Gesture-based Intention Prediction for Automatic Door Opening using Low-Resolution Thermal Sensors: A U-Net-based Deep Learning Approach

2021 IEEE 3rd Eurasia Conference on IOT, Communication and Engineering (ECICE) | 978-1-6654-4516-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/ECICE52819.2021.9645718

Sheng-Ya Chiu
Department of Management Information Systems
National Chengchi University
Taipei, Taiwan
107306079@nccu.edu.tw

Sheng-Yang Chiu
Institute of Smart Industry and Green Energy
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
syc.c@nycu.edu.tw

Yu-Ju Tu
Department of Management Information Systems
National Chengchi University
Taipei, Taiwan
tuyuju@nccu.edu.tw

Chi-I Hsu*
Department of Information Management
Kainan University
Taoyuan, Taiwan
imchsu@mail.knu.edu.tw

***Abstract*—Personal health consciousness has increased amid pandemics. The implementation of automatic doors could help stop the infection. The need for an intelligent sensor emerges for automatic doors to prevent unneeded open as well as customer privacy concerns. This research proposes a novel automatic door opening mechanism using a low-resolution thermal sensor, based on which a multi-task U-Net structure network is adopted to classify hand-raising gestures. With the aid of segmentation masking, there is 74% reduction of training steps for convergence than that of mere thermal image classification while maintaining similar classification performance. On-site deployment of this approach via constantly collecting door-opening misclassification cases for model improvement will lead to practical success in the near future.**

Keywords—Thermal Sensor, U-Net, Deep Learning with Auxiliary Task, Image Classification

I. INTRODUCTION

The covid-19 virus is transmitted between people through respiratory droplets and contact routes. Surfaces in public places that people frequently touch, including doorknobs and toilet door locks, have a high risk of spreading the virus. Hence, hand hygiene is one of the most vital parts to prevent pathogens [1]. An automatic door opens when it detects the proximity of a person or object, the implementation of this technique can thus reduce the area of potentially contaminated surfaces and keep hands clean.

Usually, the sensor installed on the top of the door controls the opening of the automatic door. Commonly used methods include infrared, ultrasonic, and other wireless sensors. An infrared sensor is simple and low-cost. It detects the thermal radiation emitted by people, and an ultrasonic sensor emits ultrasonic waves and reads the returned waves to scan the environment [2]. Although these methods may detect approaching objects and automatically open the door, one of the shortcomings is that it opens as long as objects or people pass by, even though there is no real entry intention.

Existing improvement solutions include decreasing the sensitivity of sensors and adding a wireless touch (i.e., a push pad) on the door [3]. The former could lead to the delay of open, while the latter makes the push pad another potentially contaminated surface. From this point of view, the need for an intelligent sensor that can understand people's intentions emerges. Furthermore, a sensor that can protect users' privacy rights is an important issue that needs to be considered as well.

To resolve those above issues, we propose a mechanism of gesture-based intention prediction for the automatic door opening using a low-resolution thermal sensor. A low-resolution thermal sensor gets less personal information than a high-resolution thermal sensor and camera. The risk of malicious data collection is thus reduced. The proposed system is designed to learn human segmentation from the low-resolution thermal images collected by the sensor and generate a human contour mask. Through the mask, we can see a precise contour of the person approaching the door and further detect the person's action and distinguish the intention of entering the door or not.

Our research uses Panasonic Grid-Eye infrared array sensor to collect temperature distributions in low-resolution. The pose classification model is based on U-Net [4] in a supervised autoencoder fashion, which utilized an additional reconstruction regularizer to increase generalization performance [5]. Instead of using the reconstruction loss of [5], we attempt to adopt the segmentation loss mechanism due in particular for the low-resolution images as our research subject.

