2017 IEEE-RAS 17th International Conference on
Humanoid Robotics (Humanoids)
Birmingham, UK, November 15-17, 2017

Collision-free Trajectory Planning in Human-robot Interaction through Hand Movement Prediction from Vision

Yiwei Wang*[1], Xin Ye*[2], Yezhou Yang[2], Wenlong Zhang[1]

Abstract— **We present a framework from vision based hand movement prediction in a real-world human-robot collaborative scenario for safety guarantee. We first propose a perception submodule that takes in visual data solely and predicts human collaborator's hand movement. Then a robot trajectory adaptive planning submodule is developed that takes the noisy movement prediction signal into consideration for optimization. To validate the proposed systems, we first collect a new human manipulation dataset that can supplement the previous publicly available dataset with motion capture data to serve as the ground truth of hand location. We then integrate the algorithm with a six degree-of-freedom robot manipulator that can collaborate with human workers on a set of trained manipulation actions, and it is shown that such a robot system outperforms the one without movement prediction in terms of collision avoidance. We verify the effectiveness of the proposed motion prediction and robot trajectory planning approaches in both simulated and physical experiments. To the best of the authors' knowledge, it is the first time that a deep model based movement prediction system is utilized and is proven effective in human-robot collaboration scenario for enhanced safety.**

I. INTRODUCTION

Modern household and factory robots need to conduct collaborative manipulation with human users and workers [1]. They not only need to finish their manipulation tasks but also need to maximize their chance of success while human users are collaborating with them. For example, under a factory scenario, autonomous robots are good at conducting repetitive and accurate manipulations, such as hammering a nail, while they face challenges with tasks such as pinch a nail from a box of unsorted ones. In such case, assistance from human workers become crucial. However, with the human in the loop, the robot controller has to take the human motion into consideration while planning an optimal trajectory to avoid collision and ensure safety.

This paper is motivated by observing two human workers collaborating with each other. First of all, each person is aware of the location of the other. Secondly, while human workers are conducting collaborative manipulation tasks, it is essential that the human can predict the other's movement to avoid collision. Therefore, two major capabilities are involved in developing the robot controller: 1) a perception module that can track and predict the collaborator's movement, 2) an adaptive trajectory planning module that takes into consideration the movement prediction and adjusts the robot manipulation trajectories. Moreover, these two capabilities need to be integrated seamlessly to enable real-time motion adaptation responses.

While with the motion capture system, a system can track the human collaborator's hand accurately, it is achieved with the price of attaching a marker on the human arm and hand. Moreover, the robot manipulator or human body is likely to block the marker during operation and leads to a failure of the motion capture system. In this paper, inspired by the recent work of [2], we aim at predicting human collaborator's hand movement from visual signal solely without markers. The main difficulty of implementing such a perception module lies in the huge amount of variations (such as illumination, hand poses, hand texture, object texture, manipulation pattern etc.) the system has to deal with. To tackle this difficulty, we adopt the Recurrent Neural Network architecture to enable a learning subsystem that learns the spatial-temporal relationships between the hand manipulation appearance with its next several steps of movements. To validate the effectiveness of the approach, we conduct experiments first on publicly available manipulation dataset. To further validate that the method can predict the movement with decent precision, we collect a novel set of manipulation data with readings from motion capture system to serve as the ground truth.

On the other side, our proposed vision based movement prediction module is inevitably less accurate than motion capture system. In such a case, traditional motion capture system based adaptive trajectory planning approach does not suffice. Thus, we propose a novel robot trajectory planning approach based on the human motion prediction to reach its final destination and avoid collision. To validate the proposed motion planning module, we conducted two experiments. We first tested our method on a simulation platform which takes the movement prediction from vision module as the input for trajectory planning. Then, using the Robot Operating System (ROS) [3], we integrated a Universal Robot (UR5) that can collaborate with the human worker to avoid collisions.

The main contributions of this paper are as follows: 1) we propose a perception module that takes in visual data solely and predicts human collaborator's hand movement. 2) we propose a new robot trajectory adaptive planning module that takes the noisy movement prediction signal into consideration for optimization. 3) we integrate a robot system that can collaborate with human workers on a set of trained manipulation actions, and we show it improves safety compared to a robot system without movement prediction. 4) we collect a new human manipulation dataset that can

*indicates equal contribution.

