Multimodal Deep Neural Networks for Pose Estimation and Action Recognition

Diogo C Luvizon, Hedi Tabia, David Picard

Multimodal Deep Neural Networks for Pose Estimation and Action Recognition

Diogo C. Luvizon[1] Hedi Tabia[1] David Picard[1,2]

[1] ETIS UMR 8051, Paris Seine University, ENSEA, CNRS, F-95000, Cergy, France
[2] Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, F-75005 Paris, France

{diogo.luvizon, hedi.tabia, picard}@ensea.fr

Résumé

Dans cet article, nous présentons un réseaux de neurones multimodal pour l'estimation de pose et la reconnaissance d'action à partir d'images et de vidéos RGB. Notre méthode multimodale favorise l'estimation de pose en combinant des données 3D très précises et des données 2D capturées dans de conditions réelles, ce que favorise aussi l'apprentissage de caractéristiques visuelles génériques. Nous montrons que l'optimisation multi-objectif et multimodale du réseau entraîne de meilleurs résultats que l'optimisation séparée de chaque objectif mono-modal. Ceci nous permet de rapporter des résultats au niveau de l'état de l'art pour l'estimation de pose (Human3.6M) et pour la reconnaissance d'action (NTU RGB+D).

Mots clef

Reconnaissance d'action, estimation de pose, réseaux de neurones à convolution.

Abstract

In this work, we present a unified multimodal neural network for pose estimation from RGB images and action recognition from video sequences. We show that a multimodal approach benefits 3D pose estimation by mixing high precision 3D data and "in the wild" 2D annotated images, while action recognition also benefits from better visual features. Furthermore, we demonstrate by our experiments that end-to-end optimization results in better performance for action recognition than separated learning. We reported state-of-the-art results on 3D pose estimation and action recognition respectively on Human3.6M and NTU RGB+D datasets.

Keywords

Human action recognition, human pose estimation, convolutional neural networks.

1 Introduction

Recognizing human actions is a very challenging task due to the complexity of the human body and to the high similarity between different actions. Additionally, action recognition can benefit from the high level representation of the human body, such as skeletal data. Several approaches have explored the low-cost depth sensors, such as Microsoft's Kinect and Asus' Xtion, to compute the skeletons in real-time [32]. However, such methods suffer from the low quality depth maps, resulting frequently in noisy or erroneous skeletons. On the other hand, with the recent advances in deep learning and on convolutional neural networks, many approaches have reported good results on buth 2D [36, 27] and 3D [22] pose estimation.

Recent approaches for human pose estimation are, in general, using heatmap regression [28], in a way to tackle pose estimation as a detection problem. In such cases, one heatmap is learned for each body joint and the pixel values in such heatmaps correspond to a score whether the joint is present at that position or not. In order to recover the body joint position, the non differentiable argmax function is used. For 3D pose estimation, a similar approach can be used, but instead of using 2D maps, a volumetric heatmap can be learned [25]. While traditional methods for action recognition have been based on pre-computed pose data as their primary information [19], the state of the art has also has also been improved by the advances in deep neural networks [33]. For example, Baradel *et al*. [3] recently show the importance of poses to guide visual features extraction.

Despite the fact that action recognition and human pose estimation are very related tasks, both problems are frequently handled as separated tasks, such as in [9], or action recognition is used to improve pose estimation [14], and, to the best of our knowledge, there is not method providing a single optimized solution for action recognition passing through estimated poses as an intermediate stage. Deep learning approaches have been outperforming many methods in the last yeas mainly because it allows end-to-end optimization. This is even more relevant for multimodal approaches, as appointed by Kokkinos [15], where similar tasks benefit one from another. We believe that both problems have not yet been tackled together because most part of pose estimation approaches are performing heatmap prediction, and such detection based approaches use the non-differentiable argmax function to recover body joint coordinates a posteriori. We think that, with joint optimization, action recognition could benefit from estimated poses in

a more effective way. Taking it into account, we propose a single end-to-end trainable neural network that provides human pose estimation in a first stage and human action recognition as the final result.

