HHS Public Access

Author manuscript

IPSN. Author manuscript; available in PMC 2018 November 02.

Published in final edited form as:

IPSN. 2017 April ; 2017: 297–298. doi:10.1145/3055031.3055057.

Poster Abstract: 3D Activity Localization With Multiple Sensors

Xinyu Li,
Rutgers University, Piscataway, New Jersey

Yanyi Zhang,
Rutgers University, Piscataway, New Jersey

Jianyu Zhang,
Rutgers University, Piscataway, New Jersey

Shuhong Chen,
Rutgers University, Piscataway, New Jersey

Yue Gu,
Rutgers University, Piscataway, New Jersey

Richard A. Farneth,
Children's National Medical Center, Washington, District of Columbia

Ivan Marsic, and
Rutgers University, Piscataway, New Jersey

Randall S. Burd
Children's National Medical Center, Washington, District of Columbia

Abstract

We present a deep learning framework for fast 3D activity localization and tracking in a dynamic and crowded real world setting. Our training approach reverses the traditional activity localization approach, which first estimates the possible location of activities and then predicts their occurrence. Instead, we first trained a deep convolutional neural network for activity recognition using depth video and RFID data as input, and then used the activation maps of the network to locate the recognized activity in the 3D space. Our system achieved around 20cm average localization error (in a $4m \times 5m$ room) which is comparable to Kinect's body skeleton tracking error (10–20cm), but our system tracks activities instead of Kinect's location of people.

xl264@rutgers.edu, yz593@rutgers.edu, jz549@rutgers.edu, sc1624@rutgers.edu, yg202@rutgers.edu, rfarneth@childrensnational.org, marsic@rutgers.edu, rburd@childrensnational.org.

Keywords

Activity Recognition; Activity Tracking; Deep Learning; Passive RFID; Locolization; Computing methodologies; Activity recognition and understanding; Computer systems organization; Real-time system architecture

1 INTRODUCTION

Activity recognition is the fundamental building block of many advanced AI systems for applications like decision support and smart homes. Many existing systems focus only on recognizing activities from images or videos without providing important contextual information such as the 3D location of activities. Activity tracking could significantly improve the performance and value of many existing activity recognition systems (e.g., detecting the location of a fallen elder, or providing services based on activity in progress).

Most existing activity localization systems use a two-step approach: first predict an activity's location based on the locations of people who usually perform it or objects used in its performance, and then detect whether the activity occurred that location. This approach is slow and inaccurate, because detection is based on input data, most of which give only some of the information necessary for the prediction. We propose a reverse two-step approach using passive RFID data and Kinect depth frames: first recognize the activities with a multimodal convolutional neural network (ConvNet) and then localize the activities based on activation maps of the last pooling layers.

With the approval of the hospital's Institutional Review Board, we collected passive RFID data and depth video in a trauma room during 14 actual trauma resuscitations. The preliminary results show that this system achieved average activity localization error of 20cm.

2 APPROACH

Our general approach for activity recognition and localization is to first recognize activities with a deep ConvNet structure, and then use its activation maps to delineate the region in the input where activity takes place. Specifically, our system recognizes and localizes activities in three steps:

Step 1: Train a feedforward structure to recognize activities [1] (Fig. 1). Different input layers of a ConvNet should be designed to match different input data types, but all are processed by convolutional and pooling layers. We directly used the depth frame captured by Kinect and introduced the RSS map for RFID data representation. The *RSS map* projects the recorded RSS values to their antennas' effective "field of coverage", each antenna's working area can be represented as a circle on the room floor plan.

Step 2: Find the regions in the activation maps that contribute to the activity prediction decision making. [2]. We introduced the *backpropagated activation map*, which uses a backpropagation-like strategy to generate the weighted sum all the

activation maps in last pooling layer (Fig. 1). When the activity prediction is performed, the content of softmax layer can be represented as a binary vector, and we can reverse the feedforwarded expression and calculate the contribution of each neuron from the softmax layer to the flatten layer (Fig. 1):

$$x_j = \sum_{i=1}^{M} \frac{y_i - b_i}{w_{ij}} \quad (1)$$

where x_j denotes the *jth* neuron in previous layer and y_i denotes the *ith* neuron in the current layer (M neurons in total). The w_{ij} is the trained weight connecting the *ith* neuron in current layer and *jth* neuron in the previous layer, b_i is the neuron's bias term. The flatten layer can be reshaped into activation map in the pooling layer (Fig. 1).

