Poster Abstract: CAR - A Deep Learning Structure For Concurrent Activity Recognition

Yanyi Zhang
Rutgers University
Piscataway, New Jersey
yz593@rutgers.edu

Xinyu Li
Rutgers University
Piscataway, New Jersey
xl264@rutgers.edu

Jianyu Zhang
Rutgers University
Piscataway, New Jersey
jz549@rutgers.edu

Shuhong Chen
Rutgers University
Piscataway, New Jersey
sc1624@rutgers.edu

Moliang Zhou
Rutgers University
Piscataway, New Jersey
mz330@rutgers.edu

Richard A. Farneth
Children's National Medical Center
Washington, District of Columbia
rfarneth@childrensnational.org

Ivan Marsic
Rutgers University
Piscataway, New Jersey
marsic@rutgers.edu

Randall S. Burd
Children's National Medical Center
Washington, District of Columbia
rburd@childrensnational.org

ABSTRACT

We introduce the Concurrent Activity Recognizer (CAR) — an efficient deep learning structure that recognizes complex concurrent teamwork activities from multimodal data. We implemented the system in a challenging medical setting, where it recognizes 35 different activities using Kinect depth video and data from passive RFID tags on 25 types of medical objects. Our preliminary results showed our system achieved an 84% average accuracy with 0.20 F1-Score.

CCS CONCEPTS

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

KEYWORDS

Activity Recognition, Deep Learning, Multimodel, Passive RFID

ACM Reference format:
Yanyi Zhang, Xinyu Li, Jianyu Zhang, Shuhong Chen, Moliang Zhou, Richard A. Farneth, Ivan Marsic, and Randall S. Burd. 2017. Poster Abstract: CAR - A Deep Learning Structure For Concurrent Activity Recognition. In *Proceedings of The 16th ACM/IEEE International Conference on Information Processing in Sensor Networks, Pittsburgh, PA USA, April 2017 (IPSN 2017),* 2 pages.
DOI: http://dx.doi.org/10.1145/3055031.3055058

1 INTRODUCTION

Activity recognition has been studied for decades, focusing mainly on staged, single-person activity recognition. Some approaches tracked body joints to predict simple physical activities such as standing or laying down. Others used mobile sensors, such as wearable accelerometers, to track daily activities such as walking, driving and sleeping. Many real-world applications, however, produce data containing complex concurrent activities, with a significant view occlusion and sensor noise. Many of these existing systems cannot handle such complexity. To address the challenges of concurrent activity recognition, we built a multimodal system, which has an independent classifier for every activity and these classifiers share the same feature extraction module. The system was installed in a trauma room at a level-1 trauma center to predict 35 different medical activities. We used passive RFID and depth camera to avoid obtrusiveness and preserve privacy. The RFID and depth sensors complement each other in terms of the information they provide. The depth sensor compensates the RFID's dependence on tagged objects used in activities, and RFID compensates the depth sensor's limited visibility.

With the approval of the hospital's Institutional Review Board, we collected passive RFID data and depth video in a trauma room during 50 actual trauma resuscitations to run experiments with the system.

2 APPROACH

To improve feature extraction, we introduced a new RSS map data structure that explicates the spatial information and preserves redundancy in RFID data for subsequent recognition in ConvNets, and an expert-knowledge-based region proposal method that focuses feature learning on key areas of complex multi-target depth images. To achieve efficient recognition of concurrent activities, the system has a single feature extractor shared by multiple independent classifiers, one for each activity. Our system implements 5 steps (Fig. 1 dashed-line boxes):

IPSN 2017, Pittsburgh, PA USA
© 2017 ACM. 978-1-4503-4890-4/17/04…$15.00
DOI: http://dx.doi.org/10.1145/3055031.3055058

