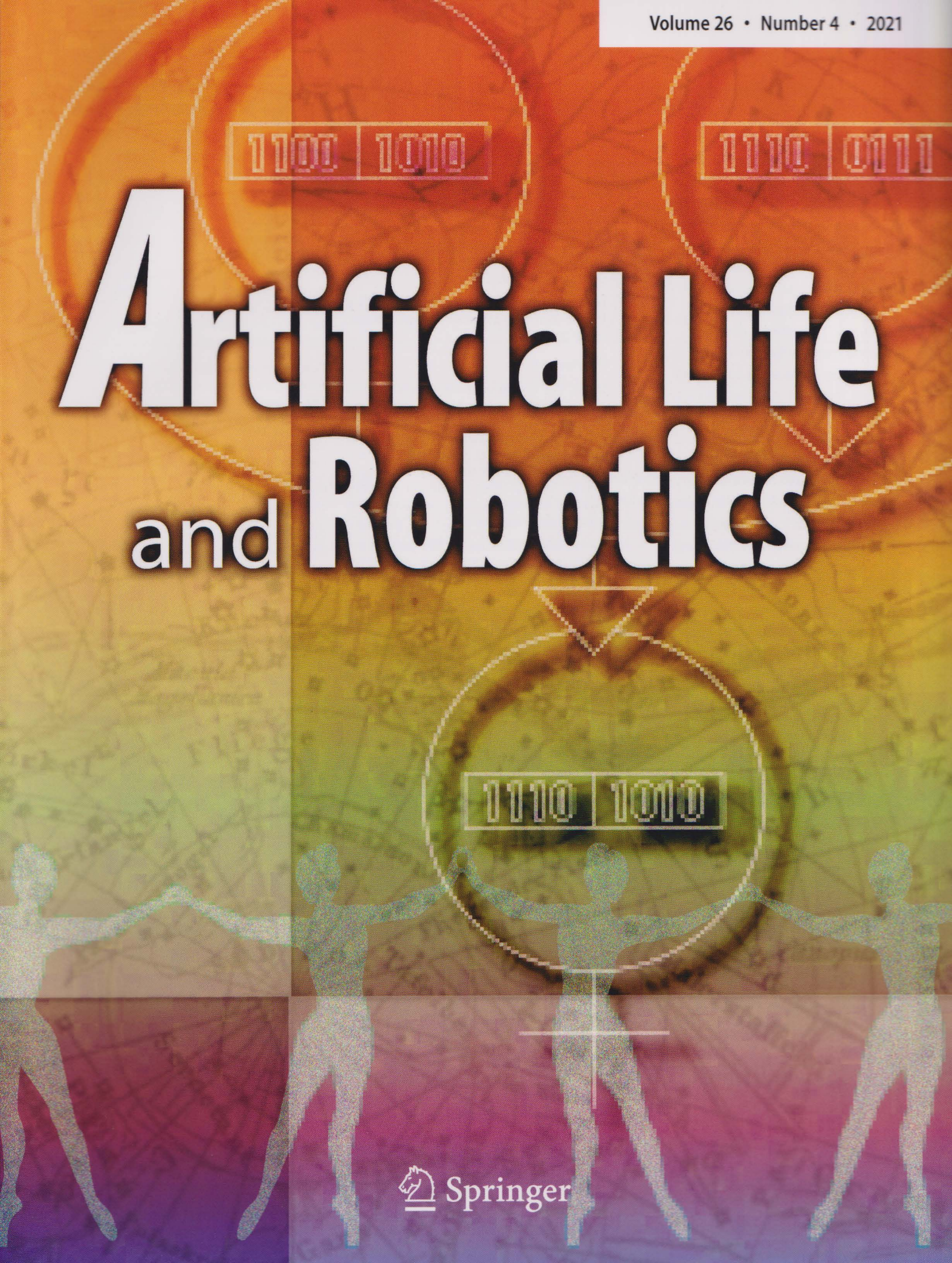
Volume 26 • Number 4 • 2021



Artificial Life and Robotics

Springer

Artificial Life and Robotics (2021) 26:390–395
https://doi.org/10.1007/s10015-021-00692-0

ORIGINAL ARTICLE

Molded article picking robot using image processing technique and pixel-based visual feedback control