Step 3: Project the activity region onto the input space using back propagated activation map, and match the overlapping regions generated by different sensors. Because sensors are prone to noise, e.g. the depth sensor can be affected by view occlusion and the passive RFID can be influenced by signal reflection induced by people movement. We used a multimodal depth-sensor and passive-RFID structure for localization of activities. Our proposed structure also works for different sensor combinations as well as a single sensor (Fig. 2).

3 PRELIMINARY RESULTS

We evaluated our system from two aspects: localization error and localization accuracy.

The estimated localization error denotes the average distance between the estimated and actual activity locations. The actual location can only be manually estimated. The proposed system achieved around 20cm average localization error (in a $4m \times 5m$ room) which is comparable to Kinect's body skeleton tracking error (10–20cm) and better than most RFID based localization system.

The activity localization accuracy denotes the percentage of points within a threshold distance (0.5 meters in this paper) of the actual activity location. We used 300 synchronized RFID RSS maps and depth frames of our medical data for testing, and calcu-lated the average accuracy and standard deviation of each activity localization (Fig. 3).

ACKNOWLEDGMENTS

This research was supported by the National Institutes of Health under Award Number R01LM011834.

REFERENCES

[1]. Li Xinyu, Zhang Yanyi, Li Mengzhu, Chen Shuhong, Austin Farneth R, Mar-sic Ivan, and Burd Randall S. 2016 Online process phase detection using multimodal deep learning. In Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), IEEE Annual. IEEE, 1–7.

[2]. Mahendran Aravindh and Vedaldi Andrea. 2016 Visualizing deep convolutional neural networks using natural pre-images. International Journal of Computer Vision 120, 3 (2016), 233–255.

CCS CONCEPTS

- Computing methodologies →Activity recognition and understanding
- **Computer systems organization** →Real-time system architecture

