Journal of Engineering Science and Technology
Special Issue on ICCSIT 2018, July (2018) 48 - 57

HUMAN ACTIVITY RECOGNITION BASED ON OPTIMAL SKELETON JOINTS USING CONVOLUTIONAL NEURAL NETWORK

NOR SURAYAHANI SURIANI*, SITI NOOR FATIHAH AHMAD, MOHD NORZALI MOHD, MOHD RAZALI TOMARI, WAN NURSHAZWANI WAN ZAKARIA

Faculty of Electrical and Electronics Engineering,
University Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia
*Corresponding Author: nsuraya@uthm.edu.my

Abstract

Recognizing human actions is a challenging task and actively research in computer vision community. The task of Human Action Recognition (HAR) has been widely used in various application such as human monitoring in a hospital or public spaces. Previous work in HAR identified the human actions by the amount of joint movement, 3D silhouettes, local occupancy pattern and spatiotemporal interest features. This paper proposed a skeletal joint points configuration captured from Kinect RGB-D camera. Skeletal points have contributed significant features to discriminate different actions. However, some joint points are irrelevant and not moving which act as noise. This problem will reduce the performance of HAR. Therefore, this paper attempt to identify informative joint point using Shannon's entropy mechanism. Then, the features will be an input to the Convolution Neural Network (CNN) classifier. Experiment results using standard benchmark CAD60 dataset have shown that the proposed optimal number of joint point able to discriminate similar movement of actions.

Keywords: Convolutional neural network, Human activity recognition, Skeletal joints.

1. Introduction

Human Action Recognition (HAR) has grown more attention in many real-world applications such as video surveillance system, health care, sports analysis and smart home system that involve interactions between persons and electronic devices [1]. Basically, the objective of HAR is to detect and analyze human activities from the information acquired from sensors such a sequence of images, either captured by RGB cameras (Kinect, webcam), range sensors, or other sensing modalities. Most of the research work mainly focuses on a spatial and temporal aspect of a video sequence of action recognition captured by RGB cameras. However, the image may suffer from illumination, various points of view, self-occlusion problems, clutter background and body segmentation error. These are the issues of color videos which influence the accuracy performance, especially when dealing with complex human actions.

Therefore, the cost-effective depth-sensing cameras like Kinect RGB have drawn much interest in HAR research community [2]. The Kinect cameras capture 3D depth scene which provides additional depth information for activity recognition. Depth map records the distance of object surface towards the camera. Hence, the human body can be detected and segmented accurately. Depth map based HAR can either be directly used for machine learning or extracted the skeleton joints for action recognition. The analysis of HAR is carried out through feature extraction, learning and classification to determine correct patterns to recognize the activities. The performance of HAR is normally evaluated through the precision matrix. However, the extensive review has been carried out in [3] found out that there is no single approach is able to guarantee for the best methods for all types of dataset. Performance of each method varies depends on their benchmark dataset.

This paper extracts the invariant characteristics of human from the 3D skeleton joints. Since skeleton joints tend to have similar skeleton points configurations for different types of actions. This paper analyses the optimal number of skeleton joints which is highly informative for actions recognition. It is organized as follows: Section 2 briefly reviews related work in action recognition over depth map. Sections 3 formulate the problems-based feature extraction using skeleton joints. In Section 4, we use Convolution Neural Network (CNN) to classify the action using CAD 60 dataset. Section 5 presents experimental results and discussions of the proposed approaches. Finally, conclusions are drawn in Section 6.

2. Previous Work

This section reviews the most relevant state of the art in HAR which utilizing depth maps and skeleton joint points collected from RGB-D devices.

2.1. Depth maps

The depth map image provided by the RGB-D cameras are robust to light, color and text variations. Ni et. al. [4] used two conventional feature descriptor STIPs and Motion History Image in the collection of RGB-D images. Xia and Aggarwal [5] used STIPs and depth cuboid similarity features to define a codebook which each sequence of depth maps associated with an action and represented as ‘bag-

of-codeword'. These features have incorporated with Support Vector Machine (SVM) classifier to describe human actions. Hadfield and Bowden [6] implement common Harris corner and Hessian point scale invariant algorithm for interest point detection from depth maps. Several approaches analyzed the effectiveness of dense information to recognize actions [7, 8]. However, extracting a correct salient feature of the subject can be hard task, even worse when there is background clutter or bad lighting conditions. Extraction of accurate silhouette can be interrupted when a person is interacting with a background object (e.g. sitting on a chair). The depth map research has extended to detect abnormal human behavior or sudden fall event [9, 10].