Kohei Miki[1] · Fusaomi Nagata[1] · Takeshi Ikeda[1] · Keigo Watanabe[2] · Maki K. Habib[3]

Received: 21 February 2021 / Accepted: 4 August 2021 / Published online: 10 August 2021

Abstract
This paper aims to develop a robotic system that is able to find and remove unwanted molded articles, which fell in a narrow metallic mold space. Currently, this task is being supported by skilled workers. The proposed robotic system has the ability to estimate the orientation of articles using transfer learning-based convolutional neural networks (CNNs). The orientation information is essential and indispensable to realize stable robot picking operations. In addition, pixel-based visual feedback (PBVF) controller is introduced by referring to the center of gravity (COG) position of articles computed by image processing techniques. Hence, it is possible to eliminate the complex calibration between the camera and the robot coordinate systems. The implementation and effectiveness of the pick and place robot are demonstrated, where the conventional calibration of such task is not required.

Keywords Convolutional neural network (CNN) · Transfer learning · Pixel-based visual feedback · Pick and place

1 Introduction

A wide range of robots is applied, organized, and controlled for the purpose to achieve the desire automation of production lines in a factory. During robotic manipulation, it is required to conduct pick and place tasks, and hence it is essential for the robot to estimate the position and orientation of target objects. To cope with this need, many research development activities have been devoted to develop the necessary technologies to abstract visual features from acquired images and control robots autonomously in real-world environments [1]. For example, Tarydi et al. [2] proposed an image processing algorithm to estimate the 3D position and orientation of an object from images taken by calibrated stereo cameras, which allowed a 6 DOF robot arm with a gripper to place it at an arbitrary position in the robot workspace.

Until around 2012, the research field of object recognition had been dominated by the use of human-designed features such as Fisher and SIFT features. However, the SuperVision team using convolutional neural networks (CNNs) successfully become the winner at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 [3, 4]. After that, the use of CNN in robot control become attractive, because visual features can be extracted through training. For example, Haochen et al. [5] designed a CNN to estimate the category, position, and orientation of PBC object, and showed that it could be processed at high speed and accuracy.

However, it is also recognized at the present stage that when a robot working on a production line fails to pick and place an article, the article is quickly removed by a worker to avoid damaging the production line. Such type of task is exhausting the worker as it is usually required to assign one person to follow up the operation of several robots during actual production.

In this paper, for the purpose to solve this problem, a basic pick and place robotic system is proposed. The system has an ability to estimate the orientation of articles with a resolution of 5° using transfer learning-based CNNs [6, 7].

This work was presented in part at the 26th International Symposium on Artificial Life and Robotics (Online, January 21–23, 2021).

✉ Fusaomi Nagata
nagata@rs.socu.ac.jp

[1] Graduate School of Engineering, Sanyo-Onoda City University, 1-1-1 Daigaku-Dori, Sanyo-Onoda 756-0884, Japan

[2] Graduate School of Natural Science and Technology, Okayama University, Okayama, Japan

[3] Mechanical Engineering Department, School of Sciences and Engineering, American University in Cairo, Cairo, Egypt

The orientation information of molded articles is essential to realize stable robotic picking operations.

In addition, a pixel-based visual feedback (PBVF) controller is designed by referring to the articles' center of gravity (COG) obtained by image processing techniques as the control quantity [7]. In the proposed PBVF controller, the end-effector's position is regulated to get the COG of an article overlapped with the center of the image frame. Consequently, the normally used complicated calibration task between the camera and the robot coordinate systems can be eliminated. The effectiveness and promise of the proposed robotic system are demonstrated through pick and place experiments.

2 CNNs for orientation detection

This section describes the design of transfer learning-based CNNs to estimate objects' orientations. CNNs, such as AlexNet [4], VGG16, VGG19 [8] or GoogLeNet [9] are used as powerful base CNN models to support wide range of development. Besides, conditions of additional training (i.e., fine tuning of weights) in the transfer learning are also presented. Then, the eight types of transfer learning-based CNNs are compared quantitatively. To determine the best among the four CNN models under consideration, the generalization performance of each CNN is evaluated using approximately 2000 test images that are not used during the additional training processes.

