate when the target point cloud pose is seriously deviated from the model and a large fitness score is presented indicating incorrect registration. Therefore, this procedure further lowers the possibility of trapping into a local minimum and improves the robustness of the model registration.

4 Experiments and analysis

A series of experiments on estimating the target pose in industrial spray painting setting were conducted to validate the accuracy and effectiveness of the proposed system and algorithms. In these experiments, a wooden chair frame, a common target with rotationally asymmetric figure in spray painting industry as shown in Fig. 7, was adopted as the target. In this section, the initial setup and performance of orientation determination network and

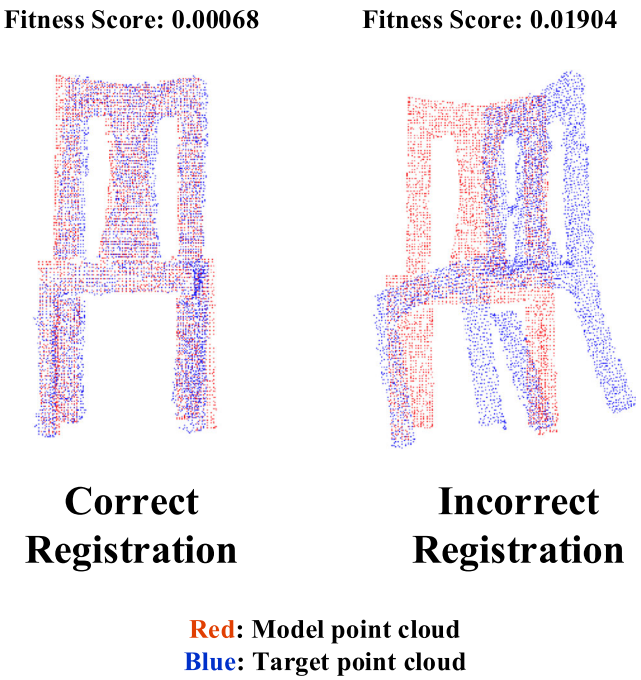

Fig. 6 The fitness score for correct and incorrect registration

Fig. 7 The target: an unpainted wooden chair frame

semantic segmentation network are introduced individually, and then the overall performance of the pose estimation system is exhibited and discussed. Moreover, a novel dataset made for training the network and validating the performance of the proposed system in industrial setting is described in this section.

4.1 Dataset

To train the deep neural networks and verify the effectiveness of the proposed system, an RGB-D dataset with following requirements is needed. First, it should contain sufficient labelled RGB images of unpainted target in industrial spray painting setting for training the classification and segmentation networks. However, the existing large-scale datasets such as ImageNet [28] and Pascal VOC [29] are mainly collected from Internet photos, which contain different targets and are located in different environment from the industrial setting. Second, the dataset requires depth data along with annotation of the target pose for the validation of final pose estimation result in the same setting, thus making the requirements harder to meet by the existing RGB-D dataset. Additionally, the size of dataset for training should be limited since the data acquisition is difficult and time-consuming in industrial application and the dataset should be able to validate the performance of the system given insufficient data. Therefore, a new dataset specialized for pose estimation system in industrial spray painting application was collected.

In this dataset, wooden chair frame in spray painting industry setting was adopted as the target. The RGB-D sensor in the proposed vision system was used for dataset collection and the RGB-D data from each frame were saved as an RGB image and a depth image that were encoded into PNG format. The depth measurement error of the sensor was 1% of its measuring distance. The chair was placed on a rotary table, while the sensor was mounted on a tripod 1.5 m away on the side from the table. So the images of different altitude angles of the sensor to the rotary table could be acquired by adjusting the height of tripod. The RGB-D images were mainly collected from three altitude angles 6°, 11.5°, and 17°, which were named groups I, II, and III respectively. In each group, the chair was rotated horizontally from 0 to 350° with 10° interval, thus producing 36 RGB images and 36 depth images that corresponded to 36 annotated rotation angles and positions for validation of pose estimation. Furthermore, the chair was rotated randomly and 68 more RGB-D images were captured and added into groups I and III according to their altitude angle for training the deep neural networks, while 50 more were added into group II for testing. In addition, RGB-D images were also collected from several random altitude angles ranging from 2 to 30° to promote the robustness of the neural network training and these data, which consisted of 100 images of randomly rotated chair, were included in group IV. Subsequently, the data in groups I, III, and IV were augmented twofold by horizontally flipping every image to enlarge the training set. Finally, the RGB data in each group were divided according to the basic orientation category for orientation determination network, while some images containing target with randomly rotated angles that were ambiguous for labelling the basic orientation were excluded from the dataset. RGB data in training set were also labelled with pixel-wise mask using LabelMe [30] for semantic segmentation. The detailed information of the dataset is presented in Tables 1 and 2.

