informatics

MDPI

Article

Deep Full-Body HPE for Activity Recognition from RGB Frames Only

Sameh Neili Boualia [1,2,*] and Najoua Essoukri Ben Amara [2,*]

[1] University of Tunis El Manar, National Engineering School of Tunis, 1002 Tunis, Tunisia
[2] Université de Sousse, Ecole Nationale d'Ingénieurs de Sousse, LATIS- Laboratory of Advanced Technology and Intelligent Systems, 4023 Sousse, Tunisie
* Correspondence: sameh.neili@gmail.com (S.N.B.); najoua.benamara@eniso.rnu.tn (N.E.B.A.)

Abstract: Human Pose Estimation (HPE) is defined as the problem of human joints' localization (also known as keypoints: elbows, wrists, etc.) in images or videos. It is also defined as the search for a specific pose in space of all articulated joints. HPE has recently received significant attention from the scientific community. The main reason behind this trend is that pose estimation is considered as a key step for many computer vision tasks. Although many approaches have reported promising results, this domain remains largely unsolved due to several challenges such as occlusions, small and barely visible joints, and variations in clothing and lighting. In the last few years, the power of deep neural networks has been demonstrated in a wide variety of computer vision problems and especially the HPE task. In this context, we present in this paper a Deep Full-Body-HPE (DFB-HPE) approach from RGB images only. Based on ConvNets, fifteen human joint positions are predicted and can be further exploited for a large range of applications such as gesture recognition, sports performance analysis, or human-robot interaction. To evaluate the proposed deep pose estimation model, we apply it to recognize the daily activities of a person in an unconstrained environment. Therefore, the extracted features, represented by deep estimated poses, are fed to an SVM classifier. To validate the proposed architecture, our approach is tested on two publicly available benchmarks for pose estimation and activity recognition, namely the J-HMDBand CAD-60datasets. The obtained results demonstrate the efficiency of the proposed method based on ConvNets and SVM and prove how deep pose estimation can improve the recognition accuracy. By means of comparison with state-of-the-art methods, we achieve the best HPE performance, as well as the best activity recognition precision on the CAD-60 dataset.

Keywords: human pose estimation; human activity recognition; deep learning; ConvNets; SVM

Citation: Neili Boualia, S.; Essoukri Ben Amara, N. Deep Full-Body HPE for Activity Recognition from RGB Frames Only. *Informatics* **2021**, *8*, 2. https://doi.org/10.3390/informatics8010002

Received: 30 November 2020
Accepted: 13 January 2021
Published: 18 January 2021

1. Introduction

Currently, the amount of available video data is explosively expanding due to the pervasiveness of digital recording devices. Estimating human poses in those videos is one of the longstanding research topics in the computer vision community, which has been extensively studied in recent years. Scientifically speaking, Human Pose Estimation (HPE) refers to the the method of localizing the human body parts (3D pose) or their projection onto a picture plane (2D pose). Video-based HPE has attracted increasing interest in recent years thanks to its wide range of applications including: human-computer interaction [1,2], sports performance analysis [3], and video surveillance [4–6]. Although the research has advanced in this field, there are still many remaining challenges such as: the high changes in human body shapes, clothing and viewpoint variations, and the conditions of system acquisition (day and night illumination variations, occlusions, etc.).

Previous works on HPE have commonly used graphical models for estimating human poses. Generally, those models are composed of joints and rigid parts. Using image-based observations, most of these classic methods follow a two step framework. The first step

is based on the extraction of hand-crafted features from raw data, and the second one consists of learning classifiers on the obtained features. In [7], the authors presented a graphical model for HPE with image-dependent pairwise relations. They used the local image measurements, not only to detect joints, but also to predict the spatial relationships between them. This aims to learn conditional probabilities for the presence of parts and their spatial relationships. After that, another approach was proposed using puppets [8]. It estimates the body poses at one frame, then checks its performance in neighboring ones using the optical flow.

Recently, following their significant progress in static image classification, Convolutional Neural Networks (CNNs/ConvNets) have been extended to take into account motion information in order to be exploited in video-based HPE. Compared with the conventional machine learning methods, deep learning techniques have a more powerful learning ability. They have shown remarkable progress due to their high precision and robustness.

In this work, we are particularly interested in estimating human poses and detecting different body parts under challenging conditions. Those human poses, which represent extracted features, will be fed to a classification stage using SVM in order to recognize daily activities. This paper presents the following novel contributions:

- We present an end-to-end CNN that exploits RGB data only for a full-body pose estimation. The estimated person poses are then considered as discriminative features to recognize different human activities.
- We extensively evaluate various aspects of our HPE architecture: We test different model parameters (including: iteration number, data augmentation techniques, and heat map size). We compare the proposed model with previous approaches on common benchmark datasets (i.e., J-HMDBand CAD-60) for which interesting results for HPE and activity recognition are reported.
- We recognize human activities using human poses rather than RGB information. We conclude that the quality of the estimated poses significantly affects the recognition performance.

The remainder of this paper is organized as follows. In Section 2, we review recent work on 2D HPE, which can be divided into two main classes: traditional HPE approaches (Section 2.1) and deep learning-based ones (Section 2.2). Recent deep learning-based HAR approaches are explored in Section 2.3. Then, we describe the proposed DFB-HPE (Deep Full-Body-HPE) approach in Section 3 where different training details are explained. In Section 4, we present the datasets used (Section 4.1) and different evaluation metrics (Section 4.2). After that, we discuss the obtained results on the benchmarks used. Finally, we conclude our work in Section 5, where potential future studies are proposed.

2. Related Work

Human poses are important cues for video analysis in a variety of tasks such as activity/action recognition [9,10], multi-object detection [11], and sign language processing and recognition [12]. Generally, HPE approaches can be divided into two main groups: traditional HPE approaches and deep learning-based ones. For more details on pose-based HAR, you can refer to our review [13].

2.1. Traditional HPE Approaches

Past work on HPE has been basically founded on hand-crafted features extracted from raw data. At first, the traditional approaches frequently utilized graphical structure models for recovering human poses using image-based observations. Generative approaches (referred to as model-based or top-down) aim to locate different body parts in video frames [7]. Based essentially on joints and rigid parts, those techniques use a priori information such as motion [14] and context [15]. Thus, an HPE process based on a generative approach is principally composed of two levels: modeling the human body explicitly and then estimating the different joint positions. Depicted as a skeleton, the

human body is represented through a collection of parts connected by a set of constraints imposed by different joints (Figure 1).

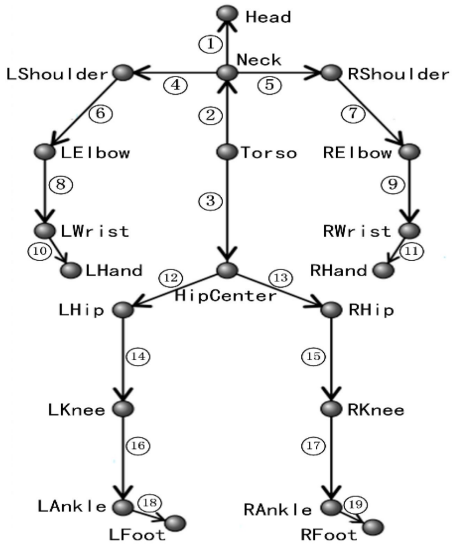

Figure 1. Example of a human skeleton body model with 19 joints.

