A Unified Deep Framework for Joint 3D Pose Estimation and Action Recognition from a Single RGB Camera

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

This work was carried out at Toulouse Computer Science Research Institute (IRIT) and Cerema, Toulouse, France.

Abstract

We present a deep learning-based multitask framework for joint 3D human pose estimation and action recognition from RGB video sequences. Our approach proceeds along two stages. In the first, we run a real-time 2D pose detector to determine the precise pixel location of important keypoints of the body. A two-stream neural network is then designed and trained to map detected 2D keypoints into 3D poses. In the second, we deploy the Efficient Neural Architecture Search (ENAS) algorithm to find an optimal network architecture that is used for modeling the spatio-temporal evolution of the estimated 3D poses via an image-based intermediate representation and performing action recognition. Experiments on Human3.6M, MSR Action3D and SBU Kinect Interaction datasets verify the effectiveness of the proposed method on the targeted tasks. Moreover, we show that our method requires a low computational budget for training and inference.

Introduction

Human action recognition from videos has been researched for decades, since this topic plays a key role in various areas. Although significant progress has been achieved in the past few years, building an accurate, fast and efficient system for the recognition of actions is still a challenging task due to a number of obstacles, *e.g.* changes in camera viewpoint, occlusions, background, etc. The rapid development of depth-sensing time-of-flight camera technology has helped in dealing with this problem, exploiting skeletal data for 3D action recognition opens up opportunities for addressing the limitations of RGB-based solutions and many skeleton-based action recognition approaches have been proposed. However, depth sensors have some significant drawbacks with respect to 3D pose estimation. For instance, they are only able to operate up to a limited distance and within a limited field of view. Moreover, a major drawback of depth cameras is the inability to work in bright light, especially sunlight.

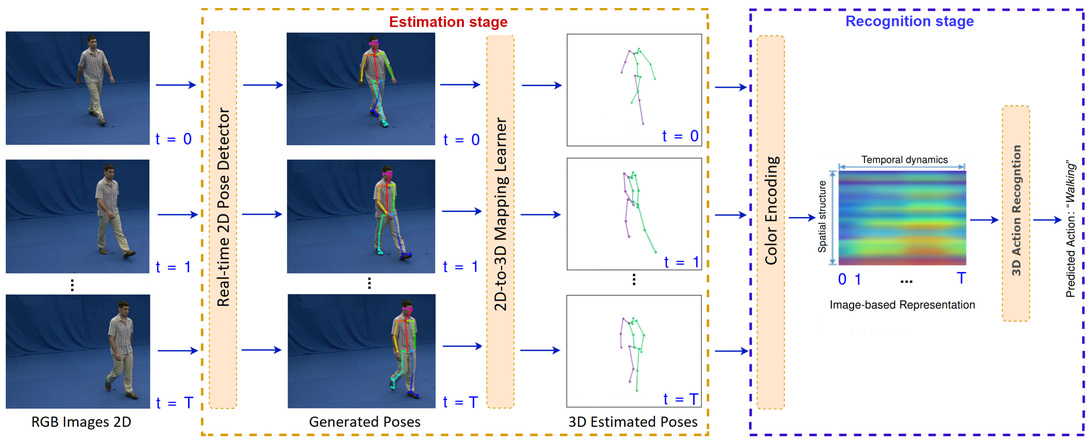

Figure 1: Overview of the proposed method. In the estimation stage, we first run OpenPose [1] – a real-time, state-of-the-art multi-person 2D pose detector to generate 2D human body keypoints. A deep neural network is then trained to produce 3D poses from the 2D detections. In the recognition stage, the 3D estimated poses are encoded into a compact image-based representation and finally fed into a deep convolutional network for supervised classification task, which is automatically searched by the ENAS algorithm [5].

Our focus in this work is therefore to propose a 3D skeleton-based action recognition approach without depth sensors. Specifically, we are interested in building a unified deep framework for both 3D pose estimation and action recognition from RGB video sequences. Our approach consists of two stages. In the first, *estimation stage*, the system recovers the 3D human poses from the input RGB video. In the second, *recognition stage*, an action recognition approach is developed and stacked on top of the 3D pose estimator in a unified framework, where the estimated 3D poses are used as inputs to learn the spatio-temporal motion features and predict action labels. The effectiveness of the proposed method is evaluated on public benchmark datasets (Human3.6M, MSR Action3D, and SBU). The experimental results demonstrate state-of-the-art performances on the targeted tasks.

