Where to Focus on for Human Action Recognition?

Srijan Das, Arpit Chaudhary, Francois Bremond, Monique Thonnat
INRIA, Sophia Antipolis
2004 Rte des Lucioles, 06902, Valbonne, France
name.surname@inria.fr

Abstract

In this paper, we present a new attention model for the recognition of human action from RGB-D videos. We propose an attention mechanism based on 3D articulated pose. The objective is to focus on the most relevant body parts involved in the action. For action classification, we propose a classification network compounded of spatio-temporal subnetworks modeling the appearance of human body parts and RNN attention subnetwork implementing our attention mechanism. Furthermore, we train our proposed network end-to-end using a regularized cross-entropy loss, leading to a joint training of the RNN delivering attention globally to the whole set of spatio-temporal features, extracted from 3D ConvNets. Our method outperforms the State-of-the-art methods on the largest human activity recognition dataset available to-date (NTU RGB+D Dataset) which is also multi-views and on a human action recognition dataset with object interaction (Northwestern-UCLA Multiview Action 3D Dataset).

1. Introduction

Human action recognition from RGB-D videos has been an important task in computer vision. It facilitates many practical applications like smart home, patient monitoring, video surveillance and so on. Challenges in this domain include actions with similar motion and appearance, for *e.g., wearing and taking off a shoe; stacking and unstacking objects*. Existing approaches based on handcrafted features [40, 29] and 2D convNet [6] based feature descriptors lack temporal structure to recognize subtle variations.
Constructing 3D models from videos is a difficult and expensive task. Initially, holistic approaches have been proposed to extract a global representation of human body structure, shape and movements [14, 19] followed by a progression of using local representation by extraction of local features [40]. The availability of multimodal information motivated the authors in [8] to propose multimodal action recognition algorithms using RGB and depth sequences. The emergence of deep learning has encouraged researchers to propose multistream networks using the fusion of appearance and motion [10, 35, 6], recurrent networks to model the evolution of 3D spatial location [32, 13] and recently, spatio-temporal networks like I3D [4] and C3D [38] to extract features from spatial and temporal dimension simultaneously. However, these existing methods are still struggling to recognize similar actions especially in Activities of Daily Living (ADL). We are particularly interested in ADL recognition. Such problem introduces specific challenges: high intra-class variance, high amount of actions which are similar, actions are performed in the same environment i.e. apartment. Some datasets such as UCF-101 [37] which are focused on action recognition from videos uploaded to the internet are different in a way that their inter-class variance is high (i.e. "ride a bike" vs. "sword exercise"). Focus on human body parts to distinguish similar actions in ADL has been shown in [6, 2, 16]. The recent evolution of 3D convNet [38] along with the mechanism of using pre-trained models on ImageNet [18] and Kinectics [4] motivate us to extend the Pose based CNN features for end-to-end action classification. Therefore, we propose a weighted aggregation of human body parts to train an end-to-end 3D model for action classification. To weight the body parts for action classification is a mutually recursive problem. Body part selection for action classification depends on action and vice-versa. So, we propose a pose based spatial attention mechanism to weight the body parts for action classification.

Fig. 1, shows a schema of our proposed network. The action "*donning*" is recognizable by looking at the motion of the object grasped by the hands (which is the *jacket*). Spatio-temporal features extracted from these body parts could be sufficient to model the action. In this work we address action recognition from clipped videos with RGB sequences and their corresponding 3D joints as input. For actions like *"jumping"*, *"running"*, and so on, simple aggregation (summation with equal weightage)

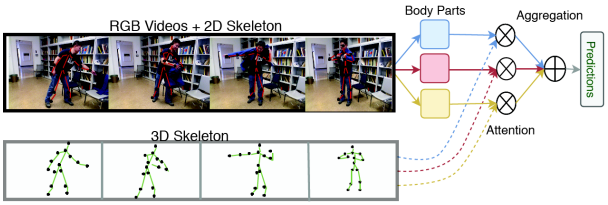


Figure 1: Schema of our proposed network for an action "donning". The 3D pose information determines the attention weights to be given to the spatio-temporal features extracted from the RGB videos corresponding to three relevant body parts of the person performing the action.

of the human body parts models the action better than using the human parts individually. But for actions like *"drinking"* and *"making a phone call"* simple aggregation of the human body parts diminishes the distinctness of the spatio-temporal features for action classification because of providing equal weightage to relevant and irrelevant body parts. So, we propose an RNN attention mechanism to provide appropriate weights to the relevant human body parts involved in the action. Such attention mechanism further improves the action classification.

In summary, we have made the following main contributions in this work.