2.2. Skeleton joints

The Human body is connected by joints and human actions present a continuous evolution of spatial configurations of joints. Early studies in skeleton based representations start with a 2D shape. Skeletal points based features can be model into two categories: joint-based [11, 12] or body-part based [13]. Since the 3D skeleton can be obtained from the depth sensor, it becomes more reliable and robust for human action recognition. In addition, skeletal tracking algorithm built-in Kinect camera offers an easy representation of skeletal joint locations. The MSR-Action 3D dataset provides standard benchmarking data of skeletal joints extracted from depth image sequences. Han et al. [14] used the conditional random field to extract logical and contextual 3D skeleton features. Liu et al. [9] present a joint based feature vector by computing position and angle information between skeleton joints. Each skeleton joint was found to vary from one activity to another.

However, implementing skeleton based method alone performs lower precision than depth map-based methods. This might be caused by the noisy data produce by skeleton tracking or some skeleton joints configurations have similar pose for different actions. Therefore, combining skeleton based features with other cue representation method has attracted many researchers in action recognition. Wang et al. [8] proposed a novel joint feature which able to describes the local depth of 3D joint. This method demonstrated the benefit of grouping skeleton joints and local occupancy to construct more discriminant features. Zhou and Ming [15] combined skeletal joint and surface cues to introduce novel sparsity of feature selection to separate different action classes. Ding et al. [16] capture motion and curvature trajectories to represent the characteristic parameters of temporal sequential patterns.

3. Proposed Framework

Deep learning methods incorporating CNN have been demonstrated to be an effective method for computer vision and pattern recognition. Figure 1 presents the overall structure of Convolutional Neural Network for action recognition. This pipeline consists of three parts; input network, feature learning process, and classification. The input of features learning process composed of 15 joint points of skeleton data and the confidence value. The confidence value indicates the success rate of skeletal point configurations. Then, the output of the first fully connected layers works as the extracted high-level features to classify different types of actions.

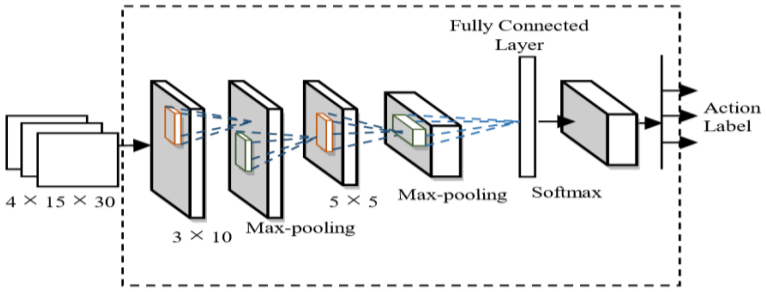

Fig. 1. Deep CNN Skeleton joint points are input to the network.

3.1. 3D skeletal joint point data

Skeleton data act as the input to the learning process of the network in Fig. 2. The network is responsible to get high-level features for the further classification task. The skeleton joint of the human body was represented by the coordinates of *x*, *y*, and *z*. The skeletal human activity represents different posture form by the sequence of the video frame. There are 15 joint points that represent different posture of activity in CAD 60 dataset as shown in Fig. 2.

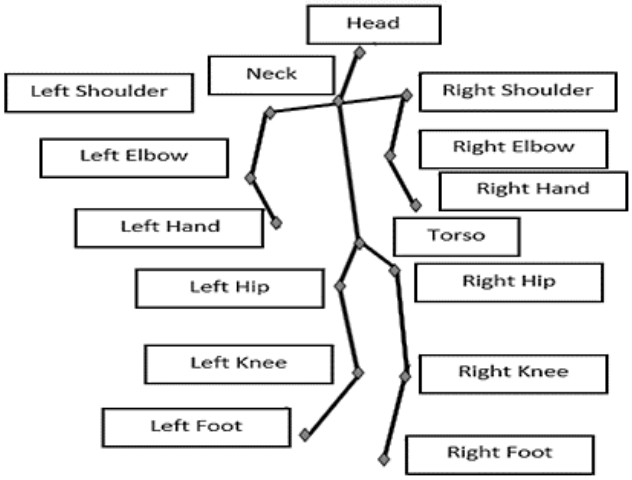

Fig. 2. Skeleton joint for 15 points in CAD 60 dataset.

