AN APPLICATION OF CONVOLUTIONAL NEURAL NETWORKS ON HUMAN INTENTION PREDICTION

Lin Zhang[1], Shengchao Li[2], Hao Xiong[2], Xiumin Diao[2] and Ou Ma[1]

[1]Department of Aerospace Engineering and Engineering Mechanics, University of Cincinnati, Cincinnati, Ohio, USA

[2]School of Engineering Technology, Purdue University, West Lafayette, Indiana, USA

ABSTRACT

Due to the rapidly increasing need of human-robot interaction (HRI), more intelligent robots are in demand. However, the vast majority of robots can only follow strict instructions, which seriously restricts their flexibility and versatility. A critical fact that strongly negates the experience of HRI is that robots cannot understand human intentions. This study aims at improving the robotic intelligence by training it to understand human intentions. Different from previous studies that recognizing human intentions from distinctive actions, this paper introduces a method to predict human intentions before a single action is completed. The experiment of throwing a ball towards designated targets are conducted to verify the effectiveness of the method. The proposed deep learning based method proves the feasibility of applying convolutional neural networks (CNN) under a novel circumstance. Experiment results show that the proposed CNN-vote method out competes three traditional machine learning techniques. In current context, the CNN-vote predictor achieves the highest testing accuracy with relatively less data needed.

KEYWORDS

Human-robot Interaction; Intentions Prediction; Convolutional Neural Networks;

1. INTRODUCTION

Robots are getting involved in many aspects of our lives, e.g. medication, education, housekeeping, rescuing, etc.. They are expected to work as collaborators and/or assistants of human beings [1]–[4]. In such context, human-robot interaction (HRI) becomes inevitable. Understanding human intentions is essential for HRI and researchers have reported how HRI tasks can be benefit from estimating human intentions [5]–[7]. Among all the intention indicators, human behaviours or motions are rich in intention information. Indeed, human intentions inference through motions has been introduced to several HRI instances. Kim et al. proposed to use the cycle of action and neural network module to classify 4 categories of human intentions [8]. Song et al. suggested using a probabilistic-based method recognize 4 classes of human intentions based on their hand movement. They even provided optimized strategies for a robot

that was involved in the same task [9]. In such studies one type of human motion can only result in one intention. Thus the human intentions can be easily distinguished from recognizing their body movement. In other scenarios, various intentions are hidden behind a single motion type, therefore are more difficult to be estimated. Vasquez et al. proposed a hidden Markov models (HMMs) to predict human intentions during walking [10]. Ziebart et al. employed an optimal control approach to predict pointing intentions of computer users [11]. Under some extreme circumstances, human intention needs to be predict ahead of the time. Wang et al. introduced a

DOI: 10.5121/ijaia.2019.10501

table tennis robot with probabilistic algorithms, which was able to predict the targets of the ball according to human player's hits [12]. In this research, predicting a table tennis player's intention requires multiple high-speed, high-resolution cameras and a fast-response robotic manipulator. Unfortunately, these apparatuses are not budget friendly and hence the experiment cannot be easily replicated. In recent years, technologies of deep learning are drastically developing. Among which, convolutional neural networks (CNN) is showing dominating power in image pattern recognition tasks [13]. CNN is also played important roles in recognizing human actions [14], [15] and human-robot interaction tasks [16], [17]. Whereas, most previous studies equalize recognizing human intentions to distinguish human actions, limited studies were able to find out subtle changes in body motion that lead to a different intention.

In this research, we propose a CNN based solution: CNN-vote, which predicts human intentions behind a single action. In our previous research, we have demonstrated that binary intentions behind a human action can be predicted by machine learning (ML) algorithms [18]. However, when human intentions expanded, traditional ML experienced difficulties on such task. With CNN's powerful pattern recognition capability, we hypothesize that CNN-vote can predict the one out of nine human intentions from the motion of tossing a ball. We propose to train our CNN-vote predictor using RGB images obtained from a RGB-D camera and labeling the images with human pitchers' intended targets. The performance of CNN-vote human intention predictor and the comparison with ML algorithms will be introduced later.

