2020 5th IEEE International Conference on Big Data Analytics

Environment Recognition Using Robot Camera

Hiroaki Sumida
Tokushima University
Tokushima City, Japan
e-mail: c501837046@tokushima-u.ac.jp

Fuji Ren
Tokushima University
Tokushima City, Japan
e-mail: ren@is.tokushima-u.ac.jp

Shun Nishide
Tokushima University
Tokushima City, Japan
e-mail: nishide@is.tokushima-u.ac.jp

Xin Kang
Tokushima University
Tokushima City, Japan
e-mail: kang-xin@is.tokushima-u.ac.jp

Abstract— **The purpose of this research is to create an environmental recognition system, specifically on emotion expressions and human states, for a humanoid robot aiming for interpersonal services. Region Convolutional Neural Networks (R-CNN) are often used for detecting objects in the environment. We employ a Mask R-CNN model for detection of emotions and states of a target person from the robot's field of view. The model was trained using various images of a human's body in several emotional states. Experiments were conducted to validate the effectiveness of the model to detect the states of surrounding people from the robot's camera. Although the set of human states assumed in the experiment was limited, the results of the experiments imply the potential of the proposed method to act as a basis of a recognition model for an intelligent humanoid robot for interpersonal services.**

Keywords- Mask R-CNN, human state recognition, image processing, humanoid robot

I. INTRODUCTION

In recent years, the mechanical engineering technology in intelligent robots has been greatly developed among many important fields. This has made it possible for making intelligent robots to serve in people's lives, with reliably designed software [1][2][3].

These robots are called service robots, which have attracted a great deal of attention in the research and industry fields. The service robots are distinctly different from traditional industrial robots which are developed mainly for speed and accuracy of operation, to reduce factory personnel and improve production rate. Indeed, they are designed to perform work closely related to human life, such as entertainment and work substitution, or perform collaborative work with humans. These service robots do not only need functions such as speed and accuracy of movement, but also need to reproduce the appearance and characteristics which are close to humans so that humans do not feel uncomfortable. Therefore, the ability of humanoid robots to interact or communicate with people is considered of high importance for doing these collaborative works.

In this study, we focus on one of the most important technologies that allow humanoid robots to recognize the communication behavior of a conversation target person [4][5][6]. Like humans in communication, the robot is required to continuously estimate the other party's condition, such as smiling or not and different body gestures, in order to make proper reactions. Besides the content of the conversation, these reactions could also greatly influence the quality of an interpersonal service. we adapt the Mask R-CNN detector based on the learning data created from a large number of human images. Using the data, we analyze images in real time while acquiring images from a camera equipped on humanoid robot and conduct an experiment to estimate the human state in the video.

The rest of the paper is composed as follows. In Section 2, related works are introduced to present the direction of this research. In Section 3, the proposed method is described. In Section 4, the results of the experiments are presented with future issues. Conclusions are presented in Section 5.

II. RELATED WORKS

There are a wide variety of researches on object recognition based on image processing in conventional studies. Such studies often aim for recognizing a variety of targets, such as people and objects. Service robots equipped with these models are expected to play important roles in many applications, such as video surveillance and monitoring.

Many methods have been proposed for object recognition such as image processing [7][8], color and distance sensor [9][10], and usage of information such as skeleton and shape [11][12]. Although the methods have shown highly reliable results in recognizing the target, these methods require special equipment, as well as a large amount of data to be input manually, which caused difficulty in dealing with unknown data. In this research, we propose a method that has high scalability and can flexibly deal with a large amount of data by combining image processing technology for video analysis with machine learning technology.

978-1-7281-4111-4/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: Auckland University of Technology. Downloaded on June 01,2020 at 22:31:21 UTC from IEEE Xplore. Restrictions apply.

III. PROPOSED METHOD

A. System Overview

Figure 1 shows the configuration of the proposed human recognition system. In order to recognize the state of the target, various human images in various conditions were collected. These images are annotated for each label to be detected, and a learning model is trained based on these images. The image of the target person is taken using the camera of a humanoid robot. Human features are analyzed from the acquired moving images by the Mask-R-CNN model and the human states and emotions are predicted.