4.2 Orientation determination results

The implementation details of the proposed orientation determination network are as follows. The network was trained and tested using NVIDIA Tesla V100 GPU. In the training process of the network, the batch size of 3 was adopted in order to accelerate the convergence and

Table 1 The dataset size and information for semantic segmentation

	Training set				Testing set
	Group I (6°)	Group III (17°)	Group IV (Random)	Total	Group II (11.5°)
Original dataset	104	104	100	308	86
Augmented dataset	208	208	200	616	86

Table 2 The dataset size and information for orientation determination

Group	Dataset	Category				
		Right	Front	Left	Back	Total
Groups I, III, and IV	Original training dataset	64	50	64	50	228
	Augmented training dataset	128	100	128	100	456
Group II	Testing dataset	36	32	36	30	134

the whole training set was divided into 186 batches. The cross entropy was adopted as the loss function. The weight parameters in the network were updated and optimized during training using Adam optimizer with a learning rate of 0.0001 to minimize the loss. It was set that the proposed network would be deemed converged if the training loss remained constant for 30 epochs. In the experiment, as shown in Fig. 8, the network converged after 109 epochs and the training duration was 1206 s in total.

The test accuracy of the trained network was 98.51%, which validated that the proposed network was able to determine the primary orientation precisely. The forward propagation time of a single image was 0.2 s.

Since the proposed network was designed based on VGG structure, the testing accuracy of the proposed network was compared with original VGG structures given in [25], including VGG-13, VGG-16, and VGG-19. As demonstrated in Table 3, it was shown that our proposed VGG structure with 18 layers outperformed the other original structures.

In addition, it was observed that the actual orientation of the test images for which the network gave wrong orientation corresponded to the secondary probabilistic output of the network, making the accuracy of primary and secondary orientations 100% accurate. Hence, the wrong primary orientation prediction would not seriously affect the subsequent pose estimation based on primary and secondary orientations.

4.3 Semantic segmentation results

The implementation details of the proposed semantic segmentation network are presented as follows. The proposed network was trained and tested using NVIDIA Tesla V100 GPU. During the training process, the batch size of 5 was adopted in order to accelerate the convergence and the whole training set was divided into 124 batches. The loss function for this network was binary cross entropy. The stochastic gradient descent optimizer was adopted to update the weight parameters. The learning rate was set to be 0.001 and a learning rate decay of 1×10^{-6} was introduced to improve the speed and accuracy of the parameters optimization. In order to reduce the fluctuations and improve convergence rate, Nesterov momentum [31] was applied and its momentum value was set to be 0.9. The total training epochs were set to 300 epochs to analyze the