For example, Ferrari et al. [16] proposed a methodology dependent on a conventional detector, which uses an upper-body pose estimation model from TV and movie video shots. Unless they exploited the belief propagation technique to refine estimated poses, the suggested detector seemed to be sensitive to self-occlusions. To solve this problem, Shotton et al. [17] proposed a new technique to reformulate the pose estimation problem into a simpler per-pixel classification task. This method was based on a human body-part segmentation from depth frames. In order to reduce the over-fitting problem, the authors exploited a randomized decision forest. Later on, Chen et al. [7] used the local image measurements, not only to detect joints, but also to predict the spatial relationships between them. The types of part relations were learned with K-means clustering in the experiments and governed spatial connections between the parts. Besides, a new approach was put forward in [8] utilizing puppets. The solution was to estimate the pose of the body only at one frame and then use the optical flow technique to check its performance in neighboring ones.

Unlike generative approaches, the discriminative ones are model-free and do not assume any particular human skeleton structure constraint. They are based essentially on learning a mapping between image observations and body poses. For example, Poppe [18] presented an example-based approach to pose recovery using histograms of oriented gradients as image descriptors. Niyogi and Freeman [19] estimated the pose of human heads using a nonlinear mapping from the input image to an output parametric description. The mapping was calculated through examples from a training set, where the output pose was presented as that of the nearest example input neighbor.

2.2. Deep HPE Approaches

The classical pipeline of HPE has shown some limitations. Recently, this domain has been greatly reshaped by new deep learning techniques. This new type of technology no longer needs hand-crafted features. They provide several layers of feature extractors, which make it easier to implicitly learn the patterns of each joint. With the introduction of "DeepPose" by Toshev et al. [20], researchers of HPE began to shift from classic approaches to deep learning. Most of the recent pose estimation systems have adopted ConvNets as their main building block, largely replacing hand-crafted features and graphical models. This technique has yielded great improvements on standard benchmarks. DeepPose was the first work that benefited from deep learning for HPE. In this approach, pose estimation

was formulated as a CNN-based regression problem towards body joints. The authors used a cascade of DNN regressors to refine the estimated pose. Gkioxari [21] used a CNN architecture for both pose estimation and action detection. In order to determine different human attributes, a collection of CNNs was trained in [22], where each one learned a poselet [23] from a set of image patches. The architecture consisted of four stages of convolution-normalization-pooling layers, one fully connected layer, and a logistic regression one utilized as a classifier of a linear nature. Poselets have been commonly used in conjunction with CNNs for people detection and pose estimation as well. Later on, in [24], the authors introduced a new top-down procedure, called iterative error feedback, which allowed error predictions to be fed back in the CNN to progressively change the initial solution. Another study [25] proposed to apply the convolutions and pooling steps in a way that would permit the image to be processed repeatedly in a bottom-up and top-down manner with intermediate supervision. Later, Belagiannis and Zisserman [26] combined feed forward and recurrent modules in a CNN-based HPE model. In [27], the authors suggested to integrate a consensus voting scheme within a CNN, where votes gathered from every location per keypoint were aggregated to obtain a probability distribution for each keypoint location. Another CNN was trained in [28] to infer 3D human poses from uncertainty maps of 2D joint estimates. To estimate human poses in videos, the authors in [29] exploited the ability of CNNs to benefit from temporal context, which was established by combining information between successive time frames using an optical flow. Recently, Nibali et al. [30] proposed some improvements in the HPE domain. They extended the heat map-based output strategies commonly used in 2D pose estimation to the task of 3D HPE. They predicted three two-dimensional marginal heat maps per joint under an augmented soft-argmax scheme. Using post-data augmentation techniques to improve the quality of extreme/wild motions' pose estimation, Toyoda et al. [31] proposed a method that augmented the input data with rotation augmentation, then applied the pose estimation technique multiple times for every frame. The most consistent pose was then selected followed by a motion reconstruction for smoothing. In [32], Kreiss et al. proposed a new bottom-up method for multi-person 2D human pose estimation that was particularly well suited for urban mobility such as self-driving cars and delivery robots. Their method was based on two parts: the PIF (Part Intensity Field) to localize different body parts and the PAF (Part Association Field) to associate body parts with each other and form full human poses. All models used were based on ImageNet pretrained base networks. Gartner et al. [33] proposed a fully trainable deep reinforcement learning-based active pose estimation architecture, which learns to select appropriate views, in space and time, to feed an underlying monocular pose estimator: "Pose-DRL". Considering the progress in computer vision for HPE, the authors in [34] showed how new deep learning architectures can influence animal pose estimation, which encourages neuroscience laboratories to leverage these tools for better quantification of behavior.

2.3. Deep HAR Approaches

For Human Action Recognition (HAR) (such as "walking", "open door", "sit down", etc.), many approaches based on deep learning techniques have been proposed in the last few years. Those approaches can be classified according to the DL model used: 2D or 3D. In the following, we present a selection of deep learning-based research works for HAR. For 2D CNN-based HAR approaches, Simonyan et al. [35] implemented a two stream ConvNet where the spatial stream recognizes the action from still frames and the temporal stream performs recognition from the motion in the form of dense optical flow. This method achieved good results on the UCF-101 and HMDB-51 datasets. However, according to the authors, the proposed model may not be suitable for real-time applications due to its computational complexity. Moreover, in [36], the authors adapted the successful deep learning architectures to the design of a two stream ConvNet for action recognition in videos, which they called "very deep two stream ConvNets". They empirically studied both GoogLeNet and VGG-16 for the design of the proposed model. In relation to [35], they presented two

novelties: (i) they extended the famous Caffe toolbox into a multi-GPU implementation with high efficiency and low memory consumption and (ii) proposed several good practices for the training of the ConvNet architecture (learning rate arrangement, data augmentation techniques, etc.). For evaluation, the UCF101 dataset was used with which they achieved a recognition accuracy of 91.4%. Later on, Ijjina et al. [37] proposed a new approach for HAR based on Genetic Algorithms (GAs) and CNNs. They demonstrated that initializing the weights of a CNN classifier based on solutions generated by the GA minimized the classification error. To demonstrate the efficacy of the proposed classification system, they evaluated their CNN-GA model on the UCF50 dataset, achieving 96.88% as the average accuracy rate.

Most of the current CNN methods use architectures with 2D convolutions, enabling shift-invariant representations in the image plane. However, the invariance to translations in the time axis is also important for HAR since the beginning and the end of the action are generally unknown. Thus, a CNN with 3D spatio-temporal convolutions addresses this issue and provides a natural extension of a 2D CNN to video. In [38], the authors developed a novel deep model for automatic activity recognition from RGB-D videos. Each human activity was presented as an ensemble of cubic-like video segments and learned to discover the temporal structures for each category of activities. Their proposed ConvNet-based model consisted of 3D convolutions and max-pooling operators over the video segments. Later, Shao et al. [39] mixed appearance and motion features for recognizing group activities in crowded scenes collected from the web. For the combination of the different modalities, the authors applied multitask deep learning. By these means, they were able to capture the intra-class correlations between the learned attributes while they proposed a novel dataset of crowed scene understanding called the "WWWcrowd" dataset. Another approach using spatio-temporal features with a 3D convolutional network was proposed in [40]. Experimentally, the authors showed that 3D CNNs are more suitable for spatio-temporal features than 2D CNNs. Furthermore, they empirically demonstrated that the CNN architecture with small $3 \times 3 \times 3$ kernels was the best choice for spatio-temporal features. Achieving 52.8% accuracy on the UCF101 dataset, their model was computationally efficient due to the fast inference of ConvNets. Just recently, Varol et al. [41] proposed the LTC-CNN model: a combination of Long-term Temporal Convolutions (LTC) with CNN in order to learn video representations. They investigated multi-resolution representations of both motion and appearance. They demonstrated the importance of high-quality optical flow estimation on action recognition accuracy. The model was tested on two recent and challenging human action benchmarks: UCF101 and HMDB51 and reported state-of-the-art performance. Shou et al. [42] also designed a novel 3D CNN model named the Convolutional-De-Convolutional (CDC) network, where CDC filters were implemented prior to a 3D ConvNet. Shou et al. were the first to combine two reverse operations (convolution and de-convolution) into a joint CDC filter. The proposed CDC conducted down-sampling in space and up-sampling in time simultaneously to infer both high-level action semantics and temporal dynamics.

