Skeleton based Human Action Recognition using a Structured-Tree Neural Networks

Muhammad Sajid Khan[1] ; Andrew Ware[2] ; Misha Karim[1]; Nisar Bahoo[1]; M.Junaid Khalid[1]
[1]Army Public College of Management & Sciences
[2]Faculty of Computing, Engineering and Science, University of South Wales, United Kingdom.

Abstract

The ability for automated technologies to correctly identify a human's actions provides considerable scope for systems that make use of human-machine interaction. Thus, automatic3D Human Action Recognition is an area that has seen significant research effort. In work described here, a human's everyday 3D actions recorded in the NTU RGB+D dataset are identified using a novel structured-tree neural network. The nodes of the tree represent the skeleton joints, with the spine joint being represented by the root. The connection between a child node and its parent is known as the incoming edge while reciprocal connection is known as the outgoing edge. The uses of tree structure lead to a system that intuitively maps to human movements. The classifier uses the change in displacement of joints and change in the angles between incoming and outgoing edges as features for classification of the actions performed.

Introduction

Human action recognition (HAR) is a vigorous and challenging research topic with the aim of observing and analyzing the actions of a person based on video observation data using 3D dataset. The system is gaining huge amount of importance due to the fact that computer vision is becoming a trend since all the old manual tasks that were carried out by human are now being taken by the machines by replacing the man-labor in various fields such as surveillance system is now moved to live CCTV cameras covering the recognition of the illegal action. HAR systems rectify and process contextual environmental, spatial, and temporal data to understand the human behavior. In general, the HAR process involves several steps starting from collecting information on human behavior out of raw sensor data to the conclusion about the currently performed actions.

Nowadays the Point Clouds System is becoming most emerging system in Human Action Recognition. In earlier time depth map sequences were widely used in 3D systems. Their greatest problem is that they take a lot of response time, feature extraction was also difficult in depth map sequence. We will use point clouds because they are easy to detect and manipulate. In the latest development most 3D scanners and image developers use point clouds. Point in point cloud are used to represent the surface of the object, they do not contain any information about internal features, color, materials, and so on

Methods and Materials

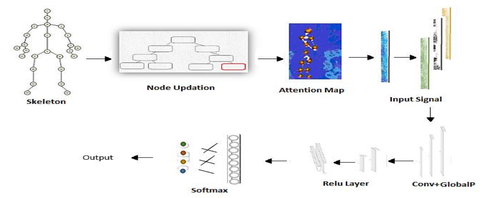
Initially, a 3D video taken from NTU-RGB-D will be given to the system as input on different Human subjects through 3 separate action view cameras. Various Preprocessing techniques break the video into multiple frames for extraction of detailed information related to our subjects while maintaining quality by removing *noise and de-shaped* subjects (Persons) for accurate and successful *subject-detection* step wise. Such noise is removed by applying *Median Filter* and the image is sharpened using *Laplacian filter*.

Using the standard *PCL* library skeleton of detected human bodies will be formed. However, till now the image will have a lot more unnecessary information.

To remove that information skeleton will be moved to a new empty 3D plane so that we have a clear visualization or movements. Certain skeletal features are extracted through a Body Pose Evolution Map. Late fusion, and Cross Setup Protocol. Mainly two features will be extracted one is *movement of nodes* and the other is *deformation of bones* within the frames.

These features are extracted by processing difference of frames and pass the collected data through *BN* and *ReLU.* Then these features are fed into Global-average pooling and SoftMax layer to give our recognized class.

Figure 1. Structured Tree Neural Networks.



Results

Python 3.7 is used to code the logic but to make GUI of the system MATLAB r2018b tool is used. The ability for automated technologies to correctly identify a human's actions provides considerable scope for systems that make use of human-machine interaction.

Thus, automatic 3D Human Action Recognition is an area that has seen significant research effort. In work described here, a human's everyday 3D actions recorded in the NTU RGB+D dataset are identified using a novel structured-tree neural network. The nodes of the tree represent the skeleton joints, with the spine joint being represented by the root.

The connection between a child node and its parent is known as the incoming edge while reciprocal connection is known as the outgoing edge. The uses of tree structure lead to a system that intuitively maps to human movements.

The classifier uses the change in displacement of joints and change in the angles between incoming and outgoing edges as features for classific

Skeleton based Human Action Recognition using a Structured-Tree Neural Network

For the implementation of this project the complete coding is being done in python language in which all the steps are followed and portrayed, in which a GUI is designed that contain six button each referring to a step from taking input to recognition and an axes is used to display the stepwise output and a table is used to display the features points that was extracted from the tree.