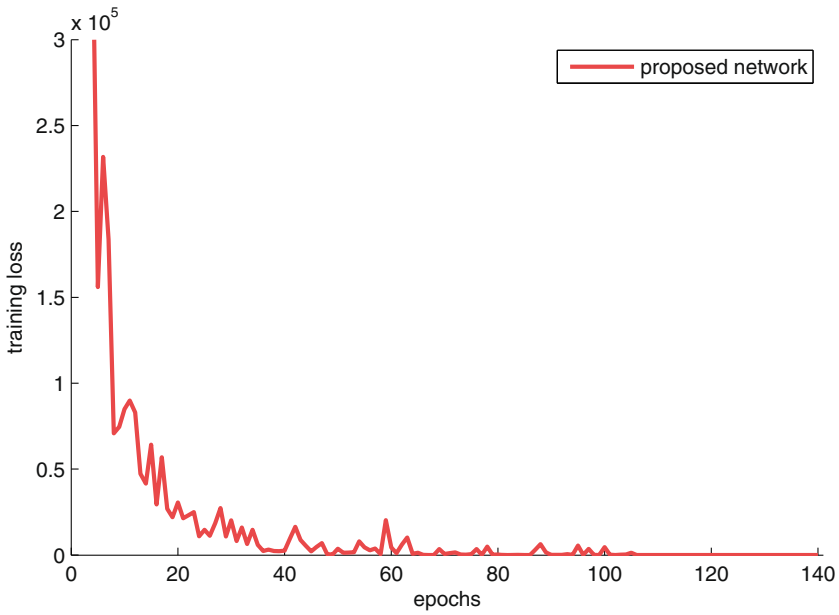

Fig. 8 The training loss curve of the proposed network

Table 3 Comparative performance of different VGG structures

Network structure	Test accuracy (%)
Proposed network	98.51
VGG-13	72.39
VGG-16	90.30
VGG-19	88.81

loss and training accuracy in the entire training process. The proposed network contained 4,837,961 trainable parameters in total and the training accuracy and loss curve is shown in Fig. 9. The graph took iterations as the horizontal ordinate to illustrate the rate of change of the loss and accuracy in the first several epochs as well as the fluctuations. The proposed network converged quickly after 5000 iterations, or 40 epochs. The forward propagation time for each image through the network was around 0.5 s on average, which was also the segmentation time of an input image.

To validate the effectiveness of the proposed segmentation network, the network was tested using the test set including 86 images and corresponding manually labelled ground truth masks, and the network output masks were compared with ground truth masks. Furthermore, we tested several different segmentation networks including FCN, SegNet, U-net, Mask R-CNN, and original U-SegNet with the same test set and made comparison within the outputs of these networks. The visualized comparison of the output segmented grayscale masks is shown in Fig. 10. From the comparison, it can be seen that the proposed modified U-SegNet structure and original U-SegNet structure resembles the ground truth closely. Comparatively, the mask from original U-SegNet structure contains outliers from the surrounding environment, which would be included in the segmented point cloud and interfere the registration, while the segmented mask from our proposed structure mainly contains the points from target region owing to the presence of additional skip connections in the structure. In contrast with U-SegNet structures, the masks from U-net and FCN network miss some part of the target region due to the inaccurate segmentation on the edge of the target. The boundary of the masks segmented by Mask R-CNN is also distorted due to the lack of shared pooling indices. The output from SegNet, however, gives inapplicable mask of the target because the insufficient training data for the VGG based deep segmentation network without skip connections.

In order to give quantitative evaluation of the test results, several metrics for segmentation were adopted. For the binary semantic segmentation, the prediction of every pixel in the segmented mask can be divided into four categories including TP, TN, FP, and FN, which are explained in Table 4. These were adopted to calculate the pixel-wise evaluation metrics for the testing performance of each network.

In this experiment, the evaluation metrics include precision, recall, F1 score (also known as Dice similarity coefficient), and IoU (intersection over union, also known as Jaccard index). The metrics are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{8}$$