3. Materials and Methods

The proposed DFB-HPE approach was basically inspired by [29]. The basic HPE architecture consists of a two stage process for upper-body pose estimation: (i) spatial layers and (ii) temporal layers. The first stage is used to calculate different upper joint positions from RGB video frames. The heat map joints are then fed to the second stage, the "temporal pooler", in order to consider the temporal dimension with the optical flow technique.

In order to take into account the full-body pose estimation, we modified the already mentioned architecture considering the fact that adding the lower-body joints should improve the pose estimation results as far as the activity recognition rate [43], opening up other possibilities for applications. Indeed, the suggested architecture consists of several convolution, pooling, and loss layers. As depicted in Figure 2, the overall network is

composed of two levels: (i) fully-convolutional layers and (ii) fusion layers. The input is a set of RGB video frames with a 320 × 240 resolution. For each frame, fifteen key joint positions are predicted. The output of the last loss layer (*loss*2) represents the 2D coordinates of the full-body joint positions. Later, those positions will be the input to the SVM classifier in order to recognize the human activity. The first stage of the proposed architecture is fully convolutional: eight convolution layers with a stride equal to 1, where the first two layers are followed by a 2 × 2 max-pooling layer with a stride equal to 2. The output of the "conv8" layer is a set of heat maps with a fixed size $i \times j \times k$, where i and j represent the heat map size and k is the number of joints to regress (here, 60 × 60 × 15). In order to learn the dependencies between the locations of human body parts, the convolution layer "conv7", which shows pre-heat map activations, is concatenated with "conv3", which represents a skip layer. In fact, training deep networks especially with a small amount of data can lead to many problems, namely vanishing and exploding gradients. In order to deal with this issue, we used a specific layer named the skip connection/layer, where activations are taken from one layer and fed to another one that is deeper in the network. This concatenation represents the input of the second stage of the fusion layers. We should note that the proposed network architecture is based on regressing heat maps for each joint instead of directly regressing the positions of the joints as this is a highly non-linear problem.

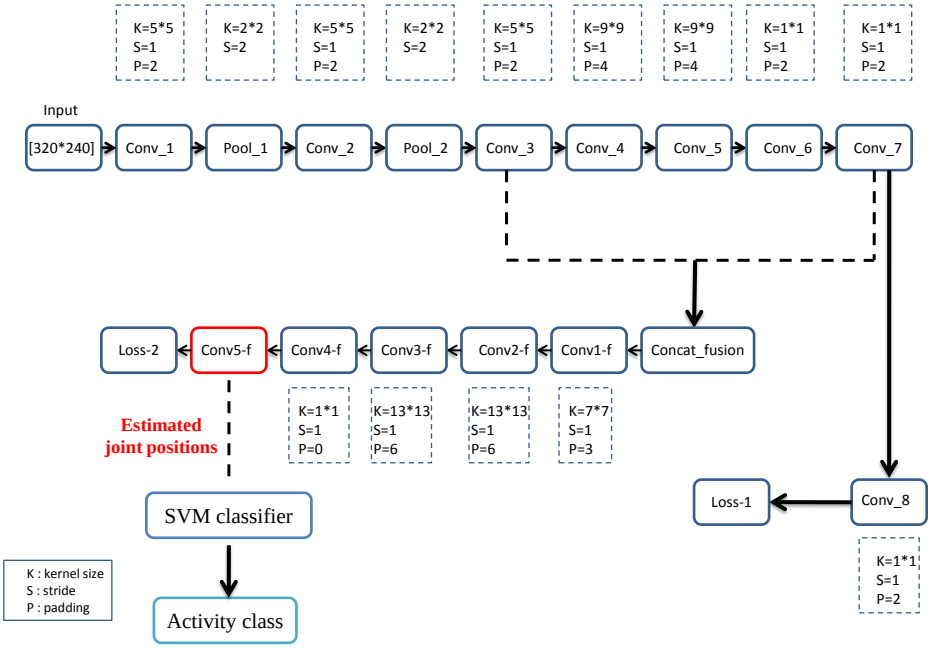

Figure 2. Overview of our Deep Full-Body (DFB)-Human Pose Estimation (HPE) architecture from RGB frames.

As a loss function, the suggested architecture uses the Euclidean loss layer, which computes the sum of squares of differences between its two inputs, as shown in Equation (1). As our network is trained to regress the location of the human full-body joints, the l_2 loss layer penalizes the l_2 distance between the predicted joint positions and the Ground Truth (GT) ones.

$$loss = \frac{1}{2N} \sum_{i=1}^{N} \|(y_i^1 - y_i^2)\|^2 \quad (1)$$

where N is the number of samples, ${y_i}^1$ represents the ith predicted joint location, and ${y_i}^2$ is the ith GT joint location.

For the classification of different human activities, we used the multi-class "one-against-one" SVM classifier. We used the LIBSVM implementation with the polynomial function as a kernel. The SVM's input is the vector of the 2D positions of all fifteen joints

calculated in the previous pose estimation stage. Each frame is associated with its fifteen 2D joint positions and its activity label. In order to have the best SVM configuration, we utilized the 10-fold cross-validation process for the training and testing splits. Then, the predicted SVM model was tested, and the accuracy rate, as well as confusion matrix were calculated.

In order to find a good enough set of weights for the specific mapping function from inputs to outputs, we used the stochastic optimization algorithm of Stochastic Gradient Descent (SGD). It is based on randomness in selecting a starting point for the search where all the weights were initialized to small random values. This process was repeated multiple times in order to have the most effective configuration. For that goal, we chose to train our network not just by fine-tuning with the pre-trained model available, but from scratch, which allowed us to control all parameters' initialization. We began with a number of iterations equal to 150 K and increased it to view its effect on the convergence of the loss to 0. The network weights were learned using a mini-batch stochastic gradient descent with the momentum set to 0.95. In each training iteration, fourteen training frames were taken randomly and used as a mini-batch. To present maximally varying input data to the network and avoid the over-fitting problem, some data augmentation techniques were used. Each frame, with a 320×240 input size, was randomly shuffled prior to training and randomly cropped to a 232×232 sub-image to be then fed forward through the network to compute human joint locations. The validation set was used for hyper-parameter estimation. At training time, the GT labels were heat maps synthesized for each joint separately by placing a Gaussian with a fixed variance at the ground truth joint position. We then utilized an l_2 loss, which penalized the squared pixel-wise differences between the predicted heat map and the synthesized ground truth one. In order to determine the best ConvNet parameter initialization, a 4-fold cross-validation was applied on the used dataset. The ConvNet training was performed on a single NVIDIA GTX Titan GPU using the Caffe framework [44].