- A method to classify actions from RGB-D videos focusing on human body parts with 3D ConvNet as the classification network.
- We introduce a novel RNN attention model. The attention model based on the temporal sequences of the articulated poses assigns soft weights to the human body parts.
- We propose a joint strategy to tightly couple the 3D ConvNet based classification network and the RNN attention model using a regularized cross-entropy loss.
- We validate our proposed method on NTU RGB-D dataset, the largest available human activity dataset and an object-interaction human action recognition dataset: the Northwestern-UCLA Multiview Action 3D Dataset outperforming the state-of-the-art results.

2. Related Work

In the past, action recognition has been dominated by local features, say dense trajectories [40] combined with fisher vector encoding [28]. The introduction of kinect sensors along with retrieving 3D poses from the depth map encouraged vision researchers to explore depth based data to recognize actions. Advances in deep learning models for image classification [5] led to the evolution of using these deep networks for spatial feature extraction followed by video aggregation techniques along the temporal domain. Authors in [35, 6, 10] propose late fusion of appearance and motion to recognize actions. These methods fail to compute tight correlation between appearance and motion. Moreover, optical flow takes care of instantaneous motion but fails to model the long-term temporal information. So, sequence models like RNNs are proposed to model long-term spatio-temporal relationships [13].

RNNs for action recognition - Authors in [9] encode temporal information by extracting spatial features from CNN network to feed LSTM. The LSTMs fail to perform efficiently on high dimensional spatial input from CNN networks. This inspires the authors in [32] to model the evolution of 3D spatial coordinates of human body joints for understanding the action dynamics. The availability of 3D data helped to boost performance for action recognition in cross-view setting as in [44]. Authors in [45] evaluated action recognition performance by optimizing the features computed on top of the 3D joints to feed the LSTM. Such a diversity of LSTM networks yielding high performance action recognition accuracy demonstrates its ability to model the body dynamics of the actor performing an action. This motivates us to use pose based RNN to estimate the importance of body parts involved in an action.

3D ConvNets for action recognition - The current studies on 3D ConvNets describe them as a good descriptor being generic, compact, simple and efficient [38]. 3D convolutional deep networks can model appearance and motion simultaneously. In 3D ConvNets, convolution and pooling operations are performed spatio-temporally while in 2D ConvNets they are performed only spatially. This study motivates us to use the recently effective I3D [4] network. Unlike [4], our proposed method computes the action recognition with attention mechanism along the tracks of human body parts.

Attention-based models for action recognition - Human perception focuses on the most relevant parts of the image to acquire information to recognize actions. This phenomenon is known as attention mechanism in artificial intelligence. Recently, two classes of attention have emerged, *hard* and *soft* attention.

Hard attention is the principle of taking hard decisions while choosing parts of the input data. This selection reduces the task (object recognition) complexity as the Region of Interest (RoI) can be placed in the center of the fixation and irrelevant features of the visual environment outside the fixed region are naturally ignored. Authors in [27] have proposed a visual hard-attention for image classification and together with a recurrent network to select the appropriate region location to be focused on. The extraction of information from local region chosen by a glimpse sensor (hard

cropping of RoI) is guided by an agent controller receiving an award for taking a correct decision. The parameters deciding where to look next are learned using Reinforcement Learning. Similar hard-attention mechanism has been used in multiple object recognition, object localization and saliency map generation [3]. [43] uses hard-attention for action detection letting attention to decide which frame to observe next and when to emit an action prediction. All these are stochastic algorithms which cannot be learned easily through gradient descent and backpropagation preventing global optimization of the network. But a major problem with Reinforcement Learning methods is that, they have a high variance (in terms of the gradient of the reward computed) which scales linearly with the number of layers in the RNN network. Thus using attention models in recent deep networks requires a differentiable loss for global optimization of the model. Moreover, the whole network can be trained end-to-end by standard back-propagation.

Soft attention weighs each part of the RoI dynamically, taking the entire image into account. Initially, authors as in [39] trained soft mask branch followed by multiplying the sigmoid normalized mask features with the original convolutional features to generate attention aware features. Then authors dealing with videos portrayed the use of soft attention mechanism in the temporal domain by refining the predictions from past instances using sequential RNN models as in [23, 34, 36]. [34] have proposed a recurrent mechanism for action recognition from RGB data, which assigns weights to different parts of a convolutional features map extracted from CNN network along time. Instead of using RGB images, authors in [23] use 3D joints with spatio-temporal attention mechanism for action recognition. They have proposed an end-to-end network with three RNN network, one for classification, one for selectively focus on discriminative joints of the skeleton (spatial attention), and one for assigning weights to the key sequences (temporal attention). Author in [36] have used similar technique as [23] replacing the input of classification RNN with patches around human hand. Their attention model soft assigns weights to the RGB hand patches taking advantage of articulated pose. Most of these methods provide spatial attention on the input spatial features extracted from 2D ConvNets fed to the RNN and temporal attention on the output latent spatio-temporal features. Recently, [3] have proposed a visual attention module that learns to predict glimpse sequences corresponding to the RoI in the image along with tracking them over time using a set of trackers, which are soft-assigned with external memory. In short, their method includes selecting the glimpses from spatio-temporal features and soft-assigning them to multiple recurrent networks (workers). However, there is no tight coupling between extracting the feature and their attention mechanism failing to globally optimize their proposed network. Thus all the spatio-temporal attention mechanisms for action recognition use recurrent networks for classification.