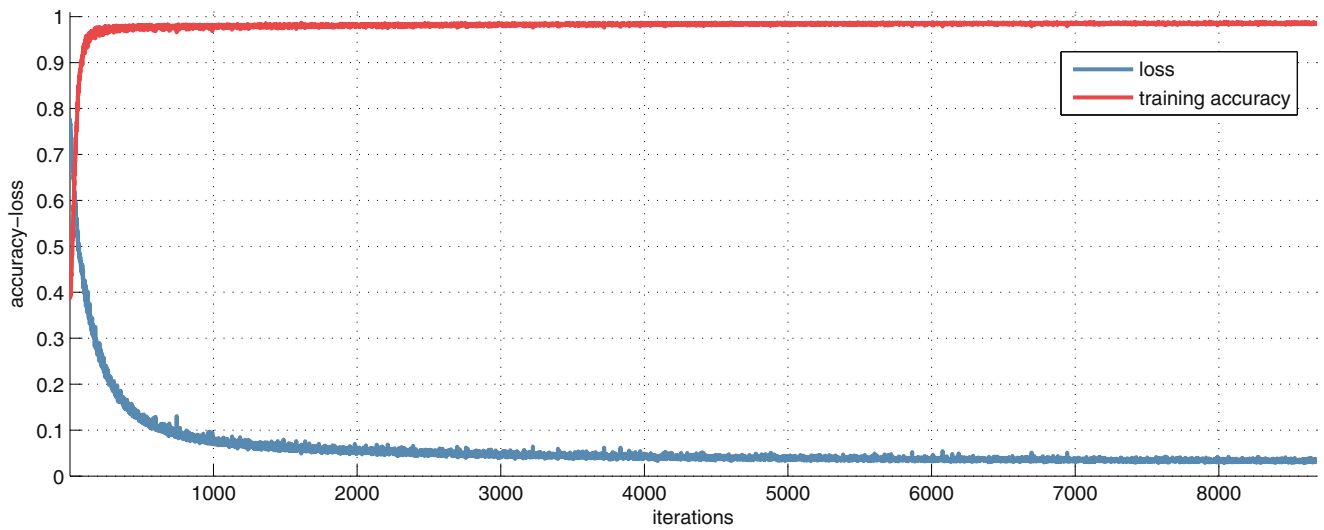

Fig. 9 The training accuracy and loss curve of the proposed segmentation network

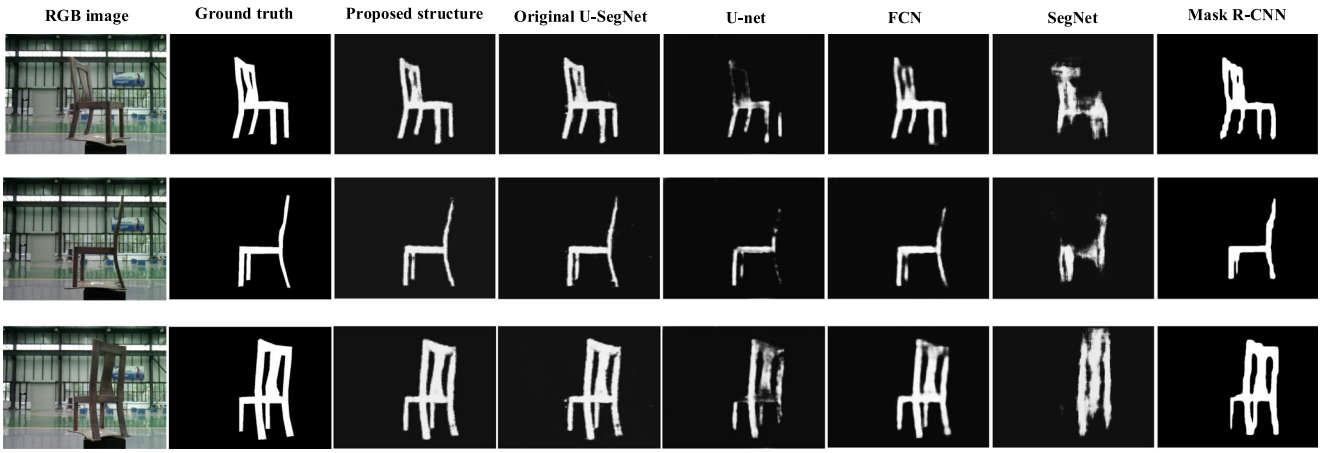

Fig. 10 Visualized comparison of the segmented masks between our proposed structure and other semantic segmentation methods

Table 4 Confusion matrix of segmented mask prediction, where positive denotes target region and negative denotes environment

Ground truth	Prediction	Category
Positive	Positive	True positive (TP)
Positive	Negative	False negative (FN)
Negative	Positive	False positive (FP)
Negative	Negative	True negative (TN)

Table 5 Comparative performance of different segmentation networks

Network structure	Precision (%)	Recall (%)	F1 score (%)	IoU (%)
Proposed network	93.37	94.21	93.75	88.30
Original U-SegNet	93.04	94.27	93.61	88.04
U-net	98.08	61.84	74.84	60.98
FCN	95.23	86.95	90.79	83.30
SegNet	66.44	66.49	66.30	50.07
Mask R-CNN	89.36	83.89	86.48	76.27