2.1 Design of eight types of CNNs using transfer learning

As for the transfer learning, the conceptual idea is focusing on making use of the knowledge obtained in one domain to improve the learning of a prediction function in another domain [10]. It is known that there are two advantages of the usage of transfer learning when designing a new CNN. The first advantage is the shortening the learning time compared to normal learning process from scratch. The other advantage is the improvement of recognition rate in spite of less training data. As for the actual design and training of a transfer learning-based CNN for a task, the output layer of the original CNN model has to be replaced with a new one that has the desired number of outputs needed by the task. Additional training of the CNN facilitated by the error back propagation method is mainly applied to the fully connected layers.

In this paper, the designs of eight types of transfer learning-based CNNs are considered and the CNNs are trained to estimate the orientation of objects within the desired resolution of 5°. The CNNs can output 36 kinds of labels, such as 0°, 5°,..., 175°, as shown in Fig. 1.

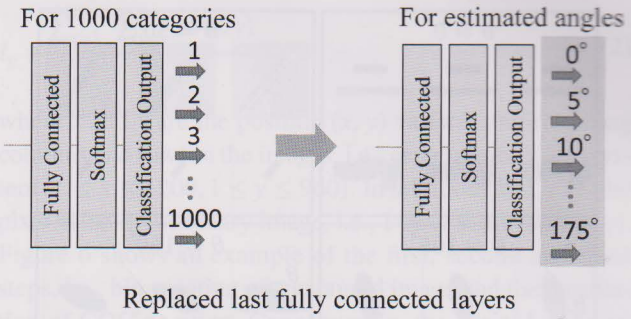


Fig. 1 Design example of transfer learning-based CNN for orientation detection, in which the desired resolution of orientation is set to 5°

In the training processes, the weight parameters of four CNN models are fine-tuned in two different training conditions. In the first training condition, the weights are trained only in the fully connected layers. On the other hand, in the second training condition, all weights in the convolution layers and the fully connected ones are trained. The superiority or inferiority of each training condition is evaluated by checking the effects of the fine-tuning. The criterion of the effect is conducted based on recognition accuracy.

The powerful four CNN models that are effectively used as the bases of transfer learning are AlexNet, VGG16, VGG19 and GoogLeNet. AlexNet, VGG16 and VGG19 have orthodox network structures called the series type. On the other hand, GoogLeNet has a different structure called DAG (Directed Acyclic Graph) type network. These are originally trained using the ImageNet data set [3]. In the design based on transfer learning, the original output layer structure for 1000 categorizations is replaced with a redesigned one that has 36 outputs, i.e., labels from 0° to 175° , as shown in Fig. 1.

As for the parameters for the additional training, the max epoch, mini batch size, desired accuracy, loss and L2 regularization coefficient are set to 500, 30, 0.999, 0.0001 and 0.004, respectively. The initial learning rate is set to 0.004 and then it is gradually decreased by multiplying 0.1 for every 2 epochs. The data set used for additional training, i.e., fine tuning, consists of 15,264 (424×36) images designed by authors, as shown in Fig. 2, in which 424 images are equally alloted to each of 36 classes. All input images are automatically resized to fit the resolutions of the input layer of the transferred CNN models. The classifiers' performances are evaluated using different features included in 2232 test images. The training and test processes of CNNs are conducted using the authors' developed CNN and support vector machine (SVM) design application implemented on MATLAB [11].

The training images are first created by rotating the twelve kinds of objects from 0° to 175° with 5° increments.

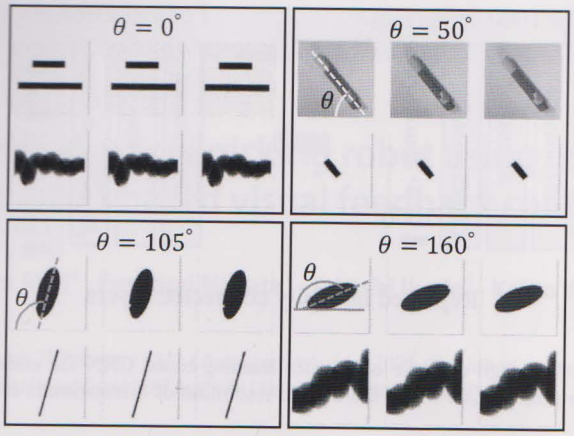


Fig. 2 Examples of training images for 0°, 50°, 105° and 160°. Note that a little bit augmented images are included in the training data