However, 3D ConvNets outperforms combination of 2D ConvNets + RNN [9] by a large margin. So, we propose a tight coupling of RNN based attention mechanism and 3D ConvNets to focus on the most important body region and for action classification. Temporal attention is computed internally by 3D ConvNets by optimizing the feature maps globally on the whole video. Up to our knowledge, this is the first time the attention mechanism weights the competitive video representation instead of weighting the image representation at each time step to feed an RNN to classify the action. The novelty lies also in joint training the RNN along with I3D subnetworks to extract the discriminative body parts.

3. Proposed Method

We propose and end-to-end 3D Conv network with soft RNN attention for action classification. We exploit the 3D articulated poses of the actor performing action to determine which part of the body can best model an action category. Fig. 2 shows its overall architecture, which consists of three I3D [4] subnetworks for extracting spatio-temporal features from human body parts (left hand, right hand and full body) and RNN attention subnetwork to assign different degrees of importance to the body parts. The input to the network being the RGB video with the sequence of corresponding 3D joints. The challenge in this task includes identifying the appropriate feature space where the spatio-temporal features from the tracks of human body parts are required to be aggregated. Another challenge includes the joint training of the attention block to weight the relevant body parts.

In the following, we discuss the body part representation establishing their importance, RNN attention from the articulated human poses and joint training these subnetworks together to model the actions.

3.1. Body Part representation

Different body parts have different degrees of importance for a particular human action. Fine-grained human action recognition can be performed by extracting cues from RGB streams. We employ a glimpse sensor to crop the tracks of human body parts, for instance - full body, left hand and right hand from the pixel coordinates detected by the middleware. Unlike [6], we restrict the glimpse sensor to crop these three body parts instead of five since the hands and the full body are of higher relevance to the actions performed in general. The cropping operation is fully-differentiable since the exact locations of the human pose are inputs to the model. We aim to aggregate the latent spatio-temporal video representation from the human body parts in order to leverage the relevant body parts for

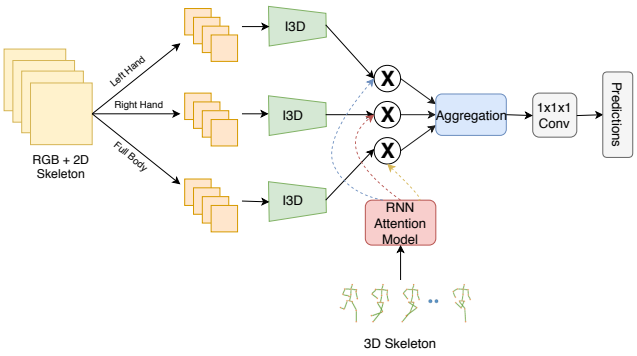


Figure 2: Proposed End to End action classification network. The input to the network is RGB videos with 3D skeleton. Actor body regions like left hand, full body and right hand are extracted from their corresponding 2D pose information. RNN based attention model takes 3D skeleton input (trained on action classification) to attend spatial attention on the spatio-temporal features from I3D (extracted from global average pooling layer after all inception blocks).

action modeling. Before aggregating, the parts based subnetworks, as depicted in fig. 2, are pre-trained on the actor body parts individually leading to generation of high-level spatio-temporal features from each parts.
Taking as input a stack of cropped images from a video V_t, the glimpse representation g_i of the body part i is computed by spatio-temporal convolutional network f_g, with parameters θ_g:

$$g_i = f_g(crop(V_{t,i}); \theta_g) \qquad i = \{1, 2, 3\} \tag{1}$$

3.2. RNN Attention Model

The action of a person can be described by a series of articulated human poses represented by the 3D coordinates of joints. We use the temporal evolution of human skeletons to model the attention to be given to different body parts. The 3D skeleton from depthmap captured by kinect sensor is exploited to pre-train the stacked pose based RNN for action classification to learn the temporal dynamics of skeleton joints for different action classes. This pre-training of the recurrent network is required to extract latent features with spatio-temporal structure for soft weighting the human body parts involved in an action. The stacked pose based RNN consists of three LSTM layers as used in the state-of-the-art with $x_t = (x_{t,1}, ..., x_{t,J})$ for $x_{t,j} \in \Re^3$ and a full set of J joints, as input. Further a dense fully connected layer is added on top of it along with $tanh$ activation to obtain the scores s depicting the importance of different body