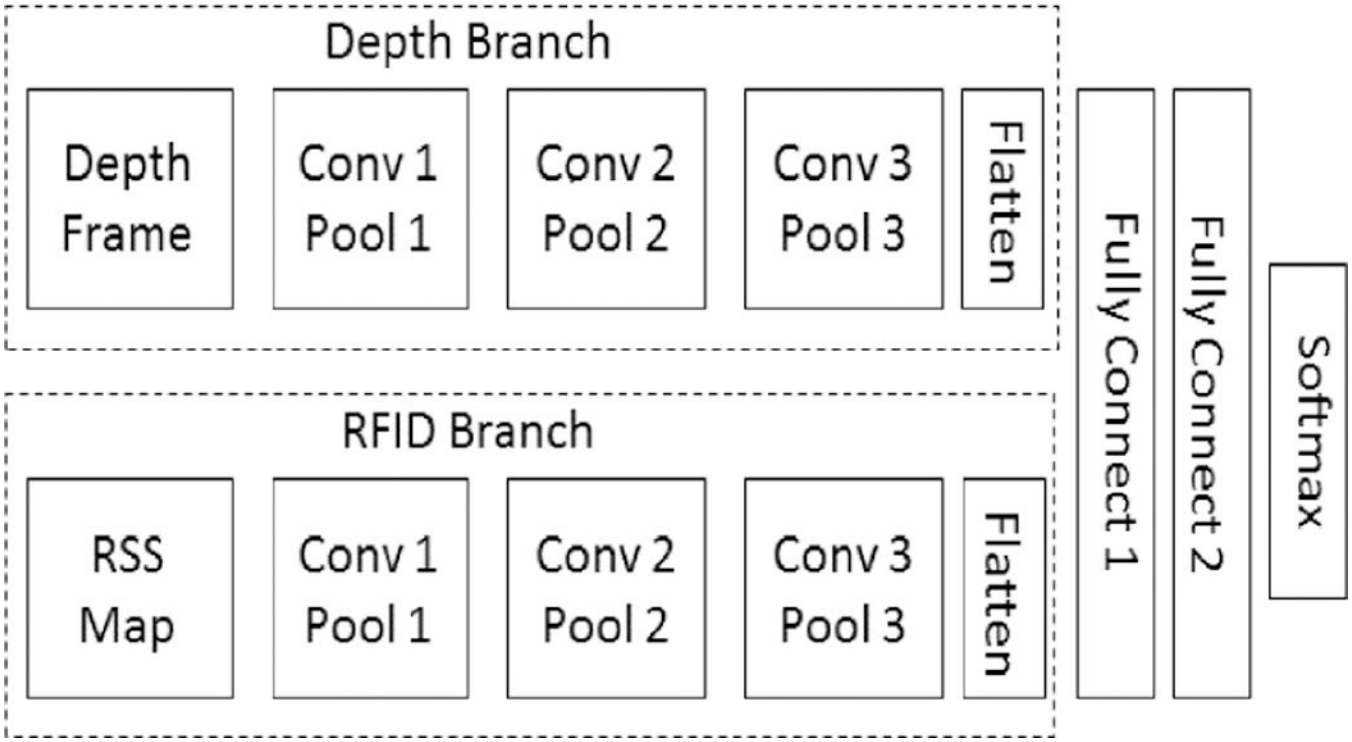

Figure 1:
The multimodal structure used for depth and RFID based activity recognition