The patterns produce by different actions are varies. This fact motivated Ofli et al. [17] to propose a new representation of human actions called Sequence of Most Informative Joints (SMIJ). SMIJ evaluated the value of joint angle trajectory sequence. However, some actions such as arm wave and draw X have similar SMIJ representation. Even in CAD 60 dataset, some of the joints can be considered as 'redundant' due to either similar joint configuration or relatively non-moving for different activity such as chopping and stirring. In most actions, many joints contribute very little changes and not significant for action recognition. Hence, they even bring noises which will affect the overall performance. This proof that certain skeleton joint configuration leads to similar action to each other. Therefore, we perform an analysis to prove the relationship between each joint and only

concentrate on some significant joints which highly contribute to and agrees with human activity. We apply Shannon entropy formula in Eq. (1) to evaluate informative skeleton joints for activity in the CAD60 dataset.

$$H(X) = -\sum_{i=0}^{N-1} p_i log_2 p_i \quad (1)$$

Shahnon entropy formula represents the higher entropy related to a maximum contribution of relative joints. For further evaluation, Fig. 3 represent the most informative joints for 12 types of activity in the dataset.

Therefore, this paper attempt to evaluate the optimal and effective number of skeleton point which has significant information to represent human activity. We neglect some of the skeleton joints by setting the joints to zero and apply suitable weight to some significant joints. This compact joint set leads to feature dimensionality reduction while giving a good improvement in terms of accuracy and precision. Figure 4 tabulates the valuable skeletal joints for different types of activities.

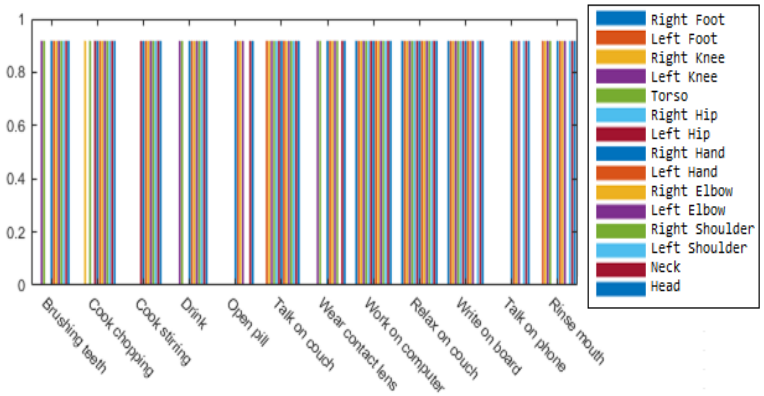

Fig. 3. Informative joints for CAD60 dataset using Shannon entropy.

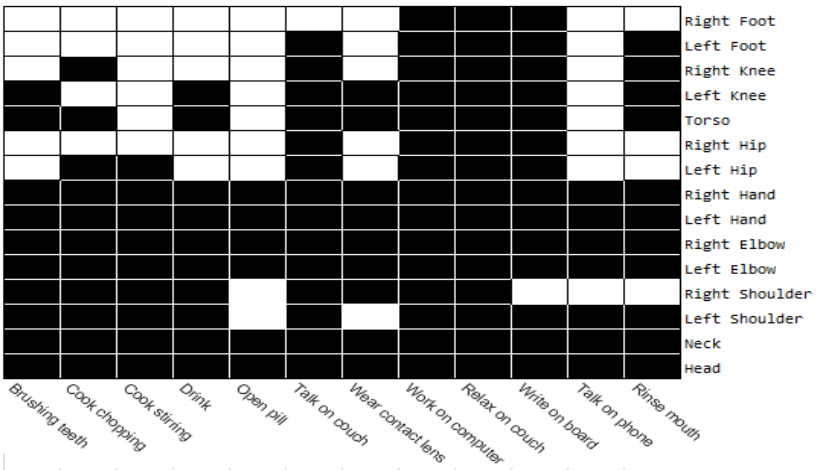

Fig. 4. The informative joints for the CAD60 dataset. The dark block indicates the selected informative joints.