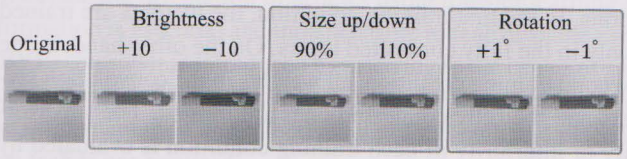


Fig. 3 Examples of several augmentation methods by changing brightness, resolution and orientation, which are applied to the original training images to increase the number

Figure 2 shows examples of training images in case of 0°, 50°, 105° and 160°. Note that a little bit augmented images are included in the training data. As for the augmentation of training images, the following image processing methods are applied.

(1) Rotating the original images every 0.1° within the range of ±1°.
(2) Changing the brightness of original images, i.e., RGB pixel values, every 1 within the range of ±10.
(3) Changing the size of images every 1% within the range of ±10%.

Figure 3 shows examples of the augmentation obtained by the above processes. The eight kinds of CNNs transfered from AlexNet, VGG16, VGG19 and GoogLeNet were additionally and finely trained using the same 15,264 images. Numerical evaluations are compared in the following subsection.

2.2 Evaluation of the transfer learning-based CNNs

After the fine training process, generalization abilities of the transfer learning-based CNNs were checked using 2232 test images as shown in Fig. 4. These test images were not included in the training data set. The classification results are

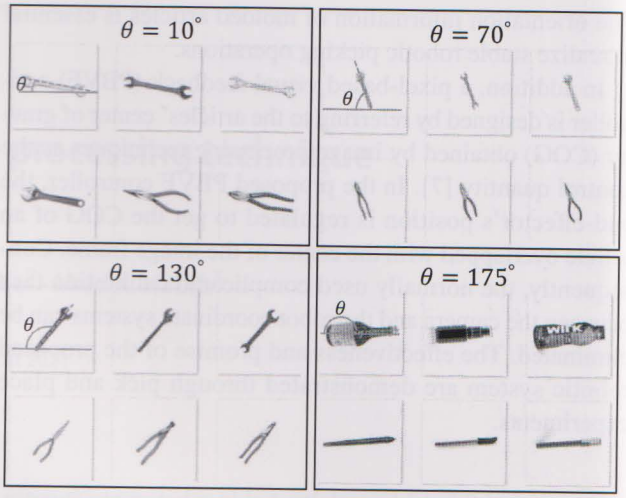


Fig. 4 Examples of test images for 10°, 70°, 130° and 175°

Table 1 Classification results of eight kinds of transfer learning-based CNNs [12]

Base models	Fine-tuned layers	Accuracy	Precision	Recall
AlexNet	ALL	0.392	0.402	0.392
AlexNet	FC-Layers	0.187	0.185	0.235
VGG16	ALL	0.683	0.713	0.683
VGG16	FC-Layers	0.196	0.220	0.196
VGG19	ALL	0.671	0.694	0.671
VGG19	FC-Layers	0.202	0.217	0.202
GoogLeNet	ALL	0.724	0.734	0.724
GoogLeNet	FC-Layers	0.078	0.080	0.078

shown in Table 1, in which the accuracy, precision and recall are used as criteria [12]. It is observed from the comparison result of the CNN models that the GoogLeNet-based CNN model obtained through all layers fine tuning process could perform the best. Besides, It is also observed from the comparison between VGG16-based and VGG19-based CNNs, VGG16-based CNN was superior with respect to the accuracy, precision and recall. Actually, the lengths of the transferred network structures of AlexNet-based, VGG16-based and VGG19-based CNNs, which are series type CNNs, are 25, 41 and 47, respectively. These results suggest that the number of training images, i.e., 15,264, may have been not sufficient to additionally train the VGG19-based CNN with the deepest structure within the three CNNs. On the other hand, although GoogLeNet-based CNN has a deeper network structure with 144 layers, the better numerical result is obtained, as shown in Table 1. It seems that the DAG type network has a superiority compared with the series type ones to be efficiently trained with less number of training images.