[1] Y. Wang and W. Zhang are with the Polytechnic School, Arizona State University, Mesa, AZ, USA, Email: {yiwei.wang.3, wenlong.zhang}@asu.edu

[2] X. Ye and Y. Yang are with the School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA, Email: {xinye1, yz.yang}@asu.edu

978-1-5386-4678-6/17/$31.00 ©2017 IEEE

supplement previous publicly available dataset with motion capture data to serve as hand location ground truth. We will make the new dataset available for future research. 5) We verify the effectiveness of the proposed motion prediction and robot trajectory planning approaches in experiments.

The remainder of this paper is organized as follows. Section II reviews some related work, and our visual movement prediction together with robot trajectory planning algorithms are introduced in Section III. Section IV demonstrates the simulation and experimental results. The conclusion and future work are summarized in Section V.

II. Related Work

Visual Movement Prediction: the problem of visual movement prediction has been studied from various perspectives. There are a number of works that aim at predicting objects movements. For example, [4] trained a Convolutional and Recurrent Network on synthetic datasets to predict object movements caused by a given force. This work focuses on passive objects motion while our study is about to predict the movements of an active subject, i.e. the human hand. Visually predicting human movement is more relevant to our work, while such works are usually called action prediction. [5] proposed a hierarchical representation to describe human movements and then used it as well as a max-margin framework to predict future action. Here, we treat our hand movement prediction as a regression process without predicting the actual action label. More recently, [6] proposed to apply a conditional variational autoencoder based human motion prediction for human robot collaboration. However, they used pre-computed skeletal data instead of raw images.

Human motion prediction using other methods: [7] predicted human motion by using Gaussian Mixture Model to ensure safety in human-robot collaboration. However, what their system predicts is the workspace occupancy, and our system predicts hand movement directly. [8] used a multivariate Gaussian distribution based method to predict the target of the human reaching motion. Beyond simple reaching motion, we consider much more complex motions during manipulation actions, such as cutting. [9] trained an encoding-decoding network from motion capture database to predict 3D human motions. Again, motion capture system is not practical in the real human-robot collaboration scenario as aforementioned in Section I.

Safety in Human-robot Interaction: the issue of generating a safe trajectory for a manipulator in human-robot interaction (HRI) has been studied for a long time, and many reported works focus on developing collision-free trajectories in HRI. Kulić and Croft defined a danger index approach to quantize the risk of safety in HRI by analyzing the relative distance and velocity between human and robot [10]. Tsai proposed a framework to generate collision-free trajectory by solving a constrained nonlinear optimization problem [11], and human motion was predicted based on the assumption of constant velocity [12]. All the aforementioned works do not emphasize on predicting human motion, which requires the robot to take fast actions based on the current measurement or unreliable prediction. In this work, we explore how to predict the human motion so that the robot trajectory planning can be proactive.

III. Our Approach

A. Visual Movement Prediction

The goal of the proposed vision submodule is to predict human hand movement from visual input. Here, we only take the video frames captured by the camera mounted above the robot as inputs. Without loss of generality, we start by assuming that the human co-worker manipulates single object with one hand on a working plane. The video frames capture the human co-worker's hand manipulation.

To represent the hand movement from current frame to the next time step, we adopt a displacement measure (dx, dy), which is at pixel level. Method from [2] uses a CNN-RNN-based model to predict human manipulation action type and force signals. Here, we adopt a similar structure but we extend it to further predict manipulation movement. The learning method includes a pre-trained Convolutional Neural Network (CNN) model to extract visual features from a patch of image input, and a Recurrent Neural Network (RNN) model is trained to predict hand movement (dx, dy).

We depict the overall visual submodule in Fig. 1 and describe different components in detail.

