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, Sergio A. Velastin, Pablo Zegers

Article

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

Huy Hieu Pham [1,2,3], Houssam Salmane [4], Louahdi Khoudour [1], Alain Crouzil [2], Sergio A. Velastin [5,6,7,*] and and Pablo Zegers [8]

1 Cerema Research Center, 31400 Toulouse, France; hieuhuy01@gmail.com (H.H.P.); louahdi.khoudour@cerema.fr (L.K.)
2 Informatics Research Institute of Toulouse (IRIT), Université de Toulouse, CNRS, 31062 Toulouse, France; alain.crouzil@irit.fr
3 Vingroup Big Data Institute (VinBDI), Hanoi 10000, Vietnam
4 Clay AIR, Software Solution, 33000 Bordeaux, France; psalmane@clayair.io
5 School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK
6 Zebra Technologies Corp., London SE1 9LQ, UK
7 Department of Computer Science and Engineering, University Carlos III de Madrid, 28270 Colmenarejo, Spain
8 Aparnix, Santiago 7550076, Chile; pablozegers@gmail.com
* Correspondence: sergio.velastin@ieee.org

Abstract: We present a deep learning-based multitask framework for joint 3D human pose estimation and action recognition from RGB sensors using simple cameras. The approach proceeds along two stages. In the first, a real-time 2D pose detector is run to determine the precise pixel location of important keypoints of the human body. A two-stream deep neural network is then designed and trained to map detected 2D keypoints into 3D poses. In the second stage, the Efficient Neural Architecture Search (ENAS) algorithm is deployed 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 the method requires a low computational budget for training and inference. In particular, the experimental results show that by using a monocular RGB sensor, we can develop a 3D pose estimation and human action recognition approach that reaches the performance of RGB-depth sensors. This opens up many opportunities for leveraging RGB cameras (which are much cheaper than depth cameras and extensively deployed in private and public places) to build intelligent recognition systems.

Keywords: human action recognition; 3D pose estimation; RGB sensors; deep learning

1. Introduction

Human Action Recognition (HAR) from videos has been researched for decades, since this topic plays a key role in various areas such as intelligent surveillance, human–robot interaction, robot vision and so on. Although significant progress has been achieved in the past few years, building an accurate, fast and efficient system for the recognition of actions in unseen videos is still a challenging task due to several obstacles, e.g., changes in camera viewpoint, occlusions, background, speed of

motion, etc. Traditional approaches on video-based action recognition [1] have focused on extracting hand-crafted local features and building motion descriptors from RGB sensors. Many spatio-temporal representations of human motion have been proposed and widely exploited with success such as Scale Invariant Feature Transform (SIFT) [2], Histograms of Optical Flow (HOF) [3] or Cuboids [4]. However, one of the major limitations of these approaches is the lack of 3D structure from the scene and recognizing human actions based only on RGB information is not enough to overcome the current challenges in the field.

The rapid development of depth-sensing time-of-flight sensor technology has helped in dealing with this problem, which is considered complex for traditional sensors. Low-cost and easy-to-use depth sensors can provide detailed 3D structural information of human motion. In particular, most of the current depth sensors have integrated real-time skeleton estimation and tracking frameworks [5], facilitating the collection of skeletal data. This is a high-level representation of the human body, which is suitable for the problem of motion analysis. Hence, 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 [6–10]. 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 low-cost depth sensors is their inability to work in bright light, especially sunlight [11].

The focus in this paper is therefore to propose a 3D skeleton-based action recognition approach without the need for depth sensors. Specifically, we are interested in building a unified deep framework for both 3D pose estimation and action recognition from RGB video sequences provided by single color sensors. As shown in Figure 1, the 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.

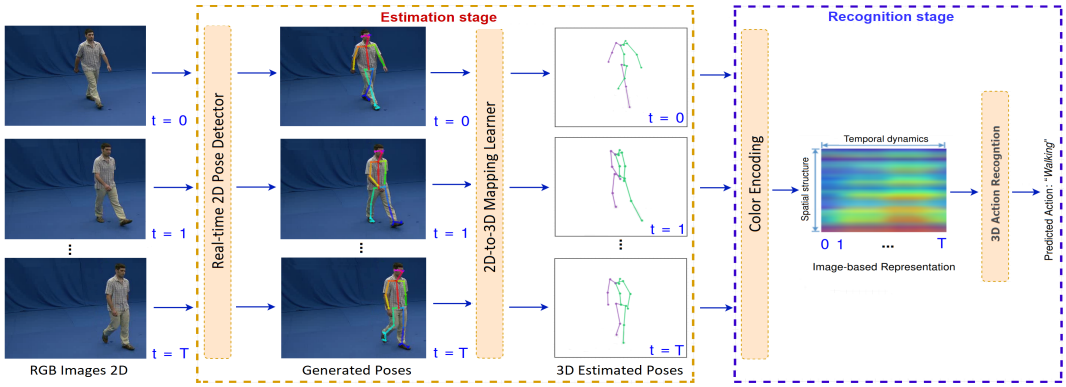

Figure 1. Overview of the proposed method. In the estimation stage, a real-time multi-person 2D pose detector, such as OpenPose [12] or AlphaPose [13], is used to generate 2D human body keypoints. A deep neural network is then trained to produce 3D poses from the 2D detections. In the recognition stage, 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 [14].

There are four hypotheses that motivate us to build a deep learning framework for human action recognition from 3D poses. First, actions can be correctly represented through 3D pose movements [15,16]. Second, the 3D human pose has a high-level of abstraction with much less complexity compared to RGB and depth streams. This makes the training and inference processes much simpler and faster. Third, depth sensors can provide highly accurate skeletal data for 3D action recognition. However, they are expensive and not always available (e.g., for outdoor scenes).

