Proceedings of the ASME 2020 International Symposium on Flexible Automation
ISFA 2020
July 5-9, 2020, Chicago, USA

ISFA-16589

DYNAMIC GESTURE DESIGN AND RECOGNITION FOR HUMAN-ROBOT COLLABORATION WITH CONVOLUTIONAL NEURAL NETWORKS

Haodong Chen*, Wenjin Tao, Ming C. Leu
Department of Mechanical and Aerospace Engineering
Missouri University of Science and Technology
Rolla, MO 65409, USA

Zhaozheng Yin
Department of Biomedical Informatics &
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794, USA

ABSTRACT

Human-robot collaboration (HRC) is a challenging task in modern industry and gesture communication in HRC has attracted much interest. This paper proposes and demonstrates a dynamic gesture recognition system based on Motion History Image (MHI) and Convolutional Neural Networks (CNN). Firstly, ten dynamic gestures are designed for a human worker to communicate with an industrial robot. Secondly, the MHI method is adopted to extract the gesture features from video clips and generate static images of dynamic gestures as inputs to CNN. Finally, a CNN model is constructed for gesture recognition. The experimental results show very promising classification accuracy using this method.

Keywords: Human-robot collaboration, Dynamic gesture recognition, Motion History Image, Convolutional Neural Networks

1. INTRODUCTION

With the development of industrial intelligence, robotic systems are becoming an essential part of factory production. Meanwhile, the concept of human-robot collaboration (HRC) has attracted more and more interest in the industrial field. Literature suggests that in the industry with a high degree of automation, the HRC system can increase human-robot collaboration efficiency and also provide more flexibility in the work environment [1]. In 2012, Shi et al. [2] proposed different degrees of work-sharing. At the lowest level, the robot and the human operator do not have any contact and they work in two different spaces, but without any barriers between them. In 2014, Morato et al. [3] designed a framework to address safety and efficiency during assembly operations involving humans and robots. In 2019, Li et al. [4] proposed a method of human-robot collaboration planning, which considered human fatigue in assigning disassembly tasks to humans and robots.

Ideally, an HRC system should be similar to human-human collaboration in the industry. However, in the application of HRC, space-separation and time-separation of workers and robots result in lower productivity. To improve this situation and realize more efficient collaboration, different communication channels between humans and robots should be established [5]. In the limited communication channels between human workers and industrial robots, gesture recognition has been effectively applied for use as an interface between humans and robots [6]. In 2010, Riek et al. [7] conducted a video-based lab experiment to measure time for a human to cooperate with a robot using gestures. Three gestures (beckon, give, shake hands) were designed in that experiment. In 2015, Chen et al. [8] proposed an approach for recognizing the gestures of a human worker during an assembly task in the HRC. In 2016, Liu et al. [9] established an interactive astronaut-robot system, which applied wearable glove and American Sign Language in the collaboration of the astronaut with a robot co-worker. In 2018, Islam et al. [10] presented a set of robust gestures for a diver to control an underwater robot in collaborative task execution.

*Corresponding author, Email address: h.chen@mst.edu

There are other prior works in gesture communication [11, 12], but most of them focus on static gestures. It is because static gestures mainly rely on the shape and poses of the fingers. But in dynamic gestures, the hand position changes continuously, and the message in a dynamic gesture is contained in the temporal sequence [13]. Therefore, dynamic gestures require more computational complexity than static gestures, and recognition of dynamic gestures is more challenging than static gestures. Besides, the number of hand poses is limited, which means static gestures are limited on semantics. But dynamic gestures can perform better in communication for their various movements [14, 15]. So in this paper, we focus on the design and recognition of a standardized dynamic gesture set.

To collaborate with human workers, robots need to understand human gestures correctly. In this regard, deep learning methods have demonstrated impressive performance in the generalization ability. For example, convolutional neural networks (CNN) have better performance in the action recognition than traditional methods [16–18]. Unlike the common feature extraction, which focuses on specific patterns, the deep learning network is trained to obtain the most discriminative features from given data. In 2018, Du et al. [19] combined the skeletonization algorithm and CNN method to realize the gesture recognition. In 2019, Wu [20] selected hand images and the edge images of a hand to design a double-channel CNN for the hand recognition task.