3.2. Convolutional neural network

3.2.1. Input vector

In this paper, the input vector represents an entire video sequence. Hence, the 3D input vector of $N_{joint\_attributes} \times N_{joint} \times N_{frames}$, which is 4×15×30. This $N_{joint\_attributes}$ vector composed of $x \times y \times z$ and confidence level component in the first dimension. While the second dimension is N_{joint} which is equal to 15 and act as the number of joints associated with each configuration in the video sequence of the dataset. Along the last dimension is N_{frames} which is the total number of frames in batch sequences of 30 frames.

3.2.2. Layers

As shown in Figure 1, the CNN composed of five layers. The first two layers were convolutional layers, followed by one fully connected layer and finally the softmax layer before the output. The first convolutional kernel size of 3 × 10 used to initially convolve with the input vector. It can be thought of as the cumulative effect of passing a local filter across the image plane to find features across the joints position and time plane. This is then passed to the activation function. Rectified Linear Unit (ReLU) is the most popular activation function for neural networks to solve gradient problems. All layers in the networks use ReLU:

$$f(z) = max(0, z) \tag{2}$$

The result of this function is max pooled over joint and frame dimensions. The second layer of the convolution processed the output of the first layer with a filter of 5×5 kernels. This layer also implements 2 × 2 max-pooling. For both convolution layer, the input was transformed by a weight function prior to the activation function. Then the input is weighted and then fed into the fully-connected layer of the convolutional net. Finally, softmax is performed and the output determines the recognized activity. The classification outputs are softmax activated and trained with cross-entropy loss.

4. Results and Analysis

We have trained the CNN to analyze human posture based on skeleton data. In this section, we evaluate the performance of optimal skeleton joints for HAR using CNN. Firstly, the test datasets and criteria of the dataset are described. Second, HAR experiment setting to examine the effect of performance accuracy using all skeletal joints. Third, comparison of the overall performance when eliminating irrelevant skeletal joints based on Shannon entropy information. Comparison measure is done based on the confusion matrix results.

4.1. Dataset

We present a comparative performance evaluation of optimal informative joints on the CAD-60 dataset. The CAD-60 dataset is made by 12 different activities, typical of indoor environments. The activities are brushing teeth, cooking (chopping), cooking (stirring), drinking water, opening pill container, relaxing on couch, rinsing mouth with water, talking on couch, talking on the phone, wearing contact lens, writing on whiteboard and working on computer and are performed in five

different environment: bathroom, bedroom, living room, kitchen and office. Fig. 5 shows the skeletal joints representations for 12 activities in the CAD-60 dataset.

Fig. 5. Sample from CAD-60 dataset from row-wise, from left: brushing teeth, cooking(chopping), cooking (stirring), drinking water, opening pill container, relaxing on the couch, rinsing mouth with water, talking on the couch, talking on the phone, wearing contact lens, working on computer and writing on a whiteboard.

4.2. Experimental setting

The CAD-60 dataset was divided into 80% for training and 20% for testing. The experiment was carried out in two parts: (1) evaluate the performance when using the information for all joints, (2) evaluate the performance when using only the informative joints as shown in Fig. 4. The performance of human activity recognition was primarily evaluated based on the percentage of accuracy of activity recognized correctly. The accuracy for each iteration is computed by comparing the equivalence of the predicted activity class and ground truth label of the activity label for each feature vector. The average equivalence value served as the accuracy metric. Then, the confusion matrix was calculated by counting the number of instances of a given prediction and labeling.

4.3. Confusion matrix

Figures 6 and 7 shows the confusion matrix result evaluation taken for all 15 joint points and selected joint points as in Fig. 4, respectively. Each column of the confusion matrix represents the predicted activity class, while each row indicates the actual activity class for the instances. The percentage of activity class in row recognize as activity class in the column represents the value in each box. The confusion matrix in Fig. 6 shows that most of the activities have been wrongly predicted. For example, brushing teeth, drink, talking to the phone and open pill container have been misclassified. By using all skeletal joint points, we tend to produce lower precision recognition rate. Hence, we evaluate for an optimal joint point to highly differentiate the activity classes. The recognition accuracy rate is 82.96%, which we may conclude that 3D skeletal joints point proposed in the methodology part is significant to classify different human postures. In contrast with previous work using a distance of joints to produce posture feature vector, incorporating x, y and z coordinates of skeletal joints achieve interesting results in CAD 60 dataset.