For that matter, we propose to extend the differentiable Soft-argmax [41] for both 2D and 3D pose estimation. This allows us to learn pose estimation using mixed 2D and 3D annotated data and to stack action recognition on top of pose estimation, resulting in a multimodal approach trainable from end-to-end. Here, we present our contributions: first, we propose a new multimodal approach for 2D and 3D human pose estimation and for action recognition, that can be trained with indistinguishably with images "in the wild", images with 3D annotated posed, and video sequences. Second, our approach for human pose achieves state-of-the-art results on 3D. Third, our full-model reached state-of-the-art results on action recognition using only still RGB images, while other methods are using images and ground truth poses.

The remaining of the paper is organized as follows. In subsection 2 we present a brief review of the most relevant related work. The proposed approach is presented in section 3. We present the experiments in section 4 and our conclusions and perspectives for future work in section 5.

2 Related work

Since our work has a dual nature, this section is divided into human pose estimation and actions recognition. Due to the limited number of pages in this paper, readers can refer to the surveys in [29, 12] for recent reviews respectively on pose estimation and action recognition.

Human pose estimation. Human pose estimation from still images is an intensively studied problem, with traditional approach from Pictorial Structures [2, 11, 26] to more recent CNN methods [24, 16]. Exploring the concepts of stacked architectures, residual connections, and multiscale processing, Newell *et al*. [23] proposed the Stacked Hourglass Network. Since then, methods are mostly based on sequential refinement of predictions. Chu *et al*. [10] proposed an attention model based on conditional random field (CRF) and Yang *et al*. [40] replaced the residual unit from the stacked hourglass by a Pyramid Residual Module (PRM). Generative Adversarial Networks (GANs) have been used to improve the capacity of learning structural information [8].

Differently from previous detection based approaches, in which the argmax function is required as a post-processing step, regression methods use a nonlinear function that maps the input images directly to poses in (x, y) coordinates. For example, Toshev and Szegedy [37] proposed a holistic solution based on cascade regression and Carreira *et al*. [5] proposed the Iterative Error Feedback. Despite of their advantage of directly predictions pose in a differentiable way, regression methods in the literature give sub-optimal solutions.

On 3D scenarios, pose estimation can be even more challenging. Some approaches first solve the body joints localization problem, then predict the 3D poses from that [7]. Another approach was presented by Sun *et al*. [34], on which the poses are converted to a bone representation, which is less variant and consequently easier to learn. However, such a structural transformation affect negatively the precision on joints in the extremities, since the prediction error is accumulated from one joint to another. Pavlakos *et al*. [25] proposed the volumetric stacked hourglass architecture which the high cost of volumetric computations. In our methods, we also propose an intermediate volumetric representation for 3D poses, but we require much lower resolution than in [25], since our approach has sub-pixel accuracy.

Action recognition. Classical methods for action recognition have explored features extraction guided by body joint locations [39]. To handle with the time dimension, 3D convolutions have been used in recent approaches [4, 6, 38], but they involve high number of parameters, since all the filters have one additional dimension. In our case, we propose a scheme to handle the temporal information with standard 2D convolutions. To cope with noisy skeletons from low-cost depth sensors, Spatio-Temporal LSTM networks have been widely used to implement attention mechanisms [17, 18].

Most previous action recognition methods explore the skeletal information as the only information or to extract local visual features. Since our architecture is able to predict very precise 3D poses, we do not have to cope with the noisy skeletons from Kinect. Additionally, in our method we can use both poses and visual features together, since we perform both pose estimation and action recognition from RGB frames.

3 Proposed method

The proposed multimodal approach takes as input a sequence of RGB images and outputs the predicted action label, as well as per frame intermediate outputs, *i.e.*, visual features, per body joint probability maps, and 2D/3D poses (see Figure 1). As follows, we detailed the multimodal CNN in section 3.1 and the action recognition part in section 3.2.

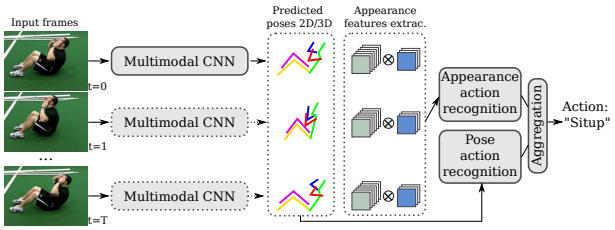

Figure 1: Overview of the proposed method. The multimodal CNN estimates 2D/3D poses for each input frame, as well as visual features and joint probability maps, which are used to extract appearance features.

