

Message Passing Framework for Vision Prediction Stability in Human Robot Interaction

Youngkyoon Jang¹ and Yiannis Demiris¹

Abstract—In Human Robot Interaction (HRI) scenarios, robot systems would benefit from an understanding of the user’s state, actions and their effects on the environments to enable better interactions. While there are specialised vision algorithms for different perceptual channels, such as objects, scenes, human pose, and human actions, it is worth considering how their interaction can help improve each other’s output. In computer vision, individual prediction modules for these perceptual channels frequently produce noisy outputs due to the limited datasets used for training and the compartmentalisation of the perceptual channels, often resulting in noisy or unstable prediction outcomes. To stabilise vision prediction results in HRI, this paper presents a novel message passing framework that uses the memory of individual modules to correct each other’s outputs. The proposed framework is designed utilising common-sense rules of physics (such as the law of gravity) to reduce noise while introducing a pipeline that helps to effectively improve the output of each other’s modules. The proposed framework aims to analyse primitive human activities such as grasping an object in a video captured from the perspective of a robot. Experimental results show that the proposed framework significantly reduces the output noise of individual modules compared to the case of running independently. This pipeline can be used to measure human reactions when interacting with a robot in various HRI scenarios.

I. INTRODUCTION

In the computer vision community, significant advances have been made in object detection and segmentation [1]–[6], hand detection [7], [8], hand pose estimation [9]–[12] and its state recognition [13], and human behavior recognition [14]–[20], which are potential sources of useful information in human-robot interaction scenarios. For example, grasping [21]–[25], picking-and-placing [26], stable human pose estimation [27], and stable object detection/discovery [28]–[37] are several topics of interest in HRI scenarios. Recent approaches are pushing the boundaries of HRI by continuously adopting advanced methods of individual vision modules to understand human behaviours, especially in cooperative [38] and safe [39] HRI. However, due to the lack of data relevance between computer vision datasets and more practical HRI tasks, the performance guaranteed by the pretrained parameters of a network trained on a specific dataset cannot be immediately redirected to similar tasks in HRI scenarios, and new data collection processes are required to use existing state-of-the-art methods.

To overcome this problem, there have been studies that reduce the noise of computer vision predictions by understanding the properties [40], [41] of objects or reasoning about their stability [31], [33], [42] in images. These approaches represent how to apply known physics knowledge and common-sense to stabilise²

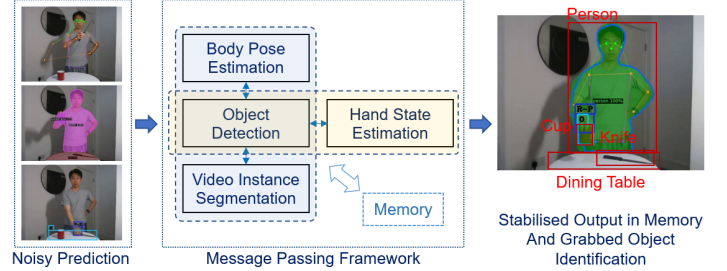


Fig. 1. Conceptual diagram of the proposed Message Passing Framework (MPF): The proposed MPF stabilises the noisy prediction output of heterogeneous state-of-the-art vision modules and stores core sources (hand grasping state, object information and human body pose) in long-term memory that can be used for human-robot interaction.

predictions in visual signal processing. For example, to estimate the destination position of a sliding object from visual information, Galileo [40] was proposed to predict the mass and friction values of an object by applying physics knowledge. To reduce the noise of 3D object detection and odd-looking pose estimation, Jia, et al. [42] adopted common-sense physics knowledge (objects can lean on or stack on other objects). Although these studies can be used to better understand object prediction in HRI scenarios, they require specific training and testing environments in which portable objects slide or lean against each others, that are less frequently observed in HRI.

Addressing the above challenges, we propose a message passing framework for vision prediction stability in HRI, as shown in Fig. 1. The proposed framework is designed to take benefits from utilising several state-of-the-art vision prediction modules, such as object detection [4], hand-object contact state prediction [13], human pose estimation [43], and video instance segmentation [44]. The main contributions are threefold:

- (1) **Message passing:** The proposed framework improves the stability with the message passing strategy between individual modules. The proposed framework does not require additional training nor additional dataset collection.
- (2) **Common-sense physics reasoning:** The proposed framework further improves the stability of vision prediction based on common-sense physics reasoning such as the law of gravity and hand-object synchronous movement rule.
- (3) **Working examples in HRI scenarios:** Real-world scenarios where state-of-the-art methods fail are discussed using the proposed pipeline.

II. RELATED WORK

In human-robot interaction (HRI), object selection (e.g. grasping and picking) is one of the first and most important tasks that initiate more intricate subsequent interactions [45]. Also, object detection/segmentation is closely related to grasp position prediction, and object detection and picking point estimation are often solved together in a single framework [29]. However,

¹The authors are with the Personal Robotics Laboratory, Department of Electrical and Electronic Engineering, Imperial College London, SW7 2BS, United Kingdom {youngkyoonjang.jang11, y.demiris}@imperial.ac.uk. This research was supported by the UKRI Node on Trust (EP/V026682/1), and a Royal Academy of Engineering Chair in Emerging Technologies.