The rest of this article is arranged as follows. In section 2, the research methods are introduced, including the experiment details, principle of CNN-vote and mechanism of three involved ML algorithms. The experiment results are revealed in section 3. Some interesting and inspiring facts are discussed in section 4. In Appendix, more experiment results are demonstrated for the sake of comparison and analysis.

2. METHOD

2.1. A Thought Experiment

Let us imagine a pitch-and-block game in a 2D plane as shown in Figure 1. A human pitcher throws a ball towards an unlimited-length rod hinged by a revolute joint in the front. The goal of the hinged rod is preventing the ball passing the horizontal line. We have following assumptions: 1) the ball and the rod are both moving in the same 2D plane; 2) the ball is moving at a constant speed of v and the rod is driven by a constant angular velocity, ω; 3) the initial distance between the ball and the horizontal line is d; 4) the angle between the ball's movement direction and the vertical line is $\alpha \in (-\frac{\pi}{2}, \frac{\pi}{2})$ the angle between the rod's initial position and the vertical line is $\theta = 0$; 5) the rod started to rotate when distance between the ball and the horizontal line is $d_f \in [0, d]$.

The condition of a successful block is revealed in (1). When v and α is predetermined, an appropriate d_f should be decided to satisfy this equation so that we guarantee a successful block. When v is increased, an unchanged d_f may not guarantee (1) is satisfied. In this case we will have

to increase d_f to maintain (1), which means the faster the ball is thrown the earlier the rod needs to be moved. In other words, we need predict the ball pitcher's intention of throwing the ball either to his/her left or right before minimal d_f is reached.

$$\frac{d_f}{v\cos\alpha} \geq \frac{\pi}{2\omega} \tag{1}$$

In our previous research [18], we successfully predicted binary human intentions with three ML algorithm (KNN, SVM, MLP). However, as long as the human intentions increased, traditional ML algorithm based predictors were experiencing difficulties.

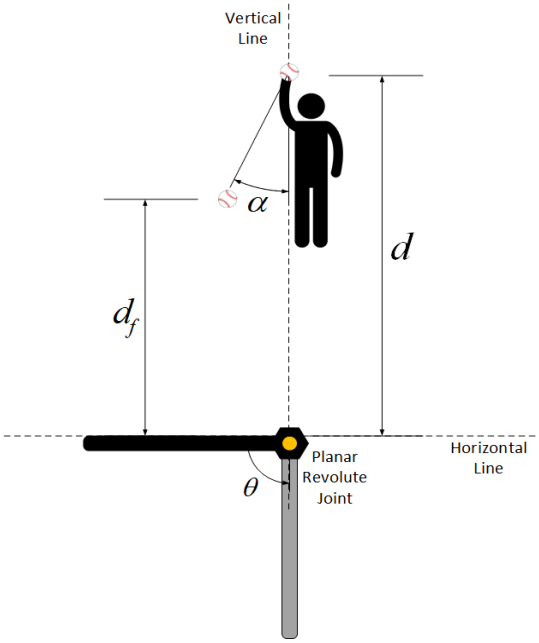


Figure 1 Thought experiment: game of pitch-and-block

2.2. CNN-vote human intention predictor

We would like to introduce a convolutional neural network (CNN) based human intention predictor. In this research 9 intentions are assigned to the action of pitching a ball. The 9 intentions are associated with 9 targets in front of the pitcher. Before each pitch, the human pitcher announces the target he/she intend. The images of the pitcher's motion are recorded and labelled with his/her intention. The CNN-vote human intention predictor can be trained with the labelled images. The data collecting and predictor training procedure can be viewed in Figure 2. The CNN- vote predictor is composed by two parts: AlexNet and voter. The AlexNet is responsible for extracting features from the images according to the specific intention [19]. Since an intention can be predicted by several images, the voter votes out the most popular intention from the involved images.