3.1 Pose estimation and visual features extraction

We propose to handle the problems of human pose estimation and visual features extraction by using a single CNN that address the pose estimation as a regression problem. In such a way, the full network is differentiable, and the pose coordinates can be used directly in the action recognition part.

Multimodal network architecture. The network architecture is based on Inception-V4 [35] and on the Stacked Hourglass [23] for prediction blocks (Figure 2). At the end of each prediction block, volumetric heatmaps are generated, on which we apply a regression method to generate 2D/3D poses, which are then supervised. These heatmaps are reinjected into the network for further refinement.

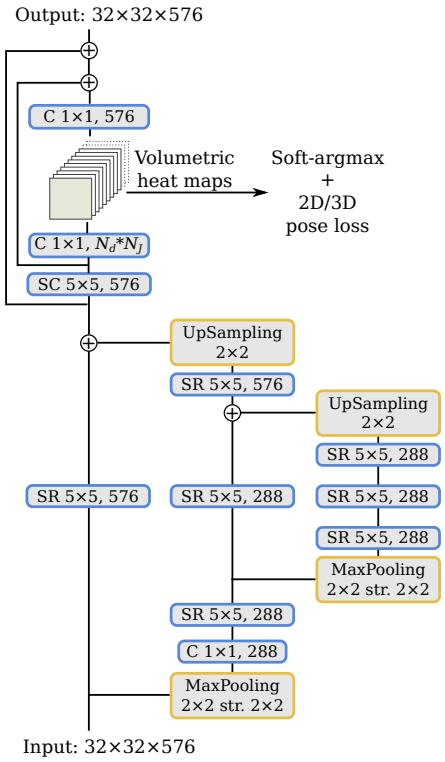

Figure 2: Prediction block architecture. Each block is stacked one after another in order to refine predictions. N_J and N_d are the number of body joints and depth heatmaps.

From the entry-flow network, we define the output as the visual features, which are used as a complementary information for action recognition, as detailed in section 3.2.

Human pose regression. A human pose is defined as a set of N_J points in a 2D or 3D space, that represent the human body joints. For the 2D case, heatmaps can be used to represent the score of a given body joint being present at some position in the image. Differently from classical approaches that use the non differentiable argmax function to extract body joint coordinates from heatmaps, we use the Soft-argmax layer, which is a differentiable way to recover the expectancy of the maximum response, given a normalized heatmap. The Soft-argmax is defined by:

$$\Psi(\mathbf{x}) = \left(\sum_{c=0}^{W_\mathbf{x}} \sum_{l=0}^{H_\mathbf{x}} \frac{c}{W_\mathbf{x}} \Phi(\mathbf{x})_{l,c}, \sum_{c=0}^{W_\mathbf{x}} \sum_{l=0}^{H_\mathbf{x}} \frac{l}{H_\mathbf{x}} \Phi(\mathbf{x})_{l,c} \right)^T, \tag{1}$$

where $\mathbf{x}$ is a heatmap with size $W_\mathbf{x} \times H_\mathbf{x}$ and Φ is the spatial Softmax. The normalized heatmaps are called *probability maps*, which are used to pool visual features localized at the body joint positions, as explained in section 3.2.

In order to extend that approach to 3D scenarios, we define the *volumetric heatmaps* as N_d stacked 2D heatmaps, where N_d is the number of depth heatmaps, since it defines the resolution on depth for regressing the third coordinate. To recover the 3D pose, the first two coordinates, *i.e.*, (x, y), are regressed by applying the Soft-argmax on averaged heatmaps on the Z axis, while the depth component (the (z) coordinate) is regressed by applying a 1D Soft-argmax on averaged heatmaps in the dimensions (x, y).

Note that the (x, y) and (z) components are independent, so we can merge 2D and 3D datasets for multimodal training in a seamlessly way, since in the first case we only backpropagate errors related to (x, y).