II. RELATED WORKS

A. Intelligent Automatic Door Opening Systems

There are several intelligent door opening systems that adopted different approaches. Sensors such as conventional camera, laser sensor, and Radio Frequency Identification (RFID) were proposed as sensing mechanisms, and the criteria for door opening includes trajectory, age, and authorized access card. Yang *et al*. detects faces by a camera and tracks the trajectory for calculating the probability of a person's entry intention [2]. In their work, conventional cameras and statistical analysis were used. Nishida *et al*. designed an innovative laser sensor to detect vertical, horizontal, and radial directions to trace people's speed and path near the door surrounding [3]. Through the path, the proposed system can predict the entry intention and determine door opening speed by monitoring approaching objects. Further, the system can

*corresponding author

ISBN: 978-1-6654-4516-0

control the opening width of the door along with the number of people in the sensing area. Kiru *et al*.. proposed to control the door with the body measurement for the approaching person. For safety reasons, an unauthorized person such as children should not exit by the door [6]. The paper trained a machine learning model to discern people by body features such as arm length, shoulder length, and so on, to decide whether the person is authorized or not. Amole *et al*. viewed doors as a spreading medium of COVID-19 and proposed a prototype automatic sliding door to curb the infection [7]. Kristyawan *et al*. designed an automatic sliding door that can only open by an authorized RFID card to strengthen building access security and enhance efficiency [6–8]. However, the automatic door system that concerns the hygiene, privacy, and intention issues is not only a novel mechanism but also an imperative facility in the COVID-19 pandemic era.

B. *Image Segmentation and Classification*

The classification model proposed in our research classifies thermal images into open and close categories. It is a U-Net-based model assigned with an auxiliary task of image segmentation to improve generalization performance. Image segmentation is a challenging part of computer vision, which helps computers gain a high-level understanding of images and multimedia. The goal of segmentation is to partition an image into multiple segments. Each segment represents an object or a meaningful set of pixels. Image segmentation is to find the pixel-level relation in an image and conduct clustering [9]. Through the segmentation process, information such as relative position, size, and other associated object information in the image can be retrieved. This would help for applications like autopilot, image search engines, and so on.

U-Net is a fully convolutional network with u-shaped architecture and is one of the most popular models used for image segmentation. The model consists of a contracting path to extract features, and expand the path to obtain precise localization, and mirroring paths that combine high-resolution features from the contracting path with the up-sampled output in the symmetric expanding path layer [4]. Through this architecture, U-Net can learn precise segmentation output.

The inclusion of U-Net is to improve the generalization performance of our proposed classification model. Some previous works have focused on improving the generalization performance, and they have proved the uniformly stable performance improved by using multiple auxiliary tasks for learning standard parameters [5,10]. Our research adopted U-Net to improve generalization performance instead of using the autoencoder mentioned in [5] as the unsupervised regularizer is due to the input thermal image. In order to distinguish humans from other heat sources in thermal images, the image segmentation process is needed instead of image reconstruction.

III. Solution Approach

This section describes the overview of the proposed method. We collected and preprocessed the ground truth data for training. Then we construct the prediction model that is a multi-task classification model based on the U-Net structure, and the main task is to classify the input, while the auxiliary task aims to generate the human masking corresponding to the input thermal image. When the sensor detects a person with a specific pose, the model classifies it as the ‘open’ class. In this paper, we specify a person with one or both hands raised to be the gesture representing ‘open.’

A. *Data Collection and Preprocessing*

The dataset used to train the model correctly predicts the input data class and generates masking consisting of low-resolution thermal images as model input data, while the corresponding human segmentation masking is the ground truth. Each data was classified into ‘open’ or ‘close.’ To our best knowledge, there is no dataset that contains thermal images with human and human segmentation masking data. As shown in Fig. 1, We set up a thermal sensor and camera in the same place to retrieve both types of images. Thermal images were retrieved directly by Panasonic Grid-eye array sensor, with a resolution of 8 x 8. In each thermal image, the absolute temperature in the sensing area was recorded, and the data first go through a temporal median filter, scaling, and interpolation as preprocessing mechanisms.

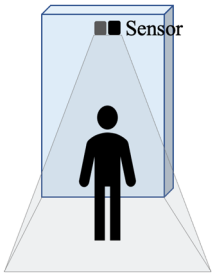

Fig. 1. Sensor setup

A camera first recorded the corresponding human masking, then generated through Mask R-CNN and implemented by the Detectron2 library [11,12]. Mask R-CNN is a flexible framework for instance, segmentation which outperforms all existing single-model, including the COCO 2016 challenge winners, on every task in 2017. This paper captured only the human class segmentation from Mask R-CNN to be our masking ground truth. Furthermore, we assumed three situations in the proposed method: no human in the sensing area, a person without entering intention, and a person with a specific gesture that shows entering intention. Figure 2 introduces the two classes of the dataset used in this paper. Each class has respectively 6,000 images of RGB, thermal, and masking in 64 x 48 pixel. All data is in range [-1, 1] by Min-Max normalization.

