A Deep Learning Approach for Real-Time 3D Human Action Recognition from Skeletal Data

Huy Hieu Pham, Houssam Salmane, Louahdi Khoudour, Alain Crouzil, Pablo Zegers, and Sergio A. Velastin

Toulouse Computer Science Research Institute (IRIT), Paul Sabatier University & Cerema Research Center, France

Waterloo, Canada, August 27, 2019

Table of contents

Introduction

Human action recognition in RGB-D videos

Research problem:

- How to recognize correctly what humans do in unknown videos?
- How to learn effectively spatio-temporal features of human motions by deep learning models (*e.g.* CNNs) ?
- How to build a real-time deep learning framework for human action recogntion from RGB-D data?

Figure 1: Some action classes of the NTU-RGB+D dataset.

Human action recognition in RGB-D videos

Objective

Developing a real-time, end-to-end deep learning approach to recognize human actions from RGB-D sequences. Application to safety/security in public transport.

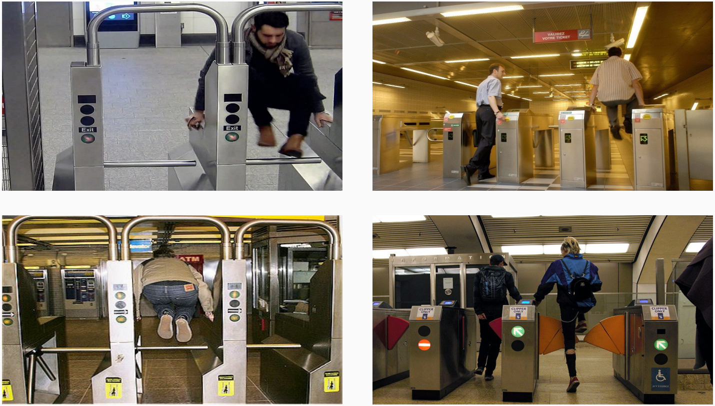

Figure 2: Detecting abnormal behaviors in surveillance videos.

Human action recognition using deep neural networks

Literature review[1]:

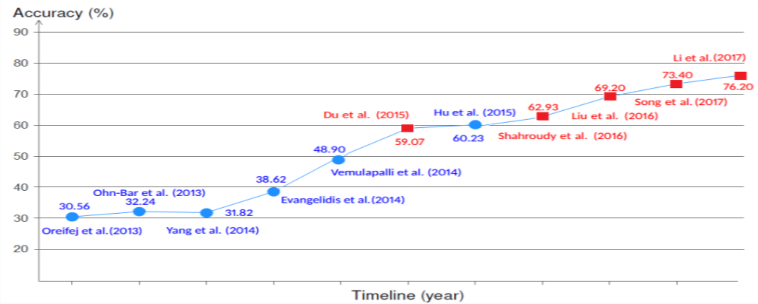

Figure 3: Recognition performance of hand-crafted and deep learning approaches reported on the NTU-RGB+D dataset. The traditional approaches are marked with circles, deep learning based approaches are marked with squares.

[1] Huy-Hieu Pham, Louahdi Khoudour, Alain Crouzil, Pablo Zegers, Sergio A. Velastin, "*Exploiting deep residual networks for human action recognition from skeletal data*" – Computer Vision and Image Understanding (CVIU 2018).

Human action recognition using deep neural networks

What is the problem?

For every 256×256 color image, there are $3 \times 256 \times 256 \approx 200k$ values that have to be stored for computation.

Meanwhile, each skeleton frame with 25 key-points just has $3 \times 25 = 75$ values.

Figure 4: Dimensionality of data: A comparison between RGB data and skeletal data.

Proposed Method

Proposed method

Building a skeleton-based action recognition method using deep neural networks. The proposed method is based on three key ideas:

- Encoding each skeleton sequence into a single RGB image via a compact image representation called SPMF.
- A color enhancement algorithm is then applied to enhance local patterns and to highlight important motions (Enhanced-SPMF).
- Exploiting state-of-the-art CNN models (e.g. DenseNet) to learn directly an end-to-end mapping between input sequences and their action labels.

Proposed method

A CNN model is able to learn effectively spatio-temporal dynamics of human motions from skeletal data via an image-based representation[2].

