Deep Learning and SVM-Based Method for Human Activity Recognition with Skeleton Data

Plamen Hristov, Agata Manolova, Ognian Boumbarov
Department of Radiocommunications and Videocommunications, Faculty of Telecommunications
Technical University of Sofia
8 Kliment Ohridski blvd., 1000 Sofia, Bulgaria
plm@tu-sofia.bg, amanolova@tu-sofia.bg, olb@tu-sofia.bg

Abstract – **In recent years, research related to the analysis of human activity has been the subject of increased attention by engineers dealing with computer vision, and particularly that which utilizes deep learning. In this paper, we propose a method for classification of human activities, composed of 3D skeleton data. This data is normalized beforehand and represented in two forms, which are fed to a neural network with parallel convolutional and dense layers. After the network is trained, the training data is propagated again to infer the output from the second last layer. This output is used for training a Support Vector Machine. All hyperparameters were found using the Bayesian Optimization strategy on the PKU-MMD dataset. Our method was tested on the UTD-MHAD dataset, achieving an accuracy of 92.4%**

Keywords – **deep learning; support vector machine; human activity recognition; skeleton; convolutional neural network;**

I. Introduction

Human Activity Recognition became a center of attention with the advent of information technologies, which provide new opportunities for application in the following areas: security systems and video surveillance, behavior analysis in high-security systems, medical systems supporting the elderly and handicapped, etc. The prerequisite for human activity recognition consists of extraction of necessary data – features, representing the location of the human body in 3D space. The appearance of depth sensors and the easy inference of skeleton data allows for implementing quick and effective algorithms with neural classifiers. The definition of skeletons as a model for the body pose in 3D space becomes the basic means for representing an action in the last decade.

The analysis of the modern algorithms for representation and analysis of skeleton data and the use of deep neural networks for activity modelling could improve their recognition rate.

According to the literary sources for human activity recognition the following problems exist:

-The sensor data doesn't always determine the body positions correctly and includes noise, which is due to occluded limbs, interaction with other individuals, sudden movements, etc.;

-The skeletons vary according to each individual and their position, making the learning process more difficult

-An intuitive method for representing skeleton activity has not yet been established, but many, which return different results depending on the used data and its preprocessing, exist;

For tackling the aforementioned problems, this paper highlights the following achievements:

-Representing the skeleton data as distance vectors in order to capture the inter-frame dynamics;

-Implementation of a supervised learning method for human activity, integrating both Convolutional Neural Networks and a Support Vector Machine;

-Finding the optimal training parameters for this combination by testing on a well-known dataset.

II. Related work

The literature in this sphere contains many methods for realization of human activity recognition. The classic ones use hand-crafted features, which are extracted frame-wise from a raw video or a processed one, which contains only the subjects, performing an action. Subjects could, for example, be represented as skeletons – vectors of points of their bodies, corresponding to their joints. In this paper the focusing is namely on this type of representation, and more specifically – the one in the three-dimensional space. In order to achieve acceptable performance, the methods above often include preliminary processing of the input data, which presents them in a unified space, suppressing the spatial and scale differences between each data sample. This process is also known as normalization.

Yang et al [1] represent an action as three types of channels of difference between the 3D joint positions in every frame. These channels correspond to separate features and are concatenated in a feature vector. These features are all types of frame differences. They are normalized in the boundary of [-1, +1] beforehand, in order to decrease the intra-class variance of a single action, which is performed by several subjects, and to become spatially invariant. The whole feature vector is processed by a PCA algorithm and the output data is fitted into SVM and NBNN classifiers, which achieve over 90% accuracy for different datasets such as MSR Action 3D and UCF Kinect.

Pazhoumand-Dar et al [4] preprocess the skeleton data, centering the middle hip joint as the beginning of the coordinate system and perform a Z-score normalization of the 3D position of every joint, using an average joint vector and a standard deviation of the same joint, which are calculated from all the actions done by the specific subject. The skeleton joints are represented as vectors in a coordinate system and serve for the creation of two handcrafted descriptors of an action. The first one describes the most informative set of joint angles (MIJA) and is