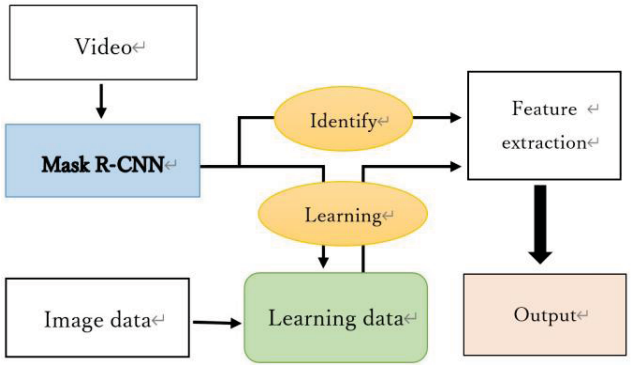


Figure 1. Model of constructed system

B. Robot System

The humanoid robot Actroid, shown in Figure 2 is used as the service robot. The camera is located in the eyeball of the Actroid. The environment was illuminated with natural light source (fluorescent light) in the laboratory, with the windows shaded with blinds to avoid the influence from sunlight. Thus, the intensity of light source is nearly uniform within the room. A chair was placed in front of the robot for the human subject to sit on and show the face to the camera. The schematic image of the scene capturing face images is shown in Fig. 3. The camera height and direction are not specified, since the robot's camera is assumed to be moving that analyzes various images acquired within the range of head movement.



Figure 2. Humanoid robot actroid

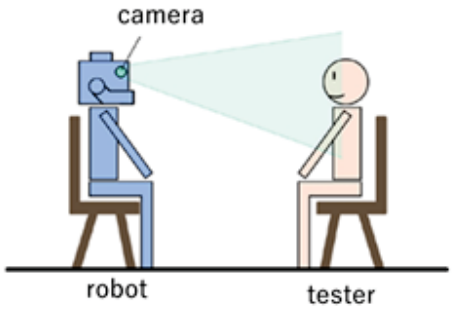


Figure 3. Schematic image of face capturing

C. Mask R-CNN

We use the Mask R-CNN model (shown in Figure 4), one of the Region DNN methods, for object detection in interpersonal communication [13]. Works on application of the model to human targets have shown the model's capability to estimate the state of the human subject [14][15]. Mask R-CNN models are created by developing the convolutional learning models "R-CNN" and "Faster R-CNN" used in many image analyses tasks [16][17]. In particular, the model detects an area in advance as a candidate area for the object region.

Usage of Mask R-CNN for image analysis calculations is much more efficient than traditional methods for real-time communication. In addition to object region detection, Mask R-CNN further detects each pixel class only in previously proposed regions by performing a segmentation process that accelerates class detection for all pixels in the image.

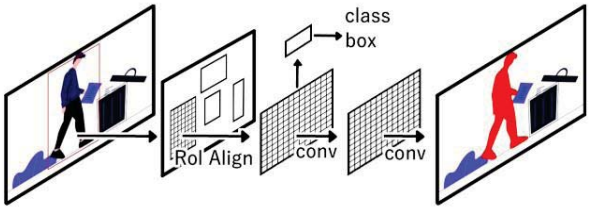


Figure 4. Mask R-CNN overview

The Mask R-CNN model can be roughly divided into three layers: the Backbone layer, the RPN (Region Proporsal Network) layer, and the Head layer. Figure 5 shows the outline of the constructed Mask R-CNN model.

Backbone layers

The Backbone in Mask R-CNN is a network which extracts the features from an input image. The network consists of multiple convolutional layers and generates a feature map by iterating the convolutions.

RPN layers

The RPN layers is designed to select a candidate area for the object in the input image, which accelerates the process of object-like area detection and class identification given the feature map generated by the Backbone layers.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on June 01,2020 at 22:31:21 UTC from IEEE Xplore. Restrictions apply.

Specifically, the output of the Backbone layer defines an area that is the basis for selecting object-like areas for RPN. For each of the areas, the correct data is defined, and information to determine the area as an object as well as its offset are obtained. Learning is performed based on the candidate areas of the objects in the training data.

Head layers

The head layers make further extraction of the region candidates for the target objects based on the RPN outputs and classify them into one of the corresponding categories. To implement the extraction and classification, the head layers consist of four parts as follows.

-RoI Pooling layer: Pooling is performed with 7 * 7 for each candidate area obtained by the RPN layers. The purpose of RoI pooling is to transform region candidates of different sizes into a fixed size tensor.

-Intermediate layer: Each tensor obtained from the RoI Pooling layer is compressed into 1024 units.

-Class identification layer: Predict the class labels with a softmax activation function for all the candidate regions based on the Intermediate layer outputs. Its outputs are a set of probability distributions, each for a candidate region.

-Area extraction layer: All units are combined into a vector of length “class number” * 4. These correspond to the offsets for the RPN region candidates.

IV. EXPERIMENT

A. Experimental Environment

In this experiment, we evaluate the performance of a humanoid robot for recognizing the emotion and state of the target person in human-to-human robot communication. The humanoid robot captures an input video from the camera and makes real time predictions of human emotions based on the human target’s face region and state based on the whole-body posture of the target person.