Proposed Method

Problem definition

Given an RGB video clip of a person who starts to perform an action at time $t = 0$ and ends at $t = T$, the problem studied in this work is to generate a sequence of 3D poses $\mathcal{P} = (\mathbf{p}_0, ..., \mathbf{p}_T)$, where $\mathbf{p}_i \in \mathbb{R}^{3\times M}$, $i \in \{0, ..., T\}$ at the estimation stage. The generated $\mathcal{P}$ is then used as input for the recognition stage to predict the corresponding action label $\mathcal{A}$ by a supervised learning model.

3D human pose estimation

Given an input RGB image $\mathbf{I} \in \mathbb{R}^{W\times H\times 3}$, we aim to estimate the body joint locations in the 3-dimensional space, noted as $\hat{\mathbf{p}}_{3D} \in \mathbb{R}^{3\times M}$. To this end, we first run the state-of-the-art human 2D pose detector, namely OpenPose [1], to produce a series of 2D keypoints $\mathbf{p}_{2D} \in \mathbb{R}^{2\times N}$. To recover the 3D joint locations, we try to learn a *direct 2D-to-3D mapping* f_r: $\mathbf{p}_{2D} f_r \hat{\mathbf{p}}_{3D}$. This transformation can be implemented by a deep neural network in a supervised manner

$$\hat{\mathbf{p}}_{3D} = f_r(\mathbf{p}_{2D}, \theta), \tag{1}$$

where θ is a set of trainable parameters of the function f_r. To optimize f_r, we minimize the prediction error over a labelled dataset of $\mathcal{C}$ poses by solving the optimization problem

$$\arg\min_{\theta} \frac{1}{\mathcal{C}} \sum_{n=1}^{\mathcal{C}} \mathcal{L}(f_r(\mathbf{x}_i), \mathbf{y}_i). \tag{2}$$

Here $\mathbf{x}_i$ and $\mathbf{y}_i$ are the input 2D poses and the ground truth 3D, respectively; $\mathcal{L}$ denotes a loss function.

Network design

Our design is based on a simple and lightweight multilayer network architecture without the convolution operations. In the design process, we exploit some recent improvements in the optimization of the modern deep learning models [2]. Concretely, we propose a two-stream network. Each stream comprises linear layers, Batch Normalization (BN), Dropout, SELU and Identity connections. During the training phase, the first stream takes the ground truth 2D locations as input. The 2D human joints predicted by OpenPose [1] are inputted to the second stream. The outputs of the two streams are then averaged.

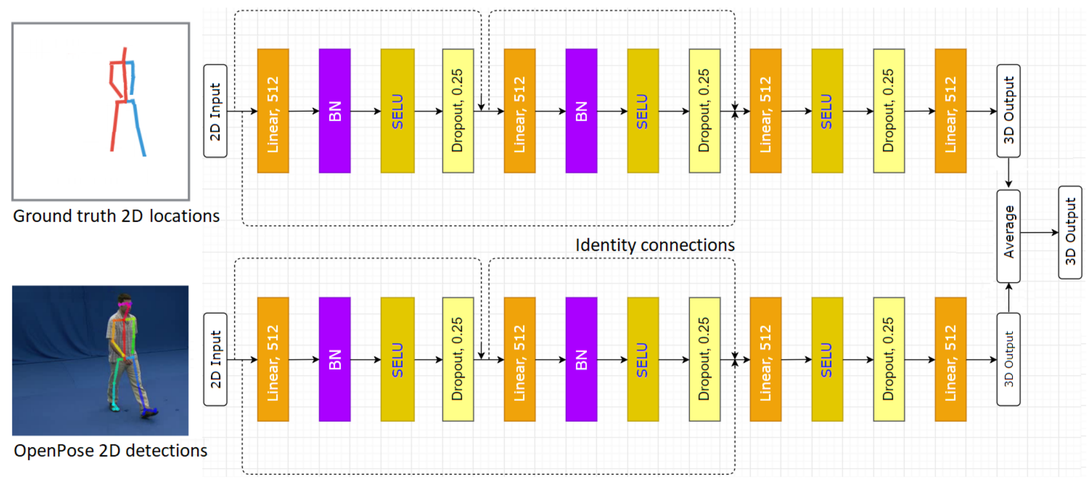

Figure 2: Diagram of the proposed two-stream network for training our 3D pose estimator.