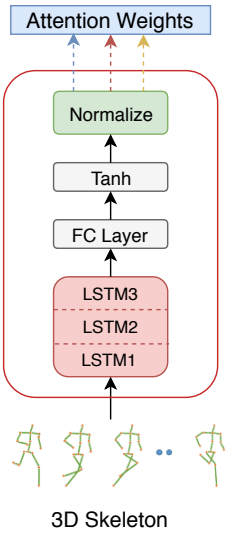


Figure 3: A detailed picture of RNN attention model which takes 3D skeleton poses input and computes weight attention on the spatio-temporal features from different body region of the actor.

parts indicated as

$$s = W_s tanh(W_h h^s + b_s) + b_{us} \tag{2}$$

where W_s, W_h are the learnable parameters, b_s, b_{us} are the bias. h^s is the concatenated hidden state vector of all the timesteps from the last LSTM layer as illustrated in fig. 3. The novelty of our spatial attention block lies in obtaining the attention scores from the latent spatio-temporal information of the whole video instead of obtaining them over time from the output of cell states at each timestep. The objective of such video based attention mechanism is to soft weight the spatio-temporal video representation which is a 4-D hypercube. The obtained scores are normalized using a softmax layer to obtain the attention probabilities. For the k^{th} body part, the activation as the part selection gate is computed as

$$\alpha_k = \frac{exp(s_k)}{\Sigma_{i=1}^{K} exp(s_i)} \tag{3}$$

The RNN attention model provides weights to the different body parts based spatio-temporal representation. The parts based features are integrated with the spatio-temporal features computed by the I3D network. A remaining question is to choose the appropriate spatio-temporal feature space in I3D to aggregate the body part features. The spatio-temporal features from the last layer of I3D are used for aggregating the body parts because these features are spatio-temporally rich and distinct with respect to action categories. The aggregation of these body part features lead

to the formation of distinguishable spatio-temporal features F as

$$F = \Sigma_{k=1}^{K} \alpha_k g_k \quad (4)$$

where g_k is the 4-D spatio-temporal representation from body part k. For aggregation, we also explore assigning attention at different levels of spatio-temporal feature space in I3D with both summation and concatenation operations for aggregation, discussed later in ablation studies. The former tends to squash feature dynamics by pooling strong feature activations in one body part with average or low activations in other body part, the latter leads to formation of highly rich, discriminative features with low generalization.

3.3. Joint training the subnetworks

Joint training the I3D subnetworks consisting of several inception blocks and RNN attention model is a challenge due to the vanishing gradient problem and different backpropagation strategy (BPTT in case of LSTM). Thus we pre-train all the subnetworks separately and joint train them freezing the RNN layers to backpropagate. This strategy along with the formulated cross entropy loss discussed below enables the network to assign weights to the body parts, thus modeling the actions.

Regularized Objective Function - We formulate the objective function of the end-to-end network with a regularized cross-entropy loss and K being the number of body parts as,

$$L = \sum_{i=1}^{C} y_i \log \hat{y_i} + \lambda_1 \sum_{k=1}^{K} (1 - \alpha_k)^2 + \lambda_2 \|W_{uv}\|_2 \quad (5)$$

where $\mathbf{y} = (y_1, ..., y_C)$ represents the groundtruth labels. y_i=1 if it belongs to i^{th} class and y_j=0 for $j \neq i$. $\hat{y_i}$ denotes the probability of the sample belonging to class i, where $\hat{y_i} = p(C_i|X)$. λ_1 and λ_2 are the regularization parameters. The first regularization item forces the model to pay attention at each human parts. This is because the model is prone to ignoring some body parts completely though they have valuable contribution in modeling the actions. So, we impose a penalty as $\alpha_k \approx 1$ encouraging the model to pay equal attentions to different tracks of human parts. The second regularization item is to reduce overfitting of the networks. W_{uv} denotes the weight matrix in connecting the layer u and v.

The optimization is difficult due to the mutual influence of the I3D subnetworks and the pose based RNN attention model. The methodology of separate pre-training of the pose based subnetworks ensures faster convergence of the networks. The training procedure is described in algorithm 1.

Algorithm 1 Joint Training of the RNN attention subnetwork with body part I3D subnetworks

Input: RGB video, 3D joint coordinates, model training parameters $N1$, $N2$ (e.g., $N1$ = 10, $N2$ = 25).

