DOI 10.31256/UKRAS19.31

An Embedded System for Real-Time 3D Human Detection

Haibin Cai, Lei Jiang, Junyi Wang, Mohamad Saada, and Qinggang Meng, Senior Member, IEEE

***Abstract*— Recent years have seen great achievements in the design of deep learning network structures and the construction of large benchmark datasets for object detection. However, it still remains a great challenge to achieve real-time performance when adapting to embedded systems with low computational ability. This paper proposes an embedded system for real-time 3D human detection. The system consists of a neural computing stick for the deploy of CNN, an intel RGBD sensor for the 3D sensing and a Raspberry Pi platform. Furthermore, a novel multi-thread based human detection framework is proposed to improve the detection speed. Experimental results show that the system can effectively detection human in real-time performance.**

***Index Terms*—Embedded System, Real-time, Human Detection.**

I. INTRODUCTION

The ability of automatically detecting the 3D location of human is crucial for many embedded robot systems. For example, an intelligent robot is expected to accurately detect the localization subjects and conduct interaction with them by recognizing their actions [1], understanding their intentions or simply follow their routines for further commands. To automatically follow a target, the robot also needs to sensing the 3D information of the environment to avoid collisions and plan the routines. Although great achievements have been made in both hardware and software technologies in the recent few decades, it still remains a great challenge for a low-cost embedded system to achieve accurate and real-time human detection performance.

There has been extensive research regarding the detection of objects in the literature. For example, Ren et al. [2] proposed the Faster R-CNN model which consists of a region proposal network (RPN) and a Fast R-CNN network [3] for object detection. The RPN shares convolutional features with the Fast R-CNN network and improves the detection speed and accuracy to a large extent. However, its detection speed is still far away from real-time performance for embedded platforms. Later, Liu et al. [4] proposed a single shot multi-box detector (SSD) which uses small convolutional filters to directly predict box offsets and relative category scores. The SSD achieves real-time performance with a modern GPU. By replacing all the regular convolution with depthwise separable convolutions, the MobileNet [5] greatly improves the detection speed and reduces the model size, paving the way for the deploy on embedded platforms.

This paper proposes an embedded system for real-time 3D human detection. Although there are some embedded platforms with GPU equipped on the market such as Nivida Jetson TX2, the expensive price hinders their board applications into the industry. Thus, this paper focuses on building a low-cost system for human detection. To this end, this proposed system utilized a raspberry pi 3B+, a neural computing stick for the deploy of light weighted convolution neural networks and the Intel RealSense sensor for the sensing of 3D environments. To make the system runs in real time, this paper further proposes a novel multi-thread based human detection framework where the data sensing, analyzing and display interactions are implemented into three separate threads.

The rest of the paper is organized as follows: Section 2 describes the detail of the proposed embedded system. Section 3 demonstrates the experimental results. Finally, section 4 summarized this paper with a conclusion.

II. THE PROPOSED EMBEDDED SYSTEM

This section firstly presents the hardware configuration of the proposed embedded system and then introduce the designed software framework.

A. Hardware Configuration

Fig. 1 shows the functionality of the three main components of the system. Released on 2017, the Intel Neural Compute Stick is a low-power consuming device that allows deploying light-weighted deep learning network. The tiny size and fanless design make it an ideal extension for most of the embedded systems. To sense the RGB and depth information of the objects, the Intel RealSense Depth Camera D415 is employed. Unlike previous RGBD sensors such as Microsoft Kinect and Asus Xtion, D415 can be used in both indoor and outdoor environments. Its valid sensing depth distance ranges from 0.16m to 10m, which is also a big improvement over previous RGBD sensors. The Raspberry Pi is a well-known embedded platform owing to its low price and stable performance. This paper utilizes the last product of Raspberry Pi 3 Model B+ for the control of the other sensors. The ARMv8 Cortex-A53 processor and 1GB LPDDR2 SDRAM Guaranteed its outstanding performance in most cases.

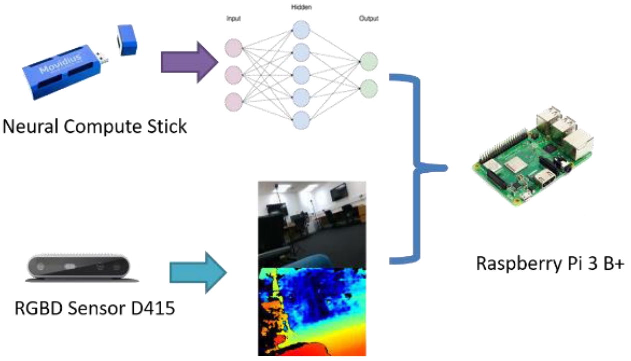


Fig. 1. The functionally of three main components of the system.