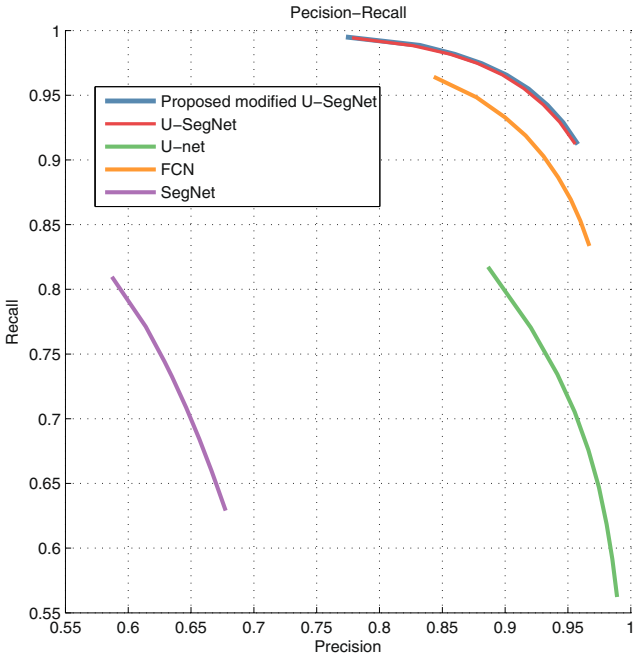

Fig. 11 Precision-recall curves of different networks

Before the evaluation, the segmented grayscale probabilistic mask of the target was binarized. The binarization threshold determines whether the pixel belongs to the region of interest and thus has influence on the metric value. This threshold is a normalized intensity value lies within the range of 0 to 1. If normalized intensity value of a pixel is over the threshold, the pixel is deemed as part of the target region and given the maximal intensity, otherwise it is allocated into the background and given the intensity value of 0. To meet the above mentioned requirement of the proposed system, a slightly higher binarization threshold of 0.4 was adopted in this experiment as well as in the application of the proposed system.

The performance of different segmentation networks was evaluated accordingly using the evaluation metrics. The results are presented in Table 5. It is shown that the proposed modified U-SegNet structure outperforms others in comprehensive metrics F1 score and IoU. Noticing that the U-net and FCN show higher precision, the reason is that the mask is too thin for the target and misses relatively large part of the target though not covering beyond the target area.

Moreover, to further evaluate the performance of the networks under different segmented mask binarization thresholds, the precision-recall curves of the networks with different thresholds were plotted. Since Mask R-CNN is an instance segmentation network which does not provide a probabilistic mask for binary segmentation like other networks, the result of Mask R-CNN is not presented in the comparison of precision-recall curves. As shown in Fig. 11, it is indicated that the proposed network structure achieves better recall while providing the same precision and vice versa.

4.4 Pose estimation results

In this experiment, using the segmented mask from the segmentation, the corresponding area of depth image was cropped and mapped into 3D point cloud. The 3D point processing and ICP registration were implemented using PCL [32]. The model candidates and the target point cloud were first downsampled using voxel grid filter in [32] with a leaf size of 0.01. The point cloud size was decreased from around 28,000 points to 5,000 points, which largely accelerated the ICP registration. For ICP registration, the maximal correspondence distance between two aligned point was set to 0.21. To judge the convergence and determine the end of the iteration, the Euclidean fitness epsilon was set to 0.01 and maximal iterations were set to 30.

In this experiment, the processing time of ICP registration between two point clouds was 0.8 s on average. For the images that obtained secondary orientation from the orientation determination network, the process time was doubled to 1.6 s. In total, the whole processing time added up to 2 to 3 s from the original images captured by sensor to an estimated pose.

The dataset annotated with pose from group II was used to validate the model registration. ICP registration gave a pose estimation of the rotation and translation matrix between the model candidates and target point cloud along with a fitness score. The matrix corresponded to smallest score was adopted as the final rotation and translation matrix and then the rotation angle and 3D position were computed and compared with annotated value. The mean absolute error values and standard deviations of the rotation angle and 3D position were calculated to validate the performance.