A fast and accurate approach of 3D pose estimation from only RGB input is highly desirable. Fourth, state-of-the-art 2D pose detectors [12,13,17] are able to provide 2D poses with a high degree of accuracy in real time. Meanwhile, deep networks have proved their capacity to learn complex functions from high-dimensional data. Hence, a simple network model can also learn a mapping to convert 2D poses into 3D. The effectiveness of the proposed method is evaluated on public benchmark datasets (Human3.6M [18] for 3D pose estimation and MSR Action3D [19] and SBU [20] for action recognition). Beyond the initial expectations, the experimental results demonstrate state-of-the-art performance on the targeted tasks (Section 4.3) and support the hypotheses above. Furthermore, we show that this approach has a low computational cost (Section 4.4). Overall, our main contributions are as follows:

- First, we present a two-stream, lightweight neural network to recover 3D human poses from RGB images provided by a monocular camera. The proposed method achieves state-of-the-art result on 3D human pose estimation task and benefits action recognition. The novelty of the study is that a very simple deep neural network could be trained effectively to learn a 2D-to-3D mapping for the task of 3D human estimation from color sensors.
- Second, we propose to put an action recognition approach on top of the 3D pose estimator to form a unified framework for 3D pose-based action recognition. It takes the 3D estimated poses as inputs, encodes them into a compact image-based representation and finally feeds to a deep convolutional network, which is designed automatically by using a neural architecture search algorithm. Surprisingly, the experiments show that we reached state-of-the-art results on this task, even when compared with methods using depth cameras.

The rest of this paper is organized as follows. A review of the related work is presented in Section 2. The proposed method is explained in Section 3. Experiments are provided in Section 4 and Section 5 concludes the paper.

2. Related Work

This section reviews two main topics that are directly related to the proposed approach, i.e., 3D pose estimation from RGB images and 3D pose-based action recognition. An extensive literature review is beyond the scope of this section. Instead, the interested reader is referred to the surveys of Sarafianos et al. [21] for recent advances in 3D human pose estimation and Presti et al. [22] for 3D skeleton-based action recognition.

2.1. 3D Human Pose Estimation

The problem of 3D human pose estimation has been intensively studied in recent years. Almost all early approaches for this task were based on feature engineering [18,23,24], while the current state-of-the-art methods are based on deep neural networks [25–30]. Many of them are regression-based approaches that directly predict 3D poses from RGB images via 2D/3D heatmaps. For instance, Li et al. [25] designed a deep convolutional network for human detection and pose regression. The regression network learns to predict 3D poses from single images using the output of a body part detection network. Tekin et al. [26] proposed to use a deep network to learn a regression mapping that directly estimates the 3D pose in a given frame of a sequence from a spatio-temporal volume centered on it. Pavlakos et al. [27] used multiple fully convolutional networks to construct a volumetric stacked hourglass architecture, which can recover 3D poses from RGB images. Pavllo et al. [28] exploited a temporal dilated convolutional network [31] for estimating 3D poses. However, this approach led to a significant increase in the number of parameters as well as the required memory. Mehta et al. [29] introduced a real-time approach to predict 3D poses from a single RGB sensor. They used ResNets [32] to jointly predict 2D and 3D heatmaps as regression tasks. Recently, Katircioglu et al. [30] introduced a deep regression network for predicting 3D human poses from monocular images via 2D joint location heatmaps. This architecture is in fact an overcomplete autoencoder that learns a high-dimensional

latent pose representation and accounts for joint dependencies, in which a Long Short-Term Memory (LSTM) network [33] is used to enforce temporal consistency on 3D pose predictions.

To the best of our knowledge, several studies [27,29,30] stated that regressing the 3D pose from 2D joint locations is difficult and not too accurate. However, motivated by Martinez et al. [34], we believe that a simple neural network can learn effectively a direct 2D-to-3D mapping. Therefore, this paper aims at proposing a simple, effective and real-time approach for 3D human pose estimation that benefits action recognition. To this end, a two-stream deep neural network that performs 3D pose predictions from the 2D human poses is designed and optimized. These 2D poses are generated by a state-of-the-art 2D detector, which can run in real time for multiple people. We empirically show that although the proposed approach is computationally inexpensive, it is still able to improve the state-of-the-art.

2.2. 3D Pose-Based Action Recognition

Human action recognition from skeletal data or 3D poses is a challenging task. The methods used in previous works on this topic can be divided into two main groups. The first group [6,9,35] extracts hand-crafted features and uses probabilistic graphical models, e.g., Hidden Markov Model (HMM) [35] or Conditional Random Field (CRF) [36] to recognize actions. However, almost all these approaches require a lot of feature engineering. The second group [37–39] considers the 3D pose-based action recognition as a time-series problem and proposes to use Recurrent Neural Networks with Long Short-Term Memory units (RNN-LSTMs) [33] for modeling the dynamics of the skeletons. Although RNN-LSTMs are able to model the long-term temporal characteristics of motion and have advanced the state-of-the-art, this approach feeds raw 3D poses directly into the network and just considers them as a kind of low-level feature. The large number of input features makes RNNs very complex and may easily lead to overfitting. Moreover, many RNN-LSTMs act merely as classifiers and cannot extract high-level features for recognition tasks [40].

In the literature, 3D human pose estimation and action recognition are closely related. However, both problems are generally considered to be two distinct tasks [41]. Although some approaches have been proposed for tackling the problem of jointly predicting 3D poses and recognizing actions in RGB images or video sequences [42–44], they are data-dependent and require a lot of feature engineering, except the work of Luvizon et al. [44]. Unlike previous studies, a multitask learning framework for 3D pose-based action recognition is proposed here by reconstructing 3D skeletons from RGB images and exploiting them for action recognition in a joint way. Experimental results on public and challenging datasets show that the framework can solve the two tasks in an effective way.

3. Proposed Method

This section presents the proposed method. First, the approach for 3D human pose estimation is presented, followed by the proposed solution for 3D pose-based action recognition.

3.1. 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, in which M denotes the number of keypoints for the pose $\mathbf{p}_i$. 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. See Figure 1 for an illustration of the problem.

3.2. 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 any state-of-the-art human 2D pose detector, in this case OpenPose [12], 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: $\mathbf{p}_{2D} \overset{f_r}{\mapsto} \hat{\mathbf{p}}_{3D}$. 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, the prediction error is minimized 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 poses, respectively; $\mathcal{L}$ denotes a loss function. Here, the robust Huber loss [45] is used to deal with outliers.