1: Initialize I3D subnetworks with model weights trained on IMAGENET and Kinetics.
//Pre-train I3D subnetworks.
2: Finetune I3D network with RGB data from different body parts individually.
//Pre-train Stacked Pose based RNN.
3: Train the three layered stacked LSTM network for action classification taking as input 3D skeleton of actors in video frames.
//Initialize other attention module parameters.
4: Add a Fully connected layer $tanh$ and a softmax layer on top of stacked LSTM and initialize the attention scores with equal values and the remaining network parameters using Gaussian.
5: Jointly train the Pose based RNN network with part-wise I3D network for $N1$ iterations to obtain the attention scores.
//Jointly Train the Whole Network
6: Fine-tune the whole network by fixing the learned Pose LSTM subnetwork for further $N2$ iterations.

Output: the learned network.

4. Experimental Analysis

Dataset Description - We performed our experiments on the following two human action recognition datasets: NTU RGB+D Dataset [32] and Northwestern-UCLA Multiview Action 3D Dataset [41].

NTU RGB+D Dataset (NTU) - The NTU dataset is currently the largest action recognition dataset containing samples with varied subjects and camera views. It was acquired with a Kinect v2 sensor and contains 56880 video samples with 4 million frames labeled with 60 distinct action classes. The actions were performed by 40 different subjects and recorded from 80 viewpoints. Each person in the frame has 25 skeleton joints which were pre-processed to have position and view invariance [32]. We followed the Cross-Subject (CS) and Cross-View (CV) split protocol from [32].

Northwestern-UCLA Multiview Action 3D Dataset (N-UCLA) - This dataset is captured simultaneously by three Kinect v1 cameras. It contains RGB, depth and human skeleton for each video sample. It contains 1194 video samples with 10 different action categories performed by 10 distinct actors. Most actions in this dataset contains interaction between human and object which is difficult to model making this dataset even more challenging as described in [3]. We performed our experiments by

following Cross-View (CV) protocol from [41], we take samples from two camera views for training our model and test on the samples from the remaining view. $V^3_{1,2}$ means that samples from view 1 and 2 are taken for training, and samples from view 3 are used for testing.

Implementation Details - For all the experiments, we have fixed $K = 3$, with the body parts being left hand, right hand and full body. The I3D network is pre-trained on ImageNet [18] and kinetics [4]. Data augmentation and training procedure for training the I3D networks on individual body parts follow [4]. For training the RNN attention network we use three layer stacked LSTM. Each LSTM layer consists of 512 and 128 LSTM neurons for NTU and N-UCLA respectively. Similar to [32], we cut the videos into sub-sequences of 20 and 5 frames and sample sub-sequences for NTU and N-UCLA respectively. We use 50% dropout to avoid overfitting. We set λ_1 to 0.00001 and 0.0001 for the NTU and N-UCLA datasets respectively, and λ_2 to 0.001 for both the datasets. For training the entire network, we use Adam Optimizer [17] with an initial learning rate set to 0.0005. This learning rate is adjusted automatically during optimization. We used minibatches of size 16 on 4 GPUs. We sample 10% of initial training set as a validation set, for hyper-parameters optimization and for early stopping. For training the model for N-UCLA we used NTU pre-trained I3D subnetworks and fine-tuned on it. During testing, 5 sub-sequences are tested and finally average their logits.

Ablation Study - Table 1 and 2 shows the performance of different image patches based on tracks of human body parts. The statistics show a considerable improvement in the classification accuracy on focusing at the individual body parts rather than using the whole images and thus including unnecessary background information. In table 1, we also quantitatively analyze the best position in the I3D [4] network to aggregate the latent spatio-temporal features from the different human body parts. By sum_r, we mean the aggregation of the spatio-temporal features after $(9-r)$ inception blocks pre-trained on individual body parts in I3D and then using r inception blocks to further extract meaningful information from the aggregated features. Our observation depicts that aggregation at the last inception block without the need of further inception blocks best models the action implying that aggregation of high-level rich features trained on individual body parts does not need further 3D convolutional operations to extract distinguishable spatio-temporal features. For aggregation, we explore the use of summation (sum_r) and concatenation ($concat\_r$) operator at the end of I3D network (since concatenation at earlier layers is not feasible because of curse of dimensionality). Experimental results (in table 1 and 2) shows the effectiveness of summation operation of spatio-temporal features unlike the usual concatenation operation of spatial features as in [2]. In addition to that table 2 also show the effectiveness of using the I3D subnetworks trained on NTU as a pre-trained subnetworks for N-UCLA.

