Robust Real-Time Hands-and-Face Detection for Human Robot Interaction

Sepehr MohaimenianPour and Richard Vaughan*

***Abstract*— We present a robust real-time systen for detecting hands and faces in RGB images. Starting with the YOLO Deep Convolutional Neural Network architecture and re-training with a mixture of third party datasets and our own original labelled data, we obtain qualitatively good results at *60Hz* on commodity GPU. We designed the detector to be a useful component for Human-Robot Interaction systems.**

I. INTRODUCTION

Humans frequently interact with each other using face engagement and hand gestures[1]. Several researchers have demonstrated Human-Robot interfaces that use either face detection, hand detection or both[2]. We present a fast and accurate detector that finds the hands and faces of multiple uninstrumented users in RGB images at frame-rate, which can be used as an input to an HRI system, either directly or through a tracker.

There is a need for this system since previous hand detection methods are either too slow[4] or use an RGBD camera[5] with limited operation distance and environment. Near-real-time multi-modal methods for hand detection and tracking[6] only work from very close range, in a controlled environment. Methods based on skin detection are not robust under illumination changes and when gloves are worn. Face detectors are common, but must be run separately[3]. Our system is a CNN based model that can be adapted to detect more objects and the network itself is re-sizable for speed/accuracy trade-off. We believe this is the first real-time hands-and-faces detector.

II. METHOD

A. Model

YOLO[7] is a state-of-the-art generic object detection system, in which a CNN predicts bounding boxes and probabilities for objects. It is very fast compared to competing systems. YOLO uses global context from the whole image to predict object bounds. This behaviour is ideal for our application and reduces false-positive rate as hand and face locations and mutual appearance are correlated with each other and with arm and torso location and appearance.

B. Dataset

We recorded a dataset consisting of 69 different length Human-UAV interaction scenarios ($16,883$ frames overall) and hand labelled $34,588$ hands and $18,838$ faces in them. We also used Mittal's hand detection dataset[4] by labelling the faces, and also selected some images from well known face detection datasets (WIDER, Helen, Faces in the Wild, FDDB, Pascal_VOC)[3] and labelled the hands in them. We slightly modified the YOLO graph, and, starting with the generic object detector weights, re-trained the network with hands and faces data.

*Autonomy Lab, School of Computing Science, Simon Fraser University
{smohaime, vaughan}@sfu.ca

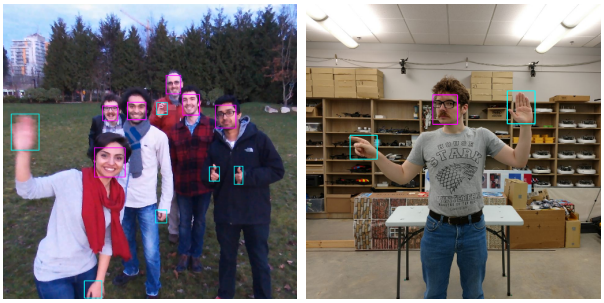

Fig. 1: Typical outputs: hands and faces detected accurately in $15msec$ per frame.

III. RESULTS AND FUTURE WORK

The detector gives qualitatively good results in video at $60FPS$ on a commodity GPU (NVIDIA GTX 970), or $15msec$ per image. Typical results are shown in Figure1. We have successfully used the detector with a real-time HRI system both with and without a tracker to compensate for occasional false negatives. The source-code of the system is available from our lab.

Our next goals are (i) to detect which hands belong to which face in images with multiple people; (ii) to track hands and faces in video directly using a recurrent network; (iii) optimize the network size so it runs at frame rates on CPU.

REFERENCES

[1] Monajjemi VM, Wawerla J, Vaughan R, Mori G. Hri in the sky: Creating and commanding teams of uavs with a vision-mediated gestural interface. In Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on 2013 Nov 3 (pp. 617-623).

[2] Rautaray SS, Agrawal A. Vision based hand gesture recognition for human computer interaction: a survey. Artificial Intelligence Review. 2015 Jan 1;43(1):1-54.

[3] Zafeiriou S, Zhang C, Zhang Z. A survey on face detection in the wild: past, present and future. Computer Vision and Image Understanding. 2015 Sep 30;138:1-24.

[4] Mittal A, Zisserman A, Torr PH. Hand detection using multiple proposals. In BMVC 2011 Sep (pp. 1-11).

[5] Tompson J, Stein M, Lecun Y, Perlin K. Real-time continuous pose recovery of human hands using convolutional networks. ACM Transactions on Graphics (ToG). 2014 Sep 23;33(5):169.

[6] Spruyt V, Ledda A, Philips W. Real-time, long-term hand tracking with unsupervised initialization. In Image Processing (ICIP), 2013 20th IEEE International Conference on 2013 Sep 15 (pp. 3730-3734).

[7] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. arXiv preprint arXiv:1612.08242. 2016 Dec 25.