Skeleton-based Human Action Recognition System with GCN

L. Benhamida, Prof. S. Larabi, Computer Science Faculty, USTHB University

1. Introduction

The skeleton-based action recognition task has been addressed using Graph Convolution Networks (GCN) by treating the sequences of skeleton movements as spatio-temporal graphs (Figure 1), where the joints are treated as nodes and the links between different joints represent the edges linking the nodes. First GCN-based method was proposed by Yan et al [2] using a successive spatial graphs and one-dimensional temporal graph convolution blocks: ST-GCN. The adjacency matrix and the feature map of the spatio-temporal graph are injected into the model's input layer. When tested on benchmark datasets, this new approach achieved state-of-the-art performance. Thus, many ST-GCN variants have been developed in the last few years, each addressing a specific limitation in the original implementation.

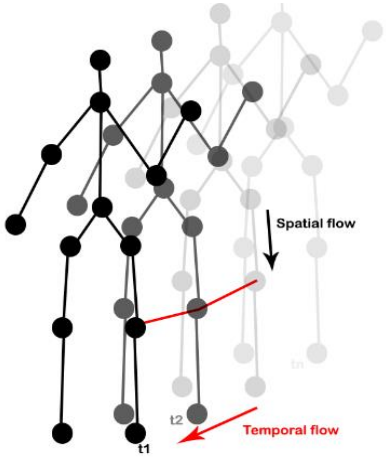

Figure1: Spatio-temporal skeleton representation: edges in black are spatial edges and red links are the temporal edges. [5]

However, the performance of these models remain unclear when applied to realistic applications with untrimmed videos for an online recognition. Most of the existing solutions for an online recognition use deep learning models in order to identify the starting point of each action in an untrimmed video. Hence, the flexibility provided by these methods comes at an enormous computational cost. As a result, the timely responses that are essential in some scenarios might not be provided. In addition, very few of the provided online HAR methods are based on skeleton data that are captured by the Kinect sensor. Most of them are based on RGB videos resulting in a relatively low performance due to the difficulty of human segmentation task.

In this work, we focus on exploiting one of the most powerful state-of-the-art Graph Convolution Network: Disentangled Unifying Multi-Scale GCN (MS-G3D)[3], using skeleton data provided by the Kinect sensor in order to develop an online human action recognition(HAR) system.

Skeleton-based Human Action Recognition System with GCN

L. Benhamida, Prof. S. Larabi, Computer Science Faculty, USTHB University

2. Method

Skeleton Data Captured by Kinect Sensor:

The Kinect sensor, developed by Microsoft, represents a groundbreaking advancement in the field of motion sensing technology. It utilizes an array of cameras and sensors to track the movements of users in three-dimensional space. Developers have leveraged the Kinect's capabilities to create interactive and immersive experiences, where users can control devices or interact with virtual environments using natural body movements. One of its key features is its ability to capture skeletal data, providing a highly detailed and accurate representation of the user's body movements. This skeletal data is provided as a set of 3D points representing different body joints as shown in the figures 2, 3.

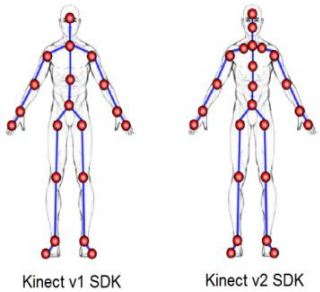

Figure 2: Skeleton joints provided by Kinect

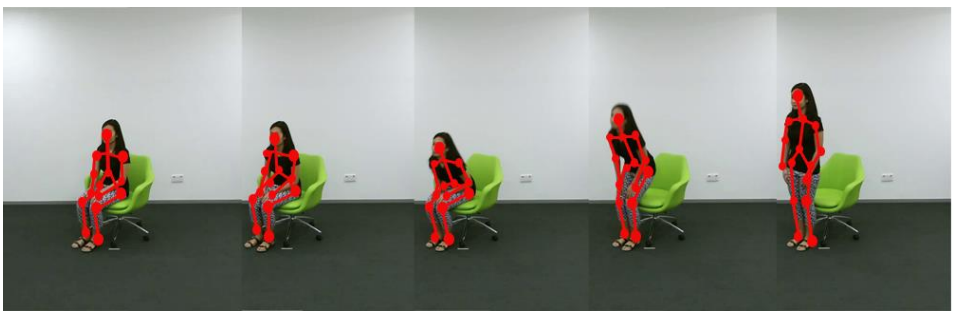

Figure3: Skeleton sequence of a person standing up.

Skeleton-based Human Action Recognition System with GCN

L. Benhamida, Prof. S. Larabi, Computer Science Faculty, USTHB University

MS-G3D

This GCN model uses a disentangled multi-scale aggregator that obtains direct information from farther nodes of the graph in input and removes redundant dependencies between node features. It also uses a unified spatial-temporal graph convolution operator to facilitate direct information flow across space and time. The combination of these two methods results in a powerful feature extraction across both spatial and temporal dimensions.