Consequently, the authors had selected the GoogLeNet-based CNN as the best one for estimating the orientation of target objects. In the subsequent section, picking experiments using a small articulated robot are demonstrated.

3 Application of transfer learning based CNN to a picking robot

In this section, the obtained transfer learning-based CNN using GoogLeNet is applied to a picking robot, as shown in Fig. 5. A Web camera is used when the PBVF controller is not applied, on the other hand, an endoscope camera is used when the PBVF controller is applied. The CNN is used to estimate objects' orientation on a working table.

3.1 Without the pixel-based visual feedback (PBVF) controller

The robotic pick and place without PBVF controller can be applied through a process consisting of five steps. First, a snapshot with the resolution of 1200×960 is captured by the Web-camera. Second, the image is binarized to black and white. Thirdly, the COG position $[I_x\ I_y]^T$ of an object is computed using Eqs. (1) and (2), assuming that the largest connected component in the image is the target object:

$$I_x = \frac{\sum_{x=1}^{1200} \sum_{y=1}^{960} xB(x,y)}{S}, \tag{1}$$

$$I_y = \frac{\sum_{x=1}^{1200} \sum_{y=1}^{960} yB(x,y)}{S}, \tag{2}$$

where x and y are the position (x, y) variables representing columns and rows in the images, i.e., camera coordinate system ($1 \le x \le 1200$, $1 \le y \le 960$). In addition, $B(x, y)$ is the pixel value in the binary image, i.e., 1 or 0, at position (x, y). Figure 6 shows an example of the first, second and third steps, i.e., binarization of a captured image and the calculation of COG position. Consequently, the desired position (x_d, y_d) in robot coordinate system that aims to move the robot's gripper just above a COG position can be obtained by

$$x_d = X_1 + I_x \frac{X_2 - X_1}{1200}, \tag{3}$$

$$y_d = Y_1 + I_y \frac{Y_2 - Y_1}{960}, \tag{4}$$

where (X_1, Y_1) and (X_2, Y_2) are the positions of left upper and right bottom of the snapshot described in robot coordinate system, as shown in Fig. 7, i.e., they correspond to image coordinates of (1, 1) and (1200, 960), respectively. Fourthly, the part of the connected component is further cropped centering the COG from the original snapshot, as shown in Fig. 7. The cropped image is resized according to the resolution of the CNN's input layer, and then it is given to the input layer. Finally, the CNN estimates the desired

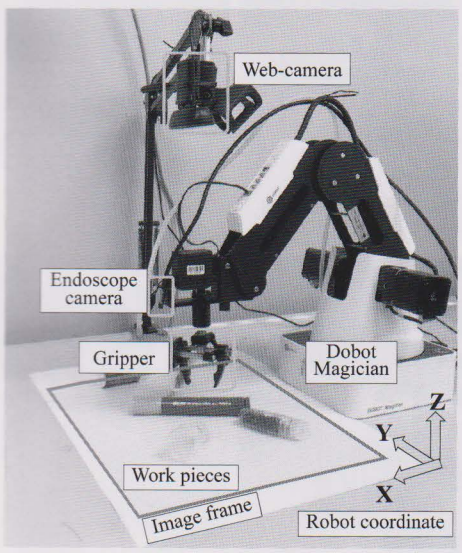


Fig. 5 Picking robot system that can move the end-effector to the position just above the COG position of a workpice by PBVF control and then recognize the orientation by CNN

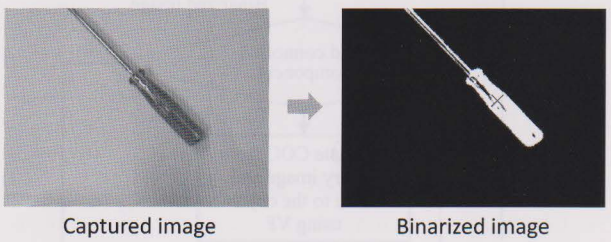


Fig. 6 Example of capturing an image, binarization, and calculation of COG position

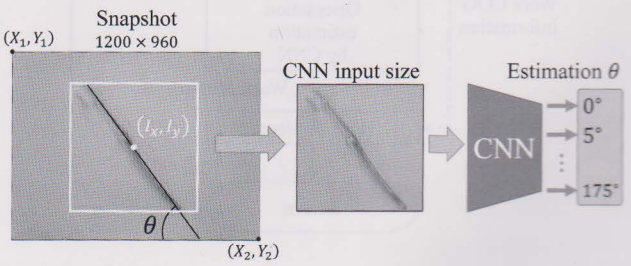


Fig. 7 Procedure to extract the orientation of a workpiece from a captured image

yaw angle θ of the object with a major axis shape, so that the robot can try to grasp the object using the estimated θ.