* This research is supported by YOBAN project funded by Newton Fund/Innovate UK, project number: 102871

H. Cai, L. Jiang, J. Wang, M. Saada, and Q. Meng are with the Computer Science, School of Science, Loughborough University, UK.

B. *Software Framework*

The software framework of the proposed embedded system is shown in Fig. 3. The system divides the functionalities into three groups and implement them in three different threads. The data retrieving thread continuously read RGB and Depth data from the RealSense D415 and update a shared memory space when new data arrives. The human detection thread takes the RGB data as input and conducts human detection on the intel neural computing stick. The detected bounding boxes serve as one of the inputs for the main thread, which is responsible for 3D human localization and display relative information on a screen for interaction purposes. The depth information of the human and the scene is obtained directly from the depth image in the data receiving thread. The transformation of depth data to the 3D human location in the RealSense D415's world coordinate is implemented via the following equation:

$$\frac{u-u_0}{f_x}=\frac{X}{Z}$$
$$\frac{v-v_0}{f_y}=\frac{Y}{Z} \qquad (1)$$
$$P\ =\ (X,Y,Z)$$

where P stands for a point in the world coordinate. (u,v) is its relative location in the color image whose resolution is set to be 640*480 to improve the detection speed. The intrinsic parameters of the color camera (fx, fy, u0, v0) are obtained via the classic chessboard-based calibration [6].

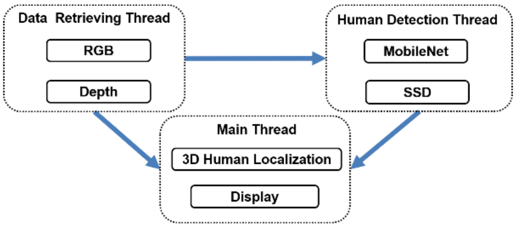


Fig. 3. The multi-thread human detection framework

To efficiently performance human detection, this paper utilized the MobileNet version of the SSD network [5]. The pre-trained model is directly used as it already includes human subjects as a target class. Trained on the coco dataset [7], this model can detect 90 classes of objects with good accuracy.

III. RESULTS

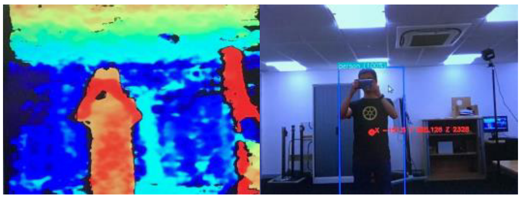
This section describes the experimental results for human detection on the embedded platform with a special focus on the speed performance.

Table 1 summarizes the time performance under different configurations. The average value of 1000 tests is used for each configuration. The first row shows the human detection speed using the Raspberry Pi 3B+ along. The 1.4 frames per second (FPS) makes it unsuitable for a robot to continuously follow a target. As shown in the second row, this system reaches around 7.3 fps when equipped with the Intel Neural Computing Stick. After further applying the proposed multi-thread framework, the system can detect human at 8.5 fps. It should also be noted that the system can still display the sensed images at 30 fps even with the low human detection speed due to the employment of the multi-thread framework. The drawback is that the detected bounding boxes will have a feeling of slow movement.

TABLE I. SPEED PERFORMANCE UNDER DIFFERENT CONFIGURATIONS

Configuration	Frames Per Second (FPS)
Pi	1.4
Pi + Stick	7.3
Pi + Stick + Multi-Thread	8.5

Fig. 4 shows an example of the human detection performance. The depth image was encoded as a color map for displaying purpose. The detected bounding box is shown directly in the RGB image. The location of the human is represented by the center point of the bounding boxes. Its 3D position is calculated via the Eq. 1.



(a) (b)

Fig. 4. An example of the human detection performance. (a) Depth image;(b) Color image.

IV. CONCLUSION

This paper describes an embedded system for real-time 3D human detection. By combining a low-cost Raspberry Pi platform, a neural computing stick and an RGBD sensor together, the proposed system can detect effectively detect human subjects. Furthermore, the proposed multi-thread human detection framework enables the system to run in real-time performance. Due to the employment of light-weighted network structures, this system has difficulty in detecting small size subjects. Thus, the further direction of this research will be on improving the network structure for a better human detection performance.

REFERENCES

[1] B. Liu, Z. Ju, and H. Liu, "A structured multi-feature representation for recognizing human action and interaction," Neurocomputing, vol. 318, pp. 287–296, 2018.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Trans. Pattern Analysis Machine Intelligence, 2017.

[3] R. Girshick, "Fast R-CNN," Proc. IEEE Int. Conf. on Computer Vision, vol. 2015 Inter, pp. 1440–1448, 2015.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," European Conference on Computer Vision, pp. 21–37, 2016.

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and C. Liang-Chieh, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2018.

[6] Z. Zhang, "A flexible new technique for camera calibration," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 11, pp. 1330–1334, 2000.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," European Conference on Computer Vision, 2014.