We train a CNN learning model based on the facial expression parts and the posture parts of human images. These parts are from the facial expression, it is labeled with "normal anger", "sadness" and "smile", and from the posture, the class label is attached to the hawk standing and sitting bag (Table I).

TABLE I. PARAMETER SETTING FOR RECOGNITION

	Facial expression				Posture	
state	normal	angry	sad	smile	standing	sitting

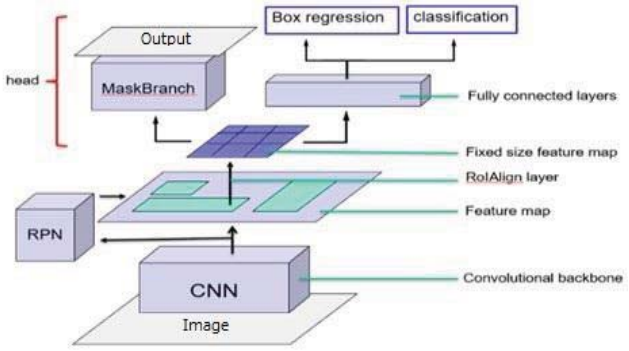


Figure 5. Mask R-CNN Network structure

Each label consists of 800 square images for each classification, 64 x 64 square images, and builds a dataset of 4800 images.

B. Experimental Result and Discussions

We verify the results of recognizing human emotions and states based on the video input of a humanoid robot as described in Chapter 3. The recognized images are shown in Fig. 6 below, and the match rate for each label class in the images are shown in Table II. All face and body parts of the target person are correctly detected through Fig. 6 Output results (a) to Fig. 11 Output results (f). Fig. 6 Output results (a) to Fig. 9 Output results (d), different output results can be confirmed for each facial expression. In addition, the standing state and sitting state of the target person are also correctly recognized in Fig. 10 Output results (e) and Fig. 11 Output results (f), respectively.

By analyzing the results for human recognition, recognizing the four facial expressions and two states of a person, we are confirmed that highly accurate emotion and state labels have been obtained in the output results.

TABLE II. MATCH RATE FOR EACH DETECTED STATE

	Facial expression		Posture	
(a)	normal	0.79	standing	0.90
(b)	angry	0.74	standing	0.79
(c)	sad	0.68	standing	0.92
(d)	smile	0.75	standing	0.87
(e)	normal	0.93	standing	0.89
(f)	normal	0.63	sitting	0.72

Authorized licensed use limited to: Auckland University of Technology. Downloaded on June 01,2020 at 22:31:21 UTC from IEEE Xplore. Restrictions apply.

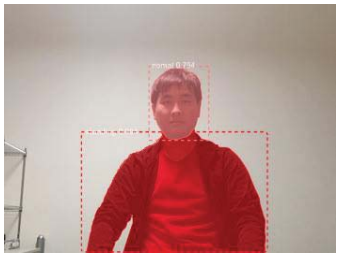


Figure 6. Output results (a)

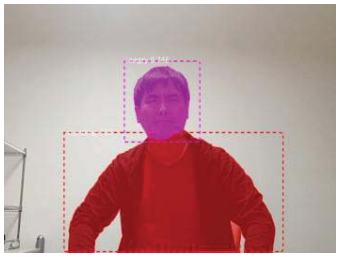


Figure 7. Output results (b)

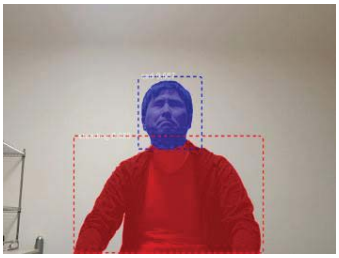


Figure 8. Output results (c)

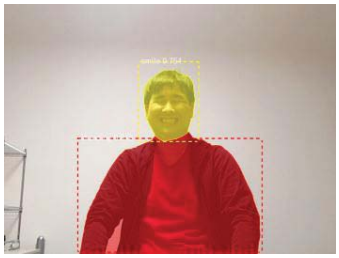


Figure 9. Output results (d)

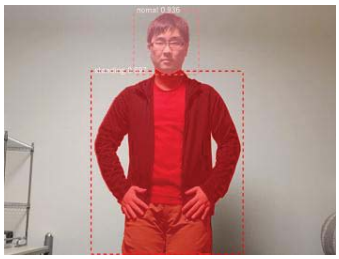


Figure 10. Output results (e)

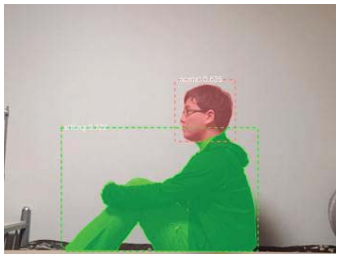


Figure 11. Output results (f)

The final goal of our research is to build humanoid robots which can understand the surrounding environment for interpersonal services. As a preliminary study, our work has revealed the possibility to train a robot with the ability to recognize simple emotions and states of the target human in interpersonal conversation environment.

