Reliable Vision-Based Grasping Target Recognition for Upper-limb Prostheses: Appendix

This appendix includes additional experiment details and results. Figure 1 shows the uncertainty measures in the *synthetic trajectory* (SyTj) dataset for Bayesian GRU. Figure 2 contains the reliability diagrams for Bayesian GRU. Figure 3 and 4 show several example frames in our simulation datasets. Section I contains the experiment details and the results of incorporating an additional sensing modality into our framework. The generalization experiment is in Section II.

I. Incorporating Sensing Modalities

This experiment demonstrated our framework's ability to incorporate additional sensing modalities for prediction. We calculated the velocity (in xyz directions) of the arm and concatenated it with the features from the second last layer—the last layer (the *Softmax* layer) generated predictions using both the vision and velocity information. We evaluated our framework on the *Human Grasping Trajectory* (GrTj) dataset which contained realistic human grasping trajectories.

Figure 5 presents the results with different training data sizes. Due to the high variability of arm motions, the framework required more training data to learn meaningful velocity patterns. As a result, velocity information lowered down the performance of the framework using insufficient training data (subplot (a)-(b)). When the training data size was increased (subplot (c)), the performances of the framework with the velocity information were similar to the ones without velocity information. As shown in subplot (b)-(c), the velocity information was beneficial for BMLP if a low NPC (e.g. NPC < 2) was desired. This is a preliminary testing of the extension of our vision framework, and deeper investigation of sensor fusion is left as future work.

II. Generalization Capability

To demonstrate the generalization capability of our approach, we applied our framework to a grasp classification task with the datasets published in [1]: the *ImageNet* dataset and the *HandCam* dataset. The datasets contained object images labeled with five types of appropriate grasps: power, three-jaw chuck, tool, pinch and key. Figure 6 (a) presents several example images in the datasets. The *HandCam* dataset was collected with a camera on a prosthetic hand. We trained our framework purely with the *ImageNet* dataset and evaluated it on the *HandCam* dataset (unseen objects).

With the same datasets and tasks, DeGol et al. [1] achieved 93.2% grasp classification accuracy while our framework (BMLP) achieved 91.2%. Our accuracy was slightly lower because we utilized a less powerful but more efficient pre-trained network MobileNetV2 [2] while DeGol et al. [1] used VGG-VeryDeep-16 [3]. Figure 6 (b) shows the histograms of the first principle component of the three uncertainty measures, shown for both correctly classified and incorrectly classified samples. The results indicate promising capability in detecting potential mistaken predictions. We did not evaluate the probability calibration because the datasets were too small to perform a valid evaluation.

References

[1] J. DeGol, A. Akhtar, B. Manja, and T. Bretl, "Automatic grasp selection using a camera in a hand prosthesis," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 431–434.