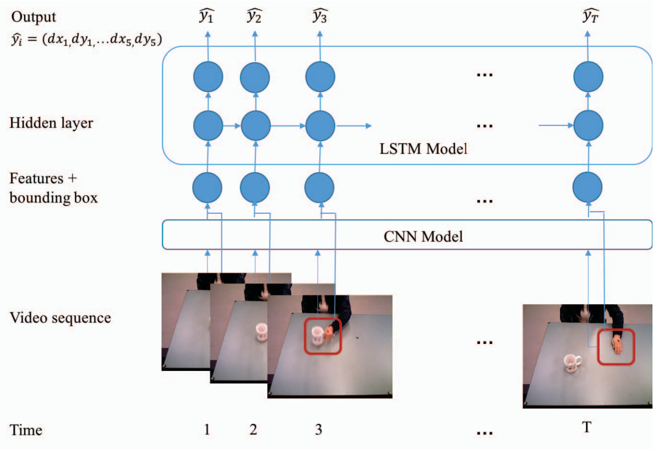


Fig. 1: Hand movement prediction model

First, to monitor human hand movement and manipulation, the system needs an attention model to focus on human hand. Given a frame of human manipulation captured by the camera, our system focuses on small patch around the human hand. This makes sense because we as human beings pay attention to the co-worker's hand as we are interested in its movement. To achieve it, our system adopts the method from [13] to track the human hand to get the corresponding bounding box of the hand at each frame. The method from [13] tracks human hand using the color distribution. Thus no additional information is required. Given the bounding box information, our system crops an image patch centered around the human hand for each frame. Then, the system treats such image patch as the input to the CNN model (here we adopt the VGG 16-layer model [14]), to extract feature representation. This preprocessing step provides a sequence of feature representations of hand patches, as well as their corresponding bounding boxes which represent the absolute positions of the hand in the frames. Our system pipelines this sequence as the input to the RNN model.

The RNN model has recurrent connection in its hidden layer, which makes it suitable for modeling time-dependent sequences. However, since each hidden state stores all the history information from the initial state, it is extremely difficult to train the traditional RNN model with back-propagation algorithm, due to the vanishing gradient problem. Thus, we adopt the LSTM model [15], in which each hidden state selectively "forgets" some history information by introducing the mechanism of memory cells and gating. To make this paper self-contained, we briefly introduce the LSTM model here.

We denote the input of the LSTM model as a sequence $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t, ..., \mathbf{x}_T\}$. In our case, $\mathbf{x}_t$ here is the extracted feature vector of the hand patches at time t together with the corresponding hand bounding box as we mentioned before. Then, by introducing memory cell $\mathbf{c}_t$, input gate $\mathbf{i}_t$, forget gate $\mathbf{f}_t$ and output gate $\mathbf{o}_t$, the LSTM model computes the hidden state $\mathbf{h}_t$ as follows:

$$\begin{aligned}
&\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i) \\
&\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f) \\
&\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o) \\
&\mathbf{c}_t = \mathbf{f}_t\mathbf{c}_{t-1} + \mathbf{i}_t \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c) \\
&\mathbf{h}_t = \mathbf{o}_t \tanh(\mathbf{c}_t).
\end{aligned} \tag{1}$$

where $\mathbf{W}_*$ denotes weight matrix, $\mathbf{b}_*$ denotes bias, and $\sigma(*)$ is the sigmoid function. Once $\mathbf{h}_t$ is computed, we connect the final model output as the hand displacement $(d\hat{x}, d\hat{y})$ at time t, which we denote here as $\hat{\mathbf{y}}_t$:

$$\hat{\mathbf{y}}_t = \mathbf{W}_{hy}\mathbf{h}_t + \mathbf{b}_y. \tag{2}$$

During the LSTM training phase, we first compute the ground truth value of hand displacement $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_t, ..., \mathbf{y}_T\}$ by estimating the hand position at each frame as the center point of the hand bounding box from the preprocessing step. Then, we define the loss function as the squared distance between $\hat{Y}$ and Y as shown in (3). We train the model by minimizing this loss function with stochastic gradient decent (SGD) method:

$$L(\mathbf{W}, \mathbf{b}) = \sum_{t=1}^{T} \|\hat{\mathbf{y}}_t - \mathbf{y}_t\|_2^2. \tag{3}$$

Last but not the least, to assist the control submodule for planning a safer and smoother trajectory, we further extend our model to predict several further steps of the hand movement. Specifically, instead of just predicting the next one step in the future, during our experiments we predict the hand movement for the next five steps into the future, namely, $\hat{\mathbf{y}}_t = (d\hat{x}_1, d\hat{y}_1, ..., d\hat{x}_5, d\hat{y}_5)$. Once the LSTM model is trained, during the testing phase, we pipeline the preprocessing step with the trained model to predict the next five steps of the human hand movement in real time.