MS-G3D with Sliding Window Strategy

MS-G3D was designed to be used with trimmed videos, not for online recognition where the boundaries of each action are unknown. Some researchers handle the problem of untrimmed videos by using sliding window strategy. This strategy consists of giving the model a starting point from the untrimmed video and a size **n** to the frame-window to be classified, and then sliding the frame-window by a stride step **p** to classify the rest of the video (Figure 4).

The performance of this strategy relies on the values of **n** and **p** chosen for the sliding window and the stride step respectively. To fix the values of these two parameters that strike balance between the classification performance and the computation cost, we conducted a study on different action sequences. However, to our knowledge, no dataset is available on the internet that contains untrimmed RGB-D action sequences. Thus, we concatenated actions from the NTU RGB+D60 human action dataset resulting in ten untrimmed videos. We tested the sliding window strategy with n = [20, 25, 30, 35], and p = [3, 6, 10] for each value of n.

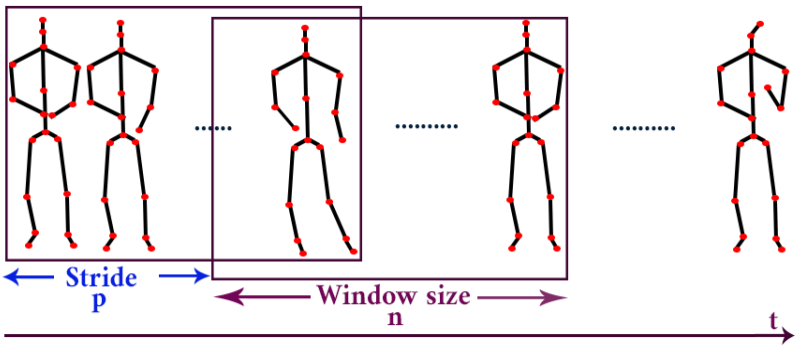

Figure 4: Sliding window strategy [1]

Skeleton-based Human Action Recognition System with GCN

L. Benhamida, Prof. S. Larabi, Computer Science Faculty, USTHB University

Action Transition Detection

We propose a method to locate actions by detecting the transition from one action to another using skeleton's joints distribution with SVM which is known for producing significant accuracy with less computation power. We use a sliding window W of n frames. Once a transition is detected in W, the recognition classifier (MS-G3D) is called to recognize the corresponding action (see figure 5).

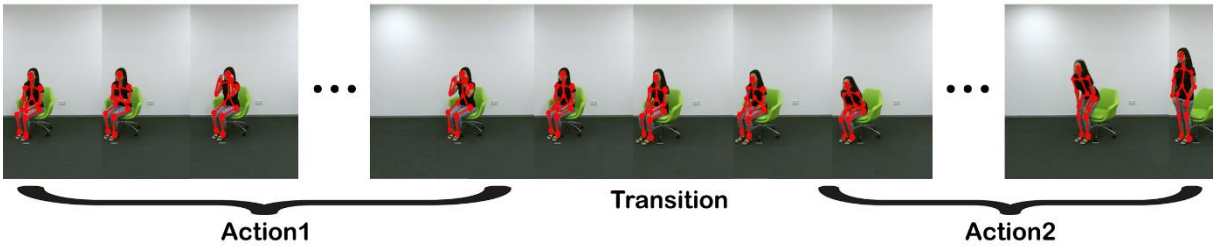

Figure5. Transition between two actions in a video sequence

$$W = F_1 F_2 \ldots F_i \ldots F_n$$

F_i is a matrix of 25×4 where 25 is the number of joints, and 4 refers to the joint's coordinates and time (x, y, z, t).

$$F_i \begin{bmatrix} j_{1,1} & \cdots & j_{1,4} \\ \vdots & \ddots & \vdots \\ j_{25,1} & \cdots & j_{25,4} \end{bmatrix}$$

First, we transform each $F_i \in W$ to a vector f_i of 25 components by reducing the dimensionality of skeleton coordinate system using Principal Component Analysis (PCA). As a result, W is transformed to a matrix of 25×n.

$$W = f_1 f_2 \ldots f_i \ldots f_n$$

Second, we apply PCA dimensionality reduction again on W in order to obtain a vector V of n points.

Finally, an SVM is used to learn the distribution of the points of different vectors and find a hyper-plane that distinctly classifies the vectors.

To train the SVM model, we generated two classes: **same-action** class and **transition** class, using skeleton sequences from NTU–RGBD60 dataset. **same-action** class was generated by calculating vectors from each NTU-RGBD action sequence, and **transition** class was generated by calculating vectors of pairwise combinations of different action classes.