²Stabilisation: Reducing noisy prediction results and maintaining consistent predictions across input sequences

this research topic is still actively studied and has not yet been solved due to difficult visual scene analysis problems caused by noisy prediction output [22] and incomplete visual features [26], or learning methods that are not applicable in other domain that appear in real-world scenarios (e.g., unstructured scene [21]). Unlike the computer vision community, which understands realistic visual signals in a given scenario, the robotics community considers realistic but disembodied or unrestricted signal adaptation scenarios, such as lifelong object discovery [36] and management, and stabilisation of performance in unexpected scenarios [23]. In this section, we discuss how conventional methods have addressed the problem of stabilising visual signal processing results in human-robot interaction scenarios.

Identifying meaningful object regions while reducing redundancy: As mentioned earlier, the robotics community considers more practical scenarios, such as unseen object discovery, because in daily life people not only interact with the known objects, but also adopt unseen objects. To this end, Osep, et al. [46] attempted to merge object proposals into meaningful regions using a category-agnostic image segmentation method. Pham, et al. [47] proposed a convolutional oriented boundary network to decompose objects into meaningful regions representing objects or scene surfaces that are used to construct region hierarchies. Chan, et al. [28] proposed an unsupervised salient object discovery method to discover arbitrary salient objects on a frame-by-frame basis. As such, detecting meaningful regions while identifying redundant regions for an object can reduce potential object prediction noise. In this paper, we adopted a similar idea of merging multiple prediction bounding boxes over time for the same object as a temporal smoothing strategy.

Common-sense reasoning: Most algorithms are designed with common sense in mind to increase the reliability of predictions and, consequently, to provide better performance. To refine object recognition output, Sun, et al. [30] proposed combining scene recognition and object recognition that reasonably predict objects in a specific context (surrounding scene). With the same concept, Li, et al. [48] proposed to extract semantic features such as pole-like objects, lane markings, and curbs that can be seen in road scenes (context) to improve global localisation in autonomous vehicle scenarios. To reduce the noisy predictions that appear in-between with other accurate predictions for the video input, Sabater, et al. [31] proposed a post-processing pipeline with learning-based similarity evaluation between detection across frames, taking into account the location, geometry, appearance and semantic of each object. To identify portable objects in a scene, Langer, et al. [33] proposed using global structure to extract object candidates and then using local verification comparing the isolated geometry with a reference reconstruction provided by a task. As discussed here, understanding the contextual geometry and scene will help identify objects and stabilise vision prediction results in the scene.

Physics-based reasoning: There have been studies on perceiving the properties of physical objects by integrating a physics engine, called Galileo [40]. Galileo predicts the mass and friction values of an object from a given video showing the object sliding. However, it is still difficult to apply this method to real-world scenarios where objects, hands, and human bodies dynamically interact, resulting in a lot of unexpected noise. Jia, et al. [42] proposed a stability-reasoning framework to improve RGB-D segmentation and scene volumetric representation. It represents how

to apply common-sense physics knowledge (such as being able to stack objects on top of other objects) to stabilise predictions. Also, there are studies on decomposing physical primitives from a visual representation to understand the physical and geometric attributes of objects [41], and a study on mapping from image observations to object representations, which is a pairwise physics interaction function [49]. However, these robust theoretical approaches cannot be redirected to the practical problem of causing unexpected random movements.

A review of the prior works discussed in this section confirms that improving/stabilising object detection/recognition performance in real-time HRI involving objects, human body, and hands remains an open and challenging problem. There are major approaches to improving datasets, context-dependent learning techniques/architecture, and post-processing. However, these approaches were limited to narrow scenarios. This paper aims to demonstrate a explainable approach that can be used in HRI when interactions involve objects, people, and their hands. We make our code publicly available³ so that other HRI researchers can potentially benefit from more stabilised outputs from their vision modules.

III. PROPOSED FRAMEWORK: MPF

The proposed Message Passing Framework (MPF) is shown in Fig. 2. The MPF consists of state-of-the-art methods of object detection [4], hand-object state prediction [13], human pose estimation [43], and video instance segmentation (VIS) methods [44] that can be used as the main sources in HRI scenarios. The input to the MPF is a sequence V of colour image I . There are image paths and video paths, depending on the type of input each module is required to process. As shown in Fig. 2, the object detection, hand-object state prediction, and human pose estimation modules use a frame image as input, whereas the VIS takes a short video snippet consisting of 64 frames as input.

Each module outputs a different result. Object detection produces bounding boxes of objects, prediction scores of individual bounding box, and segmentation region within the bounding boxes. Hand-object state recognition predicts the bounding box of each hand and interacting objects with the hands, as well as hand-contact states. Human pose estimation predicts the positions of body joints and predictive confidence scores for each joint. Video instance segmentation predicts the instance region of an object and the ID of the instance. ID confirms if the instance in the current frame is the same instance appearing in other frames in the video input. The combined and synchronised outputs of each individual module are shown in Fig. 2.