In this paper, a method of dynamic gesture recognition based on the Motion History Image (MHI) approach and CNN algorithm is proposed and demonstrated. This paper is organized as follows. Section 2 describes ten dynamic sign gestures and the corresponding robotic arm motions in our HRC system. Section 3 combines an image preprocessing algorithm and an MHI method to extract the features of dynamic gestures, and the dynamic gesture videos are converted into static images. Section 4 illustrates the construction of the CNN framework. The experimental setups and results are described in Sections 5. Section 6 provides the conclusion.

2. DESIGN OF DYNAMIC GESTURE SET

For the gestures used in the communication between human workers and robots, they should be easy to sign and remember, socially acceptable, and minimize the cognitive workload. McNeil [21] proposed a classification scheme of gestures with four categories: Iconic (gestures present images of concrete entities and/or actions), Metaphoric (gestures are not limited to depictions of concrete events), Deictic (the prototypical deictic gesture is an extended 'index' finger, but almost any extensible body part or held object can be used), and Beats (gestures use hands to generate time beats). The Metaphoric gestures put abstract ideas into a more literal and concrete form. It is not straight-forward. The Beats gestures are just used to keep the rhythm of speech, and they usually convey no semantic content whatsoever. Therefore, we design gestures that are mainly Iconic or Deictic for the HRC [22].

2.1. Gesture Set

Based on the real cases of human collaboration with a six-degrees-of-freedom (6 DoF) robotic arm (with a gripper as the end effector), we defined some essential commands to communicate with the robot. It consists of a basic set of ten gestures, which are shown in Fig. 1. All the gestures are dynamic gestures. They are more natural than static gestures and can be combined together to generate more commands.

In Fig. 1, the left image of each gesture illustrates the start position: The person stands up with arms straight down. Hands are in a natural pose and pinky fingers are to the back. Then, the person can follow the direction of yellow arrows to carry out the gestures with their hands and arms. The right image of each gesture illustrates the end position. These various gestures can be carried out as follows:

• *Start*: Fully clap in front of the chest.

• *Stop*: Raise the right arm until the hand reaches the shoulder level and extend the arm with the palm facing the front, like a 'stop' sign in the traffic direction gesture.

• *Up*: Extend the right arm straight up with the index finger pointing up.

• *Down*: Bend the left hand and raise its wrist to the chest level. Then extend the left hand straight down with the index finger pointing down.

• *Left*: Swing the left arm straight out and up to the side with the index finger extended until the arm reaches the shoulder level.

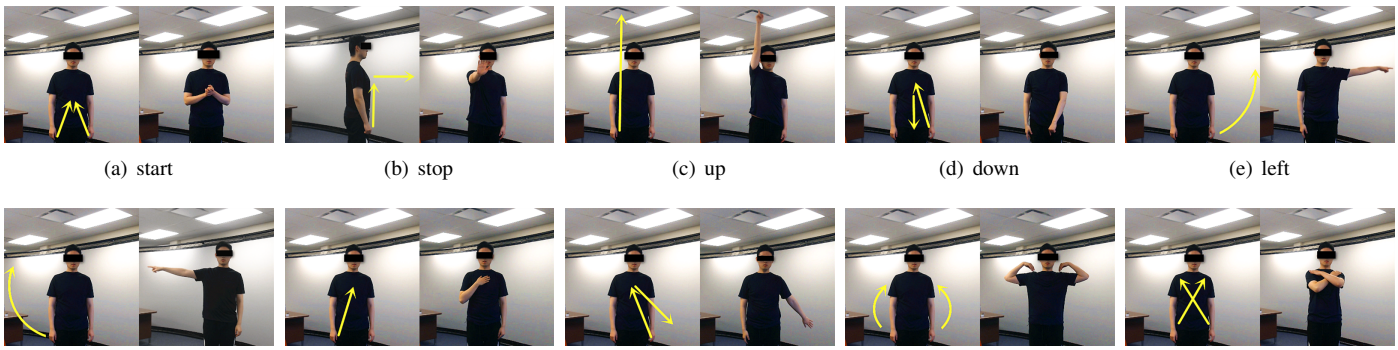
• *Right*: Swing the right arm straight out and up to the side with the index finger extended until the arm reaches the shoulder level.