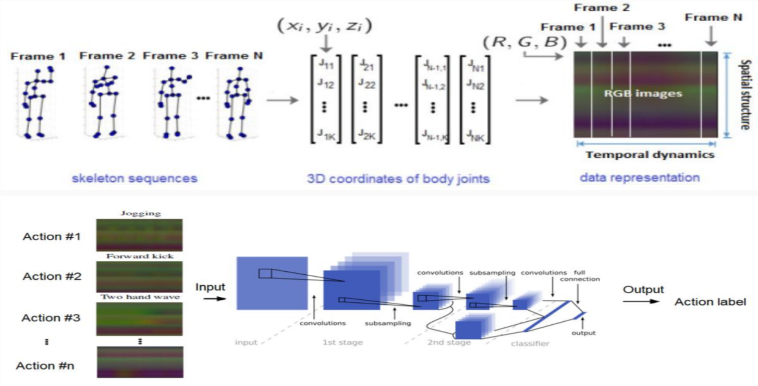

Figure 5: Human action recognition using CNNs and skeletal data.

[2] Huy-Hieu Pham *et al.* "Skeletal Movement to Color Map: A Novel Representation for 3D Action Recognition with Inception Residual Networks" – ICIP 2018.

Action recognition using deep networks

Building a more complex skeleton-based representation called **SPMF** for human action recognition in videos. Each action map contains two key components:**Pose Features** (PF) and **Motion Feature** (MF).

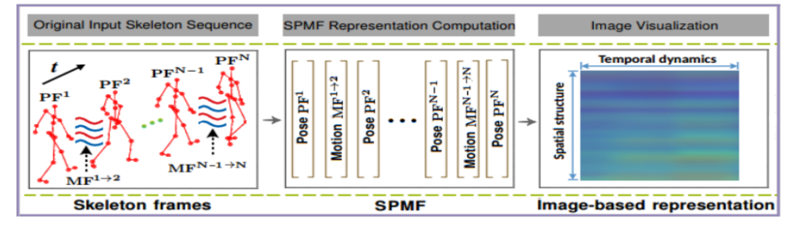

Figure 6: Encoding a skeleton sequence into a single action map.

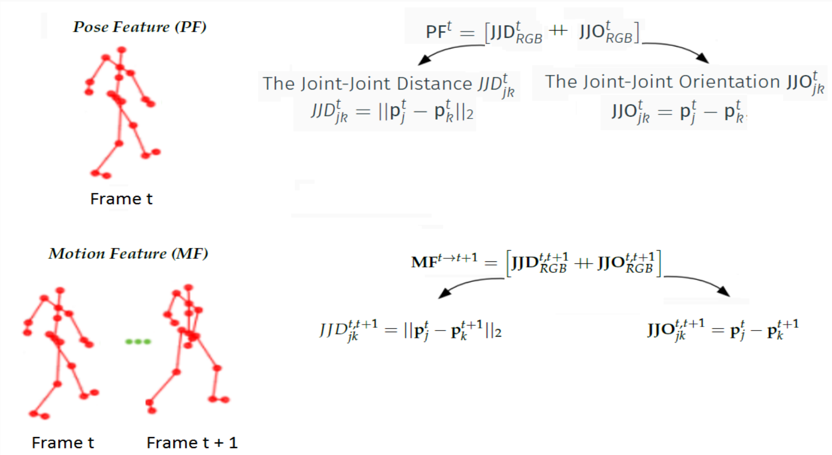

Figure 7: Computing Pose Features (PF) and Motion Features (MF) from skeletons.

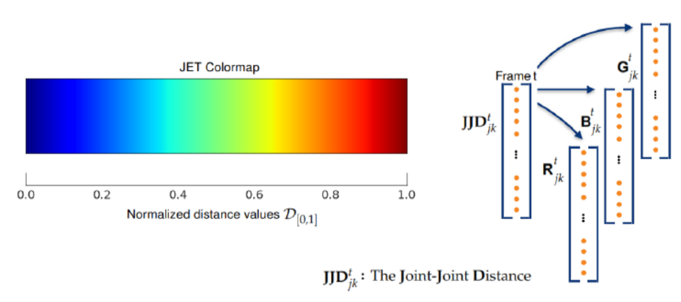

Figure 8: Mapping from Euclidean distances to 3D color vector.

Action recognition using deep networks

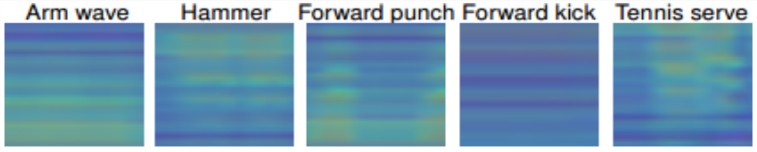

Figure 9: The SPMFs obtained from some samples of the MSR Action3D dataset.