[2] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

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

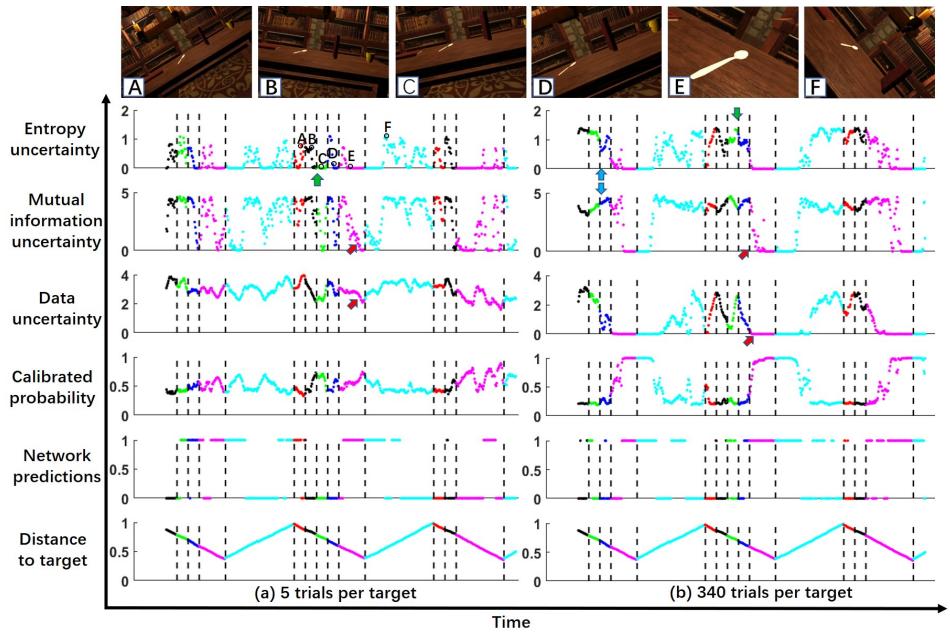

Fig. 1. Uncertainty measures in *synthetic trajectory* (SyTj) dataset with different training data sizes: (a) 5 trials per target; and (b) 340 trials per target. The six snapshots on the top are examples of what the camera saw during different segments. The first three rows of the scatter plots are the three measurements of uncertainty. The forth row is the calibrated probability from the probability calibration network. The fifth row is the network's target recognition result, 1 indicates the prediction was correct and 0 indicates an incorrect prediction. Since we assign labels to the entire trajectory, we do not expect the model to make correct predictions for the trajectory segments that are ambiguous (i.e. Segments A,B,C,D,F). The last row is the distance from the camera to the target. The lower this value is the closer the camera was to the target. The result is based on the *Clean* testing dataset with Bayesian GRU (BGRU) as the model.

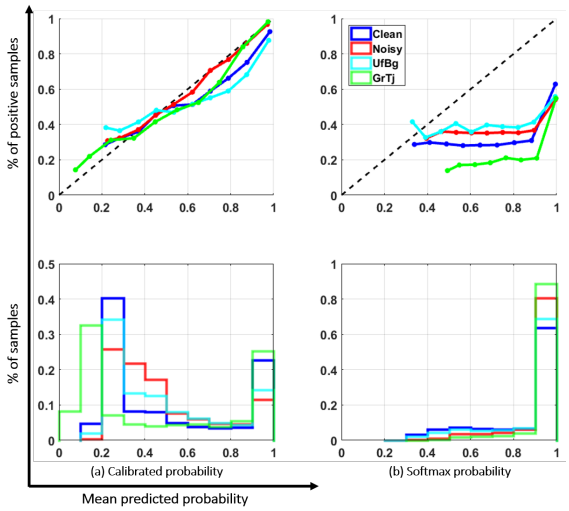

Fig. 2. Confidence histograms (bottom) and reliability diagrams (top) for calibrated probability (left) and Softmax probability (right). The results are based on the Bayesian GRU (BGRU) model with sufficient training data (340 trials per target).

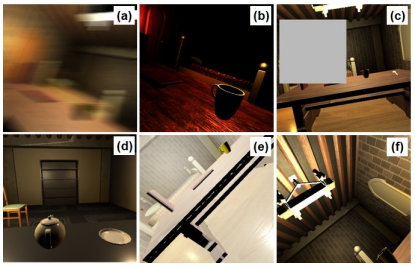

Fig. 3. Example frames of challenging scenarios: (a) motion blur, (b) dim light, (c) occluded images, (d) undefined targets, (e) overexposure, (f) abnormal camera orientation.

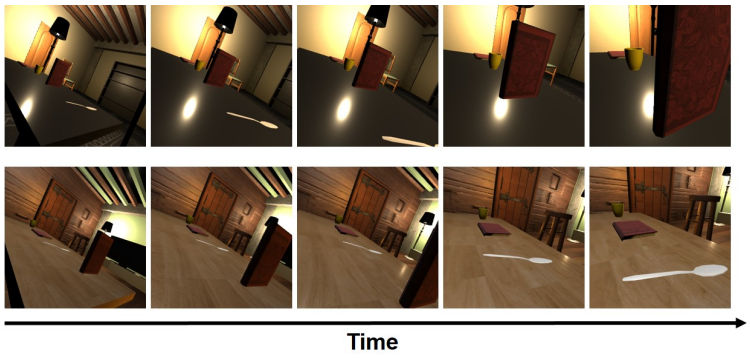

Fig. 4. Example frames in the *Human Grasping Trajectory Dataset*. The target object is the vertical book in the first row and the spoon in the second row.