• *Inward*: Rotate the right forearm up around the right elbow joint with the hand open, until the right hand reaches the chest level and the palm faces back.

• *Outward*: Rotate the left forearm up around the left elbow joint with the left hand open until the hand reaches the chest level. Then rotate the left arm down around the left elbow joint until the arm is straight with about 30 ° from the bodyline and its palm faces back.

• *Open gripper*: Bend each of the two arms up against its shoulder and the elbow until its fingers touch the same side of shoulder and its pinky finger at the front.

• *Close gripper*: Bend the two arms and cross them in front of the chest with the two hands on the different sides of shoulders. The palms face backward and the fingers are open.



(a) start (b) stop (c) up (d) down (e) left

(f) right (g) inward (h) outward (i) open gripper (j) close gripper

FIGURE 1: The ten dynamic gestures.

2.2. Robot Movement

The next step is to consider the corresponding robot movement for each gesture above. The robot is a 6 DoF robot, with six joints and a gripper as the end-effector. The six DoF is needed in order to reach a volume of space from any orientation. The robot is free to change position of its end-effector as forward/backward (surge), up/down (heave), and left/right (sway) translations in three orthogonal (x-y-z) axes, as well as changing in the orientation of the end-effector, through rotation about three perpendicular axes.

The position and orientation changes of the end-effector within the robot's workspace can be realized by converting the end-effector position and orientation changes into the changes in the linear and angular displacements of the six joints. This is an inverse kinematics problem that can be readily solved for most of industrial robots. In order to have more intuitive human-robot interaction, the movement of the robot and the movement of the human will have a mirrored relationship since the human will face the robot in signing the gestures in human-robot collaboration. Thus when the human signs the robot to move right, the robot should move left and vice versa. However, when the human signs the robot to move up, down, inward, or outward, to start or stop, and to open gripper or close gripper, the robot should move accordingly (i.e., no mirror images on these commands).

3. FEATURE ACQUISITION BASED ON MOTION HISTORY IMAGE

The Motion History Image (MHI) approach is adopted to realize the feature extraction of human movements. This approach is a view-based template method that records the temporal history of a movement and converts it into static images [23]. The MHI H_τ (x, y, t) can be obtained from an update function Ψ (x, y, t) using the following formula:

$$H_\tau(x,y,t) = \begin{cases} \tau & \text{if } \Psi(x,y,t) = 1 \\ \max(0, H_\tau(x,y,t-1) - \delta)) & \text{otherwise} \end{cases} \quad (1)$$

where x and y are the image pixel coordinates and t is time. Ψ (x, y, t) represents the movement of an object in the current video frame, the duration τ denotes the temporal extent of a movement, and δ is the decay parameter. This function Ψ (x, y, t) is called for every new video frame analyzed in the sequence. The result of this computation is a scalar-valued image where more recently moving pixels are brighter and vice-versa.

Regarding the parameters in Eq. (1), an MHI with a τ smaller than the number of frames will lose prior motion information. When the value of τ is set too high, the brightness changes (changes of pixel values) in the MHIs will be less clear. So in the generation of MHIs, τ is set as the same as the number of frames in the video clips. While loading the frames, if there is no change (or no presence) of motion in a specific pixel where earlier there was a motion, the value of pixels will be reduced by δ [24]. In the basic MHI method , the dacay parameter δ is replaced by 1 [25]. In our raw gesture videos, there is no extra movements before the target gesture, so the dacay parameter δ is set as 1.

Figs. 2 and 3 demonstrate the generation of the MHIs for the gesture representing the $left$ movement. Generally, an MHI is obtained from binary images of the sequential frames in Fig. 2. Note that only some sample frames are shown in this figure. The binary images are generated using the frame subtraction:

$$\Psi(x,y,t) = \begin{cases} 1 & \text{if } D(x,y,t) \geq \xi \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $\Psi(x,y,t)$ represents the binarized image, and ξ is a threshold. The threshold ξ is used to eliminate the background noise in the MHIs. $D(x,y,t)$ is defined as:

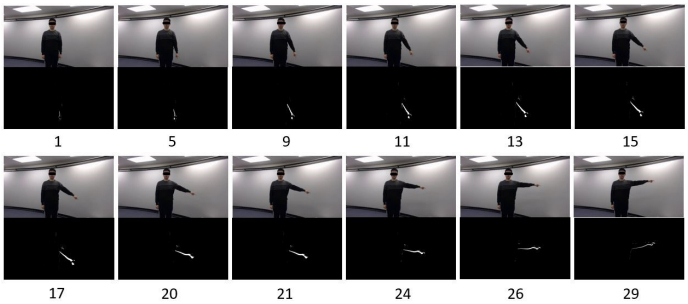


FIGURE 2: Binary images of different frames.

FIGURE 3: The MHI of the $left$ gesture.

$$D(x,y,t) = \mid I(x,y,t) - I(x,y,t-\triangle) \mid \tag{3}$$

where the $I(x,y,t)$ is the intensity value of pixel location with the coordinate (x, y) at the tth frame of the image sequence. $\triangle$ is the temporal difference between two pixels at the same location but at different times [26].

4. CONVOLUTIONAL NEURAL NETWORK MODEL

The overall architecture of our CNN model is shown in Fig. 4. The input images are MHIs. The input MHIs are resized to 32 $\times$ 32 ($width \times height$). The CNN consists of two convolution layers, each of which is followed by the max-pooling layers. The sizes of the convolution kernels, feature maps at each layer, and the pooling operators are all shown in Fig. 4. A $5 \times 5 \times 40$ feature map is obtained after the second pooling. Next, it is flattened as a 1000 feature vector. Then, a fully connected layer with 128 neurons is obtained. The output of this network is a softmax layer, which produces the class-membership probabilities for the 10 gestures. a function that takes as input a vector of K real numbers, and normalizes it into a probability distribution consisting of K probabilities proportional to the exponentials of the input numbers.

In each convolution layer, the Recitified Linear Unit (ReLU) is applied as the activation function [27]. The output of the softmax layer is computed as:

$$P(C \mid x) = \frac{exp(z_C)}{\sum_{C=1}^{10} exp(z_C)} \tag{4}$$

where $P(C \mid x)$ is the predicted probability of being class C for sample x, z_C is the weighted inputs of the softmax layer, and 10 is the number of gestures.

The dropout is carried out after the second pooling layer, which randomly drops units from the neural network during training. It has been proven to be a powerful regularization technique used to avoid overfitting [28].

5. EXPERIMENT AND RESULT

5.1. Datasets and Parameter Selection

The raw dataset includes 10 dynamic gestures signed by 6 human subjects. After data collection, there are about 4570 gesture video samples and each gesture class has about 450$\sim$460 samples. The dataset is splited into training dataset (80%) and testing dataset (20%) randomly.

The threshold ξ in the Eq. (2) is depended by the validation experiment. 1/8 of the training dataset (10% of the whole dataset) are extracted as the test dataset in the validation and the rest 7/8 data (70% of the whole dataset) are used as the training dataset in the validation. In the validation experiments, recognition accuracies of the ten gestures are calculated for each threshold ξ, and the average value of them is obtained as the final accuracy. Table 1 shows the validation results of different threshold ξ. Note that the accuracy here only shows the ability of different classifiers to classify a sample gesture x from a certain class C as class C correctly.

TABLE 1: The metrics of classification evaluation.

Threshold ξ	10	50	90	130	170	210	250
Accuracy	1.000	0.993	0.987	0.909	0.344	0.100	0.100

Based on the validation results in Table 1, the threshold ξ that results in highest accuracy should be 10. Fig. 5 shows the some examples of MHIs for the ten gestures, from which we can observe that MHIs successfully exhibit appearance differences for different dynamic gestures. Fig. 6 shows the gesture $left$ MHIs of six human subjects. All arm pixels during the movements are recorded in MHIs. The brightness of pixels in MHIs is related to its timestamp in the gesture sequence. More recently-moving pixels are brighter and vice-versa.

For deep learning, training data including a large number of samples are necessary to achieve a good performance. To build a powerful image classifier using limited training data, image augmentation is applied to boost the performance of the network model.