Also, to compare the influence of segmentation on the final pose estimation results, the model registration result from segmented target point cloud using the proposed segmentation network was compared with the segmented point cloud obtained using other networks. From Table 6, it is shown that the proposed network gives 1.83° in angle error and 0.0284 m in position error on average with relatively small standard deviations, which outperform others in the final pose estimation statistically. The distribution histograms of angle and position error are shown in Figs. 12 and 13 to illustrate the error distribution. It could be observed that the majority of data is within a relatively small error range, which is less than 5° in angle error and less than 0.05 m in position error. As shown in the illustration, 92% of the position errors and angle errors of the proposed network are within this range. This estimation error is within the spray radius of 0.08 m in the spray painting process of the chair and the pose provided by this

Table 6 Mean error and standard deviation of rotation angle (°) and 3D position (m)

Mask	Mean error (°)	Standard deviation (°)	Mean error (m)	Standard deviation (m)
Proposed network	1.830	1.044	0.0284	0.0209
Original U-SegNet	1.899	2.083	0.0294	0.0227
U-net	2.415	5.509	0.0617	0.1928
FCN	1.906	2.192	0.0286	0.0215
SegNet	4.771	13.674	0.0371	0.0326
Mask R-CNN	2.307	2.817	0.0446	0.0775

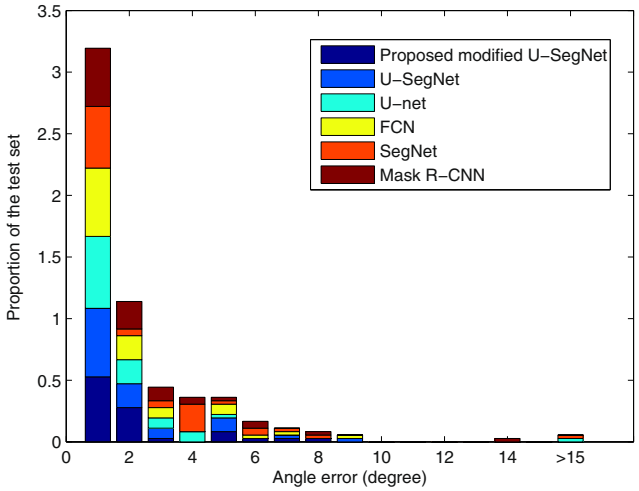

Fig. 12 Angle error distribution histogram of pose estimation based on different networks

system is sufficient for accomplishing the spray painting operation accurately.

Furthermore, the experiment results were compared with other pose estimation systems based on model registration. In [18] and [20], the overall estimation result is deemed successful when the angle error is within 15° and the position error is within 0.05 m. In this sense, the results obtained from the proposed system within this limit add up to 92%, while the best performance of [20] is 80% and [18] is 79%. Compared by the average error of rotation angle, the mean angle error of the proposed system is 1.83°, while the best performance given by [20] is 4.5° and by [18] is 3.95°. In contrast, our proposed system gives superior performance in angle estimation, and all the angle estimation errors are less than 15° in our system. When comparing the 3D position error, our system gives

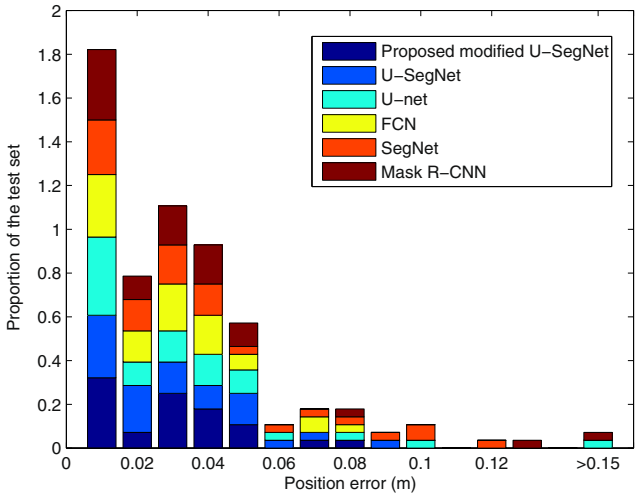

Fig. 13 Position error distribution histogram of pose estimation based on different networks