The main motivation of this paper is to stabilise the output of the currently available vision modules. Although our proposed framework consists of state-of-the-art methods in each computer vision topic, it still produces a noisy output in real-world HRI scenarios, as shown in Fig. 3(a). This is primarily due to the lack of learning contextual knowledge in computer vision task-oriented training, where methods attempt to capture short-term temporal context or local spatial attention from a visual signals. However, many HRI scenarios explore the possibility of long-term interactions in which humans and robots collaborate or support each other in various tasks. This can lead to problems (noisy outputs) caused by traditional vision approaches that do

³The code is available: www.imperial.ac.uk/personal-robotics/software.

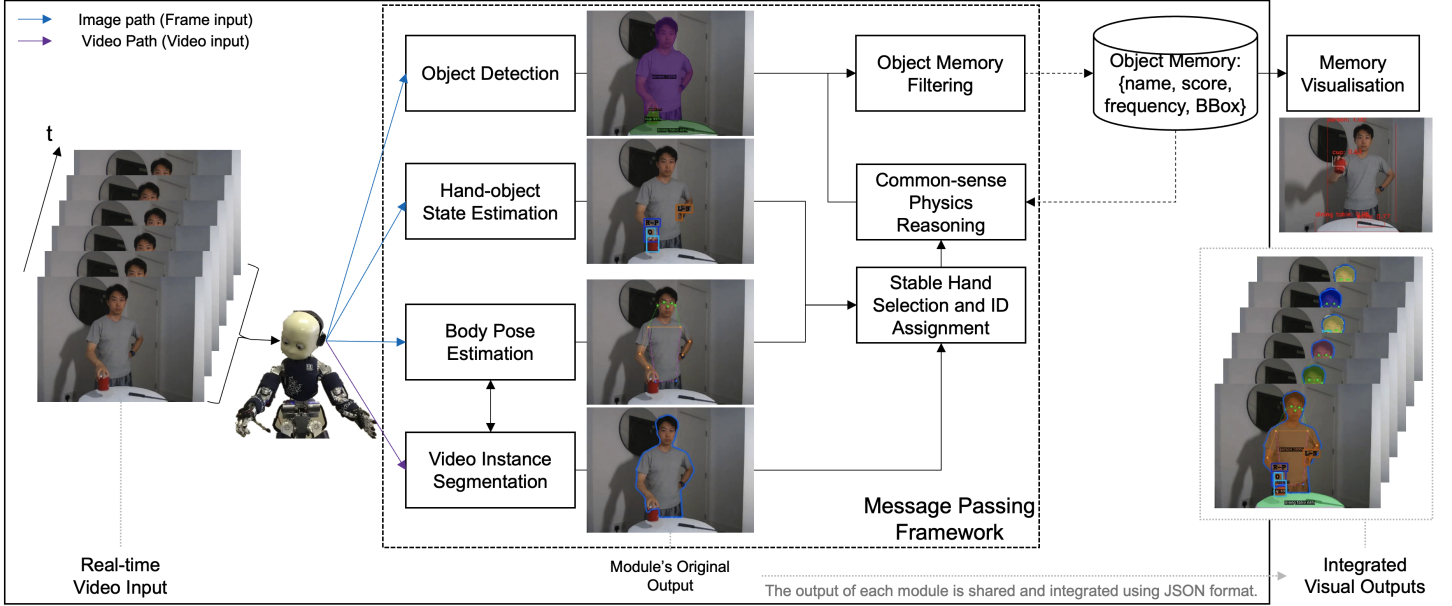


Fig. 2. Overview of the proposed Message Passing Framework (MPF): The framework adaptively constructs the appropriate input type, either frame or video. Then, the object detection outputs are refined with the object memory filtering block, and object prediction information is stored in long-term memory for further updates later. Once the refined object information is stored in memory, the common-sense physics reasoning block further removes the object detection module’s noisy output in a run-time manner. For this, MPF mainly uses a hand bounding box and its state result given from the hand-object state estimator. If the state estimator fails, the wrist joint position is chosen to be used as an alternative hand information. The body pose estimation results are also stabilised by cross-checking with the output of video instance (body) segmentation. In this way, multiple heterogeneous modules exchange messages through a shared output memory to improve overall recognition performance.

not carefully take into account capturing context in long-term interactions. However, the individual vision modules have great potential to complement each other.

The following subsections describe how the individual modules complement each other. It also describes how to adopt message passing strategy and common-sense physics reasoning in the proposed framework to make the output of individual modules more reliable.

A. Message Passing

In the proposed MPF, the outputs of the individual vision modules are first acquired in parallel. The outputs are all related to the HRI scenario, so they can be interconnected, as shown in Fig. 1. Object detection, body pose estimation, and video instance segmentation modules are interconnected to each other because all the modules share human body instances in their prediction. In addition, object detection and hand-object state estimation modules are interconnected, as the hand state represents the interaction states with an object. Object bounding boxes, except for human instances, provided by the object detection module can be stored and updated in external memory for long-term use in future HRI scenarios.