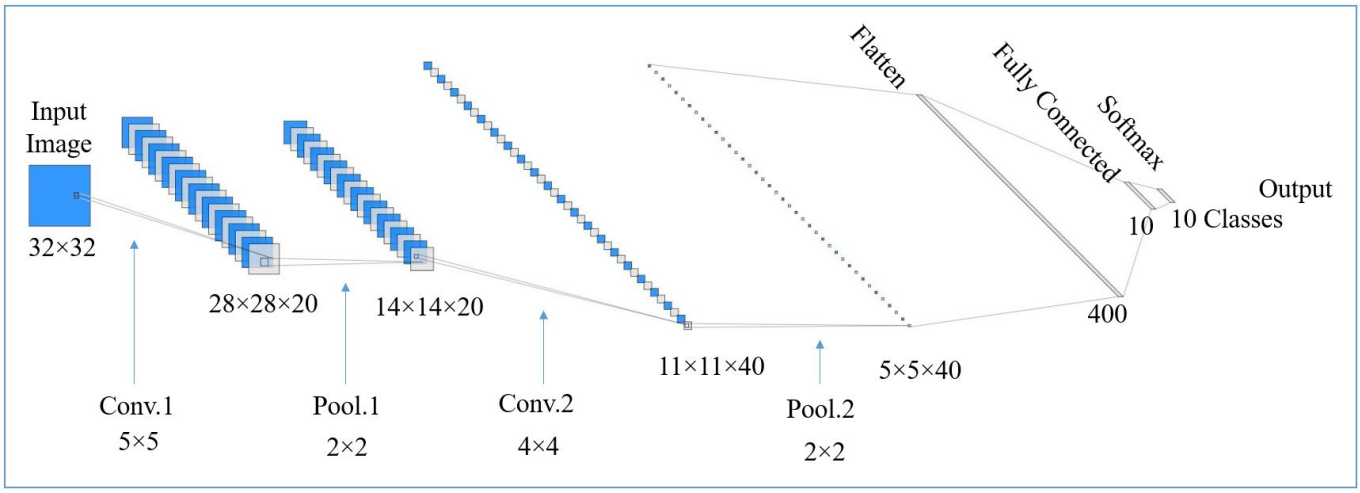


FIGURE 4: The overall architecture of the CNN model. (The 'Conv.' and 'Pool.' denote the operations of convolution and pooling, respectively).

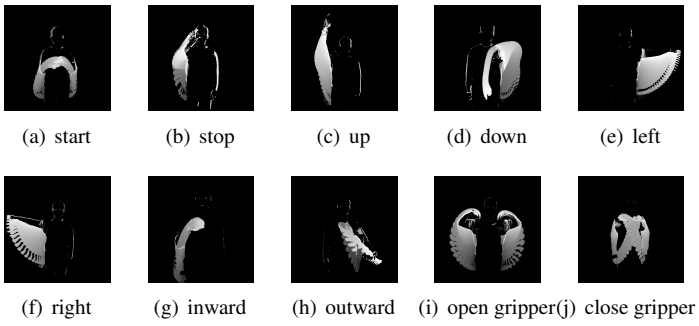


FIGURE 5: The MHIs of the ten dynamic gestures.

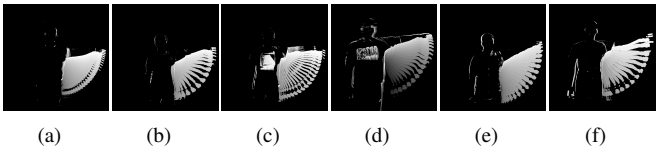


FIGURE 6: The gesture $left$ MHIs of the six human subjects.

Image augmentation artificially increases the variations of images in training data by using flips, rotation, variations in brightness and shifts, etc. [29]. In Fig. 5, it is obvious that the gestures $left$ and *right* are the same movements of the mirrored direction with a different arm. Hence the flip and rotation transformations will reduce the separability of these two images. The flips and rotation are not adopted. The brightness change and horizontal/vertical shifts are finally carried out to enlarge the size of the training dataset to about 10000 images, and there are about 1000 images in every gesture class. Fig. 7 shows some samples of the augmented images. The inputs are the first images in both Fig. 7(a) and 7(b).