3.2 Human action recognition

A very important characteristic of the proposed multimodal approach is its capability to extract both low-level visual features and high-level pose coordinates in a fully differentiable way. Thus, we can mixture both information to predict human actions. Additionally, the shared multimodal network can be trained with pose and action data, in both cases learning from "in the wild" 2D images or from controlled 3D scenarios, which allows the network to learn more robust features. That is possible thanks to the encapsulation of the pose estimation model as a *time distributed* model, which means that the same architecture can be used to handle a sequence of frames, instead of a single image.

The proposed method for action recognition can be seen as composed by two parts: one as pose-based recognition, which uses a sequence of body joints coordinates to predict the action label, and the other as appearance-based recognition, which relies on a sequence of visual features pooled at the joint regions. The predictions from each part are combined by means of one fully-connected layer that gives the final prediction.

The network architecture for action recognition is similar for both pose and appearance data, and is detailed in Figure 3. A first set of convolutions are applied to extract smaller feature maps, which are then fed to a sequence of *action prediction blocks*. On each prediction block, the model outputs one probability map for each action, given a video clip. In order to transform that probability map for actions to action predictions, we perform the *max plus min* pooling operation, followed by a Softmax, which gives a vector of per action probability. This kind of pooling is more sensitive to the strongest responses for each action,

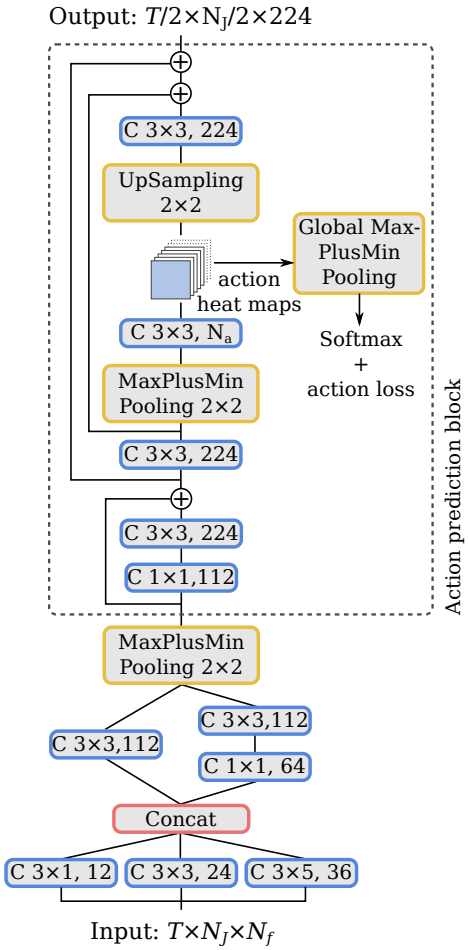

Figure 3: Network architecture for action recognition, where T and N_J are the number of video frames and body joints, respectively. N_f depends on the feature type, which can be 3 for 3D poses or an arbitrary number of visual features, and N_a corresponds to the number of actions.

resulting in more robust predictions. Finally, following the same inspiration as in the human pose estimation part, we refine action predictions by using intermediate supervision on predictions blocks and re-injecting action heatmaps into the network.

As follows, we give some additional information about the pose and appearance branches for action recognition.

Action recognition from human poses. The human body joints encode a high level representation of the human body skeleton, which is essential to recognize some actions. In order to exploit that information, we propose to transform a sequence of T body joints composed of 2D or 3D points into an image-like form, where the vertical axis encodes the time information, the horizontal axis encodes the different body joints (N_J), and the channels encode the different joint coordinates, (x, y) for 2D or (x, y, z) for 3D cases. This transformation is illustrated in Figure 4.

With the proposed approach for pose-based recognition, actions that frequently are dependent only on a few body

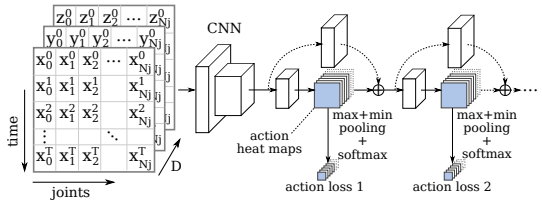

Figure 4: Overview of the action recognition method from human poses.

joints, like trowing an object, can be easily recognized by some patterns detected from the hands, for example.