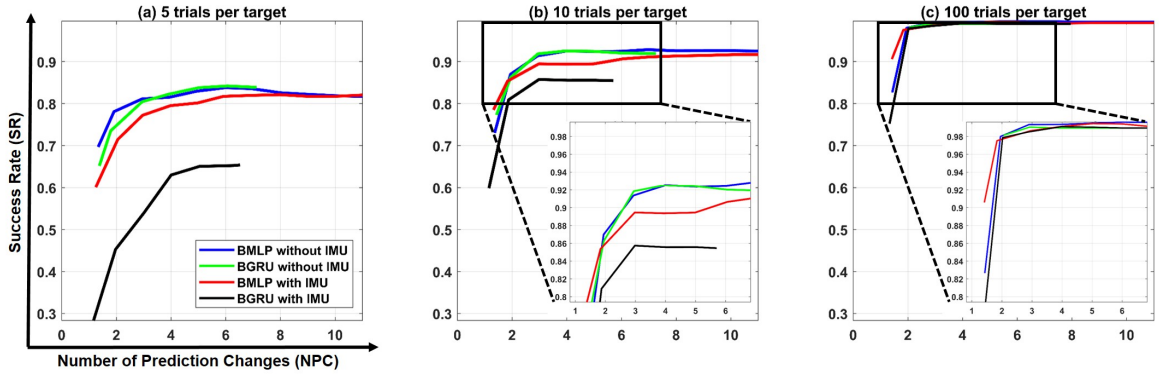

Fig. 5. The plot of Success Rate (SR) with respect to Number of Prediction Changes (NPC) for the *Human Grasping Trajectory Dataset*. The performance with different training data sizes are compared: (a) 5 trials per target; (b) 10 trials per target; (c) 100 trials per target.

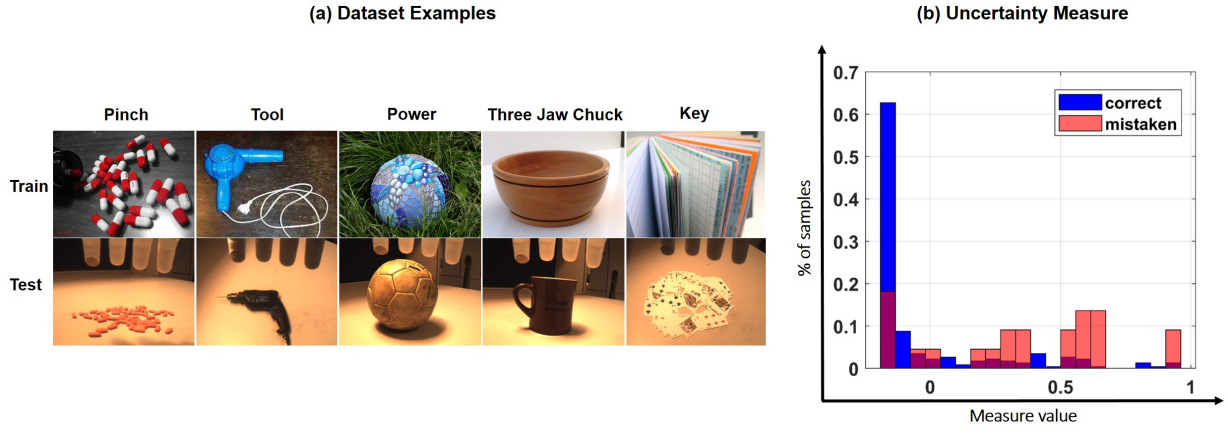

Fig. 6. (a) Example images in the grasp selection dataset. The top and bottom rows were used for training and testing respectively. (b) The histogram of the first principle component of the three uncertainty measures, shown for both correctly classified and incorrectly classified samples. This shows the potential for separability between the correct and incorrect distributions.