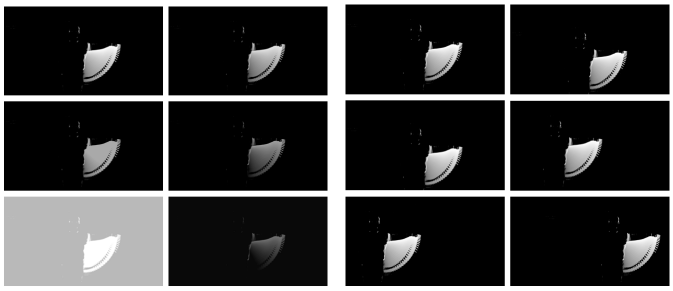


FIGURE 7: Samples of the augmented images.

5.2. Result and Discussion

First, we compute the confusion matrix of our classification, as shown in Fig. 8. The confusion matrix is also known as an error matrix. It realizes visualization of the classification performance. Each column of the matrix represents the instances in a predicted class while each row represents the instances in a ground truth class.

Some commonly-used metrics are adopted to evaluate the classification performance:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{5}$$

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

$$F-score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{8}$$

where in Eq. (5), (6) and (7), the True Positive (TP) describes a sample x from a certain class C that is correctly classified as C. The True Negative (TN) means a sample x from a 'not C' class is correctly classified as the 'not C' class. The False Positive (FP) is defined as a sample x from a 'not C' class is incorrectly classified as C. The False Negative (FN) describes a sample x of class C is misclassified as other 'not C' classes. They are the four basic combinations of actual data category and assigned category in the classification [30, 31].

In Eq. (8), the $F-score$ is a measure that considers both the precision and the recall of the test. It represents the harmonic mean of the precision and recall [32].

Table 2 shows the values of the metrics of the classification results. Taking the *start* gesture as an example, in the first cell, 93 *start* gestures (ground truth label) are predicted as the *start* (predicted label), so the TP is 93. In other diagonal cells in Fig. 8, a total of 822 samples from 'not *start*' gestures are predicted as 'not *start*' gestures, so the TN is 822. In the first column of Fig. 8, there are 1 'not *start*' gestures are predicted as *start* gesture, so the FP is 1. In the first row of Fig. 8, 0 *start* gestures is predicted as other gestures, so the FN is 0.

In the evaluation, the Precision describes the exactness or quality of the method, whereas Recall can be seen as a measure of completeness or quantity. The F-score can provide a more comprehensive measure of a test's performance by using both Precision and Recall. From Fig. 8 and Table 2, it can be seen that most gestures are recognized completely correctly. All the metrics are higher than 97% and the values of Accuracy and Precision are even higher than 99%, which shows how well the trained model could recognize different gestures.

In order to further the performance of our gesture recognition system, some possible remedies are as follows. First, the attention mechanism will be adopted. Spatial attention can find which regions in the image are more important, and temporal attention can find which frames in the temporal sequence are more important. Secondly, the gestures should be captured from multiple views. After that, more features can be obtained for each gesture and a fully view-independent movement recognition can be achieved. Thirdly, we will improve instructions for signing all the gestures and will enlarge the dataset for the extraction of more gesture features. Besides, different network architectures can be investigated and compared to improve the classification result.

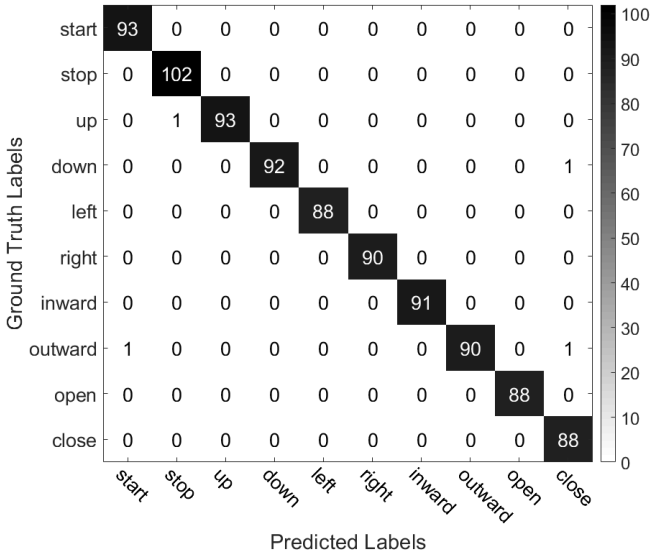


FIGURE 8: The confusion matrix of classification.