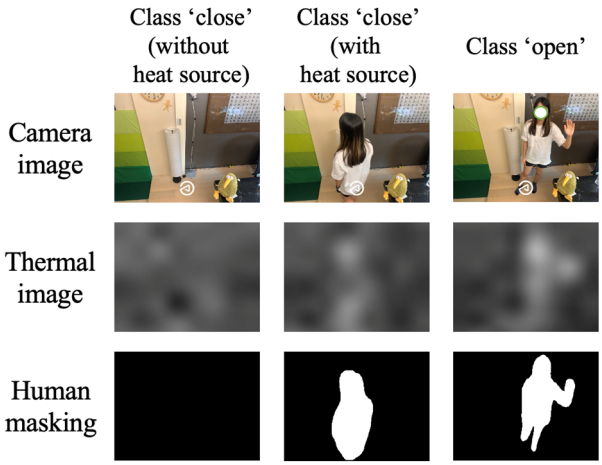

Fig. 2. Dataset

B. *Prediction Model Structure*

This paper uses the U-Net structure into a multi-task learning model for constructing human masking and

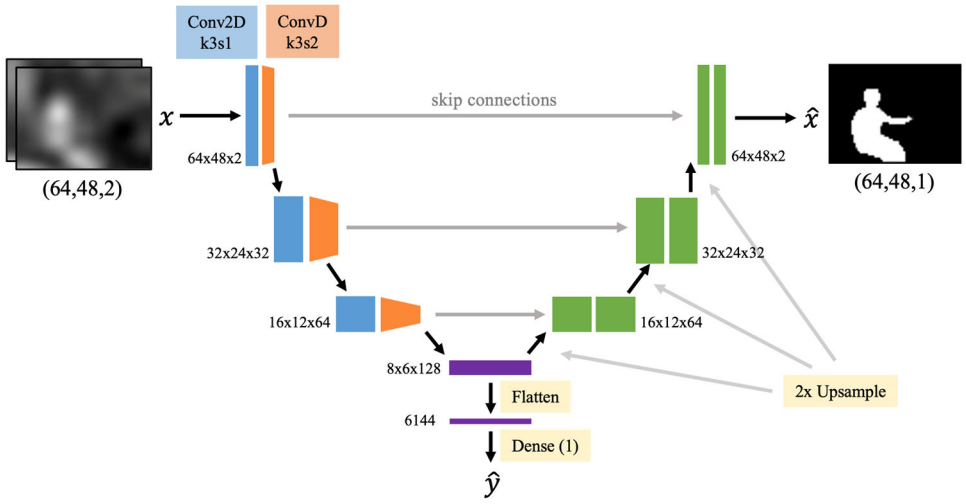

Fig. 3. Model structure.

classifying input. As shown in Fig. 3, the skip-connection layer passes the information from an encoder layer to the corresponding decoder layer, and convolution 2D is used for down-scaling and up-scaling with kernel size 3x3, and LeakyReLU the activation function. Through the encoder-decoder, the model learns how to mask the input data. This is the auxiliary task for improving generalization performance. Moreover, the model will learn the input feature from the down-scaling process. With a flattened layer and a dense layer added after the bottleneck, the model will output the input class. Binary cross-entropy is calculated as the classification loss, and mean square error is used as the loss for constructing human masking.

Figure 4 shows the finite-state machine of our proposed system. As mentioned above, the assumed situations include no heat source in the area (class 'Close'), it is human in the area but without the specific gesture (class 'Close'), and with a human in the area and posing the specific gesture (class 'Open'). The specific gesture set in this paper is raising the arm(s). The proposed system will continue to sense the area. As the class 'Open' appears, the state transfers from 'Close' to 'Open' for 5 seconds as long as the class 'Open' does not appear.