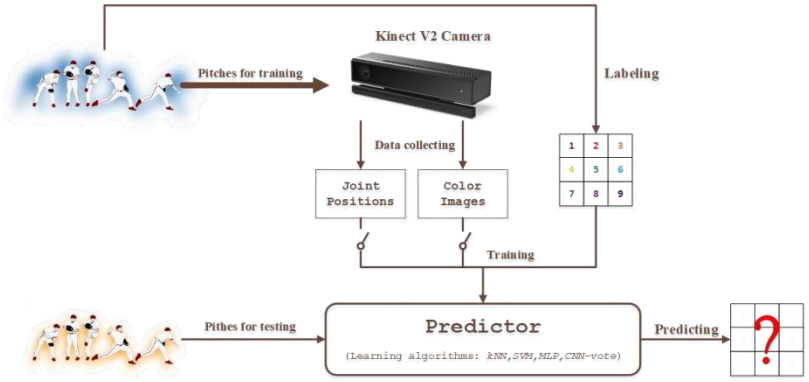


Figure 2 Workflow of using learning based methods to predict human pitchers intention

2.2.1 AlexNet Structure

The architecture of CNN-vote predictor is illustrated in Figure 3 which reveals the role of AlexNet. AlexNet takes in one image at a time and output the predicted human intention. The overall network contains eight layers. The first five are convolutional layers and the last three are fully- connected dense layers. The input of this network is a color image with a dimension of 224 × 224 × 3 pixels. The output of the last dense layer is fed to a 9-way softmax function which produces a distribution over 9 intention labels. The first layer filters the 224 × 224 × 3 input image with 96 kernels of size 11 × 11 × 3 with stride of 4. The output of the first convolutional layer is response-normalized and overlapping pooled. The second convolutional layer takes the output of the first convolutional layer as the input and is filtered by 256 kernels of size 5 × 5 × 96 with stride of 1. The output of the second convolutional layer is also response-normalized and overlapping-pooled with the same setting as the first layer. The third convolutional layer is filtered by 384 kernels of size 3 × 3 × 256 with stride of 1. The fourth convolutional layer is filtered by 384 kernels of size 3 × 3 × 384 with stride of 1. The fifth convolutional layer is filetered by 256 kernels of size 3 × 3 × 384 with stride of 1. There is no intervening normalization or pooling operation after either the third, fourth or fifth convolutional layer. The first and second fully- connected layers have 4096 neurons each. The last layer has 9 neurons which is equal to the number of possible human intentions. A total number of 58318217 parameters is included in our CNN-vote predictor.

2.2.2 Voter

For each pitching trial, N frames of color images, $\{\vec{x}_1,\ldots, \vec{x}_i, \ldots, \vec{x}_N\}$ are served as input data to the AlexNet frame by frame. These frames can be classified by AlexNet as $\{y_1, \ldots, y_i, \ldots, y_N\}$, where y_i is the classified label of ith frame and $y_i \in \{1,2,3,4,5,6,7,8,9\}$. Every single number indicates a corresponding target, and the targets layout was shown in Figure 3. Due to the low threshold of body movement detection, the initiation of the pitching can be detected early in the whole trial. Thus, a heuristic discount weight, γ is arranged to all classification results, and the prediction of the trial can be drawn as,

$$p = \underset{y_i}{\operatorname{argmax}}\left(\sum_{i=1}^{N} \gamma^{N-i}\, \operatorname{onehot}(y_i)\right) \tag{2}$$