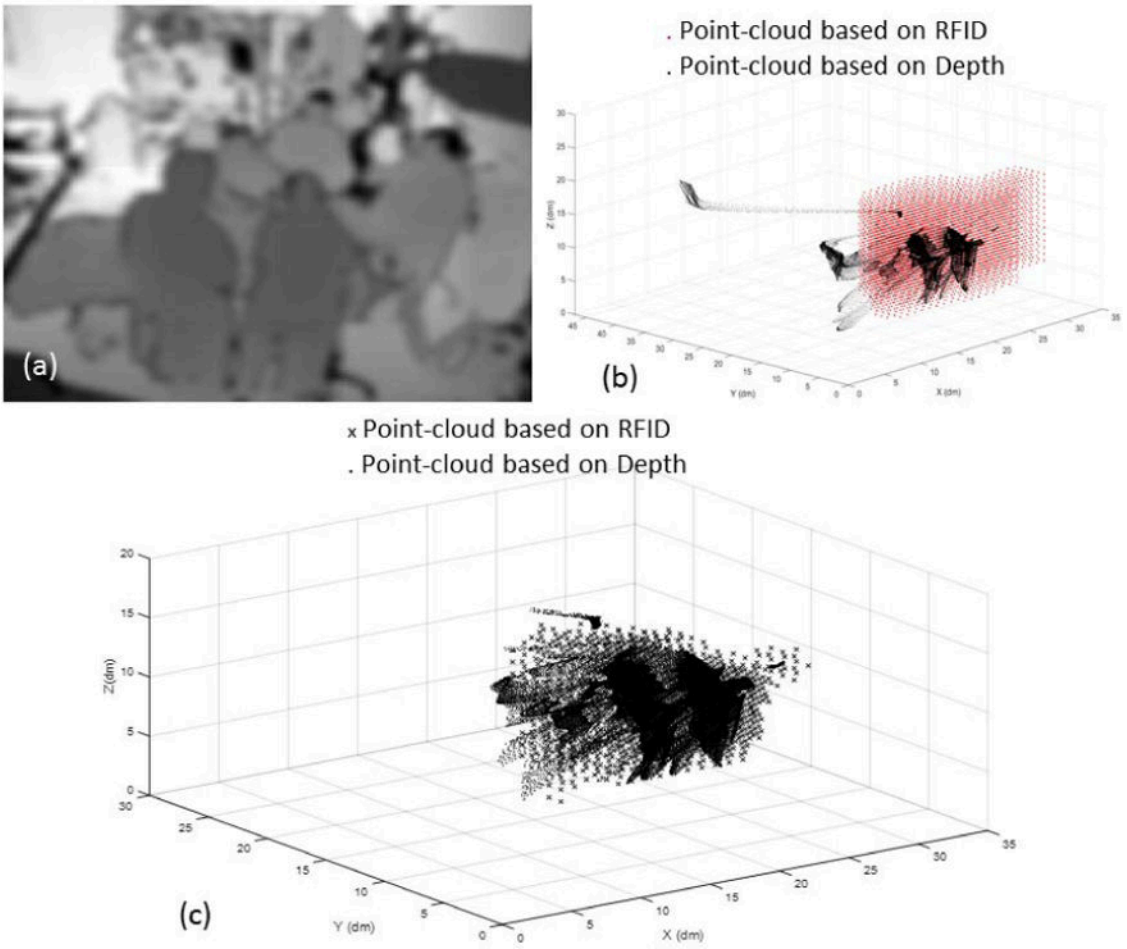

Figure 2:

(a) Ground truth depth image. (b) The activity point cloud based on RFID and depth data. (c) The point cloud representing the 3D location of an activity after combining them (Right). View digital version for colors.

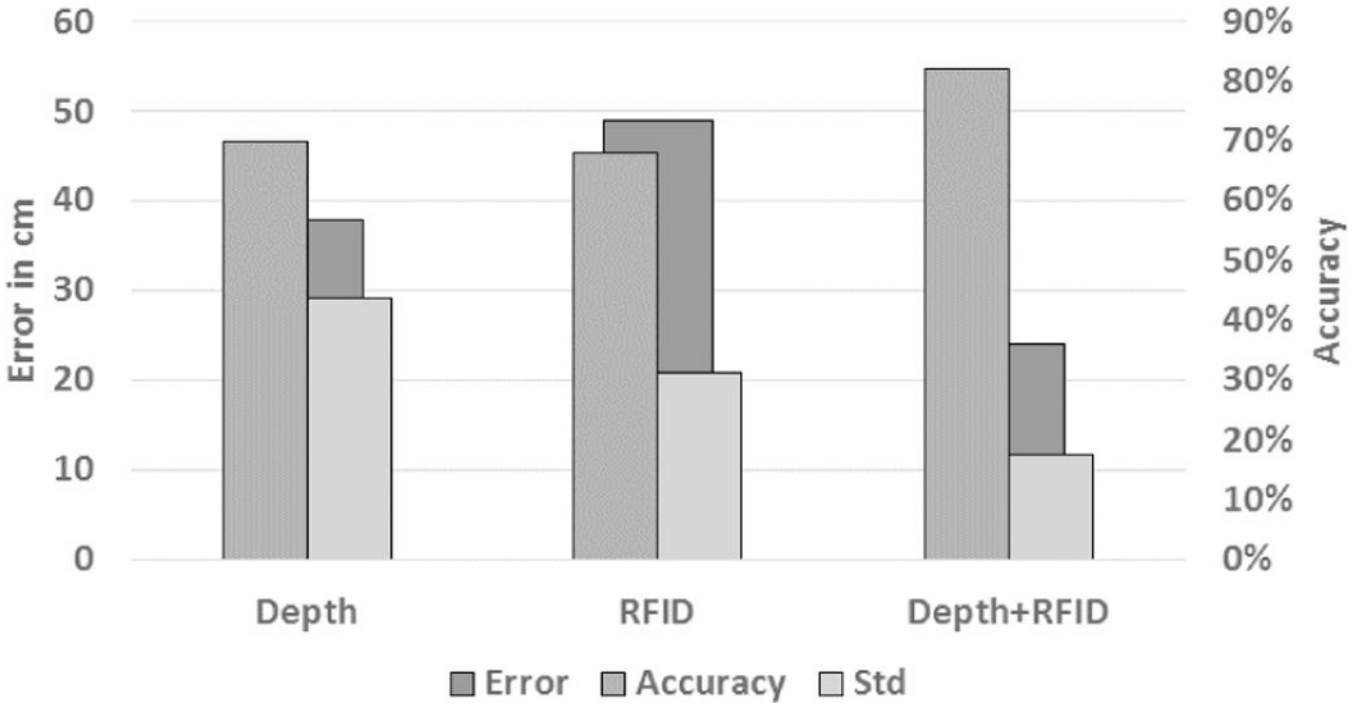

Figure 3:
The activity localization errors, localization accu-racy and standard deviation of system with different sensor types.