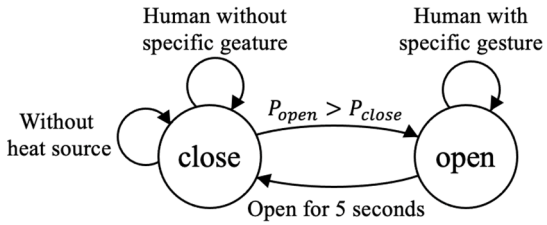

Fig. 4. System finite-state machine

IV. Experiments

A. Settings and Metrics

We partition data with a ratio of 8:2 into train & test and validation set. In the training process, 100 epochs with batch size 128 are set, and early stopping is adopted when the validation loss has not improved over 10 epochs. Adam is used as the optimizer, and the learning rate, betas are set as 0.0005, (0.9, 0.999), respectively. We implement the proposed model in TensorFlow.

This model performs classification and segmentation, with the former as the main task and the latter as the auxiliary task. Hence, in the loss function, we use BCE and MSE to evaluate the performance. The classification loss weight is set with 0 or 1, and the segmentation masking loss weight is set with 0, 0.001, 0.01, 0.1, 1, 10, 100, and 1000. Both weights are mixed as different combinations in the experiment.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (Y_{mi} - \widehat{Y_{mi}})^2, \quad (1)$$

$$BCE = -\frac{1}{n}\sum_{i=1}^{n} Y_{ci} \cdot \log(\widehat{Y_{ci}}) + (1 - Y_{ci}) \cdot \log(1 - \widehat{Y_{ci}}), \quad (2)$$

where each of Y_{mi}, Y_{ci}, n represents the segmentation ground truth, the classification ground truth, and the total number of inputs, and $\widehat{Y_{mi}}, \widehat{Y_{ci}}$ is the predicted segmentation and classification results.

B. Experiment Results

Table 1 shows the prediction performance derived from different combinations of loss weights. The left part is the losses in training while the right part is the validation losses and binary accuracy with the threshold of 0.5. The performance is not able to achieve high binary accuracy in 100 epochs when classification loss weight is set with 0. Other combinations with classification loss set with 1 can achieve high binary accuracy, even with fewer epochs for some of them.

We further conduct a comparative analysis on their convergence performance for those loss weight combinations with high binary accuracy. As shown in Fig. 5, the epochs in the training session vary among the weighted loss combination. Since we have applied early stopping, the effects of the weighted loss function are significant. Compared with the model of the loss set with 0 for the segmentation mask, our proposed approach mixing the loss weights of both classification and segmentation mask can achieve similar classification performance with quick training convergence. With the aid of an auxiliary task, the convergence can have a 74% improvement (from 100 epochs to 26 epochs). This result further illustrates that the auxiliary task mentioned in Ref. [5] is stable in improving the model performance.

TABLE I. MODEL PERFORMANCE

Loss Weight (classification, segmentation)	BCE	MSE	BCE	Binary Accuracy
(0, 1)	0.4510	0.0854	0.6560	0.7320
(1, 0)	0.0000	0.8630	0.5010	1.0000
(1, 0.001)	0.0000	0.0297	0.5010	1.0000
(1, 0.01)	0.0000	0.0407	0.5010	1.0000
(1, 0.1)	0.0000	0.0377	0.5010	1.0000
(1, 1)	0.0000	0.0601	0.5010	1.0000
(1, 10)	0.0001	0.0575	0.5010	1.0000
(1, 100)	0.0001	0.0314	0.5010	1.0000
(1, 1000)	0.0004	0.0398	0.5020	0.9990

