Pattern Recognition 76 (2018) 80–94

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition

Juan C. Núñez, Raúl Cabido, Juan J. Pantrigo, Antonio S. Montemayor, José F. Vélez*

Universidad Rey Juan Carlos, Madrid, Spain

ARTICLE INFO

Article history:
Received 2 December 2016
Revised 6 October 2017
Accepted 24 October 2017

Keywords:
Deep learning
Convolutional Neural Network
Recurrent neural network
Long Short-Term Memory
Human activity recognition
Hand gesture recognition
Real-time

ABSTRACT

In this work, we address human activity and hand gesture recognition problems using 3D data sequences obtained from full-body and hand skeletons, respectively. To this aim, we propose a deep learning-based approach for temporal 3D pose recognition problems based on a combination of a Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM) recurrent network. We also present a two-stage training strategy which firstly focuses on CNN training and, secondly, adjusts the full method (CNN+LSTM). Experimental testing demonstrated that our training method obtains better results than a single-stage training strategy. Additionally, we propose a data augmentation method that has also been validated experimentally. Finally, we perform an extensive experimental study on publicly available data benchmarks. The results obtained show how the proposed approach reaches state-of-the-art performance when compared to the methods identified in the literature. The best results were obtained for small datasets, where the proposed data augmentation strategy has greater impact.

1. Introduction

Vision-based human action recognition concerns the task of automatically interpreting an image sequence to decide what action or activity is being performed by the subjects in the scene. It is a relevant topic in computer vision, with practical applications such as video surveillance, human-computer interaction, gaming, sports arbitration, sports training, smart homes, life-care systems, among many others [1,2]. Due to the huge possibilities for practical application, human activity recognition problems have received the attention of researchers in the fields of computer vision, artificial intelligence and machine learning. Researchers of the field organize different contests as, for example, the ChaLearn Looking at People challenge [3], and provide large datasets as NTU RGB+D [4]. As a consequence, it is possible to find a significant number of related works in the literature describing an extensive variety of methods and strategies to deal with this problem. In particular, in recent years, deep neural networks have been successfully applied in human action recognition problems as a suitable approach when relatively large datasets are available.

The toolkits of many affordable RGBD devices allow the acquisition of 3D data at interactive framerates. These devices can be used to capture human movements or hand poses, offering 3D coordinates of the joints as skeletons [5]. These skeletons can capture the evolution of the pose of a human body or hand and, therefore, they can be used to classify the activities or gestures performed by subjects in the area.

In this paper, we propose the combination of a Convolutional Neural Network (CNN) and a Long-Short Term Memory (LSTM) recurrent network for handling time series of 3D coordinates of skeleton keypoints. We have tested our proposal on six publicly available datasets.

Fig. 1summarizes the proposed system, in which the input data at each time step is presented to the CNN+LSTM network. The CNN is mainly responsible for capturing relevant features from the 3D data input on every time step, while the LSTM takes into account the time evolution of the 3D data series. Finally, the CNN+LSTM model generates a classification result for the presented model sequence.

An important contribution of this paper is that the proposed network architecture does not need to be adapted to the type of activity or gesture to be recognized as well as to the geometry of the 3D time-series data as input. Nonetheless, it obtains results that are competitive to previous works that need to make assumptions on those. Additionally, we present a data augmentation method that allows us to solve the problem of overfitting. The proposed augmentation techniques provide a significant performance improvement when applied to small datasets. Finally, it is also important to note that the proposed network architecture

* Corresponding author.
E-mail address: jose.velez@urjc.es (J.F. Vélez).