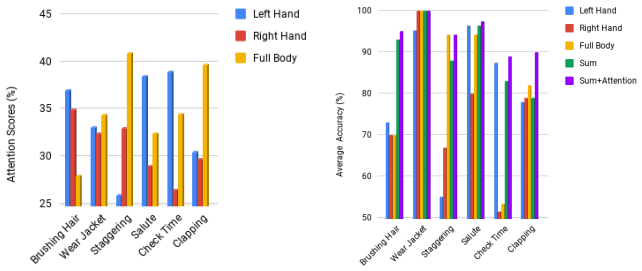
Table 1: Ablation study on NTU RGB+D dataset with Cross-Subject (CS) and Cross-View (CV) protocol.

Methods	CS	CV	Avg
Full image	70.93	80.53	75.73
Left hand	84.31	84.75	84.53
Right hand	82.94	81.83	82.38
Full body	85.47	87.26	86.36
sum_2	89.30	92.02	90.66
sum_1	90.39	92.19	91.29
sum_0	90.8	92.5	91.65
concat_0	89.05	92.07	90.56
sum+attention	93	95.4	94.2

Table 2: Ablation study on Northwestern-UCLA Multiview Action 3D with Cross-View $V^3_{1,2}$ protocol.

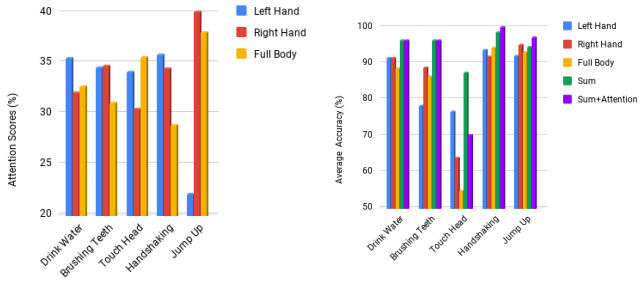
Methods	$V^3_{1,2}$	$V^3_{1,2}$ (NTU pre-trained)
Full image	83.95	87.93
Left hand	77.37	80.60
Right hand	78.50	80.38
Full body	85.99	88.79
sum_0	86.80	91.37
concat_0	86.63	90.30
sum+attention	87.50	93.10

Effectiveness of the Proposed Attention Model - In fig. 4, we illustrate the attention scores and corresponding average classification rate of human body parts, their aggregation and proposed attention model for some action categories. The body part with highest classification rate is correctly delivered with attention weights resulting in improved classification rate of our proposed *sum + attention* model. However, the other body parts may not receive meaningful attention scores because of the activity regularizer we propose to dynamically focus on all the body parts which may overlap with one another.
In fig. 5, we illustrate the statistical results of unsuccessful attention scores attained and their effect on our attention network. Failure in attaining spatial attention on human body parts does not affect the attention model and performs



(a) (b)

Figure 4: Examples of successful attention scores attained in (a) and their corresponding average classification accuracy on individual body parts, their aggregation and our proposed attention model in (b).



(a) (b)

Figure 5: Examples of unsuccessful attention scores attained in (a) and their corresponding average classification accuracy on individual body parts, their aggregation and our proposed attention model in (b).

similar to the aggregation model for actions like *"drinking water"* and *"brushing teeth"*. This is because of the dominance of all the human body parts involved in the action. For action like *"touching head"*, wrong attention delivered to the appearance based spatio-temporal features degrades the performance of the attention model. For actions like *"handshaking"* and *"jumping up"*, failure in attention still improves the performance using our attention model. This is because of uneven weights delivered to different body parts with almost similar relevance to model the action.

Comparison to Other state-of-the-art - We have shown performance comparison of our end-to-end action recognition model with the other state-of-the-art methods which use multiple modalities, including RGB, depth and pose in Table 3 and 4 for the NTU RGB+D and N-UCLA datasets, respectively. Our RNN attention model is able to extract out discriminative spatio-temporal features by efficiently weighing the relevant body parts needed for modeling an action. Its effectiveness is seen by the increase in performance for the two action recognition datasets. Sample visual results displaying the attention scores attained for each body parts can be seen in fig. 6.

Table 3: Results on NTU RGB+D dataset with cross-subject and cross-view settings (accuracies in %).

Methods	CS	CV	Avg
Lie Group [29]	50.1	52.8	51.5
Skeleton Quads [11]	38.6	41.4	40.0
Dynamic Skeletons [15]	60.2	65.2	62.7
HBRNN [16]	59.1	64.0	61.6
Deep LSTM [32]	60.7	67.3	64.0
p-LSTM [32]	62.9	70.3	66.6
ST-LSTM [23]	69.2	77.7	73.5
STA-LSTM [36]	73.2	81.2	77.2
Ensemble TS-LSTM [20]	74.6	81.3	78.0
GCA-LSTM [24]	74.4	82.8	78.6
JTM [42]	76.3	81.1	78.7
MTLN [47]	79.6	84.8	82.2
VA-LSTM [44]	79.4	87.6	83.5
view-invariant [25]	80.0	87.2	83.6
DSSCA-SSLM [33]	74.9	-	-
STA-Hands [2]	82.5	88.6	85.6
Glimpse Cloud [3]	86.6	93.2	89.9
PEM [26]	91.7	95.2	93.4
Proposed Method	**93.0**	**95.4**	**94.2**

Table 4: Results on Northwestern-UCLA Multiview Action 3D dataset with cross-view settings (accuracies in %).