4. Results

4.1. Datasets

We utilized two public well-known datasets: J-HMDB [45] and CAD-60 [46].

J-HMDB: Extracted from the HMDB51 dataset, J-HMDB contains 928 clips comprising 21 action categories. It is not only a human action dataset, but also a good benchmark for pose estimation and human detection. Each frame was annotated using a 2D articulated human puppet model [47] providing: a scale, a pose, a segmentation, a coarse viewpoint, and a dense optical flow for humans in action.

CAD-60: concerns 12 classes of daily-life actions (e.g., wearing contact glasses, opening a pill container, brushing teeth) in addition to two non-action classes relative to still and random behaviors. It was performed only by four actors and offers images relative to the RGB and depth frames, besides the skeletal streams relative to 15 body joints. Its main challenge is having one left-handed actor out of those present. The skeleton data are illustrated in Figure 3.

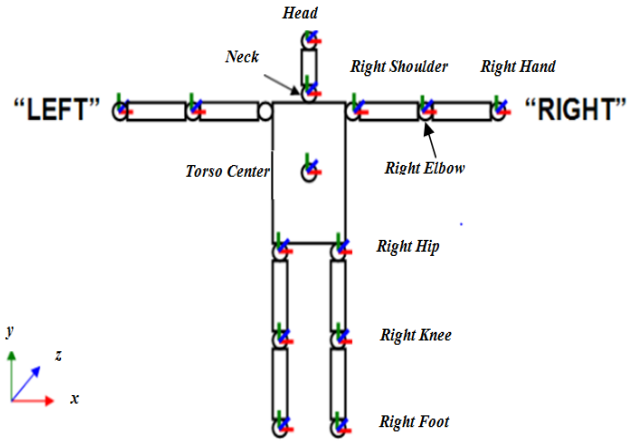

Figure 3. Key joint positions in the CAD-60dataset.

4.2. Evaluation Metrics

In all pose estimation experiments, we compared the estimated joints against the GT ones. The GT joint positions were given in a real-world coordinate system. Thus, they were converted into image-plane coordinates (x, y). For any particular joint localization precision radius r (measured in a Euclidean pixel distance), we report the percentage of correct joints in the test set within this radius. Indeed, for a test set of size N, radius r, and a particular joint i, the accuracy is given by Equation (2):

$$acc_i(r) = \frac{100}{N} \sum_{t=1}^{N} 1\left(\frac{\|(y_i^{t*} - y_i^t)\|}{h_t/100} \leq r\right) \quad (2)$$

where y_i^{t*} is the ith predicted joint location on test sample t and h_t represents the torso height of the tth sample.

In addition to the accuracy evaluation metric, the Percentage of Correct Parts (PCP), the Percentage of Correct Keypoints (PCK), and the Percent of Detected Joints (PDJ) have been commonly used in recent pose estimation work:

- PCP: It describes a broadly-adopted evaluation protocol that measures the percentage of correctly localized body parts. A candidate body part is labeled as correct if its segment endpoints lie within 50% of the length of the ground-truth annotated endpoints [20,48].
- PCK: It defines a candidate keypoint to be correct if it falls within $\alpha \times \max(h, w)$ pixels of the GT keypoint, where h and w are respectively the height and width of the bounding box and α is the relative threshold for correctness [16].
- PDJ: A joint is considered detected if the distance between the predicted joint and the true one is within a certain fraction of the torso diameter. By varying this fraction, detection rates are obtained for varying degrees of localization precision. This metric alleviates the drawback of PCP since the detection criteria for all joints are based on the same distance threshold [20].

4.3. Results of J-HMDB Dataset

Based on the work of Charles et al. [48], a joint is considered to be correctly located if it is within a set distance of d pixels from a marked joint center in the GT. Accordingly, different results are presented as graphs that plot accuracy per joint type vs. distance from the GT in pixels in Figure 4.

Those results are confirmed with those presented in Figure 5, which shows PDJ results per joint type according to the normalized precision threshold. For upper-body joints, the detection rate can achieve approximately 90% even from a 0.5 precision threshold. We note that our pose estimator performs well for almost all action classes, although it is about real-world occluded scenarios. For some actions as "brush hair" or "wave", the accuracy

rate is lower for principally the knee and ankle joints. Indeed, for those action classes, the provided RGB frames are just upper-body, which makes it difficult to estimate lower-body joints such as the ankle or the knee.

We compare the proposed approach with seven state-of-the-art methods tested on the same dataset in Table 1. The first two methods: Dense Trajectories (DTs) [49] and the Spatial Temporal "And/Or" Graph Model (STAOGM) [50] are hand-crafted. However, the remaining approaches are CNN based: Pose-CNN (P-CNN) [51], Action-tubes (A-tubes) [52], Semantic Region-based CNN (SR-CNN) [53], Motion-Salient Region CNN (MSR-CNN) [54] and Human-Related Multi-Stream CNN (HR-MSCNN) [55]. From the comparison with DTs and STAOGM, we find that the deep learned features outperform the hand-crafted ones for action recognition. For P-CNN, the pose-estimator used does not always perform well. Our method achieves close results to those A-tubes. However, the authors used an empirically selected parameter α, which is fixed as constant and might not be optimal for different kinds of videos. The two stream SR-CNN algorithm is similar to our method. It incorporates semantic regions that are detected by Faster R-CNN [56] into the original two stream CNNs. This method uses all detected regions, not only the human body, but also other foreground and background regions. The extracted features in those regions may negatively impact the performance of SR-CNN. In contrast, our method focuses on the human body region where the features are beneficial for the task of action recognition. Compared to MSR-CNN, the authors in [53] used a spatio-temporal 3D convolutional method for fusion. Thus, their network performs a little better. Regarding the HR-MSCNN results, the proposed architecture combines two traditional streams: appearance (R1) and motion (R2), in addition to the captured tubes of the human-related regions (R3), which can make the computation time a bit long. In fact, they achieve a 62.98% accuracy rate when using only one region input (R1) and 71.17% when using all of them (R1 + R2 + R3).