Our study also suggests a room for further improvements in enriching the facial expressions as well as the body postures for recognition. Evaluation of the method in complex environments such as multiple humans, human positions, non-human-body subjects are also important for a service robot to operate in a practical environment. These environments require more sophisticated recognition models which will be studied in our future work.

V. CONCLUSION

In this study, we proposed an environmental recognition method that recognizes the emotion and state of the target person in an interpersonal conversation environment using the video acquired by the humanoid robot. Specifically, the proposed method analyzes the video acquired from a camera attached to the eyeball of a humanoid robot and recognizes the facial expression and posture of the person in the field of view. The proposed model obtains head and body regions from the target human image based on the Mask R-CNN and classifies them based on the contents of these obtained regions. From the experimental results, it can be seen that in a conversation environment, the humanoid robot can correctly analyze the surrounding people and other environments, classify the emotions and states of the subject in conversation with high accuracy. For future work, we plan to improve the performance of the model in addition to evaluate the method in complex environments for practical application.

ACKNOWLEDGEMENTS

The author thanks Prof. Fuji Ren, his supervisor. Thank you for your great advice, including guidance on the experimental techniques covered in this paper.

We also thank Prof. Shun Nishide and Prof. Xin Kang for their paper corrections and advice.

This work was supported by JSPS KAKENHI Grant Number 19H04215.

REFERENCES

[1] Ricard Sole "Rise of the humanbot" arXiv:1705.05935

Authorized licensed use limited to: Auckland University of Technology. Downloaded on June 01,2020 at 22:31:21 UTC from IEEE Xplore. Restrictions apply.

[2] Giovanni Saponaro, Pedro Vicente, Atabak Dehban, Lorenzo Jamone, Alexandre Bernardino, José Santos-Victor "Learning at the Ends: From Hand to Tool Affordances in Humanoid Robots" arXiv:1804.03022

[3] Ben Goertzel, Julia Mossbridge, Eddie Monroe, David Hanson, Gino Yu "Humanoid Robots as Agents of Human Consciousness Expansion" arXiv:1709.07791

[4] Doreen Jirak, David Biertimpel, Matthias Kerzel, Stefan Wermter "Solving Visual Object Ambiguities when Pointing: An Unsupervised Learning Approach" arXiv:1912.06449

[5] Cheng Zhang, Hedvig Kjellstrom "Multi-Class Detection and Segmentation of Objects in Depth" arXiv:1301.5582

[6] Yoshikatsu Nakajima, Hideo Saito"High accuracy of camera position posture estimation based on feature quantity database selection by deep learning" Research Report Computer Vision and Image Media(CVIM)2016-CVIM-200(33)p1-5

[7] Janne Tommola, Pedram Ghazi, Bishwo Adhikari, Heikki Huttunen "Real Time System for Facial Analysis" arXiv:1809.05474

[8] Junji Satake and Jun Miura"Human Follow-Up Control of Mobile Robots Using Stereo Vision" The Journal of the Robotics Society of Japan vol.28 No.9 pp.1091-1099, 2010

[9] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, Tae-Kyun Kim "First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations" CVPR 2018.

[10] Hayato Fukuda"Interactive Object Recognition for Service Robot with RGB-D Camera" IEICE technical report. MVE, Multimedia and virtual environment basics pp.191-197, 2012

[11] YongChol Sin, MyongSong Choe, GyongIl Ryang "Pan-Tilt Camera and PIR Sensor Fusion Based Moving Object Detection for Mobile Security Robots" arXiv:1510.07390

[12] Nakahara Keita "Household behavior detection method assuming mobile sensors" Multimedia, Distributed Cooperation and Mobile Symposium 2016 Proceedings pp.751 - 762, 2016

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick "Mask R-CNN" The IEEE International Conference on Computer Vision (ICCV), pp. 2961-2969, 2017,

[14] Behzad Hasani, Mohammad H. Mahoor "Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks" The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 30-40, 2017

[15] P. Burkert, F. Trier, M. Afzal, A. Dengel, and M. Liwicki "DeXpression: Deep Convolutional Neural Network for Expression Recognition" CoRR (2015)

[16] Ross Girshick "Fast R-CNN" The IEEE International Conference on Computer Vision (ICCV), pp. 1440-1448, 2015,

[17] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks" Advances in Neural Information Processing Systems 28 NIPS 2015

Authorized licensed use limited to: Auckland University of Technology. Downloaded on June 01,2020 at 22:31:21 UTC from IEEE Xplore. Restrictions apply.