Methods	$V^3_{1,2}$
DVV [22]	58.5
CVP [46]	60.6
AOG [41]	45.2
HPM+TM [31]	91.9
Lie Group [29]	74.2
HBRNN [16]	78.5
view-invariant [25]	86.1
Ensemble TS-LSTM [20]	89.2
Hankelets [21]	45.2
nCTE [12]	75.8
NKTM [30]	85.6
Glimpse Cloud [3]	90.1
Proposed Method	**93.1**

Runtime - The model has been trained on a GPU clus-

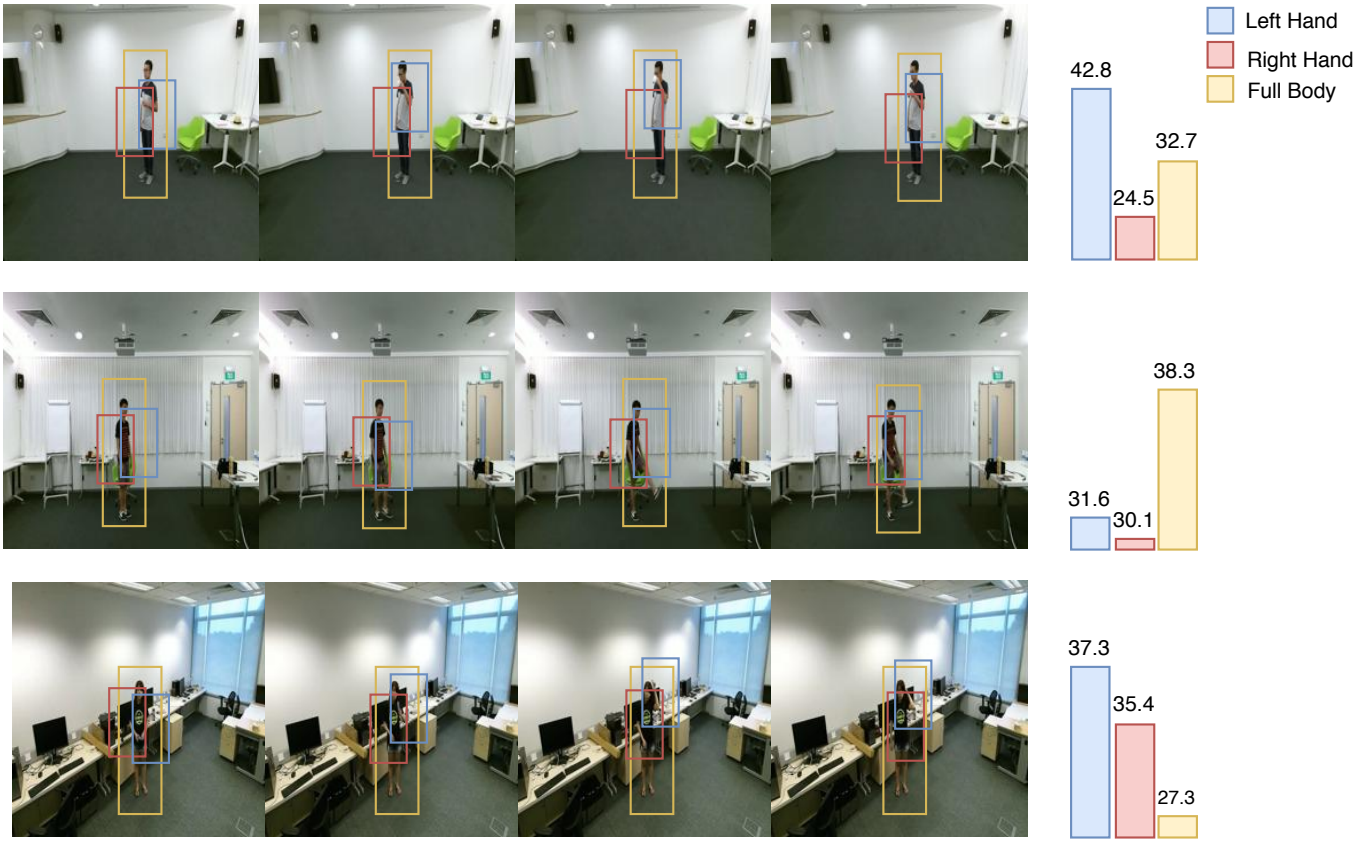


Figure 6: Example of video sequences with their respective attention scores. The action categories presented are drinking water with left hand (1st row), kicking (2nd row) and brushing hair with left hand (last row).