3D pose-based action recognition

The spatio-temporal patterns of a 3D pose sequence are transformed into a single color image as a global representation called Enhanced-SPMF [6]. For learning and classifying the obtained images, we propose to use the Efficient Neural Architecture Search (ENAS) [5] – a recent state-of-the-art technique for automatic design of deep neural networks. The following figure illustrates the entire pipeline of our approach for the recognition stage.

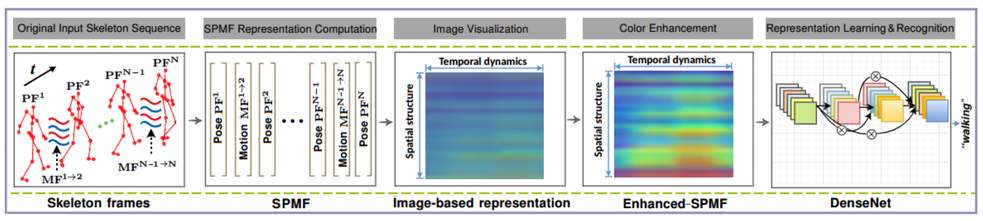

Figure 3: Illustration of the proposed approach for 3D pose-based action recognition.

Empirical Evaluation

We evaluate the effectiveness of the proposed 3D pose estimation network using the standard protocol of the Human3.6M dataset [3]. Experimental results are reported by the average error in millimeters between the ground truth and the corresponding predictions over all joints. Our method outperforms the previous best result from the literature [4] by 3.1mm, corresponding to an error reduction of 6.8% even when combining the ground truth 2D locations with the 2D OpenPose detections. Additionally, we also achieved promising results on human action recognition tasks on the MSR Action3D and SBU Kinect Interaction datasets using the predicted 3D poses.

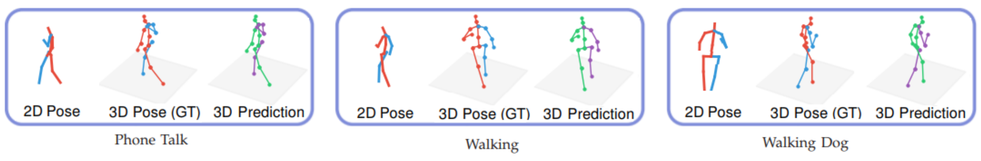

Figure 4: Visualization of 3D output of the estimation stage with some samples on the test set of Human3.6M [3]. For each example, from left to right are 2D poses, 3D ground truths and our 3D predictions, respectively.

Conclusions

- We presented a unified deep learning framework for joint 3D human pose estimation and action recognition from RGB video sequences. The proposed method first runs a state-of-the-art 2D pose detector to estimate 2D locations of body joints. A deep neural network is then designed and trained to learn a direct 2D-to-3D mapping and predict human poses in 3D space.
- We also introduced a novel action recognition approach based on a compact image-based representation and automated machine learning, in which an advanced neural architecture search algorithm is exploited to discover the best performing architecture for each recognition task. The proposed framework is able to reach state-of-the-art performance, whilst requiring less computation budget for training and inference.

References

[1] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime Multi-person 2D Pose Estimation using Part Affinity Fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017.

[2] H. Gao, L. Zhuang, M. Laurens van der, and Q. W. Kilian. Densely Connected Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[3] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2014.

[4] J. Martinez, R. Hossain, J. Romero, and J. Little. A Simple Yet Effective Baseline for 3D Human Pose Estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2640–2649, 2017.

[5] H. Pham, M. Guan, B. Zoph, Q. Le, and J. Dean. Efficient Neural Architecture Search via Parameters Sharing. In *International Conference on Machine Learning (ICML)*, pages 4095–4104, 2018.

[6] H. Pham, H. Salmane, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin. Spatio-Temporal Image Representation of 3D Skeletal Movements for View-Invariant Action Recognition with Deep Convolutional Neural Networks. *Sensors*, 19(8), 2019.