In that way, an alternative approach with fully-connected layers requires learning zeros on unrelated body joints, which can be very difficult. Conveniently, 2D convolutions with small filters enforce it automatically, resulting in an easier learning problem. Additionally, different body joints have different distributions, which facilitates some filters to get specialized to respond to very specific patterns.

Another advantage of this approach is that 2D convolutions can be used to generate action heatmaps, which are then used in subsequent prediction blocks for refining predictions, as detailed in Figure 4.

Appearance features extraction. As stated before, we profit from our multitask framework to extract low level visual features from the input video frames. Nevertheless, we are interested in very specific features localized at the human body joints, which are much more discriminant than global visual features. These localized visual features are called *appearance features*, which are extracted by multiplying the visual features $F_t \in \mathbb{R}^{W_f \times H_f \times N_f}$ (that are the output from the entry-flow network) by the joint probability maps $M_t \in \mathbb{R}^{W_f \times H_f \times N_J}$ in the last pose prediction block, where $W_f \times H_f$ is the resolution of feature maps, and N_f is the number of visual features.

In order to pool the visual features, we perform a multiplication between each channel from visual features F by each probability maps M, followed by a global sum. That process is repeated for each frame in the video clip, resulting in a new tensor of size $\mathbb{R}^{W_f \times H_f \times N_J \times N_f}$, which holds the appearance features. An illustration of this process is shown in Figure 5.

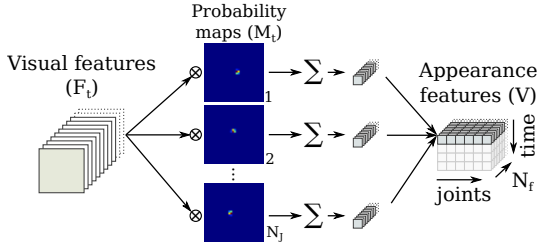

Figure 5: The visual features are pooled by the joint probability maps to produce appearance features, which are used to perform action recognition from a given video clip.

Similarly to the pose-based recognition, appearance features are used as input in an action recognition network, producing estimated action based only on visual appearance information. Finally, the vector of predicted actions from pose and appearance are combined in a fully-connected layer, resulting in the final action prediction.

4 Experiments

In this section we evaluate our method with respect to two different tasks: human pose estimation and action recognition, which shows the effectiveness of our multimodal approach. For that matter, we use three datasets, as detailed below.

4.1 Datasets

Here we provide some basic information about two datasets on which we report results, *i.e.*, Human3.6M [13] and NTU RGB+D [30], and a third dataset, MPII [1], that we use for improving learned visual features.

Human3.6M. The Human3.6M [13] is a 3D pose estimation dataset, composed by videos recorded with 11 subjects (actors) performing 17 different activities and 4 cameras placed at different positions. In total, the dataset contains more than 3 million images, and for every recorded person 32 body joints are available, from which 17 are used to compute estimation errors in millimeters.

NTU RGB+D. The NTU [30] is a action datasets recorded by Microsoft's Kinect v2, usually used for 3D action recognition. In total, 60 actions are performed by 40 actors and recorded by 3 different cameras, with 17 different setups. This results in more than 56K high resolution videos. To the best of our knowledge, this is the most recent and most challenging dataset for 3D action recognition.

MPII Human Pose Dataset. We use the MPII Human Pose [1] dataset as additional data for training, since it is composed of about 25K images collected from YouTube videos in very challenging scenarios, which is usually called "in the wild" data. For each person in the images, 16 body joints were manually annotated in 2D pixel coordinates. Thanks to our multimodal approach, 2D data can be mixed with 3D poses in order to learn better visual representations, which we show that provides a significant improvement in performance.

4.2 Implementation details

We implemented the proposed networks (detailed in Figures 2 and 3) using depth-wise separable convolutions, batch normalization and ReLu activation. We use $N_J = 17$ body joints and $N_d = 16$ heatmaps for depth prediction. In order to merge different datasets, Human3.6M and MPII for example, we convert the poses to a common layout with 17 joints. Since the MPII dataset has only 16 joints, we included one "invalid joint" on this dataset, which is not taken into account when backpropagation the loss.