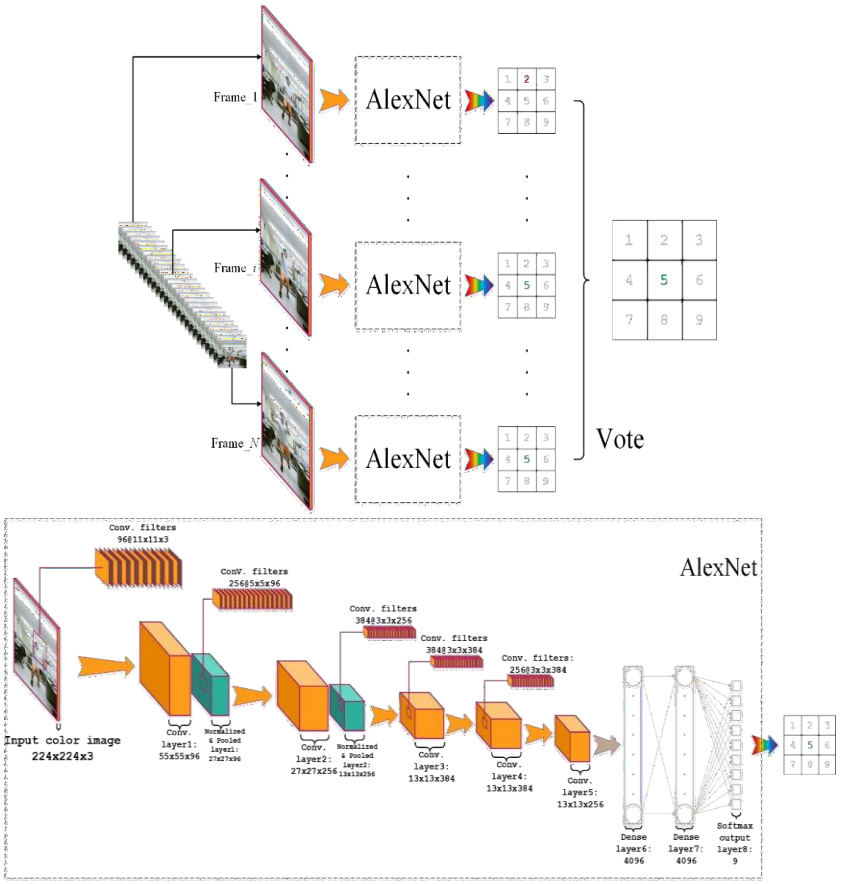


Figure 3 Architecture of CNN-vote human intention predictor

2.3. Experiment

2.3.1. Experiment settings

The whole experiment includes a human participant as the pitcher, a Microsoft Kinect V2 RGB-D camera as the sensor, a whiteboard as the target, a computer as the data receiver and processor. The experiment configuration is shown in Figure 4. A reference coordinate system is set up and the origin of it is fixed to the color camera of the Kinect V2 sensor. The y axis is pointing upward vertically, and z axis is pointing outward horizontally, thus x axis can be determined by the right- hand rule. The dimension of the targets area on the whiteboard is 1.2 meters in width, 1.2 meters in height, with its center located at (0, 0.7, -0.7) meters. Each target is a square patch with a side length of 0.4 meters. Only one human participant is involved in all the experiments, who simultaneously plays as a ball pitcher and the target recorder. This is acceptable because the method is about to train a robot to understand one person's intention. For a different person, it has to be re-trained. The sensor is placed 0.8 meters above the floor, and the HP stands 4.2 meters in front of the sensor along the positive z axis. A house-made Matlab program is coded to retrieve data from the pitching trials and the obtained data is pre-processed.

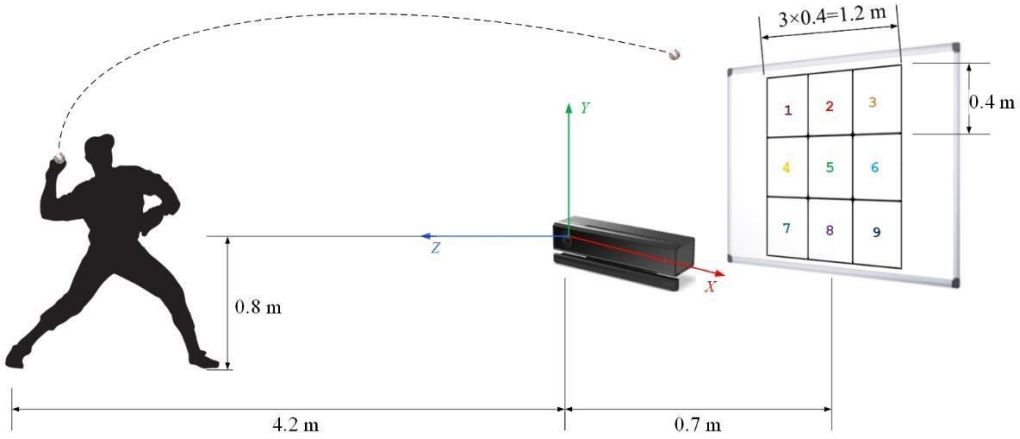


Figure 4 Experiment configuration

2.3.2. Experiment protocols

To warm up, the pitcher practices throwing ball to the target he intends to several times before the recorded trials. The pitcher has to secure at least one successful pitching for every target. The pitcher also practices for his standing pose for the phase before and after the pitching. In each recorded trial,