The Adaptive Histogram Equalization algorithm was then used to highlight the motion map and form the Enhanced-SPMF.

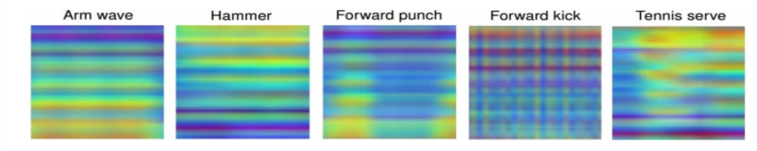

Figure 10: The proposed Enhanced-SPMF representation for human action recognition from skeleton sequences.

Action recognition using deep networks

Three different configurations of DenseNet (i.e. DenseNet-16, DenseNet-28, DenseNet-40) were designed for learning and recognition task from the proposed skeleton-based representations.

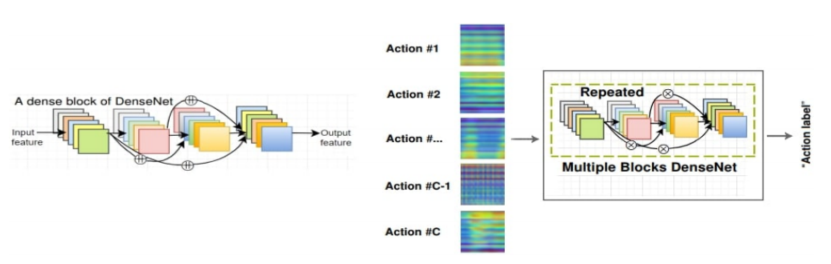

Figure 11: The proposed Enhanced-SPMFs are fed into a DenseNet for classifying action maps.

Evaluation and results

Comparison with state-of-the-art

Method (protocol of [33])	**Cross-Subject**	**Cross-View**
Lie Group [39]	50.10%	52.80%
Hierarchical RNN [6]	59.07%	63.97%
Dynamic Skeletons [13]	60.20%	65.20%
Two-Layer P-LSTM [33]	62.93%	70.27%
ST-LSTM Trust Gates [21]	69.20%	77.70%
Geometric Features [50]	70.26%	82.39%
Two-Stream RNN [40]	71.30%	79.50%
Enhanced Skeleton [24]	75.97%	82.56%
GCA-LSTM [22]	76.10%	84.00%
SPMF [27]	78.89%	86.15%
Enhanced-SPMF DenseNet-16 (**ours**)	77.89%	**86.55%**
Enhanced-SPMF DenseNet-28 (**ours**)	**79.07%**	**86.82%**
Enhanced-SPMF DenseNet-40 (**ours**)	**79.95%**	**87.52%**

Table 1: Recognition accuracy on the large-scale NTU-RGB+D dataset.

- The proposed method achieved state-of-the-art accuracy on three challenging datasets, including the largest RGB-D dataset for action recognition (*i.e.* NTU-RGB+D).
- The proposed deep learning framework requires less computation for training and inference, whilst achieving high-level performance.

Comparison with state-of-the-art

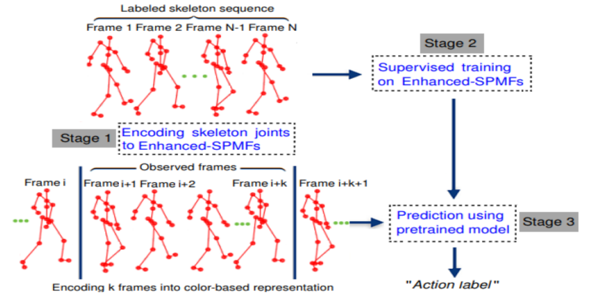

Figure 12: Three main stages of the proposed deep learning framework for recognizing human actions from skeleton sequences. The inference stage, including the stage (1) that is executed on a CPU and the stage (3), takes an average of **0.175**s per sequence without parallel processing.

CEMEST Dataset

- A new real-wold surveillance dataset containing both normal and anomalous events for studying human behaviors in public transport.
- 203 video samples containing RGB videos, depth map sequences, and 3D skeletal data.
- Three action classes: *crossing normally over the barriers*, *jumping over the ticket barriers*, and *sneaking under ticket barriers*.