For human pose estimation, the network was trained using the elastic net loss function [42]:

$$L_{\mathbf{p}} = \frac{1}{N_J} \sum_{n=1}^{N_J} \left(\|\hat{\mathbf{p}}_n - \mathbf{p}_n\|_1 + \|\hat{\mathbf{p}}_n - \mathbf{p}_n\|_2^2 \right), \quad (2)$$

where $\hat{\mathbf{p}}_n$ and $\mathbf{p}_n$ are the estimated and the truth positions of joint n. We optimize the pose human pose regression using the RMSprop optimizer with initial learning rate of 0.001, which is reduced by a factor of 0.2 when scores on validation plateaus, and batches of 24 images.

For the action recognition task, we train the network using the categorical cross entropy loss. We randomly select video clips with size $T = 16$ for training. On test, we report results on *single-clip*, which means that a single clip is cropped from a given video, or on *multi-clip*, where crop multiple clips separated by $T/2$, *i.e.*, 8 frames, one from another. In the last case, final results are computed by the average on all video clips. In that case, we train both pose and appearance models simultaneously using a pre-trained pose estimation model with weights initially frozen. In that case, we use a classical SGD optimizer with Nesterov momentum equal to 0.98 and initial learning rate of 0.0002, reduced by a factor of 0.2 when validation plateaus, and batches of 2 video clips. When validation accuracy stagnates, we divide the final learning rate by 10 and fine tune the full network for more 5 epochs.

To estimate the bounding box for action recognition on test, we do a preliminary pose prediction using the full frame. Then, we crop a bounding box around the estimated person, which is used for the final pose prediction. We use 8 prediction blocks for reporting results on pose estimation and 4 prediction blocks when using the multimodal network for action recognition. For all experiments, we use cropped RGB images of size 256×256, which resulted in feature maps (visual features) of size 32×32. We augment the training data by performing random rotations from -45° to $+45^\circ$, scaling from 0.7 to 1.3, vertical and horizontal translations respectively from -40 to $+40$ pixels, video subsampling by a factor from 1 to 3, and random horizontal flipping.

4.3 Experiments on human pose estimation

Some qualitative results from out method can be seen in Figure 6. In that case, we show the input RGB image, with the 2D predicted pose over the image, and the corresponding 3D estimated pose slightly rotated. One interesting point of our method is that it is able to predict 3D poses from 2D annotated data (see the MPII bottom row), thanks to our multimodal approach.

3D pose estimation. We evaluate the proposed approach for 3D human pose estimation on Human3.6M by measuring the mean per joint position error (MPJPE). We followed the common evaluation protocol [34, 25, 7] by training on five subjects (1, 5, 6, 7, 8) and evaluating on subjects 9 and 11, on one every 64 frames. After prediction, we project

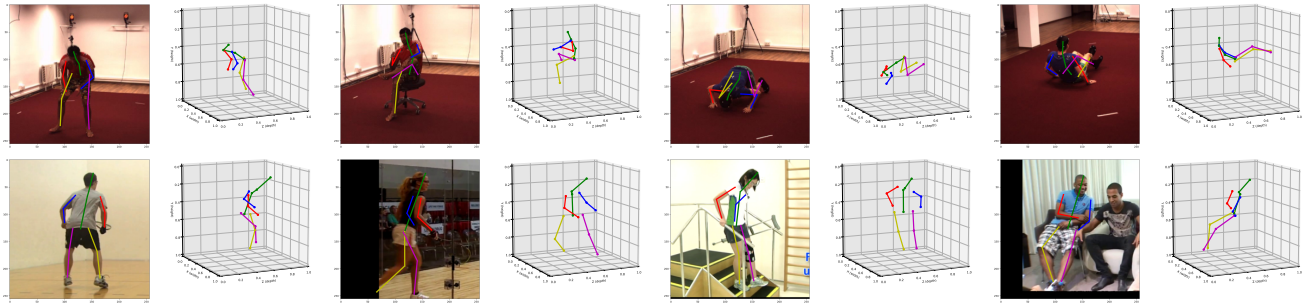

Figure 6: Pose predictions from the datasets Human3.6M (top row) and MPII (bottom row).

our estimated poses into the real world coordinate system by using the available camera calibration. Then, the error is computed between ground truth and estimated poses after alignment on the root joint.