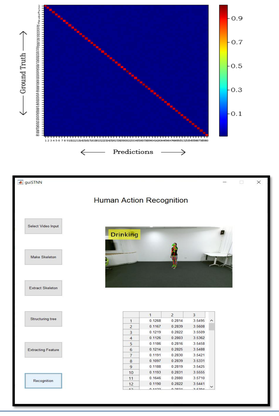
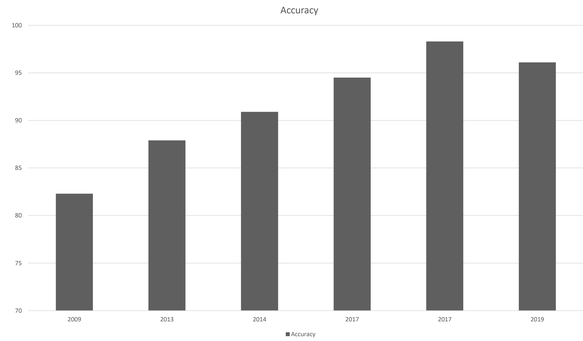


Table 1. accuracies from 2009 to 2020.

Technique	Year	Accuracy
Silhouette sequences[13]	2009	82.3%
3D Flow Estimation[10]	2013	87.9%
Eigen joints[18]	2014	90.9%
Joint spatial graphs[17]	2017	94.5%
Kinematic similarities with Depth motion maps[23]	2017	98.5%
Directed Graph Neural Network[24]	2019	96.1%
Structured Tree Neural Network (Proposed)	2020	96.3%

Chart 1. Accuracies of existing systems



Discussion

In past few years the Robot Vision have become most emerging field, so in order to make the robot to understand and respond to human actions the Human Action Recognition system was much needed.

The main objective of system is to recognize the actions performed by the human in given 3D video. Which involves following goals to reach to our objective that are:

- To take the 3D Video using UTKinect-3D Action dataset
- List of recognizing walking, running, laying, sitting, waving.
- To implement the proposed algorithm to achieve the desire actions
- More focus will be the recognition rate and performance on the system
- Response time on the system

Conclusions

A human action is the result of joint-movements of any human body such as elbows and hands for wave as well ankles and knees for running or walking. If we joint the points of these joints this forms a skeleton, if we look skeleton closer from a computer scientist point of view, it forms a tree. A tree structure constructed using skeleton, simplifies the processing and reduces the performance load as compare to graphs, because in graphs edges are also traversed and processed, but in tree only nodes are traversed and processed. The actions were classified using two features i-e joint movement and bone deformation, joint movement means that how much a joint is moving and on which axis the joint moving frame by frame, bone deformation means that how much the angle between incoming and outgoing edge of the node is changing, Yes, trees don't have edges, that's why slopes of both edges is found and then angle between these two slopes is found, the change in this angle is calculated and is used as a feature for classification. This technique uses layered approach for classification. First layer is input layer, this layer takes input in the form of skeleton data and forms a tree, then second layer extracts the features of every node in the tree, third layers removes the nodes with redundant features, then at last there is output layer is classifying the actions and places action label. This layered approach and simplicity of the structure lead to 96.3% accuracy of the system on the NTU-RGBD dataset.

Future Directions

We have seen and concluded the effective use of structured neural networks for the use of Human Action Recognition using neural networks. The tree structure can be implemented by not just the nodes of trees, but rather as a connected linked-list of arrays. But connecting and making sure that hubs of the link-list causes no problem for the whole array, i.e. for other actions is the real challenge.

Contact Information

Muhammad Sajid Khan
Army Public College of Management & Sciences
sajidpk48@yahoo.com

References

"50 years of object recognition: Directions forward." *Computer vision and image understanding* 117, no. 8 (2013): 827-891. https://doi.org/10.1016/j.cviu.2013.04.005

Shafaei, Alireza, and James J. Little. "Real-time human motion capture with multiple depth cameras." In *2016 13th Conference on Computer and Robot Vision (CRV)*, pp. 24-31. IEEE, 2016. https://doi.org/10.1109/CRV.2016.25

Shahroudy, Amir, Jun Liu, Tian-Tsong Ng, and Gang Wang. "Nturgb+ d: A large scale dataset for 3D human activity analysis." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010-1019. 2016. https://ieeexplore.ieee.org/document/7780484/

Acknowledgements

The authors are grateful to their home institutions for the support given during the collaborative research reported in the paper.

© POSTER TEMPLATE BY GENIGRAPHICS® 1.800.790.4001 WWW.GENIGRAPHICS.COM