Network Design

State-of-the-art deep learning architectures such as ResNet [32], Inception-ResNet-v2 [46], DenseNet [47], or NASNet [48] have achieved an impressive performance in supervised learning tasks with high-dimensional data, e.g., 2D or 3D images. However, the use of these architectures [32,46–48] on low-dimensional data like the coordinates of the 2D human joints could lead to overfitting. Therefore, the design is based on a simple and lightweight multilayer network architecture without the convolution operations. The design process exploits some recent improvements in the optimization of the modern deep learning models [32,47]. Concretely, a two-stream network is proposed. Each stream comprises linear layers, Batch Normalization (BN) [49], Dropout [50], SELU [51] and Identity connections [32]. During the training phase, the first stream takes the ground truth 2D locations as input. The 2D human joints predicted by OpenPose [12] are inputted to the second stream. The outputs of the two streams are then averaged. Figure 2 illustrates the network design. Please note that learning with the ground truth 2D locations for both of these streams could lead to a higher level of performance. However, training with the 2D OpenPose detections could improve the generalization ability of the network and makes it more robust during inference, when only the OpenPose's 2D output is used to deal with action recognition in the wild.

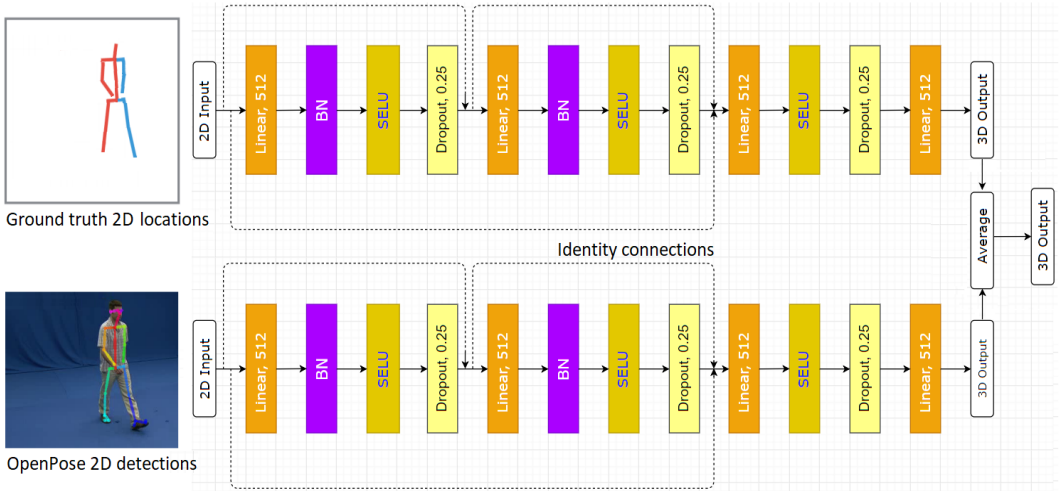

Figure 2. Diagram of the proposed two-stream network for training the 3D pose estimator.

3.3. 3D Pose-Based Action Recognition

This section describes the integration of the estimation stage with the recognition stage in a unified framework. Specifically, the proposed recognition approach is stacked on top of the 3D pose estimator.

To explore the high-level information of the estimated 3D poses, they are encoded into a compact image-based representation. These intermediate representations are then fed to a Deep Convolutional Neural Network (D-CNNs) for learning and classifying actions. This idea has been proven effective in our previous works [52–54]. Thus, the spatio-temporal patterns of a 3D pose sequence are transformed into a single color image as a global representation called Enhanced-SPMF [54] via two important elements of a human movement: 3D poses and their articulation joint motions as shown in Figure 3.

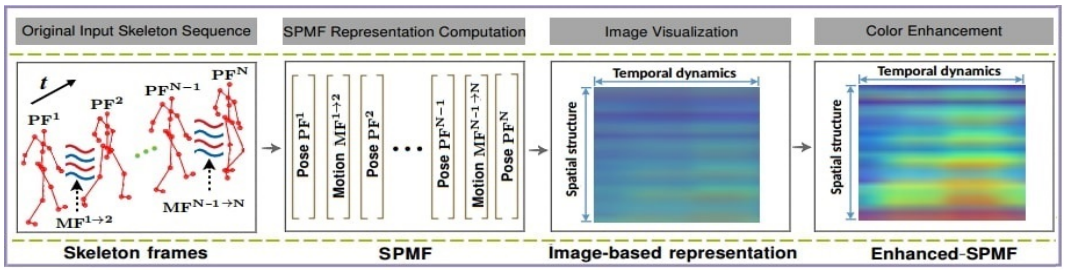

Figure 3. Illustration of the Enhanced-SPMF representation. To build an Enhanced-SPMF map from skeletal data, each skeleton sequence is first encoded as a single RGB image via a skeleton-based representation called SPMF (Skeleton Pose-Motion Feature) [53]. The SPMF is built from Pose Feature vectors (PFs) and Motion Feature vectors (MFs), which are calculated from the 3D coordinates of skeletons. Finally, we use a color enhancement technique [55] to enhance the local textures of the SPMF to form the Enhanced-SPMF. This is an image-based global representation for the whole input skeleton sequences. Figure reproduced, by permission from the publishers, from our previous work in [56].

For a detailed technical description of the Enhanced-SPMF the interested reader is referred to the work described in [54]. Figure 4 visualizes some Enhanced-SPMF representations from samples of the MSR Action3D dataset [19].

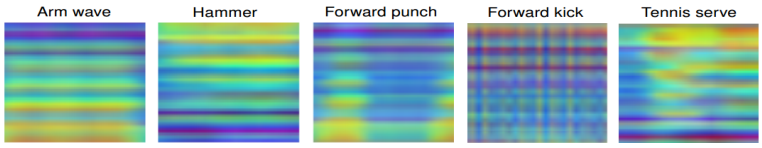

Figure 4. Immediate image-based representations for the recognition stage.