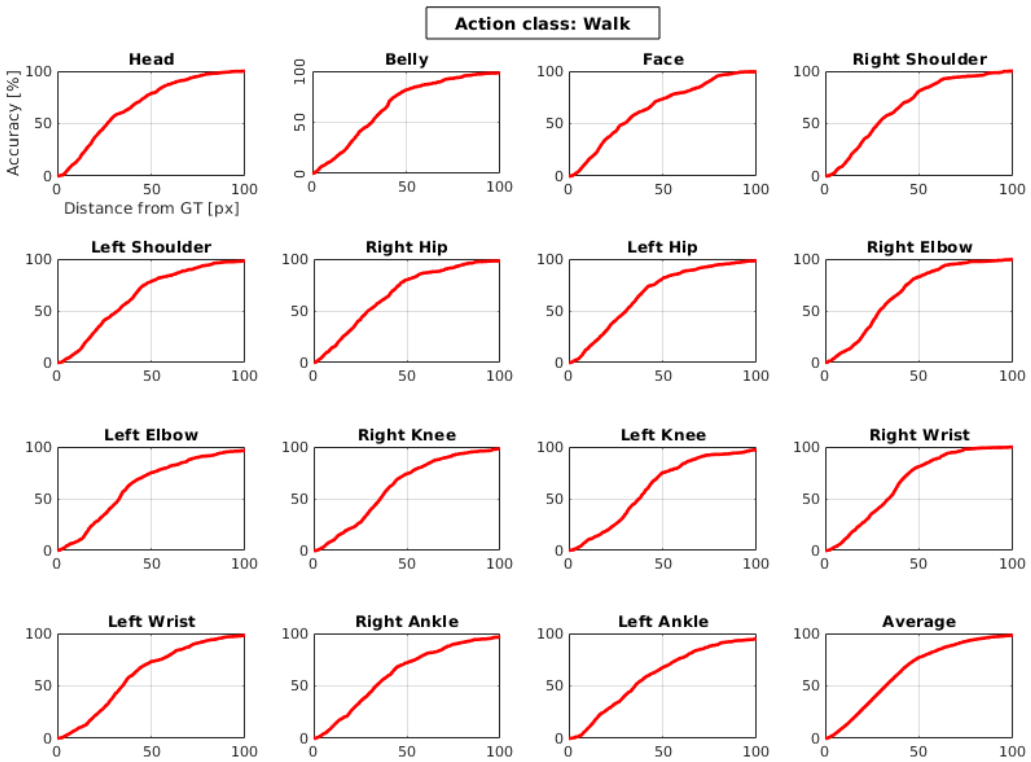

Figure 4. Pose estimation results on J-HMDB: accuracy per joint type according to the allowed distance from the GT.

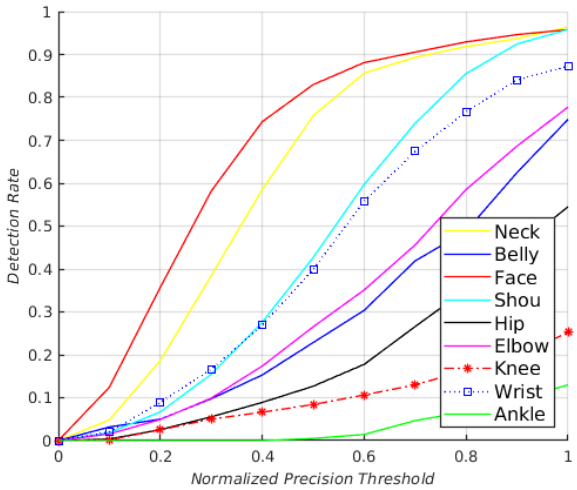

Figure 5. Percent of Detected Joints (PDJ) results on J-HMDB: detection rate per joint type according to the normalized precision threshold.

Table 1. Comparison with state-of-the-art methods on the J-HMDB dataset. DTs, Dense Trajectories; STAOGM, Spatial Temporal "And/Or" Graph Model; P-CNN, Pose-CNN; A-tubes, Action-tubes; SR, Semantic Region; MSR, Motion-Salient Region; HR-MSCNN, Human-Related Multi-Stream CNN.

Reference	Method	Accuracy (%)
Wang 2011 [49]	DTs	56.6
Nie 2015 [50]	STAOGM	55.7
Cheron 2015 [51]	P-CNN	61.1
Gkioxari 2015 [52]	A-tubes	62.5
Wang 2016 [53]	SR-CNN	65.51
Tu 2016 [54]	MSR-CNN	66.02
Tu 2018 [55]	HR-MSCNN	71.17
Ours	DFB-HPE	62.07

4.4. Results of the CAD-60 Dataset

For the CAD-60 dataset, different pose estimation results are presented in Figure 6 as accuracy graphs according to the allowed distance from the GT after applying the four-fold cross-validation process.

In Table 2, we report the different PCK-0.5results on the CAD-60 dataset.

Table 2. PCK-0.5 resultsof CAD-60 dataset.

Iteration	Head	Neck	Torso	Shoulder	Elbow	Hip	Knee	Hand	Foot	Average
k= 1	63.9	69.0	52.9	51.7	59.2	18.3	16.9	33.8	18.4	38.8
k= 2	97.2	98.5	95.5	96.0	74.7	53.1	27.8	53.3	21.4	62.9
k= 3	82.5	67.2	44.7	57.3	44.0	32.2	63.5	35.1	83.2	55.0
k= 4	71.3	19.8	14.1	26.5	29.6	39.7	83.1	47.1	12.0	38.7
Average	78.7	63.6	51.8	57.8	51.8	35.8	47.8	42.3	33.7	51.5

For the upper-body parts of the CAD-60 dataset, the pose estimation results are good enough for different joints. However, for lower-body parts, each iteration seems to be effective for a well-defined part of the human body. For example, in the forth iteration of the cross-validation process, the pose prediction reaches about an 83.1% accuracy rate for "knee". Despite being a left-handed person in the third iteration (k = 3), the estimation seems to be more effective for "foot": nearly 100% accuracy. This contrast is due mainly to the joints provided with the CAD-60 dataset. In fact, coming from the Kinect (i.e., not

manually calculated), the joints are generally sensitive to noise. In addition, their ability to detect lower parts is almost non-existent, since the distance between the camera and the person must not exceed a few meters. Those facts may explain the different failsobserved especially for lower-body parts. Accuracy results are confirmed with the PCK ones in Table 2, where the scores are reported for each key joint separately and for the whole body. HPE algorithms can be useful for various tasks in many areas, such as action recognition, human detection, human attribute recognition, and various gait processing tasks [57]. We chose the HAR task as it represents many challenges due to occlusions and overlapping scenes. For such a purpose, a multi-class one-against-one SVM classifier was used through the LIBSVM (Library for Support Vector Machines) [58] to recognize different activities. To determine the best configuration of such a classifier, a four-fold cross-validation was applied. As a kernel function, we chose the polynomial one. The SVM input is the vector of the 2D positions of all 15 joints calculated in the previous pose estimation stage. In the training stage, we used 14,294 sample frames of 21,442 and left the rest for the testing stage.