an average of 0.0284 m, which is inferior to 0.011 m of [20] and 0.007 m of [22]. Nevertheless, in comparison with the proposed system that is set in industrial spray painting environment, the vision system of [18], [20], and [22] is relatively close to the target in the laboratory setting and the target size is also smaller than the chair. Moreover, for the distance of 1.5 m in this experiment, the RGB-D sensor itself introduced an uncertainty of over 0.01 m in depth value due to its own precision limit. When compared with the real dimensions of the target, as mentioned in [22], the 3D position error compared to the dimensions of the chair (0.44 m × 0.42 m × 0.92 m) is 5.4%, which is superior to the [22] that gives 7.3%. On the other hand, the angle is scale invariant in the pose estimation, and our proposed system achieves notably higher performance.

5 Conclusions

In this paper, we present a pose estimation system based on deep neural network and ICP registration for robotic spray painting application. The system is able to meet spray painting requirements, providing an accurate pose estimation (1.83° angle error and 0.0284 m position error) of the target from a relatively large safe distance (1.5 m) under the insufficient illumination in industrial setting. Notably, the estimation performance of target rotation angle, which is scale invariant, largely surpasses the relevant pose estimation systems based on model registration. From the experiments and analysis mentioned above, the system has three properties that help to achieve high effectiveness in robotic spray painting application.

1. Firstly, the RGB-D sensor in the proposed system is placed at a relatively large safe distance (1.5 m) from the target, hence providing sufficient interspace for robotic operation, while the algorithm still guarantees enough estimation accuracy for the painting applications.
2. Secondly, a modified U-SegNet network structure is designed and applied as the semantic segmentation network. This network structure is capable of utilizing small quantity of labelled training data to achieve accurate segmentation (93.75% in F1 score metric) with few outliers in poor lighting condition.
3. Thirdly, an orientation determination network is designed to use the RGB image to determine the basic orientations precisely (98.51% accuracy for primary orientation and 100% for primary and secondary orientations) of the target and to aid the selection of model candidates accordingly. This method reduces the misalignment caused by local minima in ICP registration, making the final pose estimation result accurate.

With prepared training data and models, the proposed system can be applied to accurately estimate the pose of targets with rotationally asymmetric figure such as furniture, complex steel structure in spray painting application. Moreover, the accuracy of the proposed system could be further improved with the integration of a vision sensor of higher precision such as structured light sensor based on digital light processing technology. In the future, this proposed system is expected to be integrated with model-based robot trajectory generation methods and improve the productivity as well as the quality of the automatic spray painting manufactural lines.

Acknowledgements The authors would like to express sincere gratitude to the reviewers and the editors for their valuable suggestions.

Funding information This work was supported by the National Natural Science Foundation of China under Grant Nos. U1813208 and 61573358.

