Human Activity Recognition Based on Optimal Skeleton Joints Using Convolutional Neural Network Critique

Boyoung Zhao

From the abstract of the paper, the paper essentially recognizes that Human Action Recognition (HAR) with RGB-D cameras is a continuously growing field and goes over how RGB-D cameras can identify joints, 3D silhouettes, and skeletal body parts using convolutional neural networks.

The paper starts off with an introduction on how HAR can be applied to a variety of fields, such as health care, video surveillance, and even smart home systems. In addition, RGB-D cameras, such as the Kinect used in the paper, are relatively cheap and depth maps can be helpful in forming a human 3-D skeleton.

First, the paper goes into recent developments in HAR using depth maps and skeleton join points. It explains that depth maps are robust to light, color and text variations. There are many approaches, and extracting features can be difficult, especially with bad lighting conditions. Furthermore, interactions with background objects, such as chairs, can mess up the silhouette output, which in our case could become an issue.

The framework proposed by the paper incorporates a convolutional neural network (CNN) using deep learning to locate points on the skeleton. Essentially, the pipeline consists of an input network, feature learning process, and classification. Input of features are consisted of 15 joint points and a confidence value, which indicates the success rate of the skeletal configuration. The output would then extract high-level features to classify different types of actions.

During the feature learning process, the joints sometimes contributed very little changes for action recognition when the paper used the CAD 60 dataset for recognition (consisting of 12 different activities), and sometimes just brought additional noise. Therefore, the researchers used the Shannon Entropy formula to evaluate informative skeleton joints for activity (Equation 1 in the paper). The entropy formula represents the higher entropy related to a max contribution of relative joints, and therefore attempted to evaluate the skeletal points which would have the most significant impact on human activity by weighting more joints more heavily than others in calculating actions. For example, the hips, knee, and foot were not useful in determining whether a person was opening a pill.

As for the CNN, the input vector was a 3D vector with joint attributes (size 4), number of joints (size 15), and number of frames (30). The CNN was composed of five layers. The first two layers were convolutional layers, followed by a fully connected layer, and a softmax layer before the output. The first layer was meant to find the cumulative effects of a local filter passing through the image plane to find features across the joints and time plane. The second layer was an activation function which is used for neural networks to solve gradient problems, which incorporated the Rectified Linear Unit activation function (equation 2). This input, which is described as the max pooled over joint and frame dimensions, is then fed into a fully-connected layer the convolutional net, where afterwards softmax is performed and out determines recognized activity. This whole process was not entirely clear to me, and I think the paper could have done a better job explaining the convolutions they did because the wording used was difficult to understand.

After dividing 80% of the CAD60 data set for training and 20% for testing, the experiment was to evaluate the performance when using all joints, and evaluating performance using key joints. Overall, the data noted that the accuracy of detecting tasks when using all joints was 82.96%, while using only the important joints was 94.16%. Therefore, only using informative joints gave higher precision results compared to using all joint points.

In conclusion, the paper essentially contributes toward the recognition that some skeletal joint configurations are irrelevant when identifying human actions, and that the most informative joints approach was much more accurate.

Overall, the paper was very interesting in its findings, although there are many things that could have been better. Firstly, the paper did not explain convolutional neural networks well at all, and I had to turn to third party resources in order to fully gain the understanding that I did in this paper. Continuing off the CNN aspect of the paper, the paper also did not give reasons as to why it chose the kernel sizes for convolution (3 x 10) and why this first convolution is this size. The same goes for other convolutions. The paper does not explicitly state why kernel sizes were important and why they chose the specific sizes that they did. Furthermore, the authors did not explain why they used ReLU as opposed to other activation functions that could also work in the CNN.

Another critique I would add regarding the paper has to do with the Shannon entropy formula that was used, as well as the figure corresponding to it. It is extremely unclear how the formula was applied to joints, or how the “p” value was even calculated. No real explanation is given for readers to repeat the same procedure. In addition, Figure 3, shown below, is extremely difficult to interpret and the colors and bars are way too small for any human to get meaningful information from the figure.

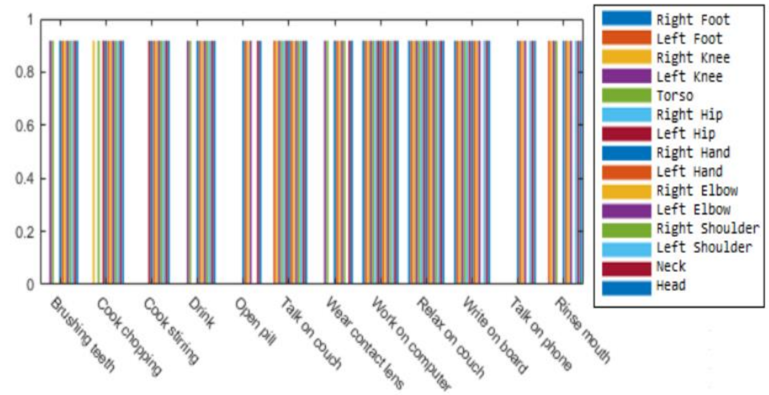

Fig. 3. Informative joints for CAD60 dataset using Shannon entropy.

Lastly, it would have been interesting if the paper compared their own results with results from other CNNs with different kernel sizes and activation functions to see whether the approach of using Shannon Entropy actually has an effect on HAR, as ultimately that is what the paper is trying to argue.

References

Xu, Wenchao, et al. “Human Activity Recognition Based On Convolutional Neural Network.” *2018 24th International Conference on Pattern Recognition (ICPR)*, July 2018, doi:10.1109/icpr.2018.8545435.