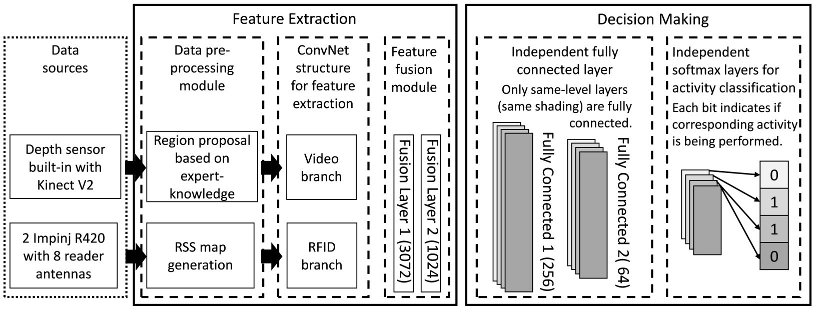
Data sources
Depth sensor built-in with Kinect V2
2 Impinj R420 with 8 reader antennas
Feature Extraction
Data pre-processing module
Region proposal based on expert-knowledge
RSS map generation
ConvNet structure for feature extraction
Video branch
RFID branch
Feature fusion module
Fusion Layer 1 (3072)
Fusion Layer 2 (1024)
Decision Making
Independent fully connected layer
Only same-level layers (same shading) are fully connected.
Fully Connected 1 (256)
Fully Connected 2(64)
Independent softmax layers for activity classification
Each bit indicates if corresponding activity is being performed.
0
1
1
0

Figure 1: The overall structure of our system for concurrent activity recognition

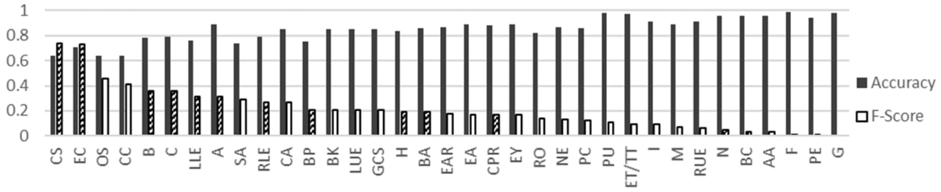

Figure 2: The recognition activity and F1-Score for 35 activities.

Step 1: We separated the depth images into three key regions: head, left and right of patient bed by asking medical experts to mark the most important regions they used for activity recognition when shown images. Since ConvNet prefers feature maps to be spatially related, we designed a novel representation of RFID data. 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: We built the model based on previous research [1]. We chose 3 convolutional layers to balance performance with the hardware constraints. We used "ReLU" as the activation function and added "dropout" to avoid overfitting.

Step 3: We designed a data fusion strategy for combining the extracted video and RFID features [2]. Considering that the RSS maps and proposed regions in depth frames are not related in a straightforward way, we chose to merge them through fully-connected fusion layers instead of stacking convolutional layers for each modality.

Step 4: We accomplished the final activity prediction through several *activity levels* of fully-connected layers, where each level is responsible for predicting the status of one activity and setting one bit of the activity code (Fig. 1). We had 35 levels (for 35 different activities) of network layers that made independent activity predictions in each time instance and achieved concurrent activity recognition. Each level in the decision-making structure shared the same input features from the fusion module.

Step 5: We implemented an individual softmax layer at the end of each activity level to produce a binary code in which each bit indicated whether the corresponding activity was currently being performed.

3 PRELIMINARY RESULTS

We used 40 trauma resuscitations for model training and the other 10 for testing. To evaluate our system performance, we first defined *recognition accuracy for concurrent activities* as the average accuracy of each activity across testing cases:

$$Accuracy = \frac{\sum_{i=1}^{N} TP_i + TN_i}{\sum_{i=1}^{N} TP_i + TN_i + FP_i + FN_i} \tag{1}$$

Our evaluation results showed 84% average accuracy with 0.20 F1-Score (Fig. 2). The striped bars indicate activities that used tagged objects, and the white bars indicate the activities without object use. We found that the activities with object use had better performance than activities without objects use, which indicates that RFID contributes independently of depth images for concurrent activity recognition. We also noticed the same activities happening on the left side of the patient bed tend to have better performance compared to those on the right, e.g. extremity assessment activities had better performance on the left side than on the right side. This is probably because the people at right side always face away from the camera, blocking the depth sensor's view of the key hand gestures.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

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