Our results with different experiments and a comparison with the state-of-the-art are shown in Table 1. When specified multi-crop, we use five cropped regions around the subject and the corresponding flipped images, then the final prediction is the averaged pose. When our method is trained with multimodal data, *i.e.*, using 50% from Human3.6M and 50% from MPII, our approach outperform the state of the art by a significant margin.

4.4 Human action recognition

We evaluate our method on action recognition using the NTU dataset, and our results compared to previous approaches are presented in Table 2. We are able to improve the state-of-the-art, despite using estimated poses, while all the other methods rely on ground truth skeletons. Considering only methods that are restricted to RGB images as input, out approach improves the best result by 9.9%.

In order to show the contribution of each part of our method on action recognition, we performed some additional experiments on NTU. If we replace the Kinect skeletons by our estimated poses, we increase the classification accuracy by 2.9%. By fine tuning the full models, from RGB to actions, we gain more 3.3%. In the aggregation stage, when combining pose-based prediction (71.7%) and appearance-based predictions (83.2%), we gain 1.2% more. And by using multiple clips from each video we also gain 1.1%.

Additionally, we show the improvement on accuracy for each successive prediction block in and the contribution of pose and appearance aggregation on Figure 7.

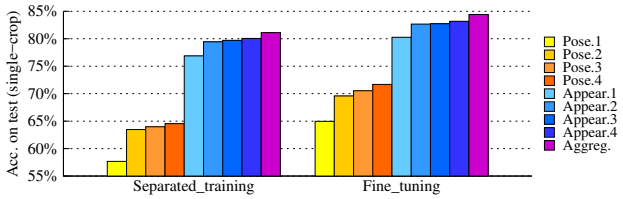

Figure 7: Accuracy on action recognition (NTU) with respect to each prediction block and to aggregated prediction, before and after fine tuning.

5 Conclusions

In this paper, we presented a multimodal neural network for both 3D human pose estimation and action recognition. The proposed approach first predict 3D poses for single images, then combine pose and visual information to predict the action label. Thanks to the sharing of weights for different tasks, our approach benefits from high precision 3D data and from "in the wild" images, resulting in very robust visual features. Additionally, with multiple prediction blocks for both pose and action, our predictions are refined at each stage. And finally, by fine tuning the fully differentiable model and by aggregating pose and appearance information we increased action recognition accuracy significantly.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1014–1021, June 2009.

[3] F. Baradel, C. Wolf, and J. Mille. Pose-conditioned spatio-temporal attention for human action recognition. *arxiv*, 1703.10106, 2017.

[4] C. Cao, Y. Zhang, C. Zhang, and H. Lu. Body joint guided 3d deep convolutional descriptors for action recognition. *CoRR*, abs/1704.07160, 2017.

[5] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback.

Table 1: Comparison with previous work on the Human3.6M dataset using the mean per joint position error (MPJPE) in millimeters on reconstructed poses. SC: single-crop, MC: multi-crop and horizontal flipping. Human3.6M only: using only Human3.6M data for training. Human3.6M + MPII: training with mixed 2D/3D data.

Methods	Direction	Discuss	Eat	Greet	Phone	Posing	Purchase	Sitting
Mehta *et al*. [21]	52.5	63.8	55.4	62.3	71.8	52.6	72.2	86.2
Martinez *et al*. [20]	51.8	56.2	58.1	59.0	69.5	55.2	58.1	74.0
Sun *et al*. [34]	52.8	54.8	54.2	54.3	61.8	53.1	53.6	71.7
Ours Human3.6M only - SC	64.1	66.3	59.4	61.9	64.4	59.6	66.1	78.4
Ours Human3.6M + MPII - SC	**51.5**	**53.4**	**49.0**	**52.5**	**53.9**	**50.3**	54.4	**63.6**
Ours Human3.6M + MPII - MC	**49.2**	**51.6**	**47.6**	**50.5**	**51.8**	**48.5**	**51.7**	**61.5**
Methods	Sit Down	Smoke	Photo	Wait	Walk	Walk Dog	Walk Pair	Average
Mehta *et al*. [21]	120.0	66.0	79.8	63.9	48.9	76.8	53.7	68.6
Martinez *et al*. [20]	94.6	62.3	78.4	59.1	49.5	65.1	52.4	62.9
Sun *et al*. [34]	86.7	61.5	67.2	53.4	47.1	61.6	53.4	59.1
Ours Human3.6M only - SC	102.1	67.4	77.8	59.3	51.5	69.7	60.1	67.3
Ours Human3.6M + MPII - SC	**73.5**	**55.3**	**61.9**	**50.1**	**46.0**	**60.2**	**51.0**	**55.1**
Ours Human3.6M + MPII - MC	**70.9**	**53.7**	**60.3**	**48.9**	**44.4**	**57.9**	**48.9**	**53.2**