For learning and classifying the obtained images, the use of the Efficient Neural Architecture Search (ENAS) [14]—a recent state-of-the-art technique for automatic design of deep neural networks, is proposed. ENAS is in fact an extension of an important advance in deep learning called NAS [48], which can automate the designing process of convolutional architectures on a dataset of interest. The method searches for optimal building blocks (called cells, including normal cells and reduction cells) and the final architecture is then constructed from the best cells. Figure 5 shows a typical CNN architecture that is generated by ENAS.

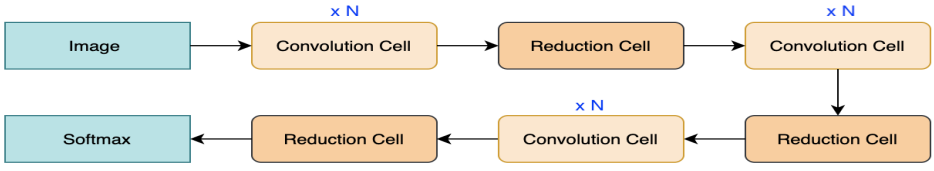

Figure 5. Illustration of a deep neural network generated by ENAS that contains 3 blocks, each with N optimal convolution cells and one reduction cell.

In NAS, an RNN is used. It first samples a candidate architecture called child model. This child model is then trained to converge on the desired task and to report its performance. Next, the RNN

uses the performance as a guiding signal to find a better architecture. This process is repeated many times, making NAS computationally expensive and time-consuming (e.g., on CIFAR-10, NAS needs 4 days with 450 GPUs to discover the best architecture). The main limitation of NAS is that the training of each child model to convergence requires a significant amount of time and computational resources as it measures model accuracy while throwing away all the trained weights. Therefore, ENAS has been proposed to improve the efficiency of NAS. Its key idea [14] is the use of shared parameters among child models, which helps reducing the training times of each child model from scratch to convergence. State-of-the-art performance has been achieved by ENAS on well-known public datasets. We encourage the readers to refer to the original paper [14] for more details. Figure 6 illustrates the entire pipeline of our approach for the recognition stage.

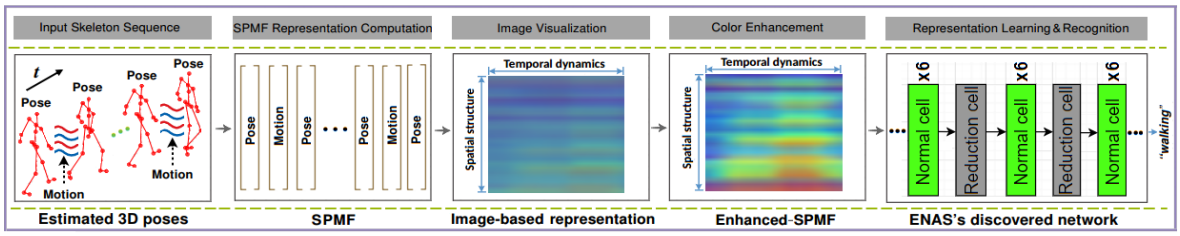

Figure 6. Illustration of the proposed approach for 3D pose-based action recognition.

4. Experiments

4.1. Datasets and Settings

The proposed method is evaluated on three challenging datasets: Human3.6M, MSR Action3D and SBU Kinect Interaction. Human3.6M is used for evaluating 3D pose estimation. Meanwhile, the other two datasets are used for evaluating action recognition. The characteristics of each dataset are as follows.

Human3.6M [18]: This is a very large-scale dataset containing 3.6 million different 3D articulated poses captured from 11 actors for 17 actions, under 4 different viewpoints. For each subject, the dataset provides 32 body joints, from which only 17 joints are used for training and computing scores. In particular, 2D joint locations and 3D poses ground truth are available for evaluating supervised learning models.

MSR Action3D [19]: This dataset contains 20 actions, performed by 10 subjects. Experiment were conducted on 557 video sequences of the MSR Action3D, in which the whole dataset is divided into three subsets: AS1, AS2, and AS3. There are 8 actions classes for each subset. Half of the data is selected for training and the rest is used for testing.

SBU Kinect Interaction [20]: This dataset contains a total of 300 interactions, performed by 7 participants for 8 actions. This is a challenging dataset as it contains pairs of actions that are difficult to distinguish such as "exchanging objects–shaking hands" or "pushing–punching". The dataset is randomly split into 5 folds, in which 4 folds are used for training and the remaining 1 fold is used for testing.

4.2. Implementation Details

The proposed networks were implemented in Python with Keras/TensorFlow backend. The two streams of the 3D pose estimator are trained separately with the same hyperparameters setting, in which mini batches of 128 poses are used with 0.25 dropout rate. The weights are initialized with He initialization [57]. Adam optimizer [58] is used with default parameters. The initial learning rate is set to 0.001 and is decreased by a factor of 0.5 after every 50 epochs. The network is trained for 300 epochs from scratch on the Human3.6M dataset [18]. For action recognition task, OpenPose is

run [12] to generate 2D detections on MSR Action3D [19] and SBU Kinect Interaction [20]. The 3D pose estimator pre-trained on Human3.6M [18] is then used to provide 3D poses. Standard data pre-processing and augmentation techniques are used, such as randomly cropping and flipping on these two datasets due to their small sizes. To discover optimal recognition networks, ENAS [14] is used with the same parameter setting as the original work. Concretely, the shared parameters ω are trained with Nesterov's accelerated gradient descent [59] using Cosine learning rate [60]. The candidate architectures are initialized by He initialization [57] and trained by Adam optimizer [58] with a learning rate of 0.00035. Additionally, each search is run for 200 epochs.

4.3. Experimental Results and Comparison

4.3.1. Evaluation on 3D Pose Estimation

The effectiveness of the proposed 3D pose estimation network is evaluated using the standard protocol of the Human3.6M dataset [18,27,29,34]. Five subjects S1, S5, S6, S7, S8 are used for training and the remaining two subjects S9, S11 are used for evaluation. Experimental results are reported by the average error in millimeters between the ground truth and the corresponding predictions over all joints. Much to our surprise, this method outperforms the previous best result from the literature [34] by 3.1mm, corresponding to an error reduction of 6.8% even when combining the ground truth 2D locations with the 2D OpenPose detections. This result proves that the network design can learn to recover the 3D pose from the 2D joint locations with a remarkably low error rate, which to the best of our knowledge, has established a new state-of-the-art on 3D human pose estimation (see Table 1 and Figure 7).