B. Generate Critical Distance from Prediction

In the last subsection, the predicted human motion are presented in pixel coordinates. Then we projected the data from the pixel locations to the points in the table plane, where the robot operated. The predicted human hand motion, denoted as $\hat{\mathbf{y}}_t = (d\hat{x}_1, d\hat{y}_1, ..., d\hat{x}_5, d\hat{y}_5)$, and the current hand position $\mathbf{p_t} = (x, y)$ form a vector $\mathbf{P_t} \in \mathbb{R}^\mathbf{n}$ where $n = 12$, which presents the location and predicted future locations of human hand at time t, as $\mathbf{P_t} = [\mathbf{p_t}, \hat{\mathbf{y}}_\mathbf{t}]^\mathbf{T}$. In order to generate a collision-free trajectory, a mapping from the robot joint angle values to the robot surface projection on the table plane is defined as follows:

$$\mathbf{R}_t(\theta_t) = \left\{b_t \in \mathbb{R}^2 | b_t \in RP_t\right\}, \tag{4}$$

where $\theta_t \in \mathbb{R}^N$ is the joint configuration of a manipulator at time t, while N is the degree-of-freedom of the manipulator. RP_t represents the projected position of the robot body on the operation plane at time t. Another projection which presents the projection of human hand location at time t is defined as follows:

$$\mathbf{H}_t(\mathbf{P}_t) = \left\{a_t \in \mathbb{R}^2 | \, ||a_t - \hat{p}_i||_2 \leq RMSE_i\right\}, \tag{5}$$

where $\hat{p}_i = [x + d\hat{x}_i, y + d\hat{y}_i]^T$, $i = 1, 2, ..., 5$. $RMSE_i$ represents the root-mean-square error (RMSE) at i^{th} predicted time step. Then, the most critical distance between these two 2D point sets is formed as follows,

$$CD(\theta_t, \mathbf{P}_t) = \inf_{a_t \in \mathbf{H}_t, b_t \in \mathbf{R}_t} ||a_t - b_t||_2. \tag{6}$$

As equation (6) shows, the most critical distance between the robot manipulator and the human hand is formed as a function of the hand motion position, prediction data and the joint angle values of the manipulator. These factors describe whether the collision will happen between the human and the robot.

C. Optimal Trajectory Generation

In this subsection, the optimization problem, which generates a collision-free trajectory for the manipulator, is carried out. The process we propose is a 2D version optimal trajectory generation method which is developed from [16]. The objective of the optimization problem is to minimize the length of path towards the goal configuration which achieves the task, while the safety constraint is fulfilled. The optimization problem is formulated as follows:

$$\min_{\theta_{t+1}} \quad ||(F(\theta_{t+1}) - x_g)||_2 \tag{7}$$

$$s.t. \quad \theta_t - \dot{\theta}_{max}Ts \leq \theta_{t+1} \leq \theta_t + \dot{\theta}_{max}Ts \tag{7.a}$$

$$CD(\theta_{t+1}, \mathbf{P}_t) \geq \Delta, \tag{7.b}$$

where $\theta_{t+1} \in \mathbb{R}^N$ is the optimal variable at time t, where N is the degree-of-freedom of the manipulator, which stands for the joint angles for the manipulator at time $t + 1$. The objective equation (7) minimizes the distance between the current robot configuration to its goal. As we define the operation space to be a plane upon a table, the goal configuration yields into a 2D vector $x_g \in \mathbb{R}^2$. It means the desired robot end effector location on the table $F(\theta_{t+1})$ is a projection $\mathbb{R}^N \rightarrow \mathbb{R}^2$, which stands for the location of the end effector on the table plane with joint angles θ_{t+1}. Equation (7.a) is the speed constraint of the optimization problem, where Ts is the sample time of the system and $\dot{\theta}_{max} \in \mathbb{R}^N$ is the maximum angular speed of all the joints. Equation (7.b) is the safety constraint which ensures a collision-free trajectory, and Δ is the minimum distance between the robot end effector to the human hand to guarantee safety.