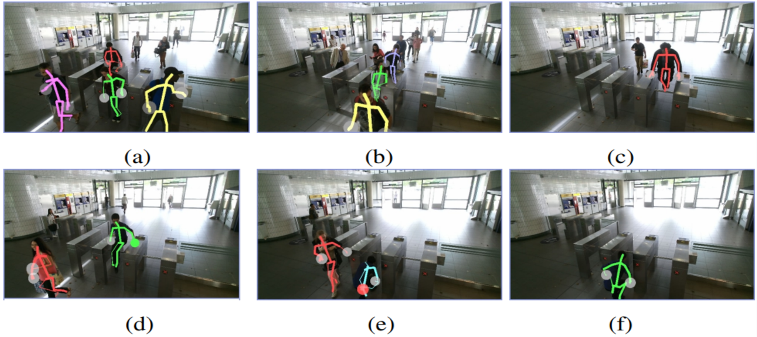

Figure 13: Some samples from the CEMEST dataset.

CEMEST Dataset

- We achieved an accuracy of **91.18**% by the DenseNet-40 when training from scratch.
- We reached an accuracy of **95.70**% with transfer learning, increasing the performance by nearly **5**% compared to the first setting.

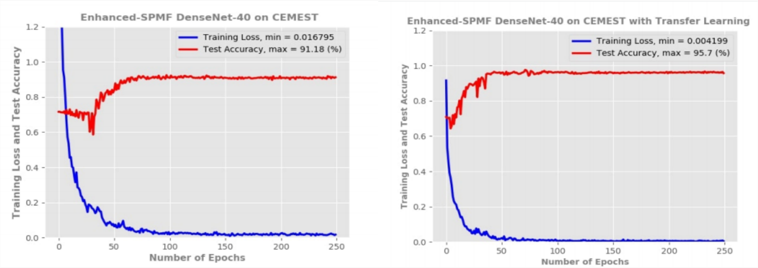

Figure 14: Learning curves of DenseNet-40 trained on the CEMEST dataset.

Current research and future works

Current research and future works

Objective: Learning for 3D human pose estimation from a single RGB image using Deep Neural Networks.

- Using a state-of-the-art 2D pose estimator (*e.g.* OpenPose or Hourglass Network) to obtain 2D human poses from RGB image sequences.
- Building Deep Learning Networks for learning and estimating 3D human poses from 2D poses.

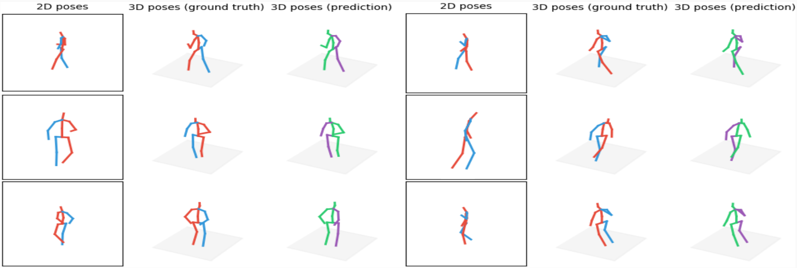

Figure 15: Experimental results on learning a 2D-to-3D mapping.

Action recognition using deep networks (current research)

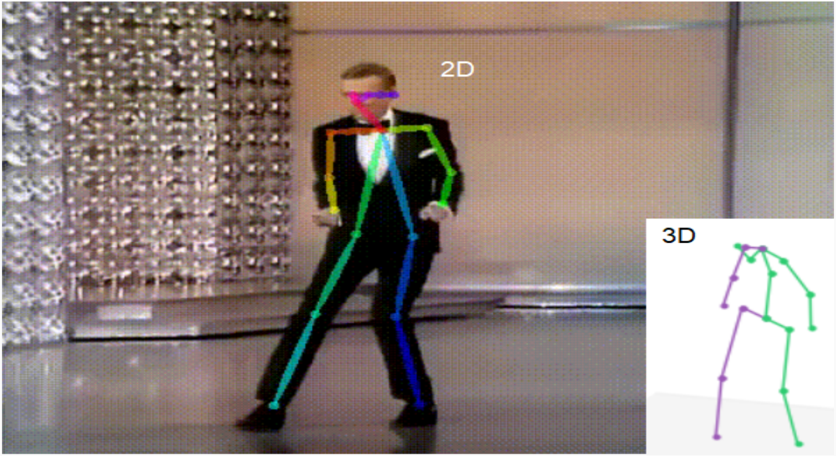

Figure 16: From 2D skeleton to 3D skeleton using deep neural networks.

Thank you for your attention!