This section describes how these links between individual modules can help improve the reliability of computer vision predictions. The proposed message passing aims to improve the analysis of human-robot interaction sources appearing in a video captured by a humanoid robot’s eye (camera) in a collaborative work environment. In a collaborative work environment, robots need to analyse what humans pick and how they behave as a feedback on the robot’s behavior. Several sources (computer vision outputs) that aim to be stabilised for human behavior analysis in HRI include:

Objects: The proposed framework has a memory to store the location of the detected objects. A pipeline from object detection

module to object memory is used to update object position and its prediction confidence scores during interaction. First, when objects appear in the first frame of a video, the proposed MPF stores in memory the bounding box position, name, and confidence score of all detected objects. In the following frames, the proposed MPF then uses Eq. 1 to check whether an object belongs to one of the original objects stored in the memory. If it belongs to one of the original objects in the memory, it updates the memory by replacing the bounding box information and keeping the object name with the highest prediction score for the corresponding object name. Otherwise, it updates the memory by appending the object information.

$$IoU = \frac{B_A \cap B_B}{B_A \cup B_B}, \quad (1)$$

where B_A and B_B represent the bounding boxes of A and B objects, respectively and IoU is a score in the range $[0 : 1]$, indicating how much the two object bounding boxes overlap.

If there are multiple predicted names for the same object bounding box, the MPF uses Eq. 2 to choose a representative name for the object. Due to the common challenges in computer vision, such as lighting, scale, and orientation changes, noisy results indicative of the problematic situation mentioned above easily occur, despite the presence of stably positioned objects in the video. The stabilised results using the proposed MPF are shown in Fig. 3.

$$OS_i = \frac{\#O_i}{\sum_{i=1}^N \#O_i} \cdot \alpha + P_i \cdot \beta, \quad (2)$$

where $\#O_i$ represents the frequency count of object label i predicted on the given bounding box location. $\sum_{i=1}^N \#O_i$ represents the total number of predictions made at the same location in a given bounding box. P_i denotes the highest prediction score

corresponding to the object label i at the bounding box location. α and β both represent the weights for each term, and we set both values to 0.5. OS_i is an i -object score which is continuously updated through run time processing.

Hand: Hands are used to identify interacting objects in the video. For this, we adopt a hand-object state estimator in MPF. The state estimator is more suitable than using wrist joint information extracted from the body pose estimator because it can directly correlate the hand contact state with the object. The hand-object state estimator also provides the contacted object bounding box information, but the noisy output is still inherited from the state estimator’s object detector (here, the estimator uses Region Proposal Network [50]). Therefore, we decided to use the object bounding box information stored in MPF memory to select an interacting object when the hand grasping state is recognised. An example of hand contact state recognition and an interacting object identification results is shown in Fig. 5.

Body: In theory, the use of the hand-object state estimator in conjunction with the object detector is expected to be sufficient to analyse basic interactions such as picking and placing objects with which a human interacts. However, the human hand is often not detected due to the same problems as in object detection (e.g. fast movement, low resolution, occlusion, etc.). In this case, we use body joint (especially wrist joint) location information to supplement the hand detection module. This information is revisited later in Sec. III-B to describe one type of common-sense physics reasoning: Coupled hand-object rule.

Video Instance: Now we know which hand is holding which object using the hand-object state estimation modules and the memory of the proposed framework. To track a person’s interaction history, information extracted about objects, hands, bodies, and interaction states over time can also be linked to a person ID. To this end, the proposed MPF optionally adopts a video instance segmentation (VIS) module to get an instance (specifically human) region and its ID from a given video input. Then, based on the Even-odd algorithm [51], a person ID corresponding to the original human instance could be assigned to an object interacting with a hand located inside the human instance region. Assigning an individual’s ID can be used to track an individual’s interaction history, which can help to provide adaptive services to the individual, the subject of personalisation. We leave this issue for future work.

B. Common-sense Physics Reasoning

As an example of the use of common-sense physics reasoning, we use the law of gravity, and a coupled hand-object rule; more can be added when dealing with specific application scenarios. We limit the application of this proposed common-sense physics reasoning to images captured by humanoid robots such as iCub [52] that stand upright. In an upright-standing robot scenario, objects are stacked on top of other objects unless held by a human hand [42]. Otherwise the object will only stay in the air while falling. This cannot be longer than one second depending on the mass (weight) of the object (if it is expected to interact with a kitchen or office object).

From Sec. III-A, the output of individual modules has been once stabilised. Now we further stabilise it based on common-sense physics reasoning: the law of gravity, and the objects humans grasp does not typically fall.

The law of gravity: We first categorised objects into two types:

portable, and furniture – difficult to move. Then, from the given memory stored in the proposed MPF, the objects were filtered out when they were not on furniture. To do so, we use Eq. (1) to find out how much the bounding box of a portable object overlaps with the bounding box of a furniture. In addition, the MPF double checks if the object is located on top of the furniture. Applying this rule, the proposed MPF could further stabilise the noisy outputs as shown in Fig. 3(c).

Since it is difficult to clearly capture an object without motion blur when it falls from a general robot camera at a capture rate of 30Hz, the case of falling object is not considered in this paper.