Table 1. Experimental results (average error in mm) and comparison with previous state-of-the-art 3D pose estimation approaches on the Human3.6M dataset [18]. The symbol * denotes that a 2D detector was used and the symbol † denotes the ground truth 2D joint locations were used.

Method	Direct.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg
Ionescu et al. [18]†	132.7	183.6	132.3	164.4	162.1	205.9	150.6	171.3	151.6	243.0	162.1	170.7	177.1	96.6	127.9	162.1
Du et al. [61]*	85.1	112.7	104.9	122.1	139.1	135.9	105.9	166.2	117.5	226.9	120.0	117.7	137.4	99.3	106.5	126.5
Tekin et al. [26]	102.4	147 2	88.8	125.3	118.0	182.7	112.4	129.2	138.9	224.9	118.4	138.8	126.3	55.1	65.8	125.0
Park et al. [62]*	100.3	116 2	90.0	116.5	115.3	149.5	117.6	106.9	137 2	190.8	105.8	125.1	131.9	62.6	96 2	117.3
Zhou et al. [63]*	87.4	109.3	87.1	103 2	116 2	143.3	106.9	99.8	124.5	199 2	107.4	118.1	114.2	79.4	97.7	113.0
Zhou et al. [64]*	91.8	102.4	96.7	98.8	113.4	125.2	90.0	93.8	132 2	159.0	107.0	94.4	126.0	79.0	99.0	107.3
Pavlakos et al. [27]	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63 2	71.9
Mehta et al. [65]*	67.4	71.9	66.7	69.1	71.9	65.0	68.3	83.7	120.0	66.0	79.8	63.9	48.9	76.8	53.7	68.6
Martinez et al. [34]*	51.8	56.2	58.1	59.0	69.5	55 2	58.1	74.0	94.6	62.3	78.4	59.1	49.5	65.1	52.4	62.9
Liang et al. [66]	52.8	54.2	54.3	61.8	53.1	53.6	71.7	86.7	61.5	53.4	67 2	54.8	53.4	47.1	61.6	59.1
Luvizon et al. [44]	49 2	51.6	47.6	50.5	51.8	48.5	51.7	61.5	70.9	53.7	60.3	48.9	44.4	57.9	48.9	53.2
Martinez et al. [34]†	37.7	44.4	40.3	42.1	48 2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Ours†,*	**36.6**	**43.2**	**38.1**	**40.8**	**44.4**	**51.8**	**43.7**	**38.4**	**50.8**	**52.0**	**42.1**	**42.2**	**44.0**	**32.3**	**35.9**	**42.4**

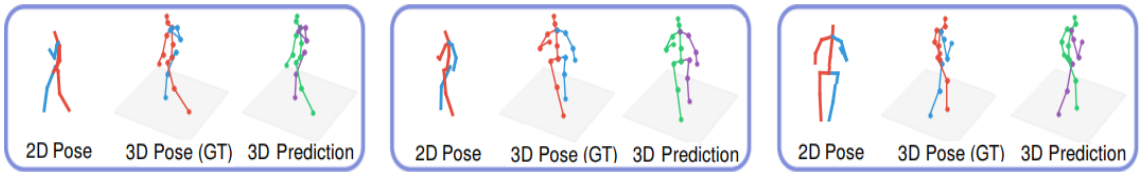

Figure 7. Visualization of 3D output of the estimation stage with some samples on the test set of Human3.6M [18]. For each example, from left to right are 2D poses, 3D ground truths and the 3D predictions, respectively.

4.3.2. Evaluation on Action Recognition

Table 2 reports the experimental results and comparisons with state-of-the-art methods on the MSR Action3D dataset [19]. The ENAS algorithm [14] is able to explore a diversity of network architectures and the best design is identified based on its validation score. Thus, the final architecture achieved a total average accuracy of 97.98% over three subset AS1, AS2 and AS3. This result outperforms many previous studies [9,19,37,38,67–71], and among them, many are depth sensor-based approaches. Figure 8 provides a schematic diagram of the best cells and optimal architecture found by ENAS on the AS1 subset [19]. For the SBU Kinect Interaction dataset [20], the best model achieved an accuracy of 96.30%, as shown in Table 3. The results reported here indicated an important observation that by using only the 3D predicted poses, it was possible to outperform previous works reported in [37,72–77] and reach state-of-the-art results provided in [54,78], which deploy accurate skeletal data provided by Kinect v2 sensor.

Table 2. Test accuracies (%) on the MSR Action3D dataset [19]. Please note that many previous methods are based on depth sensors.

Method	AS1	AS2	AS3	Aver.
Li et al. [19]	72.90	71.90	71.90	74.70
Chen et al. [67]	96.20	83.20	92.00	90.47
Vemulapalli et al. [9]	95.29	83.87	98.22	92.46
Du et al. [38]	99.33	94.64	95.50	94.49
Liu et al. [37]	N/A	N/A	N/A	94.80
Wang et al. [68]	93.60	95.50	95.10	94.80
Wang et al. [69]	91.50	95.60	97.30	94.80
Xu et al. [70]	99.10	92.90	96.40	96.10
Lee et al. [71]	95.24	96.43	100.0	97.22
Pham et al. [54]	98.83	99.06	99.40	99.10
Ours	**97.87**	**96.81**	**99.27**	**97.98**

Table 3. Test accuracies (%) on the SBU Kinect Interaction dataset [20]. Please note that many previous methods are based on depth sensors.