3.2 Pixel-based visual feedback (PBVF) controller

Figure 8 shows the flowchart of the robotic pick and place operation, in which the proposed CNN and PBVF controller are implemented. It is expected that complicated camera configuration is no more required due to the contribution of the developed PBVF controller. This has the advantage that is different from the system using a Web camera introduced in the previous subsection. A lightweight endoscope camera is attached close to the gripper, as shown in Fig. 5, so that real-time snapshot images viewed from the gripper can be obtained. Manipulated variable $\boldsymbol{v}(k) = [v_x(k)\ v_y(k)]^T$ for visual feedback is generated by a simple PI-action given by

$$\boldsymbol{v}(k) = K_p \boldsymbol{e}(k) + K_i \sum_{n=1}^{k} \boldsymbol{e}(n), \tag{5}$$

where k is the discrete time. K_p and K_i are the gains of proportional and integral actions, respectively. $\boldsymbol{e}(k) = [e_x(k)\ e_y(k)]^T$ is the error vector in image coordinate system measured by

$$\boldsymbol{e}(k) = \boldsymbol{X}_d - \boldsymbol{I}(k), \tag{6}$$

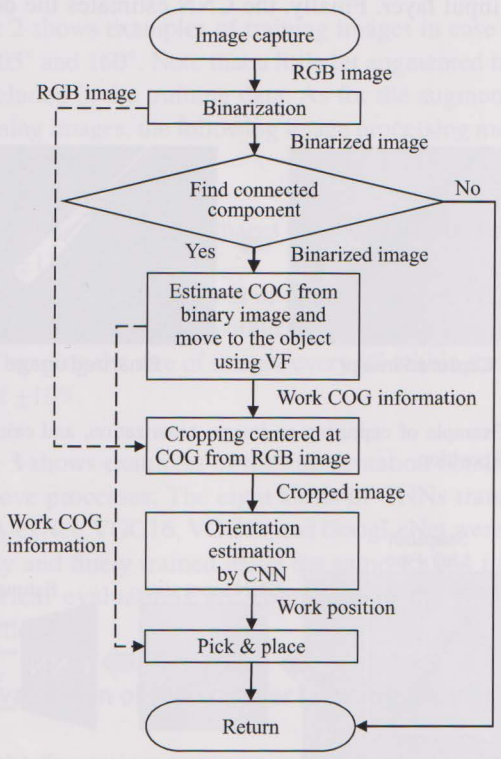


Fig. 8 Flowchart of the robotic pick and place operation, in which the trained CNN and proposed PBVF controller using an endoscope camera are incorporated

where $\boldsymbol{X}_d = (600, 480)$ and $\boldsymbol{I}(k) = [I_x(k), I_y(k)]$ are the desired position (center of image) and the measured object's COG position in image coordinate system, respectively. The PBVF controller controls the gripper position so that the object's COG position can overlap with the center of captured image as shown in the left photo in Fig. 7. In addition, the PBVF controller allows the robot to execute the pick and place task without using Eqs. (3) and (4), so that the measuring of (X_1, Y_1) and (X_2, Y_2) is not needed.

Finally, several pick and place experiments were conducted to confirm the effectiveness of the proposed pick and place robot. Figure 9 shows the experiment scenes, in which four kinds of different shapes of workpieces are successfully picked and placed to the desired position while skillfully gripping the just center of the long axis. Another important advantage of the PBVF controller is that the recalibration between camera and robot coordinate systems is not required for the recovery even in the trouble case such that they become misaligned without noticing.