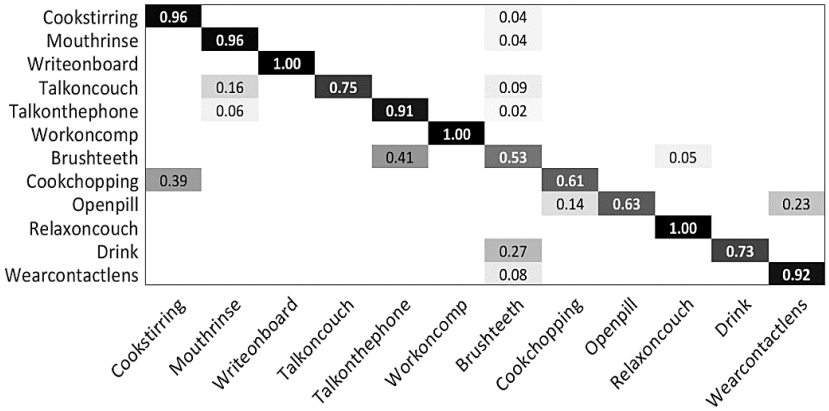

Fig. 6. Confusion matrix for all 15 skeletal joint points with accuracy 82.96%.

However, there is still a problem in discriminating interclass similarity and variability. For example, some actions are very similar pose to each other like *talk on the phone*, *mouth rinse* and *brush teeth*. The arm movement to the ear and mouth. Same goes to *cook chopping*, *cook stirring* and *opening pill container* which performed activities using either left hand, right hand or both hands. Although we achieved good recognition result, we try to not consider all skeleton joints from the CAD 60 dataset to overcome high interclass similarity and variability problems.

Based on the confusion matrix results in Fig. 7, most of the activities are highly discriminated by using only the informative joints to build the feature vector for CNN. However, some activities are confused with each other, such as drink and brush teeth activity due to having similar selected joint points. Same goes for opening pill container, cooking (chopping) and cooking (stirring) which also slightly have similar selected joint point configuration. Other than that, most of the activities were predicted very well since the contributions of the selected joints are dependent on the other joints and highly discriminated. We get the best accuracy at 91.16% for selected informative joints compared to all skeletal joints. We manage to prove than select only the informative joints gives high precision results compared to using all joint points.

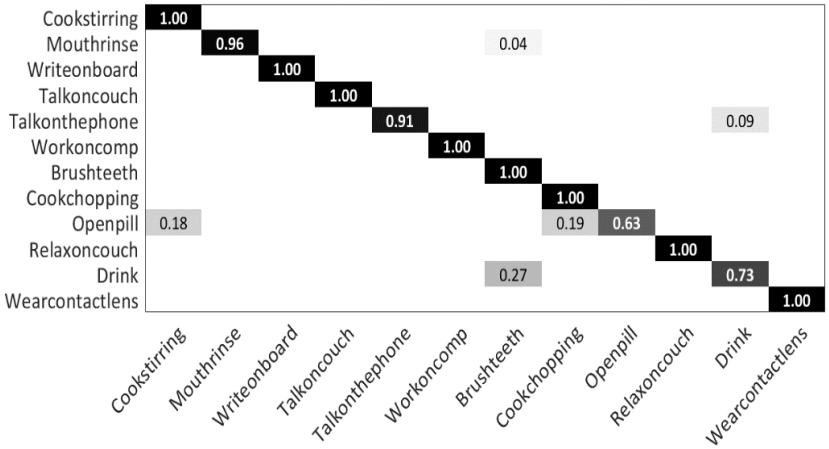

Fig. 7. Confusion matrix for selected skeletal joint points with accuracy of 94.16%.

5. **Conclusion**

This paper manages to present the result of detecting human activity using CNN. Only one standard benchmark dataset used is a CAD-60 dataset. We have proposed a skeletal informative joint evaluated by Shannon entropy formula and select only the most significant points for further classification using CNN. The main contributions of this paper are to represent that some of skeletal joints configuration are irrelevant and only the most informative joints could discriminate the activity class accurately. Overall, our proposed method achieved 94.16% for the CAD-60 dataset.

In future research, this study can be extended towards recognition of semantic event or identifying sub-activity in continuous video frames. While to validate the effectiveness of identifying a set of informative joint, more dataset of different types of activities will be carry out in our future work. Hence, the effectiveness can be measured by comparing with the state-of-the-art implementation of human activity recognition using skeletal joints. Furthermore, some surrounding cues like object near with the informative joints might be useful to achieve higher performance accuracy.