https://doi.org/10.1016/j.patcog.2017.10.033
0031-3203/

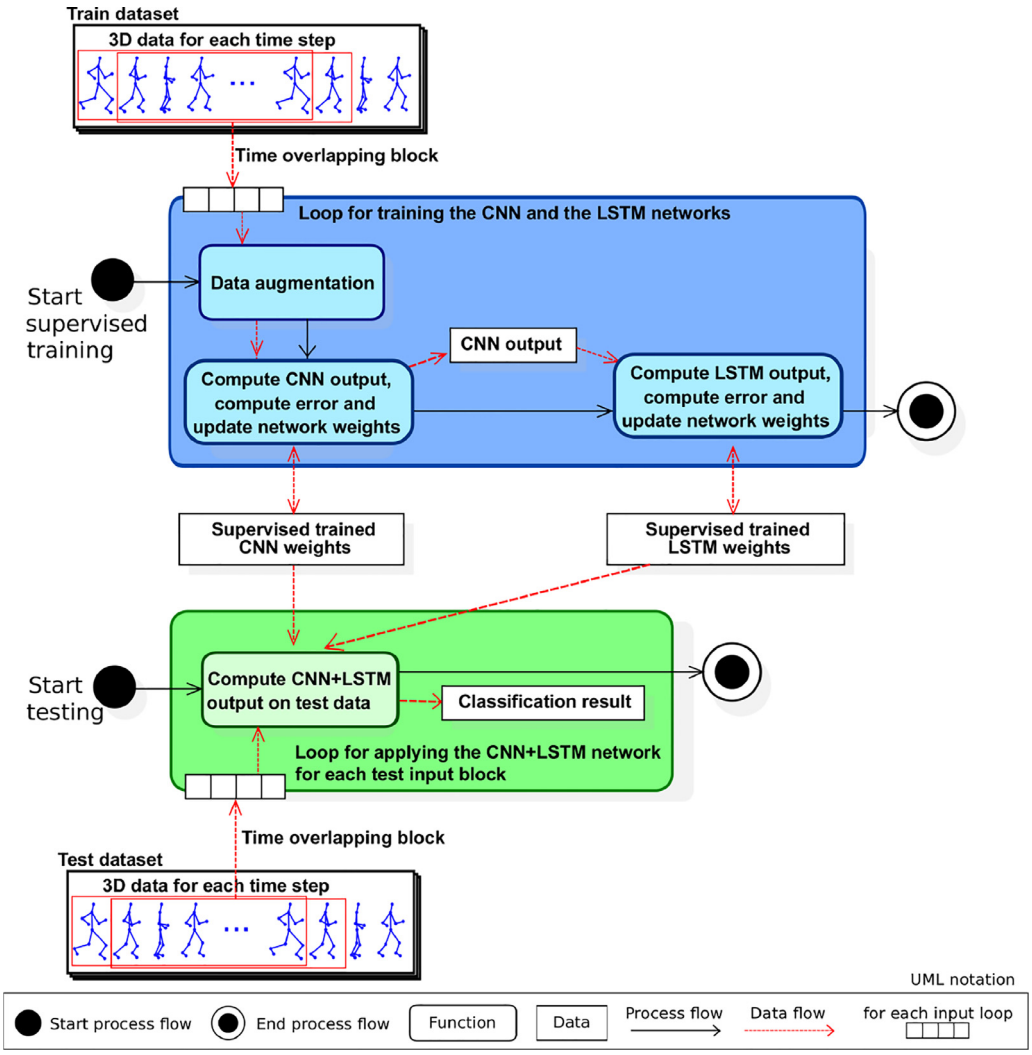

Fig. 1. UML Activity Diagram for the proposed system. The diagram shows two process flows, the upper one for the training process and the lower one for the testing process.

is lightweight enough to allow real time processing on embedded platforms.

The rest of the paper is organized as follows: Section 2 deals with the presentation of the related works in this research area. Section 3 details the proposed neural network architecture. Section 4 presents and discusses the results obtained. Finally, conclusions are outlined in Section 5.

2. Related work

This section reviews the state-of-the-art methods for the considered problems, specifically, in skeleton-based human activity recognition and hand gesture recognition problems.

2.1. Skeleton-based human activity recognition

Recently, a significant number of approaches to the problem of human activity recognition have been proposed, based on human skeleton kinematics [6]. However, although the success of deep learning techniques in the computer vision community started around 2012, most of the related works in this topic are based on classic machine learning methods. Xia et al. [7] presented an approach which uses histograms of 3D joint locations (HOJ3D) as a compact representation of postures and, afterwards, these HOJ3D are reprojected using linear decomposition analysis and clustered into k posture visual words, representing the prototypical poses of actions. Zanfir et al. [8] proposed a moving pose descriptor that considers position, speed and acceleration of the human body joints within a short time window around the current frame, and they established the classification using a modification of the k-nearest neighbour classifier. The work of Devanne et al. [9] is based on the use of trajectories, which consists of a motion channel corresponding to the evolution of the 3D position of all joint coordinates within frames of action sequence. The classification stage is addressed by a k-nearest neighbour classifier. Chrungoo et al. [10] represented the action as a histogram of direction vectors, obtaining a scale and speed invariant descriptor which is also computationally efficient. Vemulapalli et al. [11] uses geometric relationships between several body parts by means of rotations and translations in the 3D space. In this representation, human actions can be modelled as curves in a Lie group, and then, these action curves are mapped as a vector space. The classification stage includes a combination of dynamic time warping, Fourier temporal pyramid representation and linear SVM. Evangelidis et al. [12] propose a local skeleton descriptor to encode the relative position of joint quadruples and a Fisher vector representation to describe the skeletal quads contained in a (sub)action, based on a learned GMM and multi-level representation of those vectors. Zhang & Parker