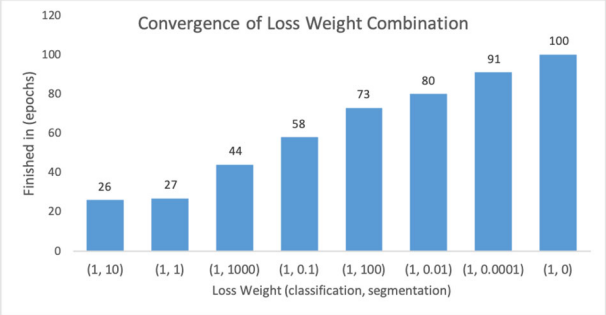

Fig. 5. Convergence performance

V. DISCUSSIONS & CONCLUSIONS

In this paper, we propose a multi-task U-Net structure network to classify hand-raising gestures in the form of thermal images. Through the aiding mechanism of segmentation masking, there is a 74% reduction of training steps than that of the pure main task of thermal image classification. We also found that for the segmentation masking, loss weight set among the range of [1..10] produces the best computation performance. Regarding the classification performance, this approach recognizes the specific posed gesture of a human thermal image. However, in a real-world environment, certain interference surrounding the automatic door could result in model misclassification.

An extended, more comprehensive dataset regarding gestures from multiple people or complex environments could help increase the model robustness. In addition, model optimization with the recent advances in neural architect search (NAS) methods would potentially lead to a better model design, besides the standard U-Net structure with improved efficiency, that might be beneficial in resource-limited edge applications. Finally, deploying the model on the actual site and improving the model performance via constantly collecting door-opening misclassification cases will help achieve practical success.

ACKNOWLEDGMENT

This research is supported by the Ministry of Science and Technology, Taiwan, R.O.C. under the grant numbers MOST 110-2813-C-004-013-H & MOST 110-2635-E-424-001.

REFERENCES

[1] Centers for Disease Control and Prevention (CDC) - Coronavirus Disease 2019 (COVID‐19): FAQ on Hand Hygiene. Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/infection-control/hcp-hand-hygiene-faq.html (Accessed 12th September 2021).

[2] Yang, J. C., Lai, C. L., Sheu, H. T., and Chen, J. J. 2013. An Intelligent Automated Door Control System Based on a Smart Camera. Sensors, vol. 13(5), pp. 5923-5936. https://doi.org/10.3390/s130505923.

[3] Nishida, D., Tsuzura, K., Kudoh, S., Takai, K., Momodori, T., Asada, N., Mori, T., Suehiro, T., and Tomizawa, T. 2014. Development of Intelligent Automatic Door System. 2014 IEEE International Conference on Robotics and Automation (ICRA), pp. 6368-6374, doi: 10.1109/ICRA.2014.6907799.

[4] Ronneberger, O., Fischer, P., and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015, pp. 234--241, isbn=978-3-319-24574-4.

[5] Lei, L., Patterson, A., and White, M. 2018. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. Advances in Neural Information Processing Systems, vol. 31.

[6] Kiru, M. U., Belaton, B., Mohamad, S. M. S., Usman, G. M., and Kazaure, A. A. 2020. Intelligent Automatic Door System based on Supervised Learning. 2020 IEEE Conference on Open Systems (ICOS), pp. 43-47, doi: 10.1109/ICOS50156.2020.9293673.

[7] Amole, A., Oyediran, M. O., Olusanya, O. O., Elegbede, W. A., Olusesi, A. T., & Adeleye, A. O. 2020. Design and implementation of a prototype active infrared sensor controlled automatic sliding door for mitigation of coronavirus disease 2019 (COVID-19). Journal of Electrical, Control and Technological Research, vol. 2, pp. 1-17. https://doi.org/10.37121/jectr.vol2.122.

[8] Kristyawan, Y., and Achmad, R. 2020. An Automatic Sliding Doors Using RFID and Arduino. International Journal of Artificial Intelligence & Robotics (IJAIR), vol. 2(1), pp. 13. doi:10.25139/ijair.v2i1.2706.

[9] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., and Garcia-Rodriguez, J. 2017. A Review on Deep Learning Techniques Applied to Semantic Segmentation. Arvix: 1704.06857.

[10] Liu, T., Tao, D., Song, M., and Maybank, S. J. 2017. Algorithm-Dependent Generalization Bounds for Multi-Task Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 2, pp. 227-241, doi: 10.1109/TPAMI.2016.2544314.

[11] He. K., Gkioxari, G., Doll, P., and G, R. B. 2017. Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980-2988, doi: 10.1109/ICCV.2017.322.

[12] Wu, Y., Kirillov, A., Massa. F., Lo, W. Y., and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.