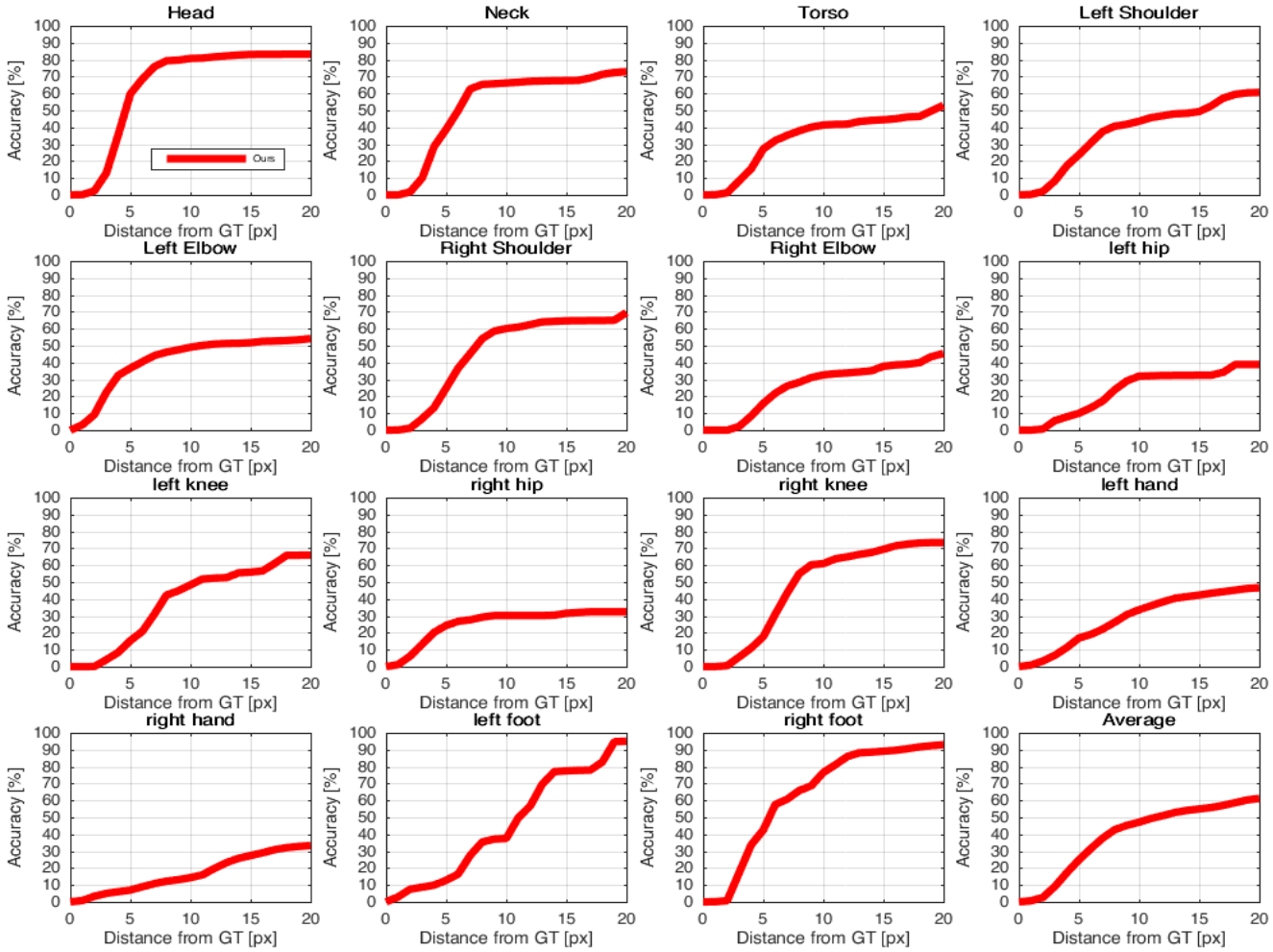

Figure 6. HPE results on CAD-60 with four-fold cross-validation: accuracy per joint type according to the allowed distance from the GT.

As the HAR results, we show the confusion matrix for the CAD-60 dataset in Figure 7. In fact, we have some confusion errors between the "drinking water" and "talking on phone" and between "rinsing mouth with water" and "talking on the phone" activities with 0.02% and 0.03%, respectively. This is due to the great similarity existing between the different activity classes (such as "drinking water" and "talking on phone"). We remember that in our work, we estimate a full-body pose directly from RGB images and then recognize the corresponding activity. Table 3 proves the competitiveness of our approach with the

CAD-60 dataset. Using the accuracy measure, our solution ranks in the first position and demonstrates a robust precision/recall ratio (95.4% and 95.6%, respectively). It reaches a higher value of 95.5% for the accuracy in terms of correctly labeled samples. Note that it admits the highest recall of 95.6%, as shown on the confusion matrix in Figure 7. Our approach achieves promising performance even in challenging cases (left-handed actor in the CAD-60 dataset) and using only RGB frames as the system input.

Table 3. Comparison with state-of-the-art results on CAD-60. DBN, Dynamic Bayesian Network; MRF, Markov Random Field; BOW, Bag Of Words; GMM, Gaussian Mixture Modeling; HMM, Hidden Markov Model; STIP, Spatio-Temporal Interest Point. (* means which input data: Skeleton, RGB or Depth is used)

Algorithm	Precision	Recall	Input Data			Method
			Skeleton	*RGB*	*Depth*	
Sung 2012 [59]	67.9	55.5	*	*	*	DBN
Koppula 2012 [60]	80.8	71.4	*	*	*	MRF
Zhang 2012 [61]	86	84	*	*	*	BOW + SVM
Yang 2013 [62]	71.9	66.6	*			Eigenjoints
Piyathilaka 2013 [63]	70	78	*	*	*	GMM+ HMM
Ni 2013 [64]	75.9	69.5		*	*	Latent SVM
Gupta 2013 [65]	78.1	75.4			*	Codewords + Ensemble
Wang 2014 [66]	74.70	-	*	*	*	Fourier temporal pyramid
Zhu 2014 [67]	93.2	84.6	*	*	*	STIP+ skeleton
Faria 2014. [68]	91.1	91.9	*			Dynamic Bayesian, mixture model
Shan 2014 [69]	93.8	94.5	*			Keypose, random forest, HMM
Gaglio 2015 [70]	77.3	76.7	*			SVM, HMM
Parisi 2015 [71]	91.9	90.2			*	Self-organizing neural
Cippitelli 2016 [72]	93.9	93.5	*			Atomic motion, naive Bayes, nearest neighbor
Seddik 2017 [73]	92.4	93.6	*	*	*	Bags of visual words, Fisher vectors, and SVM
Ours	**95.4**	**95.6**		*		DFB-HPE (ConvNets + SVM)

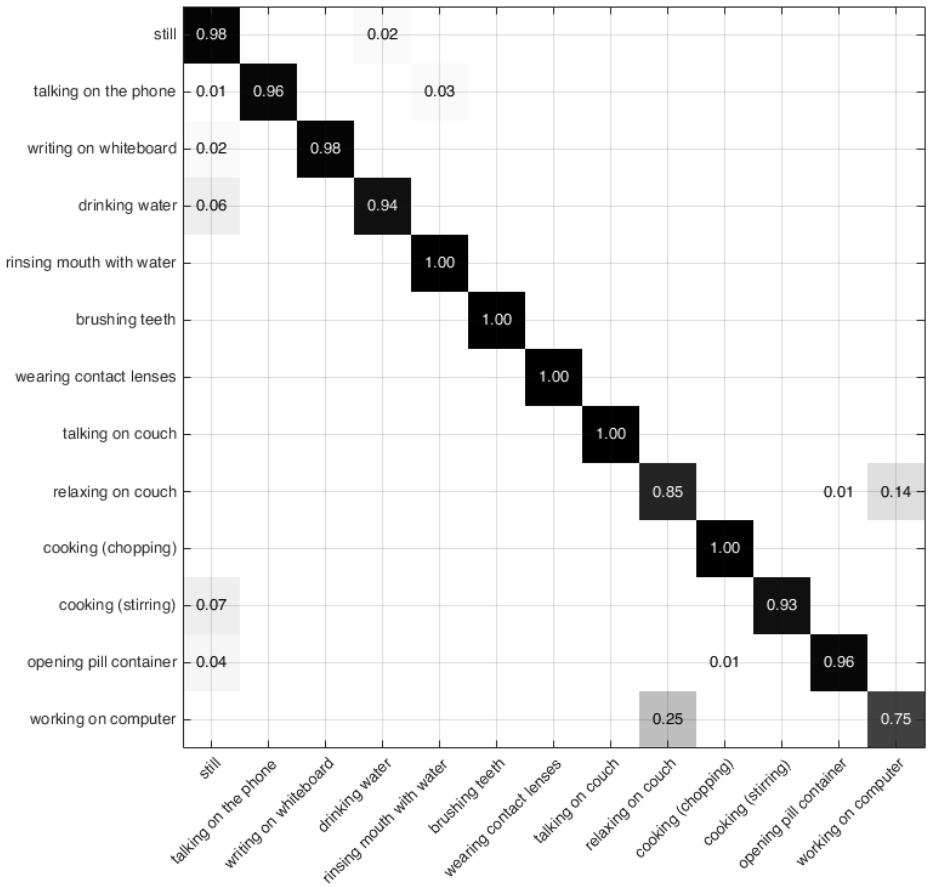

Figure 7. CAD-60 confusion matrix for 12 activities.