By solving this optimization problem iteratively at every time step of the system, a collision-free trajectory of the manipulator can be generated in real time, while achieving the goal of robot for task execution. The objective equation (7) ensures that the trajectory always tracks its goal while (7.a) and (7.b) guarantee the smooth and safe trajectory.

D. *System Integration for Real-time Execution*

In this subsection, structure of the proposed system is demonstrated in Fig. 2, which enables real-time hand tracking, prediction and optimal collision-free trajectory generation. The image of human motion is captured by an Xtion PRO LIVE RGBD camera from ASUS. The hand tracking node in ROS subscribes the image frames, recognizes the hand pattern and delivers the hand patch to the neural network nodes.The CNN and RNN nodes generate a message vector, which contains the current hand location and predicted hand motion, and publish it to ROS. A node named UR5 Controller solves the optimization problem, which is described in detail in Section III.D, with sequential quadratic programming (SQP) solver from scipy.optimization toolbox. The result of the optimization problem forms a command of the desired angular position UR5' joints, which is sent to UR5 robot.

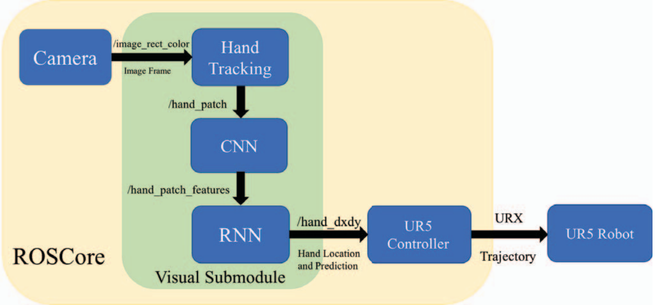


Fig. 2: Structure of the entire system for safe HRI

IV. Experiments

The theoretical and practical descriptions of the proposed system suggest three hypotheses that need empirical validation: a) the proposed vision module is able to predict human hand movement with reasonable accuracy; b) the proposed control module is able to plan a better robot movement trajectory to avoid collision during collaboration, given the hand movement prediction from the vision submodule; c) these two modules can be integrated together to enable a physical robot manipulator to collaborate with the human co-worker in a real-time fashion. To validate hypotheses (a) and (b), we conducted experiments on both publicly available and new data collected from our lab. Section IV-A describes the datasets and Section IV-B reports the performance. To validate hypothesis (c), we integrated a robotic system within the ROS framework. We evaluate the performance of the system both in simulations (Section IV-C) and on an actual robot in experiments (Section IV-D and the supplemental video).

A. *Datasets*

To validate the previous proposed hand movement prediction module, we need a test bed with various users conducting different manipulation actions with several objects. The recent work from [2] provides such a collection of human manipulation data. Though the purpose of their data is to validate manipulation action label and force prediction, the same set of actions contain significant amount of hand movements on a table plane. Thus, it also suits our need for training and validating the hand movement prediction module. The dataset includes side-view video recordings of five subjects manipulating five different objects with five distinct actions, and each is repeated five times (a total number of 625 recordings). TABLE I lists all object-action pairs. For further details about the public available dataset, one can also refer to [2].