Skeleton-based Human Action Recognition System with GCN

L. Benhamida, Prof. S. Larabi, Computer Science Faculty, USTHB University

MS-G3D with the Action Transition

By combining the obtained results from both experiments on MS-G3D with a simple sliding window method and the action transition detection method, we implemented the following system for online HAR: Once a transition is determined using a 20-frame window using the proposed method, the next 35-frame window is fed to the MS-G3D model to recognize the following action, and then we slide the window by 10 frames (Figure6).

3. Results

After analyzing the findings of the conducted statistics, we found that a sliding window of n = 35 with a stride step of 10 frames is the most effective so far to guarantee better classification with the least computation time. However, even with the chosen parameter values, the computation time is still relatively high, due to the use of the MS-G3D model every 10 frames.

To locate actions by detecting the transition from one action to another, we trained the SVM with different values of n, the results showed that SVM obtained the best accuracy with 20-frame window. We conclude that the best size for best action transition detection is n=20.

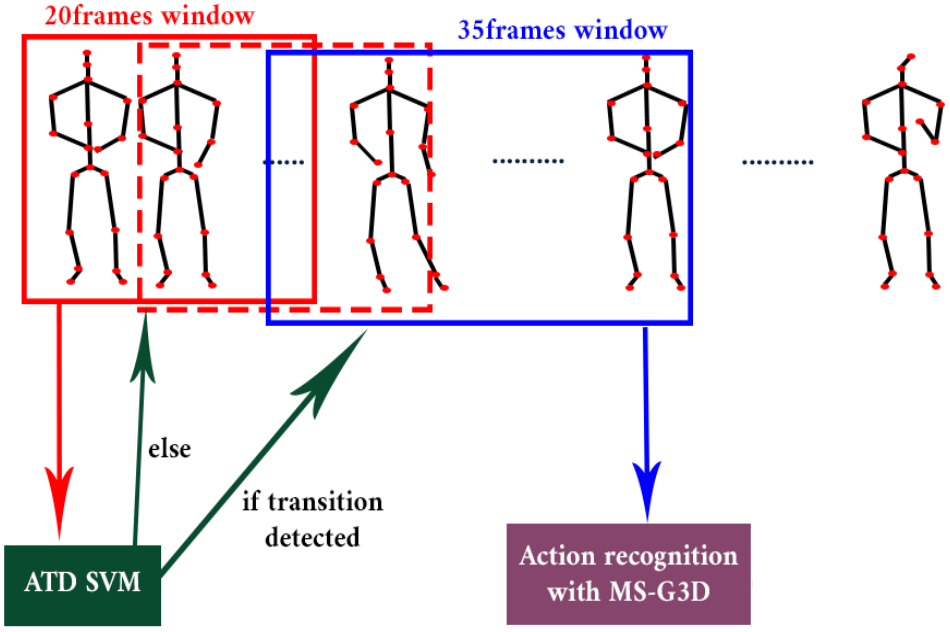

Figure 6: Action Transition method with MS-G3D for an online HAR [1]

Skeleton-based Human Action Recognition System with GCN

L. Benhamida, Prof. S. Larabi, Computer Science Faculty, USTHB University

After a comparison study between the proposed online HAR system and the simple sliding window, we observed a huge difference in computation time with quite similar and sometimes better recognition performance. The minimum time cost with a simple sliding window was 7 seconds, whereas the maximum time cost using the proposed system is 6 seconds. This proves that our system is effective at reducing computation time, which is crucial when using a deep learning model for real-time HAR applications, while preserving the same recognition performance of the model.

Conclusion

This work provides a method that can be employed with any offline skeleton-based HAR deep learning model for real-time applications. The method is based on the concept of sliding window, but rather than calling the model to classify the window at each stride step, we do so only when a transition between two actions is detected. We were able to determine the ideal values for both parameters: window size n and stride step n, which ensure the best recognition performance of MS-G3D with the lowest computation cost. This helps to reduce the computation time of the recognition system while maintaining the model's performance.

References

[1] Benhamida, Leyla, and Slimane Larabi. "Human Action Recognition and Coding based on Skeleton Data for Visually Impaired and Blind People Aid System." 2022 First International Conference on Computer Communications and Intelligent Systems (I3CIS). IEEE, 2022.
[2] Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.
[3] Liu, Ziyu, et al. "Disentangling and unifying graph convolutions for skeleton-based action recognition." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[4] Shahroudy, Amir, et al. "Ntu rgb+ d: A large scale dataset for 3d human activity analysis." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[5] Benhamida, Leyla, and Slimane Larabi. "Theater Aid System for the Visually Impaired Through Transfer Learning of Spatio-Temporal Graph Convolution Networks." arXiv preprint arXiv:2306.16357 (2023).