Acknowledgment

This work was supported by the Ministry of Higher Education Malaysia under FRGS grant of Universiti Tun Hussein Onn Malaysia (No. 1584).

Nomenclatures

N_{frames}	Number of frames
N_{joint}	Total number of skeletal joints
$N_{joint\_attributes}$	3D vector of skeletal joints
p_i	Probability of symbol

Abbreviations

CNN	Convolution Neural Network
HAR	Human Action Recognition
ReLU	Rectified Linear Unit
SMIJ	Sequence of Most Informative Joints
SVM	Support Vector Machine
STIP	Spatiotemporal Interest Point

References

1. Avci, A.; Bosch, S.; Marin-Perianu, M.; Marin-Perianu, R.; and Havinga. P. (2010). Activity recognition using inertial sensing for healthcare, Wellbeing and Sports Applications: A Survey. *23rd International Conference on. Computer System Architecture*, 1-10.
2. Cippitelli, E.; Gasparrini, S.; Gambi, E.; and Spinsante, S. (2016). A human activity recognition system using skeleton data from RGBD sensors. *Computing Intelligent and Neuroscience.*

3. Han, F.; Reily, B.; Hoff, W.; and Zhang, H. (2017). Space-time representation of people based on 3D skeletal data: A review. *Computer Vision Image Understanding*, 158, 85-105.
4. Ni, B., Wang, G.; and Moulin, P. (2011). RGBD-HuDaAct: A color-depth video database for human daily activity recognition. *Proc. IEEE Int. Conf. Computer Vision*, 1147-1153.
5. Xia, L.; and Aggarwal, J.K. (2013). Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. *Proc. IEEE Computer Society Conference Computer Vision Pattern Recognition*, 2834-2841.
6. Hadfield, S.; and Bowden, R. (2013). Hollywood 3D: recognizing actions in 3D natural scenes. *Proc. IEEE Computer Society Conference Computer Vision Pattern Recognition*, 3398-3405.
7. Jaeyong S.; Ponce, C.; Selman, B.; and Saxena, A. (2012). Unstructured human activity detection from RGBD images. *IEEE Int. Conf. Robot Automation*, 842-849.
8. Wang, J.; Liu, Z.; Wu, Y.; and Yuan, J. (2012). Mining action let ensemble for action recognition with depth cameras. *IEEE Conference on Computer Vision and Pattern Recognition, 1290-1297.*
9. Liu, Z., Zhang, C.; and Tian, Y. (2016). 3D-based deep convolutional neural network for action recognition with depth sequences. *Journal Image Vision Computing*, 55, 93-100.
10. Suriani, N.S.; Hussein, A.; and Zulkifley, M. A. (2013). Sudden Event Recognition: A Survey. *Sensors*, 13(8).
11. Mohamed, E. H.; Marwan, T.; Mohammad, G. A.; and Motaz, E. S. (2013). Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. *Int. Joint Conf. on Artificial Intelligence*, 2-7.
12. Xia, L.; Chen, C.C.; and Aggarwal, J. (2012). View-Invariant human action recognition using histograms of 3D joints. *Computer Vision and Pattern Recognition Workshop*, 2-7.
13. Yang, X.; and Tian, Y. (2014). Effective 3D action recognition using eigen joints. *Journal Vision Communication Image Representation*, 25(1), 2-11.
14. Han, L.; Wu, X.; Liang, X.; Hou, G.; and Jia, Y. (2010). Discriminative human action recognition in the learned hierarchical manifold space. *Image Vision Computing*, 28(5), 836-849.
15. Zhou, Y.; and Ming, A. (2016). Human action recognition with skeleton induced discriminative approximate rigid part model. *Pattern Recognition Letter*, 83(3), 261-267.
16. Ding, W.; Liu, K.; Cheng, F.; and Zhang, J. (2015). STFC: Spatio-temporal feature chain for skeleton-based human action recognition. *Journal Vision Communication Image Representation*, 26, 329-337.
17. Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; and Bajcsy, R. (2012). Sequence of the most informative joints (SMIJ): a new representation for human skeletal action recognition. *Computer Vision and Pattern Recognition Workshop*, 2-7.