1. Computer picks target randomly and inform the target to the pitcher before the trial started.
2. The pitcher listens to the auditory cue and initiate throwing. The pitcher has to stand still before initiating the throwing. The initial pose is not mandatory, hence can be determined at pitcher's preference.
3. The pitcher throws the ball toward the pre-noticed target within 5 seconds. The pitcher is not allowed to perform any irrelevant action during and after throwing. We recommend the pitcher to return to the initial pose.
4. Unless the pitcher or computer operator was not satisfied with the trial, recorded data will be saved. Repeat step 1 through 4 until the data collection is finished.

In some trials, the pitcher throws the ball into the target that was not intended. We save both the intended target and actual hit target. However, in this research, we only investigate the trials that intended target match the actual hit target. Due to the large amount of data is required, the pitcher is allowed to separate experiment into subsets because of fatigue.

2.4. Data Processing

A total of 292 trials of pitching were recorded, every trial recorded data in 5 seconds. Kinect V2 camera works under the frequency of 30 Hz, hence each trial contains 150 data points. A data point is either a 75-dimentional (25 × 3) vector when it is in the form of joint positions, or it is a tensor with shape of 224 × 224 × 3 when it is in the form of color image. The joint position data represents the 3-D coordinates of 25 joints on the skeleton tracking model [20]. For every data point, one in nine possible labels is arranged according to the pitching result of the trial.

The whole dataset is divided into two groups, 256 trials were used for training and the rest 36 trials were served for evaluation test. In the training dataset, intention #1 has 28 trials; intention #2 has 30 trials; intention #3 has 26 trials; intention #4 has 31 trials; intention #5 has 28 trials;

intention #6 has 27 trials; intention #7 has 27 trials; intention #8 has 31 trials, and intention #9 has 28 trials. In the test dataset, each intention has 4 trials.

Every pitch generates 5 seconds that includes150 frames of data, but large amount of which are irrelevant and useless (before and post pitching). Hence, we need to extract effective data from all 150 frames. We proposed a heuristic way to do so. First, we need to detect pitching initiation in every trial according to the change of the skeleton model. Suppose a pitcher's joint positions vector at frame t (where $0 \leq t \leq 149$) is $\vec{p}_t$, then $\vec{p}_t$ stands for the initial joint positions. By calculating L-2 distance between every $\vec{p}_t$ and $\vec{p}_0$, we can track how pitchers body move against his initial pose and obtained a body movement curve as shown in Figure 5. Scan this curve with 20 frames sized window, if the curve is increasing monotonically and all the values are greater than a threshold=1 in the window, then the first frame of the window will be defined as the cutting- started frame. If we cut out the data from the cutting-started frame to a certain frame (up to the 40th) after the cutting-started frame to form eight datasets based on subsets of these 40 frames of data. These datasets were using first 5 frames; first 10 frames; first 15 frames; first 20 frames; first 25 frames; first 30 frames; first 35 frames, and all 40 frames of data. A good predictor was expected to predict the intention of the pitching trial with higher accuracy with less frames of data recruited. The kNN and SVM predictors are trained with a Python module of scikit-learn [21], MLP and CNN-vote predictor is trained with the assistance of a Python library: tensorflow [22].