4 Conclusions

In this paper, transfer learning-based CNN models were developed and implemented to estimate the orientation of objects with a resolution of 5 degrees on a working table for stable picking operation. Actually, the transfer learning-based CNNs for orientation estimation were produced

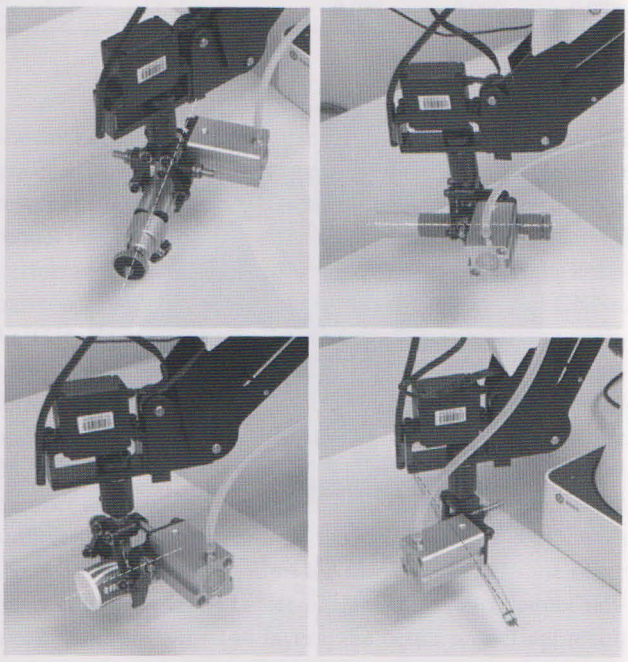


Fig. 9 Pick and place experiments using the proposed CNN and PBVF controller

by finely tuning four existing CNN architectures named AlexNet, VGG16, VGG19, and GoogLeNet. It was confirmed from the classification experiments using test images, the GoogLeNet-based CNN model obtained through all layers' fine-tuning could have the highest recognition accuracy. In addition, to reduce the calibration load between the camera and robot coordinate systems, a simple pixel-based visual feedback controller with PI actions was successfully implemented. The robot could have a higher ability to pick and place operation due to the proposed CNN model and the PBVF controller.

References

1. Kragic D, Christensen HI (2002) Survey on visual servoing for manipulation. In: Computational vision and active perception laboratory technical report, Department of Numerical Analysis and Computing Science, Stockholms University, p 59
2. Taryudi, Wang MS (2017) 3D object pose estimation using stereo vision for object manipulation system. In: Proceedings of 2017 international conference on applied system innovation (ICASI), Sapporo, Japan, 13–17 May 2017, pp 1532–1535
3. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115:211–252
4. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proceedings of advances in neural information processing systems, Lake Tahoe Nevada, USA, 3–6 Dec 2012, pp 1097–1105
5. Haochen L, Bin Z, Xiaoyong S, Yongting Z (2017) CNN-based model for pose detection of industrial PCB. In: Proceedings of international conference on intelligent computation technology and automation (ICICTA), vol 1, Changsha, China, 9–10 Oct 2017, pp 390–393
6. Miki K, Nagata F, Watanabe K (2020) Defective article picking robot in narrow metal mold space using image processing technique. In: Proceedings of the 2020 JSME conference on robotics and mechatronics (ROBOMECH2020), Kanazawa, Japan, 27–30 May 2020, 2P2-B03, p 4 (**in Japanese**)
7. Miki K, Nagata F, Watanabe K, Habib MK (2021) Picking robot of defective molded articles using image processing technique and visual feedback control. In: Proceedings of 26th international symposium on artifical life and robotics (AROB 26th 2021), Oita, Japan, 21–23 Jan 2021, pp 498–502
8. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Proceedings of international conference on learning representations 2015 (ICLR2015), San Diego, CA, USA, 7–9 May 2015, p 14
9. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of conference on computer vision and pattern recognition (CVPR), Boston, MA, USA, 7–12 June 2015, pp 1–9
10. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
11. Nagata F, Miki K, Otuka A, Yoshida K, Watanabe K, Habib MK (2020) Pick and place robot using visual feedback control and transfer learning-based CNN. In: Proceedings of IEEE international conference on mechatronics and automation (ICMA), Beijing, China, 13–16 Oct 2020, pp 850–855
12. Tharwat A (2020) Classification assessment methods. Appl Comput Inf 17(1):168–192