Table 2: Comparison results on the NTU for 3D action recognition. Results given as the percentage of correctly classified actions

Methods	Kinect poses	RGB	Estimated poses	Acc. cross subject
Song *et al*. [33]	X	-	-	73.4
Liu *et al*. [18]	X	-	-	74.4
Shahroudy *et al*. [31]	X	X	-	74.9
Baradel *et al*. [3]	X	-	-	77.1
	*	X	-	75.6
	X	X	-	84.8
Ours	-	X	-	**84.6**
	-	X	X	**85.5**

* GT poses were used on test to select visual features.

In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4733–4742, June 2016.

[6] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[7] C.-H. Chen and D. Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[9] G. Ch'eron, I. Laptev, and C. Schmid. P-CNN: Pose-based CNN Features for Action Recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[10] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] M. Dantone, J. Gall, C. Leistner, and L. V. Gool. Human Pose Estimation Using Body Parts Dependent Joint Regressors. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3041–3048, June 2013.

[12] S. Herath, M. Harandi, and F. Porikli. Going deeper into action recognition: A survey. *Image and Vision Computing*, 60(Supplement C):4 – 21, 2017. Regularization Techniques for High-Dimensional Data Analysis.

[13] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, jul 2014.

[14] U. Iqbal, M. Garbade, and J. Gall. Pose for action - action for pose. *FG-2017*, 2017.

[15] I. Kokkinos. Ubernet: Training a 'universal' convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] I. Lifshitz, E. Fetaya, and S. Ullman. *Human Pose Estimation Using Deep Consensus Voting*, pages 246–260. Springer International Publishing, Cham, 2016.

[17] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *European Conference on Computer Vision (ECCV)*, pages 816–833, Cham, 2016.

[18] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot. Global context-aware attention lstm networks for 3d action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[19] D. C. Luvizon, H. Tabia, and D. Picard. Learning features combination for human action recognition from skeleton sequences. *Pattern Recognition Letters*, 2017.

[20] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

[21] D. Mehta, H. Rhodin, D. Casas, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation using transfer learning and improved CNN supervision. *CoRR*, abs/1611.09813, 2016.

[22] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. In *ACM Transactions on Graphics*, volume 36, 2017.

[23] A. Newell, K. Yang, and J. Deng. Stacked Hourglass Networks for Human Pose Estimation. *European Conference on Computer Vision (ECCV)*, pages 483–499, 2016.

[24] G. Ning, Z. Zhang, and Z. He. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Transactions on Multimedia*, PP(99):1–1, 2017.

[25] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[26] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet Conditioned Pictorial Structures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 588–595, June 2013.

[27] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[28] U. Rafi, I. Kostrikov, J. Gall, and B. Leibe. An efficient convolutional network for human pose estimation. In *British Machine Vision Conference (BMVC)*, volume 1, page 2, 2016.

[29] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152(Supplement C):1 – 20, 2016.

[30] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[31] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang. Deep multimodal feature analysis for action recognition in rgb+d videos. *TPAMI*, 2017.

[32] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time Human Pose Recognition in Parts from Single Depth Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '11, pages 1297–1304, 2011.

[33] S. Song, C. Lan, J. Xing, W. Z. (wezeng), and J. Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI Conference on Artificial Intelligence*, February 2017.

[34] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[35] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016.

[36] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 648–656, June 2015.

[37] A. Toshev and C. Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1653–1660, 2014.

[38] G. Varol, I. Laptev, and C. Schmid. Long-term Temporal Convolutions for Action Recognition. *TPAMI*, 2017.

[39] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu. Joint action recognition and pose estimation from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[40] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[41] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. LIFT: Learned Invariant Feature Transform. *European Conference on Computer Vision (ECCV)*, 2016.

[42] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.