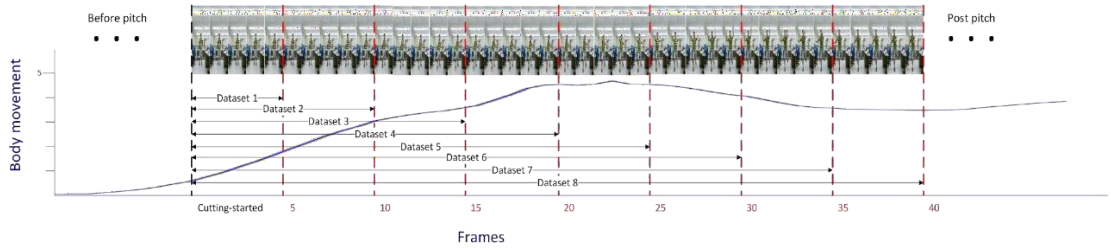


Figure 5 Pitch initiation detection and dataset segmentation

3. EXPERIMENTAL RESULTS

We trained four types of predictors using two types of data on eight sub-segmented datasets. Figure 6 illustrated the performance of all kinds of predictors with respect to different data formats and various segmented datasets. Because The CNN-vote predictor announced a 0.7222 testing accuracy trained with 35 frames of color image data. Despite this, CNN-vote won all the competitions against other predictors no matter what kind of data they relied on. Moreover, CNN- vote even claimed a 0.5833 test accuracy with only 5 frames of color images which was surprising and against the trend of more accurate predictor needs more data fed. The kNN, SVM and MLP predictors showed better performance when fed with joint data than color image data. The SVM predictor achieved over 0.5 of accuracy when 30 or more frames of joint data were fed.

We can take a closer look at the CNN-vote predictor by Figure 7(a). Higher training accuracies indicated that overfitting issue existed. In fact, by looking at all other predictors in Appendix, overfitting was a global issue in this research. Fortunately, our later work claimed that data augmentation depressed overfitting problem while slightly boosted performance of CNN-vote predictor [23]. In Figure 7(b), a confusion matrix of CNN-vote predictions on test dataset is illustrated. We found that even the mis-predicted intentions were adjacent to the true intention. However, taking a peek at the appendix, the second best predictor, SVM predictor had some mis- predictions two blocks away from the true intentions.

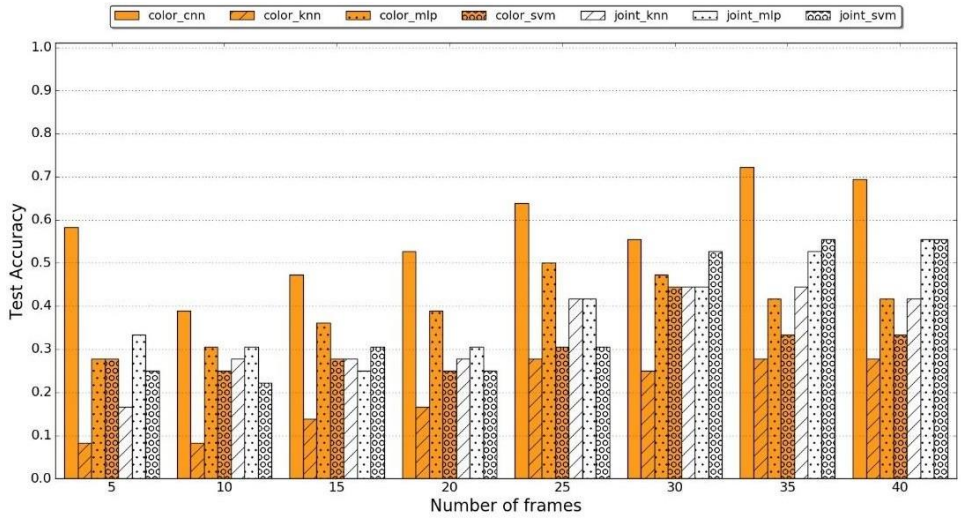
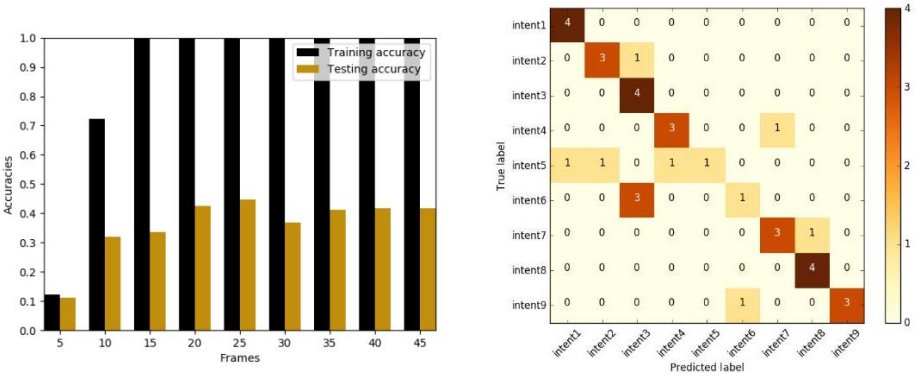


Figure 6 Performance of all types of predictors using different input data format, trained and tested on different datasets. Orange bars indicate predictors trained with color image data, white bars indicate predictors trained with joint data.