TABLE 2: The metrics of classification evaluation.

Classes	TP	TN	FP	FN	Accuracy	Precision	Recall	F-score
start	93	822	1	0	0.999	0.989	1.000	0.995
stop	102	813	1	0	0.999	0.990	1.000	0.995
up	93	822	0	1	0.999	1.000	0.989	0.995
down	92	823	0	1	0.999	1.000	0.989	0.995
left	88	827	0	0	1.000	1.000	1.000	1.000
right	90	825	0	0	1.000	1.000	1.000	1.000
inward	91	824	0	0	1.000	1.000	1.000	1.000
outward	90	825	0	2	0.998	1.000	0.978	0.989
open	88	827	0	0	1.000	1.000	1.000	1.000
close	88	827	2	0	0.998	0.978	1.000	0.989

6. CONCLUSION

In this paper, we design a new set of dynamic gestures for human-robot collaboration and construct a Convolutional Neural Network (CNN) model for dynamic gesture recognition. Ten hand gestures are designed, each represents a different

instruction to the robot. The Motion History Images (MHIs) approach is applied to convert dynamic gestures in video clips into a single image, and an image dataset is established from the video-based dataset. The gesture dataset involves six subjects. The developed CNN model is evaluated on the test dataset and achieves a recognition accuracy of higher than 99%, which shows that our method has very good practicability in classification.

In the future, we will apply our dynamic gesture system to real-time human collaboration with robots and design more dynamic gestures to improve the performance of our HRC system.

ACKNOWLEDGMENT

This research work was supported by the National Science Foundation via CPS Synergy project CMMI-1646162 and National Robotics Initiative project NRI-1830479, and also by the Intelligent Systems Center at Missouri University of Science and Technology. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Wenjin Tao, Ze-Hao Lai, Ming C Leu, and Zhaozheng Yin. Worker activity recognition in smart manufacturing using imu and semg signals with convolutional neural networks. *Procedia Manufacturing*, 26:1159–1166, 2018.

[2] Jane Shi, Glenn Jimmerson, Tom Pearson, and Roland Menassa. Levels of human and robot collaboration for automotive manufacturing. In *Proceedings of the Workshop on Performance Metrics for Intelligent Systems*, pages 95–100. ACM, 2012.

[3] Carlos W Morato, Krishnanand N Kaipa, Jiashun Liu, and Satyandra K Gupta. A framework for hybrid cells that support safe and efficient human-robot collaboration in assembly operations. In *ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 1. American Society of Mechanical Engineers Digital Collection, 2014.

[4] Kai Li, Quan Liu, Wenjun Xu, Jiayi Liu, Zude Zhou, and Hao Feng. Sequence planning considering human fatigue for human-robot collaboration in disassembly. *Procedia CIRP*, 83:95–104, 2019.

[5] Hao-dong Chen, Yi-fan Wang, Zheng Guo, Wen-xiu Chen, and Ping Zhao. A gui software for automatic assembly based on machine vision. In *2018 IEEE International Conference on Mechatronics, Robotics and Automation (ICMRA)*, pages 105–111. IEEE, 2018.

[6] Hongyi Liu and Lihui Wang. Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics*, 68:355–367, 2018.

[7] Laurel D Riek, Tal-Chen Rabinowitch, Paul Bremner, Anthony G Pipe, Mike Fraser, and Peter Robinson. Cooperative gestures: Effective signaling for humanoid robots. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 61–68. IEEE Press, 2010.

[8] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. Effects of robot motion on human-robot collaboration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 51–58. ACM, 2015.

[9] Chenguang Yang, Chao Zeng, Peidong Liang, Zhijun Li, Ruifeng Li, and Chun-Yi Su. Interface design of a physical human–robot interaction system for human impedance adaptive skill transfer. *IEEE Transactions on Automation Science and Engineering*, 15(1):329–340, 2017.

