HW – GRAD ONLY

Submitted By: Nafis Neehal

RIN: 661990881

Paper Title: Making the Invisible Visible: Action Recognition Through Walls and Occlusions

Authors: Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, Dina Katabi (all of them are from MIT CSAIL)

Venue, Publishing Date: ICCV 2019

Citation and URL:

Li, Tianhong, et al. "*Making the Invisible Visible: Action Recognition Through Walls and Occlusions*", ICCV 2019, pp. 872–81. *openaccess.thecvf.com*, http://openaccess.thecvf.com/content_ICCV_2019/html/Li_Making_the_Invisible_Visible_Action_Recognition_Through_Walls_and_Occlusions_ICCV_2019_paper.html

Problem Definition:

The authors have proposed a neural network model that, with help of radio frequency (RF) signals, can detect human actions through occlusions such as walls and it also works in poor lightning conditions.

Core Contributions of the paper:

- First model proposed in which RF signals has been used for action recognition.
- Account for Occlusions and poor lighting conditions (this part is done solely using RF signals)
- Uses skeletons as an intermediate representation for knowledge transferring with empirical proof of this approach achieving better performance
- Introduces a new spatio-temporal self-attention module which helps in improving the performance of skeleton-based action recognition regardless of the skeleton generation procedure (Vision-based/RF)
- Proposes a novel multi-proposal module which detects simultaneous actions as well as interactions of multiple people

Algorithm Summary:

- In first layer, 3D skeleton of each subject is extracted from raw wireless signals -
    - Input in this layer is horizontal and vertical heatmaps
    - A spatio-temporal CNN extract features at first from RF signals
    - Output from the CNN is passed to a Region Proposal Network (RPN) for obtaining multiple proposals for several skeleton bounding boxes
    - Finally, a 3D pose estimation sub-network is used to extract 3D-skeletons from each of the skeleton bounding boxes proposed earlier
- In second layer, action detection and recognition on the extracted 3D skeleton sequences are performed -
    - Input to this layer 3D skeletons generated from the earlier layer
    - Each input skeleton sequence is a matrix of 4xTxNj where 4 is spatial dimensions and confidence, T is the number of frames in a sequence and Nj is number of keypoints in a skeleton. Action recognition is done using 3 modules.

    - First module is an attention-based feature learning network which extracts high level features from skeleton sequences
    - Second Module is for extracting proposal from the multi-proposal module. Two kinds of proposals are generated – single person actions, and multiple people interactions
    - Final module is to use the generated proposals to crop and resize the corresponding latent features and forward them to a classification network for training.

<u>Data Set:</u>

- RF-MMD (RF Multi-Modality Dataset):
    - It has 25 hours of data with 30 volunteers
    - 10 different environments
    - 20 hours of data from training, 5 for testing
- 35 action set (29 single actions and 6 interactions) from PKU-MMD (one of the existing vision-based 3D skeleton dataset)

<u>Experimental Protocols and Results:</u>

- The whole evaluation was done using mean average precision (mAP) at different IoU (intersection-over-union) thresholds θ
- Results for $\theta = 0.1$ and $\theta = 0.5$ were reported only
- Compared with HCN (SOTA for skeleton based action detection) and Aryokee (SOTA for RF based action recognition). RF-Action outperformed both (specially with a large margin while comparing with Aryokee) in both for visible scenes and occluded scenes.
- Comparison was also made to see which one was better between RF and Vision generated skeletons. From the result they achieved, it was inferred that RF Based action recognition system can achieve a performance close to a carefully calibrated camera system with 10 viewpoints. But RF based system didn’t perform significantly well for occlusion-based action prediction in compared to vision based.
- Effectiveness of the attention module was evaluated by reporting results by both using and without using the attention module on RF-MMD and PKU-MMD. The effectiveness of using attention module was particularly prominent in RF-MMD dataset, but not so much for the other one. It was also tested on the NTU-RGB+D dataset for classification only (detection excluded) and failed to show much impact on this dataset.
- While evaluating the multi-proposal module, the authors evaluated the overall network performance with and without using the multi-proposal module, and as a result showed that using this module improve the performance from 65.6 to 87.8. Although, the model did not perform well for single proposals as the RF-MMD dataset mostly contained multiple actions being performed in same image.
- The authors also experimented whether multimodal training achieved better accuracy comparing to single training dataset and empirically showed that training the model with both RF-Action dataset and PKU-MMD dataset ended achieving better performance as naturally it supplied with more training data.

<u>Strength:</u>

- Detection and classification work even if there is occlusion (and this work is the first of its kind)
- Uses an intermediate step (generating 3D skeleton) before final prediction which results in robustness of the model (generalizes to new environment easily)

- Model can learn from both RF-based datasets and vision-based datasets (more training data, better performance) which attributes to robustness of the model
- Can deal with multi-person multi-action scenarios in same image

Weakness:

- RF Signals, while travelling through walls, will attenuate, and as a result will tend to end up generating erroneous results! Authors have acknowledged this problem and stated that they dealt with this problem by calculating a time-varying confidence score for each skeleton joint. But they didn't give any detailed procedure about how they calculated this confidence score. There was also no mention of any error margin or any kind of threshold for calculating the confidence score.
- As RF signals that traverse through walls have lower spatial resolution, it becomes hard to distinguish between gestures that have almost same position (hand shaking/hair brushing – acknowledged by the authors). But they didn't provide any countermeasure for how to deal with these kinds of hard-to-classify gestures.
- In the dataset section, they acknowledged that their model was trained to detect and classify 35 actions in total. But no list of these actions was given and no separate experimental data on the performance of the model on these actions individually were given for better visualization and assessment.
- The performance of attention module was good in the RF-MMD dataset which the authors collected/built, but the module failed to generalize its performance in other datasets which leaves room for improvement.
- The model only worked better for multi-proposal module as the dataset was skewed with images of multiple interactions in same image.