(a)CNN-vote: training accuracies vs. testing accuracies (b) CNN-vote: confusion matrix

Figure 7 Performance of CNN-vote human intention predictor.

4. CONCLUSION

This research proposed a novel application of convolutional neural networks, CNN-vote, which is able to predict a human pitcher's intention of throwing a ball to a specific target among various targets. The CNN-vote predictor outcompetes other traditional ML predictors and achieves the highest test accuracy with relatively less information fed. Albeit CNNs are frequently applied to classify images or videos containing various human activities, this study cuts in from a new angle that introduces the CNNs to further understand a single human action by subdividing it to learn the subtle changes of patterns behind the action.

As Alexnet only investigating spatial features in images, temporal features of the whole action should be considered in the future. Deep neural networks with CNNs combined with recurrent neural networks (RNNs) has been reported with some encouraging results [24]. We have not introduced the RNNs into this study at the current stage due to the limitation of the computational resource. However, a deeper predictor of CNN working together with RNN has been already included in our blueprint. As more advanced deep learning architectures, such as VGG-19 [25], Inception-V4 [26], etc. were invented after AlexNet was first published in 2012,

we are definitely looking forward to upgrade the Alexnet into a more advanced level. In addition, overfitting

Problem requires more data to be collected, data augmentation and further tuning of the predictors. Besides human intentions, human activities are also worth to be predicted [27]–[29]. We also expect the framework of our human intention predictor to be applied to such fields. Although our current techniques of action detection are rudimentary, we are making progress of building a more intelligent robot and we are definitely going to keep improving the techniques in the next stages of research.

REFERENCES

[1] J. Forlizzi and C. DiSalvo, "Service robots in the domestic environment," in Proceeding of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction - HRI '06, 2006, p. 258.

[2] J. Bodner, H. Wykypiel, G. Wetscher, and T. Schmid, "First experiences with the da VinciTM operating robot in thoracic surgery☆," Eur. J. Cardio-Thoracic Surg., vol. 25, no. 5, pp. 844–851, May 2004.

[3] M. J. Micire, "Evolution and field performance of a rescue robot," J. F. Robot., vol. 25, no. 1–2, pp. 17–30, Jan. 2008.

[4] F. Mondada et al., "The e-puck , a Robot Designed for Education in Engineering," in Robotics, 2009, vol. 1, no. 1, pp. 59–65.

[5] K. Wakita, J. Huang, P. Di, K. Sekiyama, and T. Fukuda, "Human-Walking-Intention-Based Motion Control of an Omnidirectional-Type Cane Robot," IEEE/ASME Trans. Mechatronics, vol. 18, no. 1, pp. 285–296, Feb. 2013.

[6] K. Sakita, K. Ogawam, S. Murakami, K. Kawamura, and K. Ikeuchi, "Flexible cooperation between human and robot by interpreting human intention from gaze information," in 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566), 2004, vol. 1, pp. 846–851.

[7] Z. Wang, A. Peer, and M. Buss, "An HMM approach to realistic haptic human-robot interaction," in World Haptics 2009 - Third Joint EuroHaptics conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2009, pp. 374–379.

[8] S. Kim, Z. Yu, J. Kim, A. Ojha, and M. Lee, "Human-Robot Interaction Using Intention Recognition," in Proceedings of the 3rd International Conference on Human-Agent Interaction, 2015, pp. 299–302.

[9] D. Song et al., "Predicting human intention in visual observations of hand/object interactions," in 2013 IEEE International Conference on Robotics and Automation, 2013, pp. 1608–1615.