[10] Md Jahidul Islam, Marc Ho, and Junaed Sattar. Understanding human motion and gestures for underwater human–robot collaboration. *Journal of Field Robotics*, 36(5):851–873, 2019.

[11] Wenjin Tao, Ming C Leu, and Zhaozheng Yin. American sign language alphabet recognition using convolutional neural networks with multiview augmentation and inference fusion. *Engineering Applications of Artificial Intelligence*, 76:202–213, 2018.

[12] Giulio Marin, Fabio Dominio, and Pietro Zanuttigh. Hand gesture recognition with jointly calibrated leap motion and depth sensor. *Multimedia Tools and Applications*, 75(22):14991–15015, 2016.

[13] Pramod Kumar Pisharady and Martin Saerbeck. Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding*, 141:152–165, 2015.

[14] Xingyan Li. Gesture recognition based on fuzzy c-means clustering algorithm. *Department of Computer Science. The University of Tennessee Knoxville*, 2003.

[15] Sushmita Mitra and Tinku Acharya. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3):311–324, 2007.

[16] Thomas Wesley Holmes, Kevin Ma, and Amir Pourmorteza. Combination of ct motion simulation and deep convolutional neural networks with transfer learning to recover agatston scores. In *15th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, volume 11072, page 110721Z. International Society for Optics and Photonics, 2019.

[17] Haodong Chen, Zhiqiang Teng, Zheng Guo, and Ping

Zhao. An integrated target acquisition approach and graphical user interface tool for parallel manipulator assembly. *Journal of Computing and Information Science in Engineering*, 20(2), 2020.

[18] Md Al-Amin, Ruwen Qin, Wenjin Tao, and Ming C Leu. Sensor data based models for workforce management in smart manufacturing. In *Proceedings of the 2018 Institute of Industrial and Systems Engineers Annual Conference (IISE 2018)*, 2018.

[19] Du Jiang, Gongfa Li, Ying Sun, Jianyi Kong, and Bo Tao. Gesture recognition based on skeletonization algorithm and cnn with asl database. *Multimedia Tools and Applications*, 78(21):29953–29970, 2019.

[20] Xiao Yan Wu. A hand gesture recognition algorithm based on dc-cnn. *Multimedia Tools and Applications*, pages 1–13, 2019.

[21] David McNeill. *Gesture and thought*. University of Chicago press, 2008.

[22] Judith Holler and Katie Wilkin. Communicating common ground: how mutually shared knowledge influences the representation of semantic information in speech and gesture in a narrative task. *Language and cognitive processes*, 24:267–289, 2009.

[23] Zhaozheng Yin and Robert Collins. Moving object localization in thermal imagery by forward-backward mhi. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 133–133. IEEE, 2006.

[24] Md Atiqur Rahman Ahad, Joo Kooi Tan, Hyoungseop Kim, and Seiji Ishikawa. Motion history image: its variants and applications. *Machine Vision and Applications*, 23(2):255–281, 2012.

[25] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence*, 23(3):257–267, 2001.

[26] Zhaozheng Yin and Robert Collins. Moving object localization in thermal imagery by forward-backward motion history images. In *Augmented Vision Perception in Infrared*, pages 271–291. Springer, 2009.

[27] Hasib Zunair, Aimon Rahman, and Nabeel Mohammed. Estimating severity from ct scans of tuberculosis patients using 3d convolutional nets and slice selection. *CLEF2019 Working Notes*, 2380:9–12, 2019.

[28] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[29] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

[30] Md Al-Amin, Wenjin Tao, David Doell, Ravon Lingard, Zhaozheng Yin, Ming C Leu, and Ruwen Qin. Action recognition in manufacturing assembly using multimodal sensor fusion. In *The 25th International Conference on Production Research (ICPR'19).*, 2019.

[31] Wenjin Tao, Ze-Hao Lai, Ming C Leu, Zhaozheng Yin, and Ruwen Qin. A self-aware and active-guiding training & assistant system for worker-centered intelligent manufacturing. *Manufacturing letters*, 21:45–49, 2019.

[32] Ingrid Visentini, Lauro Snidaro, and Gian Luca Foresti. Diversity-aware classifier ensemble selection via f-score. *Information Fusion*, 28:24–43, 2016.