Coupled hand-object rule: As an additional example, we additionally added one more rule that grasped objects don’t fall, even if they are not on furniture. To investigate it, we first use the hand-object state estimator that provides the bounding box of the hand. But still, hand detection fails when interacting with an object in a dynamic HRI scenario. In that case, we used the body pose result to update the pseudo hand location. As a result, in this rule, only the overlap between the hand bounding box and the object bounding box is examined instead of checking the state due to the noisy state prediction result of the hand-object contact state recognition. When both gravity and coupled hand-object rules are satisfied, the object is kept in memory in the proposed MPF.

As an example, by applying the common-sense physics reasoning, the proposed framework was able to filter out major noise induced by the state-of-the-art vision modules as shown in Fig. 3(d).

IV. IMPLEMENTATION

In this paper, we take advantages of using currently available state-of-the-art vision modules, specifically Mask R-CNN [4] for object detection, 100 Days of Hands (DOH) [13] for hand-object state estimation, HRNet [43] for human body pose estimation, and VisTR [44] for video instance segmentation. We used the official source code publicly released on an online sharing platform (e.g., Github) for Mask R-CNN, 100 DOH, VisTR. For HRNet, we used the MMPose [53] library which allows us to test many recent methods with a simple installation. As shown in Fig. 2, we were able to integrate all module outputs. The final visualization may be different from the visualization in the original module, as it writes the prediction output to a json file format [54] and then reads and integrates the individual files to reconstruct the results.

Although the individual modules work as intended, if the modules are running with slow processing times, they cannot be a feasible solution in HRI scenarios where instant responsive feedback is required. Thus, we implemented a parallel processing pipeline. The input is captured from iCub’s camera and the captured image and video sequence are published as a ROS topic. Then, the individual modules read the updated image from the topic, process it, and finally share a single json output as an individual ROS topic. Then the final integration module collects and visualise each module’s outputs as a integrated version of output in a single frame, as shown in Fig. 2.

During this process, we allocated two modules each to each GPU on our test PC so that the loaded vision module had enough memory space to process after it was loaded in the initial phase. More specifically, the object detection and hand-object state estimation modules are assigned to the first GPU.

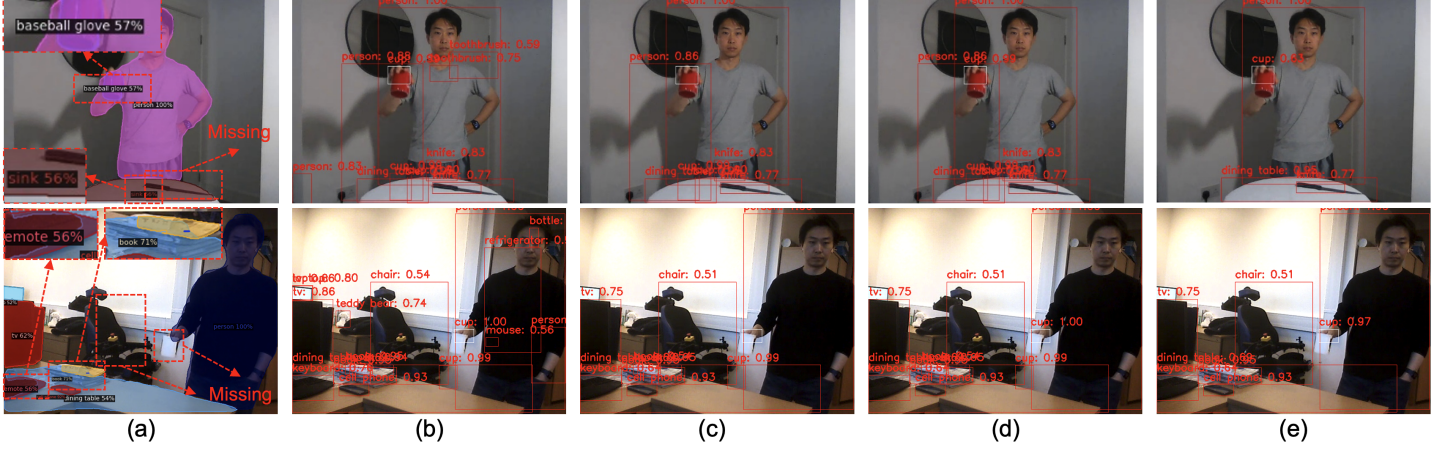


Fig. 3. Example of vision module stabilisation using message passing framework (MPF): (a) The original output of the object detection module (Mask R-CNN [4]) shows incorrect prediction (e.g. name and boundary of an object) and missing objects, (b) Visualisation of object bounding boxes stored in memory – MPF uses the temporal contextual information to display the representative object labels among the multiple predictions in the same bounding box, (c) The law of gravity – MPF removes the stored object bounding boxes that are not located on top of furniture objects, (d) Hand-object coupled movement rule – MPF maintains an object bounding box when an object is not on furniture but a person is holding it, and (e) Spatial redundancy check and elimination – Multiple bounding boxes for the same object are examined and only the closest objects with object detection results are kept. Even if the object detection results are noisy and missing at runtime, the memory of the MPF is reliably preserved in HRI scenarios.