[10] D. Vasquez, T. Fraichard, O. Aycard, and C. Laugier, "Intentional motion on-line learning and prediction," Mach. Vis. Appl., vol. 19, no. 5–6, pp. 411–425, Oct. 2008.

[11] B. Ziebart, A. Dey, and J. A. Bagnell, "Probabilistic pointing target prediction via inverse optimal control," in Proceedings of the 2012 ACM international conference on Intelligent User Interfaces - IUI '12, 2012, p. 1.

[12] Z. Wang et al., "Probabilistic movement modeling for intention inference in human–robot interaction," Int. J. Rob. Res., vol. 32, no. 7, pp. 841–858, 2013.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, p. 436, 2015.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

[15] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-Stream 3-D convNet Fusion for Action Recognition in Videos With Arbitrary Size and Length," IEEE Trans. Multimed., vol. 20, no. 3, pp. 634–644, Mar. 2018.

[16] P. Barros, C. Weber, and S. Wermter, "Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction," in 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), 2015, pp. 582–587.

[17] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, and H. Ishiguro, "Show, attend and interact: Perceivable human-robot social interaction through neural attention Q-network," in 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017, pp. 1639–1645.

[18] L. Zhang, X. Diao, and O. Ma, "A Preliminary Study on a Robot's Prediction of Human Intention," 7th Annu. IEEE Int. Conf. CYBER Technol. Autom. Control. Intell. Syst., pp. 1446– 1450, 2017.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Adv. Neural Inf. Process. Syst. (NIPS 2012), p. 4, 2012.

[20] J. Shotton et al., "Real-time human pose recognition in parts from single depth images," in CVPR 2011, 2011, vol. 411, pp. 1297–1304.

[21] F. Pedregosa et al., "Scikit-learn: Machine Learning in Pythons," J. Mach. Learn. Res., vol. 12, no. 6, pp. 2825–2830, May 2011.

[22] M. Abadi et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," 2016.

[23] S. Li, L. Zhang, and X. Diao, "Improving Human Intention Prediction Using Data Augmentation," in HRI 2018 WORKSHOP ON SOCIAL HUMAN-ROBOT INTERACTION OF HUMAN-CARE SERVICE ROBOTS, 2018.

[24] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634.

[25] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Inf. Softw. Technol., vol. 51, no. 4, pp. 769–784, Sep. 2014.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," Pattern Recognit. Lett., vol. 42, pp. 11–24, Feb. 2016.

[27] P. Munya, C. A. Ntuen, E. H. Park, and J. H. Kim, "A BAYESIAN ABDUCTION MODEL FOR EXTRACTING THE MOST PROBABLE EVIDENCE TO SUPPORT SENSEMAKING," Int. J. Artif. Intell. Appl., vol. 6, no. 1, p. 1, 2015.

[28] J. A. Morales and D. Akopian, "Human activity tracking by mobile phones through hebbian learning," Int. J. Artif. Intell. Appl., vol. 7, no. 6, pp. 1–16, 2016.

[29] C. Lee and M. Jung, "PREDICTING MOVIE SUCCESS FROM SEARCH QUERY USING SUPPORT VECTOR REGRESSION METHOD," Int. J. Artif. Intell. Appl. (IJAIA)., vol. 7, no. 1, 2016.

AUTHORS

Lin Zhang as Sr. Research Associate is currently working in Intelligent Robotics and Autonomous System Lab at University of Cincinnati. His major research interest is reinforcement learning and its application in robotics.

Shengchao Li as Ph.D student is doing research in DeePURobotics Lab at Purdue University. His major research interest is deep neural networks and image processing.

Hao Xiong as Ph.D candidate is doing research in DeePURobotics Lab at Purdue University. His major research interest is dynamics and control of cable-driven robot and its application in rehabilitation robots.

Xiumin Diao as Assistant Professor is supervising and directing DeePURobotics Lab at Purdue University. His research interests are dynamics and control of cable-driven robot and intelligent robotics.

Ou Ma as Professor is supervising and directing Intelligent Robotics and Autonomous System Lab at University of Cincinnati. His research interests are multibody dynamics and control, impact-contact dynamics, intelligent control of robotics, autonomous systems, human-robot interaction,etc..