Forecasting Hand Gestures for Human-Drone Interaction

Jangwon Lee, Haodan Tan, David Crandall, Selma Šabanović
School of Informatics, Computing, and Engineering
Indiana University
Bloomington, Indiana
{leejang,haodtan,djcran,selmas}@indiana.edu

ABSTRACT

Computer vision techniques that can anticipate people's actions ahead of time could create more responsive and natural human-robot interaction systems. In this paper, we present a new human gesture forecasting framework for human-drone interaction. Our primary motivation is that despite growing interest in early recognition, little work has tried to understand how people experience these early recognition-based systems, and our human-drone forecasting framework will serve as a basis for conducting this human subjects research in future studies. We also introduce a new dataset with 22 videos of two human-drone interaction scenarios, and use it to test our gesture forecasting approach. Finally, we suggest follow-up procedures to investigate people's experience in interacting with these early recognition-enabled systems.

CCS CONCEPTS

• **Human-centered computing** → **Gestural input**; • **Computer systems organization** → *Robotic autonomy*;

KEYWORDS

Human-Drone-Interaction, Early recognition, Gesture recognition

ACM Reference Format:
Jangwon Lee, Haodan Tan, David Crandall, Selma Šabanović. 2018. Forecasting Hand Gestures for Human-Drone Interaction. In *HRI '18 Companion: 2018 ACM/IEEE International Conference on Human-Robot Interaction Companion, March 5–8, 2018, Chicago, IL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3173386.3176967

1 INTRODUCTION

Early recognition is an emerging computer vision research area that tries to forecast future events as early as possible [2, 4]. The objective is to build intelligent systems that can anticipate what *will* happen based on current observations and prior knowledge. Early recognition of human activities may be important for natural human-robot interaction since it enables robots to respond quickly to a human partner [3, 4]. Most papers so far present just the technical approaches without investigating the effects of early recognition on human-robot interaction using actual subjects. Our goal in this paper is to develop a practical early recognition system in a constrained (but nevertheless interesting and challenging) scenario that can serve as a basis for future such human studies.

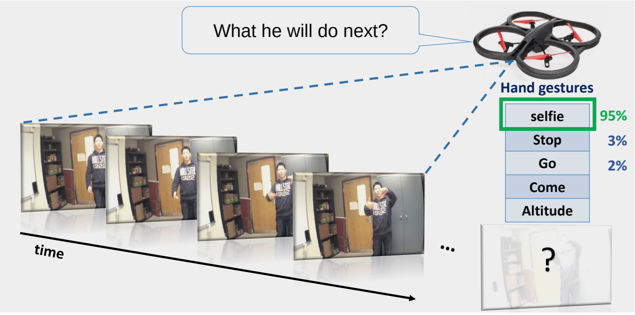

Figure 1: We propose a technique allowing a drone to forecast future hand gestures (1 second ahead of time). We introduce a new human-drone interaction dataset and use it to test our approach on five hand gestures in two application scenarios (taking a selfie and performing a delivery).

In particular, we present a novel vision-based approach to forecast human hand gestures in a Human-Drone Interaction (HDI) application. Our technique builds on the activity learning system proposed by Lee and Ryoo [3]. We introduce a dataset of humans interacting with a Parrot AR.Drone 2.0 through gestures, and use it to train our activity learning approach. We consider two interaction scenarios with different interaction distances: (1) drone delivery, in which the person directs a drone to deliver a small object, and (2) self portrait, in which the person directs the drone to take a photo. We define five hand gestures (selfie, stop, come, go, change altitude) for the human to use to command the drone, since gestures are a natural interaction modality [1]. Finally, we experimentally confirm that our approach enables a drone to forecast future human gestures and suggest follow-up work to investigate people's attitudes towards early recognition using the proposed system.

2 APPROACH

Our goal is to accurately identify a gesture one second before it is actually completed, based just on the first few video frames of the gesture. To do this, we employ a deep learning technique based on the fully convolutional future regression network proposed by Lee and Ryoo [3]. Fig. 2 illustrates our adaption of that network for our problem. Very briefly, the approach consists of three deep neural networks. The first row extracts visual features from the frames observed so far, the second row uses these to predict the visual features of a frame *one second in the future*, and then the third

HRI '18 Companion, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-5615-2/18/03.
https://doi.org/10.1145/3173386.3176967

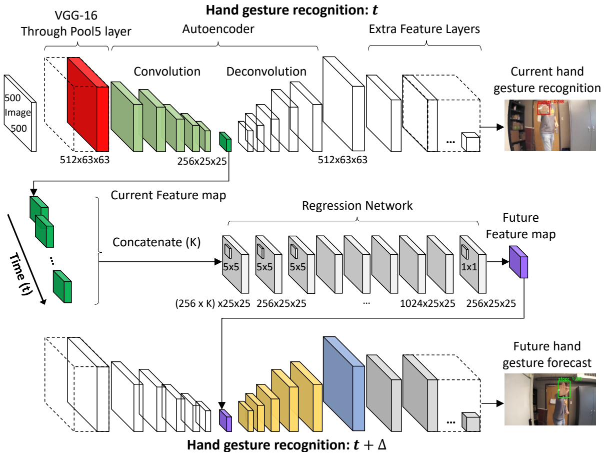

Figure 2: Data flow of our early gesture recognition systems. We employ the future regression network proposed by Lee and Ryoo [3] to forecast a human partner's future gestures.