And the rest of the modules are assigned to the second GPU. Our proposed MPF is developed in an experimental environment using two laptops, each configured with an Intel Eight Core i7-11700 (2.5GHz) 16 MB Cache CPU processor and an NVIDIA GeForce RTX 3070 8GB GPU with 32GB of DRAM. We used CUDA 11.1 and PyTorch 1.8 for implementation.

For VIS modules that require a video input, it is executed every 64 frames without using overlapping input. Computer vision modules produce relatively stable results for predictions of large instances such as the human body than other small objects such as hands or cups, so it may be reasonable to examine them every 64 frames. We plan to investigate further the effects of more intensive message passing between frame-based prediction and video-based prediction.

V. EXPERIMENTAL RESULTS

Since the currently available non-egocentric datasets do not primarily contain interaction scenarios where people, hands, and objects interact with each other, we will experiment with our collected sample videos instead of using existing datasets. Example test videos were recorded in kitchen and office environments with several objects that appeared naturally in these environments. Please see the accompanying demo video for the run-time stabilisation results. The example scenario shows a person is picking up an object located on a furniture (e.g. dining table).

For the first experiment, we test the effect of using the proposed MPF framework. When using Mask R-CNN [4], a state-of-the-art object detection method, we confirm that it still generates and stores noisy prediction results despite using a simple temporal filtering method, as shown in Fig. 4. However, our proposed framework significantly stabilised the prediction results. Every step (temporal filtering, gravity law, coupled hand-object rule, common-sense spatial redundancy check (see Fig. 3(e)) we followed helped to stabilise the final prediction result. Because object detection results still depend on the same object detection module [4], sometimes an error result is produced. However, it soon restores and stabilises the memory of the MPF, as shown in Fig. 4.

The second experiment shows improved stability of hand-object contact state estimation, as shown in Fig. 5. Original state-of-the-art hand-object state estimators often fail to identify target objects due to the Region Proposal Network [50] failing to detect motion-blurred objects. However, our proposed framework was able to store and maintain the object detection results regardless of noisy output and missing prediction results. This ultimately improves the hand-object contact state estimation performance as well as the target object identification stability.

While the experiments we described clearly indicate substantial stabilisation benefits by using our framework, these benefits will be best demonstrated in HRI scenarios that do contain contextual elements that allow our framework to use the stabilisation rules we apply; for example, the gravity rule will only show benefits when furniture or other similar surfaces exist for the objects to rest on. Similarly, we assume that objects in the scene can be detected by the individual modules in the first place.

VI. CONCLUSION

In this paper, we proposed a Message Passing Framework for vision prediction stability that can be used in Human Robot Interaction scenarios. To this end, we designed a message passing pipeline to improve the prediction performance of state-of-the-art vision modules publicly available. In addition, we further improved the results of individual prediction module by adopting common-sense reasoning e.g. rules of physics. Experimental results confirm that the proposed MPF significantly reduces the prediction noise while consistently maintaining stable output for potential long-term HRI scenarios. We plan to continue improving this framework (continuously adopting state-of-the-art vision modules) for future work and making it available to the HRI research community to eliminate one of the real challenges in deploying interactive robots.

REFERENCES

- [1] Q. Zhou, C. Yu, Z. Wang, Q. Qian, and H. Li, “Instant-teaching: An end-to-end semi-supervised object detection framework,” in *2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 4081–4090.