Method	Acc.
Song et al. [72]	91.51
Liu et al. [37]	93.30
Weng et al. [73]	93.30
Ke et al. [74]	93.57
Tas et al. [75]	94.36
Wang et al. [76]	94.80
Liu et al. [77]	94.90
Zang et al. [78] (using VA-RNN)	95.70
Zhang et al. [78] (using VA-CNN)	97.50
Pham et al. [54]	97.86
Ours	**96.30**

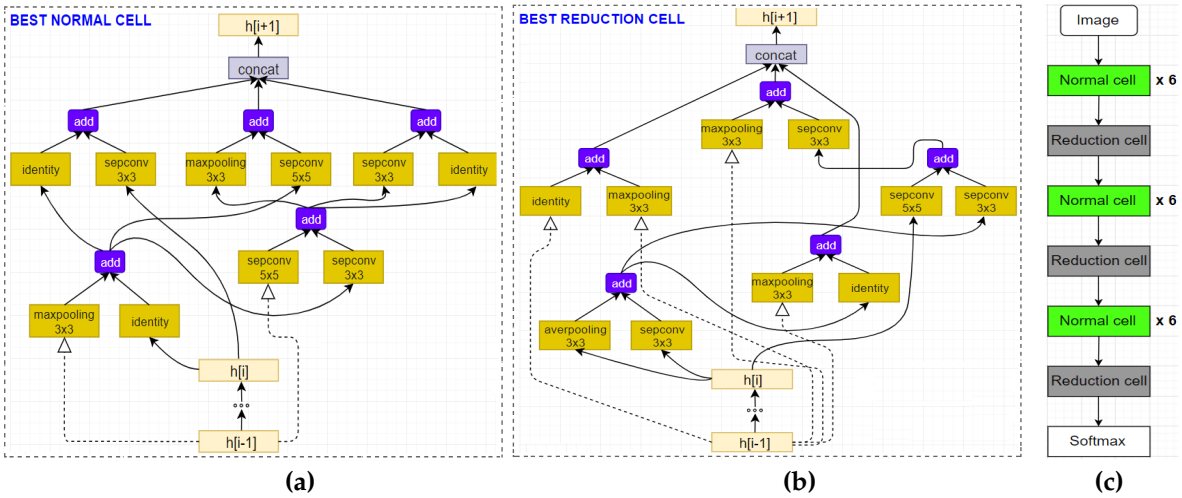

Figure 8. Diagram of the top performing normal cell **(a)** and reduction cell **(b)** discovered by ENAS [14] on AS1 subset [19]. They were then used to construct the final network architecture **(c)**. We recommend the interested readers to see [14] to better understand this procedure.

4.4. Computational Efficiency Evaluation

On a single GeForce GTX 1080Ti GPU with 11GB memory, the runtime of OpenPose [12] is less than 0.1*s* per frame for an image size of 800 $\times$ 450 pixels. On the Human3.6M dataset [18], the 3D pose estimation stage takes around 15*ms* to complete a pass (forward + backward) through each stream with a mini batches of size 128. Each epoch was done within 3 min. For the action recognition stage, our implementation of the ENAS algorithm takes about 2 h to find the final architecture ($\sim$2.3M parameters) on each subset of MSR Action3D dataset [19], while it takes around 3 h on the SBU Kinect Interaction dataset [20] to discover the best architecture ($\sim$3M parameters). With small architecture sizes, the discovered networks require low computing time for the inference stage, making the approach more practical for large-scale problems and real-time applications.

5. Conclusions

In this paper, a unified deep learning framework for joint 3D human pose estimation and action recognition from RGB video sequences has been presented. The proposed method first runs a state-of-the-art 2D pose detector to estimate 2D locations of body joints from a monocular RGB sensor, although the approach is not limited to a particular 2D pose detector. A deep neural network was then designed and trained to learn a direct 2D-to-3D mapping and predict human poses in 3D space. Experimental results demonstrated that the 3D human poses can be effectively estimated by a simple network design and training methodology over 2D keypoints. A novel action recognition approach was also introduced based on a compact image-based representation and automated machine learning, in which an advanced neural architecture search algorithm was exploited to discover the best performing architecture for each recognition task. The experiments on public and challenging action recognition datasets indicated that the proposed framework was able to reach state-of-the-art performance, while requiring less computation budget for training and inference. Despite that, this method naturally depends on the quality of the output of the 2D detectors. Hence, a limitation is that it cannot recover 3D poses from 2D failed output. To tackle this problem, we are currently expanding this study by adding more visual evidence to the network to further gains in performance. The preliminary results are encouraging. Codes and models will be made available on our GitHub project at https://github.com/huyhieupham/.

Author Contributions: Conceptualization, methodology, software, validation, investigation, resources, visualization, writing—original draft preparation, H.H.P.; Data curation, formal analysis, writing—review and editing, all authors; Supervision; project administration; funding acquisition, H.S., A.C., P.Z., L.K. and S.A.V. All authors have read and agreed to the published version of the manuscript.

Funding: Sergio A. Velastin is grateful for funding received from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement N 600371, el Ministerio de Economía, Industria y Competitividad (COFUND2013-51509) el Ministerio de Educación, Cultura y Deporte (CEI-15-17) and Banco Santander.