5. Conclusions

In this work, we put forward a new approach for 2D full-body HPE. As pose estimation is a key step for a wide range of applications, the more precise it is, the more effective the recognition will be. That is why we took advantage of a deep architecture: ConvNet, given its precision and robustness. The main contribution of our work is to estimate full-body human poses via a ConvNet architecture adapted to a regression problem. From RGB frames only, we extracted deep features represented by 15 key joint positions of the human body. In order to evaluate the proposed HPE model, we applied it to recognize daily activities of a person in an unconstrained environment. Therefore, deep estimated poses were fed to an SVM classifier. The evaluation on challenging datasets (J-HMDB and CAD-60) and the comparison with the state-of-the-art demonstrate that our method achieves competitive ranking for the benchmarks used. The obtained results show the efficiency of using the ConvNet-based pose estimation technique to improve the activity recognition rate.

However, the proposed approach can be further improved. First, an interesting direction is the investigation of more data augmentation techniques such as image translation, color contrasting, and temporal variation [74,75]. Second, a straightforward perspective is to use better performing methods to improve the pose estimation level. Therefore, we can explore the temporal dimension of input videos via 3D CNNs, which show a better adaptability to the data with continuous temporal and spatial domain characteristics of the video [40].

Author Contributions: Conceptualization, S.N.B.; Formal analysis, N.E.B.A.; Investigation, S.N.B.; Methodology, S.N.B.; Project administration, N.E.B.A.; Software, S.N.B.; Supervision, N.E.B.A.; Validation, N.E.B.A.; Writing—original draft, S.N.B.; Writing—review and editing, S.N.B. and N.E.B.A. All authors read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Qiang, L.; Zhang, W.; Hongliang, L.; Ngan, K.N. Hybrid human detection and recognition in surveillance. *Neurocomputing* **2016**, *194*, 10–23.
2. D'Eusanio, A.; Simoni, A.; Pini, S.; Borghi, G.; Vezzani, R.; Cucchiara, R. Multimodal hand gesture classification for the human–car interaction. *Informatics* **2020**, *7*, 31. [CrossRef]
3. Unzueta, L.; Goenetxea, J.; Rodriguez, M.; Linaza, M.T. Dependent 3D human body posing for sports legacy recovery from images and video. In Proceedings of the 2014 22nd European Signal Processing Conference (EUSIPCO), Lisbon, Portugal, 1–5 September 2014; pp. 361–365.
4. Chen, C.; Yang, Y.; Nie, F.; Odobez, J.M. 3D human pose recovery from image by efficient visual feature selection. *Comput. Vis. Image Underst.* **2011**, *115*, 290–299. [CrossRef]
5. Rahimi, M.; Alghassi, A.; Ahsan, M.; Haider, J. Deep Learning Model for Industrial Leakage Detection Using Acoustic Emission Signal. *Informatics* **2020**, *4*, 49. [CrossRef]
6. Konstantaras, A. Deep Learning and Parallel Processing Spatio-Temporal Clustering Unveil New Ionian Distinct Seismic Zone. *Informatics* **2020**, *4*, 39. [CrossRef]
7. Chen, X.; Yuille, A.L. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2014; pp. 1736–1744.
8. Zuffi, S.; Romero, J.; Schmid, C.; Black, M.J. Estimating human pose with flowing puppets. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3312–3319.
9. Seddik, B.; Gazzah, S.; Essoukri Ben Amara, N. Hybrid Multi-modal Fusion for Human Action Recognition. In Proceedings of the International Conference Image Analysis and Recognition, Montreal, QC, Canada, 5–7 July 2017; pp. 201–209.

10. Seddik, B.; Gazzah, S.; Essoukri Ben Amara, N. Hands, face and joints for multi-modal human-action temporal segmentation and recognition. In Proceedings of the 2015 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 31 August–4 September 2015; pp. 1143–1147.
11. Mhalla, A.; Chateau, T.; Maamatou, H.; Gazzah, S.; Essoukri Ben Amara, N. SMC faster R-CNN: Toward a scene-specialized multi-object detector. *Comput. Vis. Image Underst.* **2017**, *164*, 3–15. [CrossRef]
12. Seddik, B.; Gazzah, S.; Essoukri Ben Amara, N. Modalities combination for Italian sign language extraction and recognition. In *International Conference on Image Analysis and Processing*; Springer: Cham, Swizerland, 2015; pp. 710–721.
13. Boualia, S.N.; Essoukri Ben Amara, N. Pose-based Human Activity Recognition: A review. In Proceedings of the 2019 15th International Wireless Communications Mobile Computing Conference (IWCMC), Tangier, Morocco, 24–28 June 2019; pp. 1468–1475. [CrossRef]
14. Daubney, B.; Gibson, D.; Campbell, N. Estimating pose of articulated objects using low-level motion. *Comput. Vis. Image Underst.* **2012**, *116*, 330–346. [CrossRef]
15. Ning, H.; Xu, W.; Gong, Y.; Huang, T. Discriminative learning of visual words for 3D human pose estimation. In Proceedings of the 2008 Computer Vision and Pattern Recognition—CVPR 2008, Anchorage, AK, USA, 24–26 June 2008; pp. 1–8.
16. Ferrari, V.; Marin-Jimenez, M.; Zisserman, A. Progressive search space reduction for human pose estimation. In Proceedings of the Computer Vision and Pattern Recognition—CVPR 2008, Anchorage, AK, USA, 24–26 June 2008; pp. 1–8.
17. Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Real-time human pose recognition in parts from single depth images. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 20–25 June 2011; pp. 1297–1304.
18. Poppe, R. Evaluating example-based pose estimation: Experiments on the humaneva sets. In Proceedings of the CVPR 2nd Workshop on Evaluation of Articulated Human Motion and Pose Estimation, Minneapolis, MN, USA, 22 June 2007; pp. 1–8.
19. Niyogi, S.; Freeman, W.T. Example-based head tracking. In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, Killington, VT, USA, 14–16 October 1996; pp. 374–378. [CrossRef]
20. Toshev, A.; Szegedy, C. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660.
21. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
22. Zhang, N.; Paluri, M.; Ranzato, M.; Darrell, T.; Bourdev, L. Panda: Pose aligned networks for deep attribute modeling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1637–1644.
23. Pishchulin, L.; Andriluka, M.; Gehler, P.; Schiele, B. Poselet conditioned pictorial structures. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 588–595.
24. Carreira, J.; Agrawal, P.; Fragkiadaki, K.; Malik, J. Human pose estimation with iterative error feedback. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4733–4742.
25. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Swizerland, 2016; pp. 483–499.
26. Belagiannis, V.; Zisserman, A. Recurrent human pose estimation. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 468–475.
27. Lifshitz, I.; Fetaya, E.; Ullman, S. Human pose estimation using deep consensus voting. In Proceedings of the European Conference on Computer Vision , Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Swizerland, 2016; pp. 246–260.
28. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 4966–4975.
29. Pfister, T.; Charles, J.; Zisserman, A. Flowing convnets for human pose estimation in videos. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1913–1921.
30. Nibali, A.; He, Z.; Morgan, S.; Prendergast, L. 3d human pose estimation with 2d marginal heat maps. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1477–1485.
31. Toyoda, K.; Kono, M.; Rekimoto, J. Post-Data Augmentation to Improve Deep Pose Estimation of Extreme and Wild Motions. *arXiv* **2019**, arXiv:1902.04250.
32. Kreiss, S.; Bertoni, L.; Alahi, A. PifPaf: Composite Fields for Human Pose Estimation. *arXiv* **2019**, arXiv:1903.06593.
33. Gärtner, E.; Pirinen, A.; Sminchisescu, C. Deep Reinforcement Learning for Active Human Pose Estimation. *arXiv* **2020**, arXiv:2001.02024.
34. Mathis, M.W.; Mathis, A. Deep learning tools for the measurement of animal behavior in neuroscience. *Curr. Opin. Neurobiol.* **2020**, *60*, 1–11. [CrossRef] [PubMed]
35. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 568–576.
36. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y. Towards good practices for very deep two stream convnets. *arXiv* **2015**, arXiv:1507.02159.
37. Ijjina, E.P.; Chalavadi, K.M. Human action recognition using genetic algorithms and convolutional neural networks. *Pattern Recognit.* **2016**, *59*, 199–212. [CrossRef]