TABLE I: Object-action pairs in the public dataset from [2]

Object	Actions
cup	drink, pound, shake, move, pour
stone	pound, move, play, grind, carve
sponge	squeeze, flip, wash, wipe, scratch
spoon	scoop, stir, hit, eat, sprinkle
knife	cut, chop, poke a hole, peel, spread

Additionally, to validate that our vision based movement prediction module is able to provide accurate enough predictions for the robot control module, and further to validate our integrated system in simulation, the aforementioned dataset does not suffice. To enable simulation, we further complement the dataset with a set of newly collected data. The complementary set records the actual hand position in world coordinate during their manipulation actions through the motion capture system, as well as the camera matrices. We started from several real-world HRI scenarios and designed three actions, each with one target object under manipulation (shown in TABLE II and Fig. 4). For each action object pair, the human subject was asked to repeat the same action five times. The total number of 60 recordings serve as the test bed to 1) further validate the movement prediction module and 2) validate the integration of vision and robot movement planning modules in simulation. We intend to make the supplemental set of data also publicly available for future research.

B. *Experiment I: Visual Movement Prediction*

To evaluate the performance of our proposed movement prediction module, we need a performance metric. Here, we

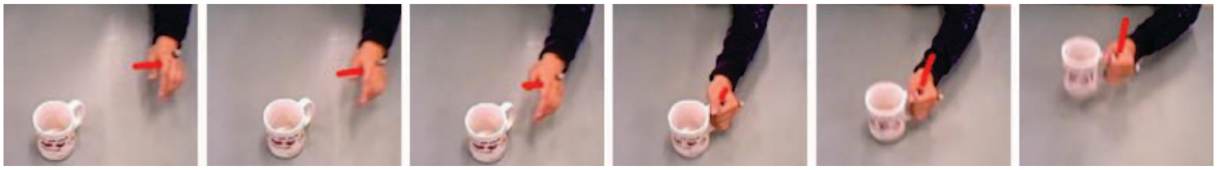
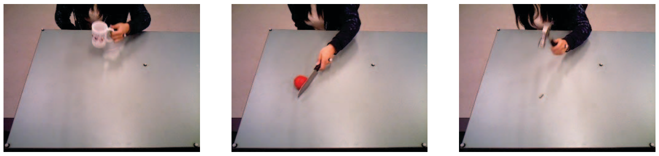


Fig. 3: An example of hand movement prediction result.

TABLE II: Object-action pairs in the supplemental dataset

Object	Action
cup	drink water
knife	cut tomato
hammer	pound



(a) Drinking (b) Cutting (c) Pounding

Fig. 4: Illustration of the performed actions

adopt the widely accepted performance metric of RMSE. It is defined in equation (8), where N denotes the total number of testing videos, T denotes the number of frames with each testing video. Here, $\hat{\mathbf{y}}_{it}$ and $\mathbf{y}_{it}$ are predicted value and ground truth value of the hand displacement on the i^{th} video sample at time t respectively. Both $\hat{\mathbf{y}}_{it}$ and $\mathbf{y}_{it}$, as well as the RMSE are measured by the number of pixels. The total number of pixels is determined by the resolution of the video frames (640×480 pixels).

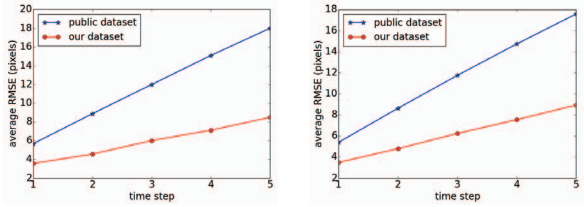
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{\mathbf{y}}_{it} - \mathbf{y}_{it})^2}{NT}}. \tag{8}$$

The training and testing protocol we used is leave-one-out cross-validation. On both testing beds, we report the average RMSEs as the performance of the trained models. Fig. 5 shows the average RMSE of predicted hand displacements range from one step in the future to five steps in the future. To demonstrate how well our movement prediction module perform, in Fig. 3 we show examples of the prediction outputs, where the orientation and length of the red arrows overlaying on the human hand depict the in-situ hand movement prediction at that specific frame.

From the experimental results, it is worth mentioning the following: 1) our prediction module is able to predict human hand movement within an RMSE of about 18 pixels, which empirically validates our hypothesis (a); 2) with the increasing number of steps to predict, the RMSE tends to increase, which aligns well with our expectation.