Acknowledgments: This research was carried out at the Cerema Research Center and Informatics Research Institute of Toulouse, Paul Sabatier University, France. The authors would like to express our thanks to all the people who have made helpful comments and suggestions on a previous draft. S.A. Velastin is grateful to funding received from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 600371, el Ministerio de Economía y Competitividad (COFUND2013-51509) and Banco Santander.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Weinland, D.; Ronfard, R.; Boyer, E. A Survey of Vision-based Methods for Action Representation, Segmentation and Recognition. *CVIU* **2011**, *115*, 224–241. [CrossRef]
2. Lowe, D. Distinctive Image Features from Scale-invariant Keypoints. *IJCV* **2004**, *60*, 91–110. [CrossRef]
3. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning Realistic Human Actions from Movies. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AL, USA, 24–26 June 2008; pp. 1–8.
4. Dollár, P.; Rabaud, V.; Cottrell, G.; Belongie, S. Behavior Recognition via Sparse Spatio-temporal Features. In Proceedings of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), Breckenridge, CO, USA, 7 January 2005; pp. 65–72.
5. Ye, M.; Yang, R. Real-time Simultaneous Pose and Shape Estimation for Articulated Objects using a Single Depth Camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 2345–2352.
6. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Rhode Island, USA, 18–20 June 2012; pp. 1290–1297.
7. Xia, L.; Chen, C.; Aggarwal, J.K. View-Invariant Human Action Recognition using Histograms of 3D Joints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Rhode Island, USA, 18–20 June 2012; pp. 20–27.
8. Chaudhry, R.; Ofli, F.; Kurillo, G.; Bajcsy, R.; Vidal, R. Bio-inspired Dynamic 3D Discriminative Skeletal Features for Human Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 471–478.
9. Vemulapalli, R.; Arrate, F.; Chellappa, R. Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 588–595.
10. Ding, W.; Liu, K.; Fu, X.; Cheng, F. Profile HMMs for Skeleton-based Human Action Recognition. *Signal Process. Image Commun.* **2016**, *42*, 109–119. [CrossRef]
11. Zhang, Z. Microsoft Kinect Sensor and Its Effect. *IEEE Multimed.* **2012**, *19*, 4–10. [CrossRef]
12. Cao, Z.; Simon, T.; Wei, S.; Sheikh, Y. Realtime Multi-person 2D Pose Estimation using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–25 July 2017; pp. 7291–7299.
13. Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C. RMPE: Regional Multi-person Pose Estimation. ICCV, 2017. Available online: https://github.com/MVIG-SJTU/AlphaPose (accessed on 23 March 2020).
14. Pham, H.; Guan, M.; Zoph, B.; Le, Q.; Dean, J. Efficient Neural Architecture Search via Parameters Sharing. In Proceedings of the International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; pp. 4095–4104.
15. Johansson, G. Visual Motion Perception. *Sci. Am.* **1975**, *232*, 76–89. [CrossRef] [PubMed]
16. Gu, J.; Ding, X.; Wang, S.; Wu, Y. Action and Gait Recognition from Recovered 3D Human Joints. *IEEE Trans. Syst. Man Cybern.* **2010**, *40*, 1021–1033.

17. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 483–499.
18. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2014**, *36*, 1325–1339. [CrossRef] [PubMed]
19. Li, W.; Zhang, Z.; Liu, Z. Action Recognition Based on a Bag of 3D Points. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 9–14.
20. Yun, K.; Honorio, J.; Chattopadhyay, D.; Berg, T.L.; Samaras, D. Two-person Interaction Detection using Body-pose Features and Multiple Instance Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 18–20 June 2012; pp. 28–35.
21. Nikolaos, S.; Bogdan, B.; Bogdan, I.; Ioannis, A.K. 3D Human Pose Estimation: A Review of the Literature and Analysis of Covariates. *CVIU* **2016**, *152*, 1–20.
22. Presti, L.; La Cascia, M. 3D Skeleton-based Human Action Classification: A Survey. *Pattern Recognit.* **2016**, *53*, 130–147. [CrossRef]
23. Sminchisescu, C. 3D Human Motion Analysis in Monocular Video Techniques and Challenges. In Proceedings of the IEEE International Conference on Video and Signal Based Surveillance (ICVSBS), Sydney, Australia, 22–24 November 2006; p. 76.
24. Ramakrishna, V.; Kanade, T.; Sheikh, Y. Reconstructing 3D Human Pose from 2D Image Landmarks. In Proceedings of the European Conference on Computer Vision (ECCV), Florence, Italy, 7–13 October 2012; pp. 573–586.
25. Li, S.; Chan, A.B. 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network. In Proceedings of the Asian Conference on Computer Vision (ACCV), Singapore, 1–5 November 2014; pp. 332–347.
26. Tekin, B.; Rozantsev, A.; Lepetit, V.; Fua, P. Direct Prediction of 3D Body Poses from Motion Compensated Sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 991–1000.
27. Pavlakos, G.; Zhou, X.; Derpanis, K.G.; Daniilidis, K. Coarse-to-fine Volumetric Prediction for Single-image 3D Human Pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–25 July 2017; pp. 7025–7034.
28. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-supervised Training. *arXiv* **2018**, arXiv:1811.11742.
29. Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H.P.; Xu, W.; Casas, D.; Theobalt, C. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *ACM TOG* **2017**, *36*, 44. [CrossRef]
30. Katircioglu, I.; Tekin, B.; Salzmann, M.; Lepetit, V.; Fua, P. Learning Latent Representations of 3D Human Pose with Deep Neural Networks. *IJCV* **2018**, *126*, 1326–1341. [CrossRef]
31. Fisher, Y.; Vladlen, K. Multi-scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
33. Sepp, H.; Jürgen, S. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780.
34. Martinez, J.; Hossain, R.; Romero, J.; Little, J. A Simple Yet Effective Baseline for 3D Human Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2640–2649.
35. Lv, F.; Nevatia, R. Recognition and Segmentation of 3D Human Action Using HMM and Multi-class AdaBoost. In Proceedings of the European Conference on Computer Vision (ECCV), Graz, Austria, 7–13 May 2006; pp. 359–372.
36. Han, L.; Wu, X.; Liang, W.; Hou, G.; Jia, Y. Discriminative Human Action Recognition in the Learned Hierarchical Manifold Space. *Image Vis. Comput.* **2010**, *28*. [CrossRef]

37. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal LSTM with Trust Gates for 3D Human Action Recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 816–833.
38. Du, Y.; Wang, W.; Wang, L. Hierarchical Recurrent Neural Network for Skeleton based Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015; pp. 1110–1118.
39. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+ D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 26 June–1 July 2016; pp. 1010–1019.
40. Sainath, T.N.; Vinyals, O.; Senior, A.; Sak, H. Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 4580–4584.
41. Chéron, G.; Laptev, I.; Schmid, C. P-CNN: Pose-based CNN Features for Action Recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 3218–3226.
42. Yao, B.; Fei-Fei, L. Modeling Mutual Context of Object and Human Pose in Human-object Interaction Activities. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 17–24.
43. Nie, B.X.; Xiong, C.; Zhu, S. Joint Action Recognition and Pose Estimation from Video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015; pp. 1293–1301.
44. Luvizon, D.C.; Picard, D.; Tabia, H. 2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5137–5146.
45. Huber, P.J. Robust Estimation of a Location Parameter. In *Breakthroughs in Statistics*; Springer: New York, NY, USA, 1992; pp. 492–518.
46. Christian, S.; Sergey, I.; Vincent, V. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Phoenix, AR, USA, 12–17 February 2016.
47. Gao, H.; Zhuang, L.; Laurens van der, M.; Kilian, Q.W. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
48. Barret, Z.; Quoc, V.L. Neural Architecture Search with Reinforcement Learning. *arXiv* **2017**, arXiv:1611.01578.
49. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015.
50. Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Improving Neural Networks by Preventing Co-adaptation of Feature Detectors. *arXiv* **2012**, arXiv:1207.0580.
51. Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. Self-Normalizing Neural Networks. *Adv. Neural Inf. Process. Syst. (NIPS)* **2017**, 971–980.
52. Pham, H.; Khoudour, L.; Crouzil, A.; Zegers, P.; Velastin, S.A. Exploiting Deep Residual Networks for Human Action Recognition from Skeletal Data. *CVIU* **2018**, *170*, 51–66. [CrossRef]
53. Pham, H.; Khoudour, L.; Crouzil, A.; Zegers, P.; Velastin, S.A. Skeletal Movement to Color Map: A Novel Representation for 3D Action Recognition with Inception Residual Networks. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 3483–3487.
54. Pham, H.; Salmane, H.; Khoudour, L.; Crouzil, A.; Zegers, P.; Velastin, S.A. Spatio-Temporal Image Representation of 3D Skeletal Movements for View-Invariant Action Recognition with Deep Convolutional Neural Networks. *Sensors* **2019**, *19*. [CrossRef]
55. Pizer, S.M.; Amburn, E.P.; Austin, J.D.; Cromartie, R.; Geselowitz, A.; Greer, T.; ter Haar Romeny, B.; Zimmerman, J.B.; Zuiderveld, K. Adaptive Histogram Equalization and Its Variations. *Comput. Vision, Graph. Image Process.* **1987**, *39*, 355–368. [CrossRef]

56. Pham, H.H.; Salmane, H.; Khoudour, L.; Crouzil, A.; Zegers, P.; Velastin, S.A. A Deep Learning Approach for Real-Time 3D Human Action Recognition from Skeletal Data. In Proceedings of the International Conference on Image Analysis and Recognition. Springer, Waterloo, Canada, 27–29 August 2019; pp. 18–32.
57. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 1026–1034.
58. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
59. Yurii, N. A Method for Solving a Convex Programming Problem with Convergence Rate $O(1/K2)$. *Sov. Math. Dokl.* **1983**, 372–377.
60. Ilya, L.; Frank, H. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv* **2016**, arXiv:1608.03983.
61. Du, Y.; Wong, Y.; Liu, Y.; Han, F.; Gui, Y.; Wang, Z.; Kankanhalli, M.; Geng, W. Marker-less 3D Human Motion Capture with Monocular Image Sequence and Height-maps. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 20–36.
62. Park, S.; Hwang, J.; Kwak, N. 3D Human Pose Estimation using Convolutional Neural Networks with 2D Pose Information. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 156–169.
63. Zhou, X.; Zhu, M.; Leonardos, S.; Derpanis, K.G.; Daniilidis, K. Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4966–4975.
64. Xingyi, Z.; Xiao, S.; Wei, Z.; Shuang, L.; Yichen, W. Deep Kinematic Pose Regression. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 186–201.
65. Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; Theobalt, C. Monocular 3D Human Pose Estimation in the Wild using Improved CNN Supervision. In Proceedings of the International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 506–516.
66. Shuang, L.; Xiao, S.; Yichen, W. Compositional Human Pose Regression. *Comput. Vis. Image Underst.* **2018**, *176–177*, 1–8.
67. Chen, C.; Liu, K.; Kehtarnavaz, N. Real-time Human Action Recognition based on Depth Motion Maps. *J. -Real-Time Image Process.* **2016**, *12*. [CrossRef]
68. Wang, P.; Yuan, C.; Hu, W.; Li, B.; Zhang, Y. Graph Based Skeleton Motion Representation and Similarity Measurement for Action Recognition. In Proceedings of the British Machine Vision Conference (BMVC), York, UK, 19–22 September 2016.
69. Weng, J.; Weng, C.; Yuan, J. Spatio-Temporal Naive-Bayes Nearest-Neighbor (ST-NBNN) for Skeleton-Based Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–26 July 2017.
70. Xu, H.; Chen, E.; Liang, C.; Qi, L.; Guan, L. Spatio-temporal Pyramid Model based on Depth Maps for Action Recognition. In Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP), Xiamen, China, 19–21 October 2015; pp. 1–6.
71. Lee, I.; Kim, D.; Kang, S.; Lee, S. Ensemble Deep Learning for Skeleton-based Action Recognition using Temporal Sliding LSTM Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1012–1020.
72. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), San Francisco, CA, USA, 4–9 February 2017.
73. Weng, J.; Weng, C.; Yuan, J.; Liu, Z. Discriminative Spatio-Temporal Pattern Discovery for 3D Action Recognition. *IEEE Trans. Circuits Syst. Video Technol. (TCCVT)* **2019**, *29*, 1077–1089. [CrossRef]
74. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. A New Representation of Skeleton Sequences for 3D Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21-26 July 2017; pp. 4570–4579.
75. Yusuf, T.; Piotr, K. CNN-based Action Recognition and Supervised Domain Adaptation on 3D Body Skeletons via Kernel Feature Maps. In Proceedings of the British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018, p. 158.

76. Wang, H.; Wang, L. Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3633–3642.
77. Liu, J.; Wang, G.; Duan, L.; Abdiyeva, K.; Kot, A.C. Skeleton-Based Human Action Recognition With Global Context-Aware Attention LSTM Networks. *IEEE Trans. Image Process. (TIP)* **2018**, *27*, 1586–1599. [CrossRef]
78. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View Adaptive Neural Networks for High Performance Skeleton-based Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2019**, 1963–1978. [CrossRef]