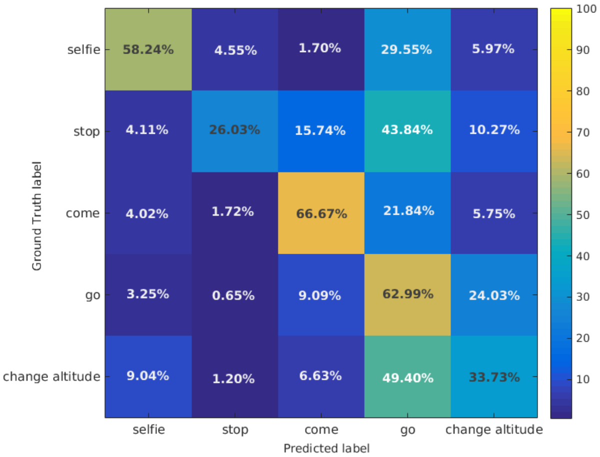

Figure 3: Confusion matrix of our gesture forecasting

row uses these predicted features to classify the gesture in that "hallucinated" frame. The network in the second row of Fig. 2 thus can be trained without gesture labels, although they are required to classify hand gestures (first row, see below and [3] for details).

Human-Drone Interaction (HDI) Videos. We train the network on a new dataset of human-drone interaction videos consisting of 5 participants in two human-drone interaction scenarios (drone delivery and taking a self-portrait). We used the front-facing camera of a Parrot AR.Drone 2.0 for collecting the videos with a resolution of 1280 × 720 at 30fps. Both scenarios were one-on-one interactions in which participants were told to direct the drone using five pre-defined hand gestures (selfie, stop, come, go, and change altitude). Each participant had 2-3 opportunities to interact with the drone for each scenario, and each interaction lasted about 1-3 minutes, yielding a total of 22 videos with 57,097 frames. We manually annotated 3,020 frames with ground-truth gesture labels, creating around 600 frames per gesture for training the network.

3 EXPERIMENTAL RESULTS

Hand Gesture Forecasting. We first trained the gesture recognition system on our HDI videos to recognize hands in current frames, using only the annotated frames and randomly spliting them into training (2,014 frames) and test (1,006). This achieved 99.1% accuracy in recognizing the current gesture on the test set.

Once the network in the first row of Fig. 2 had been trained, we used it to extract scene representations from all frames of the videos. We then trained the future regression network in the second row of Fig. 2 using the extracted scene features. Since this training process does not require any ground truth labels, we used all frames of the videos (two-thirds for training and the rest for evaluation). In the test phase, the regression network was coupled with the hand gesture recognition network (third row) to predict *future* hand gestures given observations.

Fig. 3 shows the confusion matrix of our hand gesture forecasting. We observe that our approach gives about 53.86% accuracy in predicting future human hand gestures one second ahead of time (compared to a random baseline of 20.0%), although it performed poorly on forecasting the 'stop' gesture. This result may be explained by the fact that participants used the 'stop' gesture in many different situations (i.e., when the drone was moving too far or approaching too close). The system also often confused 'change altitude' with the 'go' gesture, since pre-motions of these two gestures look similar.

4 DISCUSSION AND CONCLUSION

The specific objective of this study is to prepare an early gesture recognition system for conducting a user study to determine the effects of early recognition for human-drone interaction. Our experimental results confirmed that our gesture recognition system successfully predicts future human hand gestures (1 second in advance) for two human-drone interaction scenarios over half of the time. Since the gesture recognition system was trained on a dataset collected by a drone, we also can overcome the knowledge transfer problem and easily generate drone control commands according to a human partner's gestures. This research will serve as a base for future studies to investigate the effects of early recognition with human subjects for human-drone interaction. Future research should therefore concentrate on investigating how people respond to a drone during interaction, comparing two different conditions (with early recognition vs. without early recognition).

Acknowledgments. This work was supported in part by the Air Force Office of Scientific Research under award FA9550-13-1-0225, and by the National Science Foundation (CAREER IIS-1253549).

REFERENCES

[1] Jessica R Cauchard, Kevin Y Zhai, James A Landay, et al. 2015. Drone & me: an exploration into natural human-drone interaction. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 361–365.

[2] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. 2012. Activity forecasting. In *European Conference on Computer Vision (ECCV)*. 201–214.

[3] Jangwon Lee and Michael S Ryoo. 2017. Learning Robot Activities from First-Person Human Videos Using Convolutional Future Regression. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[4] MS Ryoo, Thomas J Fuchs, Lu Xia, Jake K Aggarwal, and Larry Matthies. 2015. Robot-centric activity prediction from first-person videos: What will they do to me?. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 295–302.