38. Wang, K.; Wang, X.; Lin, L.; Wang, M.; Zuo, W. 3D human activity recognition with reconfigurable convolutional neural networks. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 97–106.
39. Shao, J.; Kang, K.; Change Loy, C.; Wang, X. Deeply learned attributes for crowded scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4657–4666.
40. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
41. Varol, G.; Laptev, I.; Schmid, C. Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1510–1517. [CrossRef] [PubMed]
42. Shou, Z.; Chan, J.; Zareian, A.; Miyazawa, K.; Chang, S.F. CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1417–1426.
43. Neili, S.; Gazzah, S.; El Yacoubi, M.A.; Essoukri Ben Amara, N. Human posture recognition approach based on ConvNets and SVM classifier. In Proceedings of the 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Fez, Morocco, 22–24 May 2017; pp. 1–6.
44. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; ACM: New York, NY, USA, 2014; pp. 675–678.
45. Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; Black, M.J. Towards understanding action recognition. In Proceedings of the International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 3192–3199.
46. Sung, J.; Ponce, C.; Selman, B.; Saxena, A. Human Activity Detection from RGBD Images. *Plan Act. Intent Recognit.* **2011**, *64*, 47–55.
47. Zuffi, S.; Freifeld, O.; Black, M.J. From pictorial structures to deformable structures. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 3546–3553.
48. Sapp, B.; Taskar, B. Modec: Multimodal decomposable models for human pose estimation. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 3674–3681.
49. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.L. Action recognition by dense trajectories. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 3169–3176.
50. Xiaohan Nie, B.; Xiong, C.; Zhu, S.C. Joint action recognition and pose estimation from video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1293–1301.
51. 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, Santiago, Chile, 7–13 December 2015; pp. 3218–3226.
52. Gkioxari, G.; Malik, J. Finding action tubes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 759–768.
53. Wang, Y.; Song, J.; Wang, L.; Van Gool, L.; Hilliges, O. Two-Stream SR-CNNs for Action Recognition in Videos. In Proceedings of the BMVC, York, UK, 19–22 September 2016.
54. Tu, Z.; Cao, J.; Li, Y.; Li, B. MSR-CNN: Applying motion salient region based descriptors for action recognition. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 3524–3529.
55. Tu, Z.; Xie, W.; Qin, Q.; Poppe, R.; Veltkamp, R.C.; Li, B.; Yuan, J. Multi-stream CNN: Learning representations based on human-related regions for action recognition. *Pattern Recognit.* **2018**, *79*, 32–43. [CrossRef]
56. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
57. Petrov, I.; Shakhuro, V.; Konushin, A. Deep probabilistic human pose estimation. *IET Comput. Vis.* **2018**, *12*, 578–585. [CrossRef]
58. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27. [CrossRef]
59. Sung, J.; Ponce, C.; Selman, B.; Saxena, A. Unstructured human activity detection from rgbd images. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation (ICRA), Saint Paul, MN, USA, 14–18 May 2012; pp. 842–849.
60. Koppula, H.S.; Gupta, R.; Saxena, A. Learning human activities and object affordances from rgb-d videos. *Int. J. Robot. Res.* **2013**, *32*, 951–970. [CrossRef]
61. Zhang, C.; Tian, Y. RGB-D camera-based daily living activity recognition. *J. Comput. Vis. Image Process.* **2012**, *2*, 12.
62. Yang, X.; Tian, Y. Effective 3d action recognition using eigenjoints. *J. Vis. Commun. Image Represent.* **2014**, *25*, 2–11. [CrossRef]
63. Piyathilaka, L.; Kodagoda, S. Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features. In Proceedings of the 2013 8th IEEE Conference on Industrial Electronics and Applications (ICIEA), Melbourne, Australia, 19–21 June 2013; pp. 567–572.
64. Ni, B.; Pei, Y.; Moulin, P.; Yan, S. Multilevel depth and image fusion for human activity detection. *IEEE Trans. Cybern.* **2013**, *43*, 1383–1394. [PubMed]
65. Gupta, R.; Chia, A.Y.S.; Rajan, D. Human activities recognition using depth images. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21 October 2013; ACM: New York, NY, USA, 2013; pp. 283–292.
66. Wang, J.; Liu, Z.; Wu, Y. Learning actionlet ensemble for 3D human action recognition. In *Human Action Recognition with Depth Cameras*; Springer: Cham, Swizerland, 2014; pp. 11–40.

67. Zhu, Y.; Chen, W.; Guo, G. Evaluating spatiotemporal interest point features for depth-based action recognition. *Image Vis. Comput.* **2014**, *32*, 453–464. [CrossRef]
68. Faria, D.R.; Premebida, C.; Nunes, U. A probabilistic approach for human everyday activities recognition using body motion from RGB-D images. In Proceedings of the 2014 RO-MAN: The 23rd IEEE International Symposium on Robot and Human Interactive Communication, Edinburgh, UK, 25–29 August 2014; pp. 732–737.
69. Shan, J.; Akella, S. 3D human action segmentation and recognition using pose kinetic energy. In Proceedings of the 2014 IEEE Workshop on Advanced Robotics and Its Social Impacts (ARSO), Evanston, IL, USA, 11–13 September 2014; pp. 69–75.
70. Gaglio, S.; Re, G.L.; Morana, M. Human activity recognition process using 3-D posture data. *IEEE Trans. Hum. Mach. Syst.* **2015**, *45*, 586–597. [CrossRef]
71. Parisi, G.I.; Weber, C.; Wermter, S. Self-organizing neural integration of pose-motion features for human action recognition. *Front. Neurorobotics* **2015**, *9*, 3. [CrossRef]
72. Cippitelli, E.; Gasparrini, S.; Gambi, E.; Spinsante, S. A human activity recognition system using skeleton data from RGBD sensors. *Comput. Intell. Neurosci.* **2016**, *2016*, 4351435. [CrossRef]
73. Seddik, B.; Gazzah, S.; Essoukri Ben Amara, N. Human-action recognition using a multi-layered fusion scheme of Kinect modalities. *IET Comput. Vis.* **2017**, *11*, 530–540. [CrossRef]
74. Rogez, G.; Schmid, C. Mocap-guided data augmentation for 3d pose estimation in the wild. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3108–3116.
75. Peng, X.; Tang, Z.; Yang, F.; Feris, R.S.; Metaxas, D. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2226–2234.