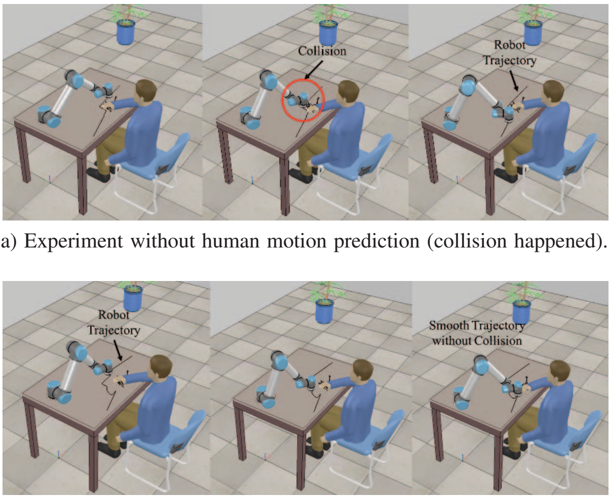
C. Experiment II: Planning with Prediction

To validate the optimal trajectory generation method, we conducted a simulation test in Virtual Robot Experimentation Platform (V-REP). We set up a scene where the human worker and the robot manipulator worked upon the same table in V-REP environment. The motion capture data, which were recorded in our complementary dataset, were adopted to create the animation of the human hand movement in the scene. Here we demonstrate the simulation results in Fig. 6, where the generated trajectories with or without the motion prediction safety constraints are compared. As Fig. 6a indicates, without the motion prediction output from vision module, the trajectory failed to guarantee safety, which leads to a collision between the human hand and the robot end effector. Fig. 6b presents the trajectory generated with the human motion prediction, which presents a detour to ensure adequate safety as well as trajectory smoothness while fulfilling the task.



(a) Average RMSE of dx (b) Average RMSE of dy

Fig. 5: Average RMSE of predicted hand displacements (dx, dy) from one step to five steps in the future.



(a) Experiment without human motion prediction (collision happened).

(b) Experiment with human movement prediction.

Fig. 6: Snapshots from a simulated experiment in V-REP.

D. Experiment III: An Integrated Robotic System

In this part, we conducted an experiment with a UR5 manipulator, a host PC and an Xtion PRO LIVE RGBD camera. Fig. 7 shows the experimental setup. The human co-worker

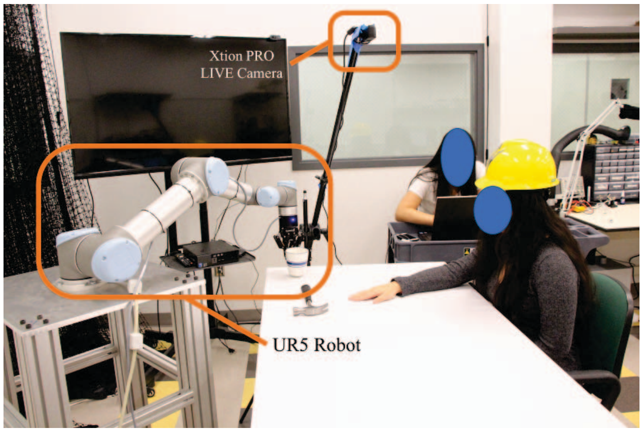
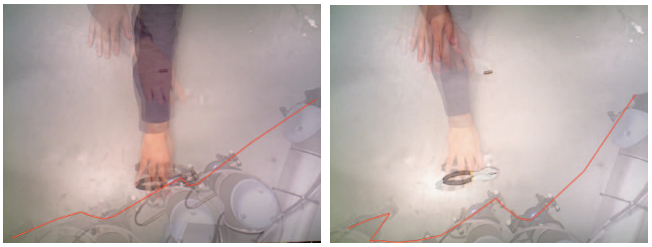


Fig. 7: Experiment setup



(a) Without Prediction

(b) With Prediction

Fig. 8: Robot Gripper Trajectories of Experiments with or without Human Motion Prediction

and the robot arm operated on the table, while the camera had a top-down view of the table for coordinate alignment. Both the UR5 and camera are connected to the host PC, where the ROS core and all the nodes are executed. A lab assistant was asked to perform the hammer pounding motion on the table, while the UR5's task is to move diagonally across the table. The original route of the robot to achieve the goal was obstructed by the human pounding motion. Fig. 8b shows the trajectory of the robot end-effector with a solid red line, which demonstrates a successful detour to avoid the human hand with the predicted human motion. While in Fig. 8a, the gripper fails to avoid the human hand without the human motion prediction. This overall integration of the system empirically validates the hypothesis that our system is capable of generating an optimal collision-free trajectory to ensure the safety in HRI with vision-based hand movement prediction. We attach a live recording of the experiments in the supplemental file of this submission.