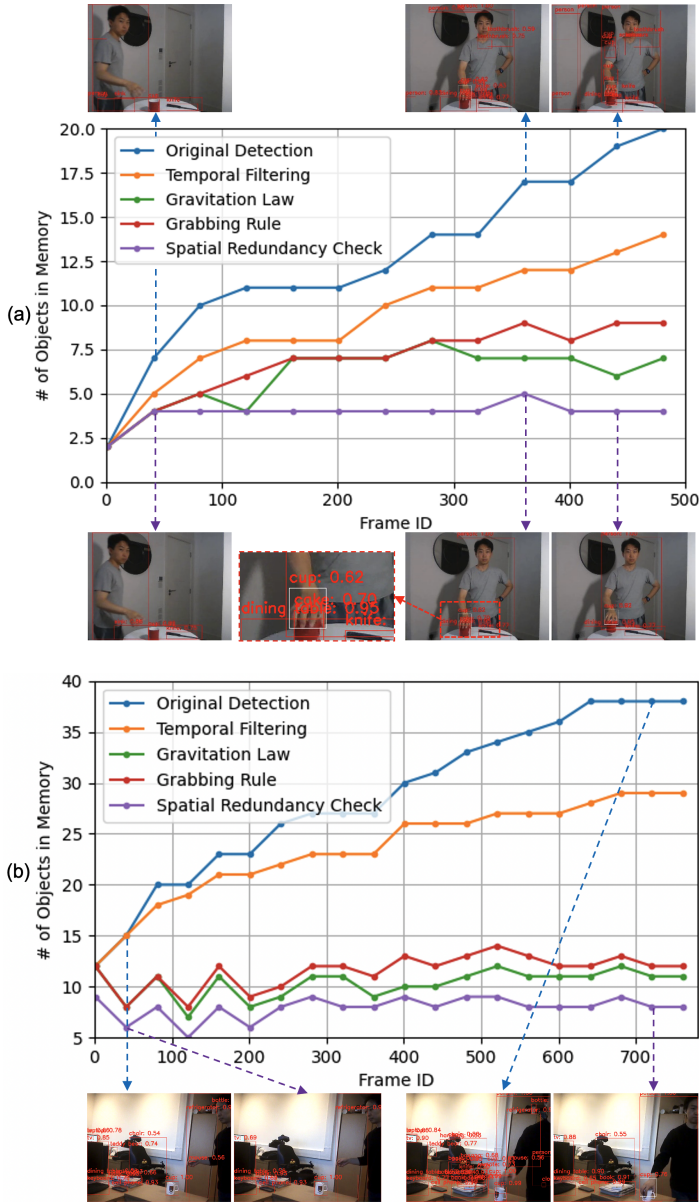


Fig. 4. Effect of using the proposed message passing framework: The noisy object detection output still produces noisy memory results, but soon recovers and stabilises the memory of MPF. (a) There are four objects (table, person, cup, knife) in the kitchen scene. (b) In the scene of the office, eight objects (table, tv, chair, keyboard, mobile phone, book, cup, and person) were kept stable in memory.

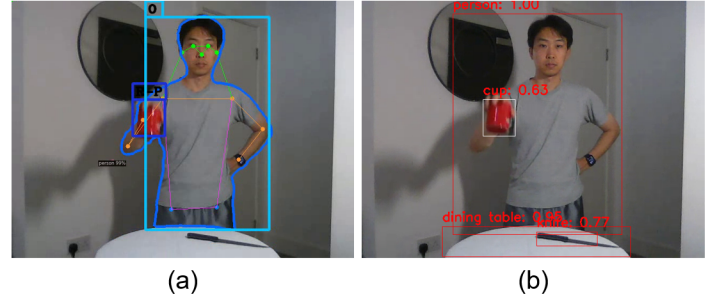


Fig. 5. Example of hand-object contact state stabilisation result: (a) The original hand object state estimation fails (the identified object here is the person) because it relies on the Region Proposal Network, which fails when motion blur occurs and (b) Because the proposed MPF stabilises the object detection output and stores it in the memory of the proposed framework, it improves the target object identification as well as the hand-object contact state estimation result. White bounding box represents the identified grasped object.

[2] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, and L. Zhang, "Dynamic head: Unifying object detection heads with attentions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 7373–7382.

[3] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-yolov4: Scaling cross stage partial network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 13 029–13 038.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[5] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, December 2015, pp. 91–99.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision*, 2016. [Online]. Available: <http://arxiv.org/abs/1512.02325>

[7] D. Kim, K. I. Kim, and S. Baek, "End-to-end detection and pose estimation of two interacting hands," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[8] S. Bambach, S. Lee, D. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[9] F. Xiong, B. Zhang, Y. Xiao, Z. Cao, T. Yu, J. T. Zhou, and J. Yuan, "A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[10] W. Huang, P. Ren, J. Wang, Q. Qi, and H. Sun, "Aw: Adaptive weighting regression for 3d hand pose estimation," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[11] S. Baek, K. I. Kim, and T.-K. Kim, "Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] D. Tang, T.-H. Yu, and T.-K. Kim, "Real-time articulated hand pose estimation using semi-supervised transductive regression forests," December 2013.

[13] D. Shan, J. Geng, M. Shu, and D. Fouhey, "Understanding human hands in contact at internet scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, and B. Gong, "Movinets: Mobile video networks for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 16 020–16 030.

[15] L. Wang, Z. Tong, B. Ji, and G. Wu, "Tdn: Temporal difference networks for efficient action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 1895–1904.

[16] A. Furnari and G. M. Farinella, "Rolling-unrolling lstrms for action anticipation from first-person video," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2020.

[17] H. Li, Y. Cai, and W.-S. Zheng, "Deep dual relation modeling for egocentric interaction recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[18] M. Zolfaghari, K. Singh, and T. Brox, "ECO: efficient convolutional network for online video understanding," in *ECCV*, 2018.

[19] S. Singh, C. Arora, and C. Jawahar, "Trajectory aligned features for first person action recognition," *Pattern Recognition*, vol. 62, pp. 45–55, 2017.

[20] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017*, pp. 4724–4733.

[21] R. Grimm, M. Grotz, S. Ottenhaus, and T. Asfour, "Vision-based robotic pushing and grasping for stone sample collection under computing resource constraints," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[22] M. Kokic, D. Kragic, and J. Bohg, "Learning task-oriented grasping from human activity datasets," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3352–3359, 2020.

[23] Y. Lin, C. Tang, F.-J. Chu, and P. A. Vela, "Using synthetic data and deep networks to recognize primitive shapes for object grasping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 10 494–10 501.

[24] H. Zhang, X. Lan, S. Bai, X. Zhou, Z. Tian, and N. Zheng, "Roi-based robotic grasp detection for object overlapping scenes," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 4768–4775.