References

1. Zhang BB, Wu J, Wang LP, Yu ZY, Fu P (2018) A method to realize accurate dynamic feedforward control of a spray-painting robot for airplane wings. Ieee-Asme T Mech 23(3):1182–1192
2. Ren SN, Xie Y, Yang XD, Xu J, Wang GL, Chen K (2017) A method for optimizing the base position of mobile painting manipulators. IEEE T Autom Sci Eng 14(1):370–375
3. Trigatti G, Boscariol P, Scalera L, Pillan D, Gasparetto A (2018) A new path-constrained trajectory planning strategy for spray painting robots - rev.1. Int J Adv Manuf Tech 98(9–12):2287–2296
4. Chen HP, Xi N (2008) Automated tool trajectory planning of industrial robots for painting composite surfaces. Int J Adv Manuf Tech 35(7–8):680–696
5. Andulkar MV, Chiddarwar SS, Marathe AS (2015) Novel integrated offline trajectory generation approach for robot assisted spray painting operation. J Manuf Syst 37:201–216
6. Wang G, Cheng J, Li R, Chen K (2015) A new point cloud slicing based path planning algorithm for robotic spray painting. In: IEEE international conference on robotics and biomimetics, pp 1717–1722
7. Chen H, Fuhlbrigge T, Li X (2008) Automated industrial robot path planning for spray painting process: a review. In: IEEE international conference on automation science and engineering, pp 522–527
8. Kharidege A, Ting D, Yajun Z (2017) A practical approach for automated polishing system of free-form surface path generation based on industrial arm robot. Int J Adv Manuf Tech 93(9–12):3921–3934
9. Chen R, Wang GL, Zhao JG, Xu J, Chen K (2018) Fringe pattern based plane-to-plane visual servoing for robotic spray path planning. IEEE-Asme T Mech 23(3):1083–1091
10. Xu Z, He W, Yuan K (2011) A real-time position and posture measurement device for painting robot. In: International conference on electric information and control engineering, pp 1942–194
11. Lin CY, Abebe ZA, Chang SH (2015) Advanced spraying task strategy for bicycle-frame based on geometrical data of workpiece. In: International conference on advanced robotics, pp 277–282
12. Lin W, Anwar A, Li Z, Tong M, Qiu J, Gao H (2019) Recognition and pose estimation of auto parts for an autonomous spray painting robot. IEEE T Ind Inform 15(3):1709–1719
13. Besl PJ, Mckay ND (1992) A method for registration of 3-D shapes. IEEE T Pattern Anal 14(2):239–256
14. Hodan T, Zabulis X, Lourakis M, Obdrzalek S, Matas J (2015) Detection and fine 3D pose estimation of texture-less objects in RGB-D images. In: IEEE/RSJ international conference on intelligent robots and systems, pp 4421–4428
15. Schwarz M, Schulz H, Behnke S (2015) RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features. In: 2015 IEEE international conference on robotics and automation, pp 1329–1335
16. Xiang Y, Schmidt T, Narayanan V, Fox D (2018) PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes. arXiv:1711.00199
17. Collet A, Martinez M, Srinivasa S (2011) The MOPED framework: object recognition and pose estimation for manipulation. I J Robot Res 30(10):1284–1306
18. Zeng A, Yu K, Song S, Suo D, Walker E, Rodriguez A, Xiao J (2017) Multi-view self-supervised deep learning for 6D pose estimation in the Amazon picking challenge. In: 2017 IEEE international conference on robotics and automation, pp 1386–1383
19. Shelhamer E, Long J, Darrell T (2017) Fully convolutional networks for semantic segmentation. IEEE T Pattern Anal 39(4):640–651
20. Wong JM, Kee V, Le T, Wagner S, Mariottini GL, Schneider A, Hamilton L, Chipalkatty R, Hebert M, Johnson DMS (2017) SegICP: integrated deep semantic segmentation and pose estimation. In: 2017 IEEE/RSJ international conference on intelligent robots and systems, pp 5784–5789
21. Badrinarayanan V, Kendall A, Cipolla R (2017) SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE T Pattern Anal 39(12):2481–2495
22. Lin CM, Tsai CY, Lai YC, Li SA, Wong CC (2018) Visual object recognition and pose estimation based on a deep semantic segmentation network. IEEE Sens J 18(22):9370–9381
23. Yang G, Wang S, Yang J, Shen B (2018) Active pose estimation of daily objects. In: 2018 IEEE international conference on mechatronics and automation, pp 837–842
24. Kumar P, Nagar P, Arora C, Gupta A (2018) U-Segnet: fully convolutional neural network based automated brain tissue segmentation tool. In: 2018 25th IEEE international conference on image processing, pp 3503–3507
25. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
26. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention – MICCAI, pp 234–241
27. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. In: IEEE international conference on computer vision, pp 2980–2988

28. Deng J, Dong W, Socher R, Li LJ, Li K, Li FF (2009) ImageNet: a large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition, pp 248–255
29. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The pascal visual object classes (VOC) challenge. Int J Comput Vis 88(2):303–338
30. Russell BC, Torralba A, Murphy KP, Freeman WT (2008) LabelMe: a database and web-based tool for image annotation. Int J Comput Vis 77(1–3):157–173
31. Sutskever I, Martens J, Dahl G, Hinton G (2013) On the importance of initialization and momentum in deep learning. In: International conference on machine learning, pp 1139–1147
32. Rusu RB, Cousins S (2011) 3D is here: point cloud library (PCL). In: 2011 IEEE international conference on robotics and automation, pp 1–4