V. Conclusion and Future Work

In this paper, we presented a HRI system that features two novel improvements from previous ones: 1) we proposed and implemented a vision based human hand movement prediction system through a CNN+LSTM architecture; 2) we put forward a corresponding robot trajectory planning algorithm that utilized the prediction from the vision module for trajectory optimization calculation. Experiments conducted on both public and new manipulation dataset, in simulation and on a physical robot executing challenging collaborative tasks validated the proposed system.

The study presented in this paper also opens several avenues for future studies. First, human robot collaborative manipulation may require the robot end-effector to approach the human hand rather than avoiding it. In such a case, recognition of the actual action that the co-worker is trying to finish becomes essential. Further research on a joint short-term and long-term prediction model of manipulation movement, action label, human gesture and applied forces from visual data is needed. On the other side, for motion planning, how to incorporate both short-term and long-term, and a variety of predictions from vision module requires an adaptive and hierarchical planner that combines both task and motion planning together.

References

[1] T. S. Tadele, T. de Vries, and S. Stramigioli, "The safety of domestic robotics: A survey of various safety-related publications," *IEEE robotics & automation magazine*, vol. 21, no. 3, pp. 134–142, 2014.

[2] C. Fermüller, F. Wang, Y. Yang, K. Zampogiannis, Y. Zhang, F. Barranco, and M. Pfeiffer, "Prediction of manipulation actions," *International Journal of Computer Vision (IJCV)*, p. Preprint, 2017.

[3] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng, "Ros: an open-source robot operating system," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA) Workshop on Open Source Robotics*, Kobe, Japan, May 2009.

[4] R. Mottaghi, M. Rastegari, A. Gupta, and A. Farhadi, "what happens if... learning to predict the effect of forces in images," in *European Conference on Computer Vision*. Springer, 2016, pp. 269–285.

[5] T. Lan, T.-C. Chen, and S. Savarese, "A hierarchical representation for future action prediction," in *European Conference on Computer Vision*. Springer, 2014, pp. 689–704.

[6] J. Bütepage, H. Kjellström, and D. Kragic, "Anticipating many futures: Online human motion prediction and synthesis for human-robot collaboration," *ArXiv e-prints*, Feb. 2017.

[7] J. Mainprice and D. Berenson, "Human-robot collaborative manipulation planning using early prediction of human motion," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 299–306.

[8] C. Pérez-D'Arpino and J. A. Shah, "Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 6175–6182.

[9] J. Bütepage, M. Black, D. Kragic, and H. Kjellström, "Deep representation learning for human motion prediction and classification," *arXiv preprint arXiv:1702.07486*, 2017.

[10] D. Kulic and E. A. Croft, "Real-time safety for human - robot interaction," in *ICAR '05. Proceedings., 12th International Conference on Advanced Robotics, 2005.*, pp. 719–724.

[11] C. S. Tsai, J. S. Hu, and M. Tomizuka, "Ensuring safety in human-robot coexistence environment," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4191–4196.

[12] C.-S. Tsai, "Online trajectory generation for robot manipulators in dynamic environment an optimization-based approach," Ph.D. dissertation, University of California, Berkeley, 2014.

[13] G. R. Bradski, "Real time face and object tracking as a component of a perceptual user interface," in *Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on*. IEEE, 1998, pp. 214–219.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] Y. Wang, Y. Sheng, J. Wang, and W. Zhang, "Optimal collision-free robot trajectory generation based on time series prediction of human motion," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 226–233, 2018.