- [25] Z. Zhou, T. Pan, S. Wu, H. Chang, and O. C. Jenkins, "Glassloc: Plenoptic grasp pose detection in transparent clutter," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 4776–4783.
- [26] M. Gualtieri and R. Platt, "Robotic pick-and-place with uncertain object instance segmentation and shape completion," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1753–1760, 2021.
- [27] T. Ohashi, Y. Ikegami, K. Yamamoto, W. Takano, and Y. Nakamura, "Video motion capture from the part confidence maps of multi-camera images by spatiotemporal filtering using the human skeletal model," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 4226–4231.
- [28] D. M. Chan and L. D. Riek, "Unseen salient object discovery for monocular robot vision," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1484–1491, 2020.
- [29] R. Araki, T. Onishi, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Mtdssd: Deconvolutional single shot detector using multi task learning for object detection, segmentation, and grasping detection," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 10487–10493.
- [30] H. Sun, Z. Meng, P. Y. Tao, and M. H. Ang, "Scene recognition and object detection in a unified convolutional neural network on a mobile manipulator," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 5875–5881.
- [31] A. Sabater, L. Montesano, and A. C. Murillo, "Robust and efficient post-processing for video object detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10536–10542.
- [32] Y. Liu, L. Liu, H. Zhang, H. Rezatofighi, Q. Yan, and I. Reid, "Meta learning with differentiable closed-form solver for fast video object segmentation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 8439–8446.
- [33] E. Langer, T. Patten, and M. Vincze, "Robust and efficient object change detection by combining global semantic information and local geometric verification," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 8453–8460.
- [34] A. Dehban, J. Borrego, R. Figueiredo, P. Moreno, A. Bernardino, and J. Santos-Victor, "The impact of domain randomization on object detection: A case study on parametric shapes and synthetic textures," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 2593–2600.
- [35] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf, "Dropout sampling for robust object detection in open-set conditions," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 3243–3249.
- [36] J. Young, L. Kunze, V. Basile, E. Cabrio, N. Hawes, and B. Caputo, "Semantic web-mining and deep vision for lifelong object discovery," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 2774–2779.
- [37] N. Massouh, F. Babiloni, T. Tommasi, J. Young, N. Hawes, and B. Caputo, "Learning deep visual object models from noisy web data: How to make it work," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 5564–5571.
- [38] E. Matheson, R. Minto, E. G. G. Zampieri, M. Faccio, and G. Rosati, "Human–robot collaboration in manufacturing applications: A review," *Robotics*, vol. 8, no. 4, 2019. [Online]. Available: <https://www.mdpi.com/2218-6581/8/4/100>
- [39] A. Zacharaki, I. Kostavelis, A. Gasteratos, and I. Dokas, "Safety bounds in human robot interaction: A survey," *Safety Science*, vol. 127, p. 104667, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925753520300643>
- [40] J. Wu, I. Yildirim, J. J. Lim, W. T. Freeman, and J. B. Tenenbaum, "Galileo: Perceiving physical object properties by integrating a physics engine with deep learning," in *NIPS*, 2015.
- [41] Z. Liu, W. T. Freeman, and J. B. Tenenbaum, "Physical primitive decomposition," in *ECCV*, 2018.
- [42] Z. Jia, A. C. Gallagher, A. Saxena, and T. Chen, "3d reasoning from blocks to stability," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 905–918, May 2015.
- [43] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [44] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.
- [45] S. Katyara, F. Ficuciello, F. Chen, B. Siciliano, and D. G. Caldwell, "Vision based adaptation to kernelized synergies for human inspired robotic manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.
- [46] A. Ošep, W. Mehner, P. Voigtlaender, and B. Leibe, "Track, then decide: Category-agnostic vision-based multi-object tracking," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3494–3501.
- [47] T. T. Pham, T.-T. Do, N. Sünderhauf, and I. Reid, "Scenecut: Joint geometric and object segmentation for indoor scenes," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 3213–3220.
- [48] K. Li, X. Zhang, K. Li, and S. Zhang, "Vision global localization with semantic segmentation and interest feature points," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 4581–4587.
- [49] M. Janner, S. Levine, W. T. Freeman, J. B. Tenenbaum, C. Finn, and J. Wu, "Reasoning about physical interactions with object-oriented prediction and planning," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: <https://openreview.net/forum?id=HJx9EhC9tQ>
- [50] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99. [Online]. Available: <http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf>
- [51] M. Shimrat, "Algorithm 112: Position of point relative to polygon," *Communications of the ACM*, vol. 5, no. 8, Aug. 1062.
- [52] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The icub humanoid robot: an open platform for research in embodied cognition," in *The 8th Workshop on Performance Metrics for Intelligent Systems*, 2008, pp. 50–56.
- [53] M. Contributors, "Openmmlab pose estimation toolbox and benchmark," <https://github.com/open-mmlab/mmpose>, 2020.
- [54] F. Pezosa, J. L. Reutter, F. Suarez, M. Ugarte, and D. Vrgoč, "Foundations of json schema," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 263–273.