ter having 4 GTX 1080 Ti GPUs. Pre-training the part-wise I3D network on the NTU dataset with CS setting takes 15 hours. Pretraining the Stacked pose based LSTM takes 1 hour. Pre-training the RNN Network for developing attention for the human body parts take 19 hours and further fine-tuning takes 9 hours. At test time, a single forward pass of an image frame over the full model takes 17ms on a single GPU. We use Keras [7] with tensorflow [1] as back-end for the implementation.

5. Conclusion

We present an end-to-end network for human activity recognition leveraging spatial attention on human body parts. We propose an RNN attention mechanism to obtain an attention vector for soft assigning different importance to human body parts using spatio-temporal evolution of the human skeleton joints. We designed the joint training strategy to efficiently combine the spatial attention model with the spatio-temporal video representation by formulating a regularized cross-entropy loss to achieve fast convergence. The proposed method outperforms the state-of-the-art performance on the NTU and N-UCLA datasets. We also demonstrated the attention learned and its effect on the classification performance for different action classes. So in future work, we plan to extend this method which is currently working with a fixed number of body parts (K) by an automated learning of the attention parts.

Acknowledgement

The authors are grateful to Sophia Antipolis - Mediterranean "NEF" computation cluster for providing resources and support.

References

[1] M. Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] F. Baradel, C. Wolf, and J. Mille. Human action recognition: Pose-based attention draws focus to hands. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 604–613, Oct 2017.

[3] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[6] G. Cheron, I. Laptev, and C. Schmid. P-cnn: Pose-based cnn features for action recognition. In *ICCV*, 2015.

[7] F. Chollet et al. Keras, 2015.

[8] S. Das, M. Koperski, F. Bremond, and G. Francesca. A Fusion of Appearance based CNNs and Temporal evolution of Skeleton with LSTM for Daily Living Action Recognition. *ArXiv e-prints*, Feb. 2018.

[9] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[10] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 1933–1941. IEEE, 2016.

[11] E. G., S. G., and H. R. Skeletal quads: Human action recognition using joint quadruples. In *2014 22nd International Conference on Pattern Recognition*, pages 4513–4518, Aug 2014.

[12] A. Gupta, J. Martinez, J. J. Little, and R. J. Woodham. 3d pose from motion for cross-view action recognition via nonlinear circulant temporal encoding. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2601–2608, June 2014.

[13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.

[14] D. Hogg. Model-based vision: a program to see a walking person. *Image and Vision Computing*, 1(1):5 – 20, 1983.

[15] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for RGB-D activity recognition. In *CVPR*, 2015.

[16] J. F. Hu, W. S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2186–2200, Nov 2017.

[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[19] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[20] I. Lee, D. Kim, S. Kang, and S. Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[21] B. Li, O. I. Camps, and M. Sznaier. Cross-view activity recognition using hankelets. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1362–1369, June 2012.

[22] R. Li and T. Zickler. Discriminative virtual views for cross-view action recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2855–2862, June 2012.

[23] 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, *Computer Vision – ECCV 2016*, pages 816–833, Cham, 2016. Springer International Publishing.

[24] J. Liu, G. Wang, P. Hu, L. Duan, and A. C. Kot. Global context-aware attention lstm networks for 3d action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3671–3680, July 2017.

[25] M. Liu, H. Liu, and C. Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.

[26] M. Liu and J. Yuan. Recognizing human actions as the evolution of pose estimation maps. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent models of visual attention. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2204–2212, Cambridge, MA, USA, 2014. MIT Press.

[28] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer, 2010.

[29] V. R., A. F., and C. R. Human action recognition by representing 3d skeletons as points in a lie group. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, June 2014.

[30] H. Rahmani and A. Mian. Learning a non-linear knowledge transfer model for cross-view action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2458–2466, June 2015.

[31] H. Rahmani and A. Mian. 3d action recognition from novel viewpoints. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1506–1515, June 2016.

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

[33] A. Shahroudy, T. T. Ng, Y. Gong, and G. Wang. Deep multimodal feature analysis for action recognition in rgb+d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.

[34] S. Sharma, R. Kiros, and R. Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015.

[35] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[36] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI Conference on Artificial Intelligence*, pages 4263–4270, 2017.

[37] K. Soomro, A. Roshan Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. 12 2012.

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4489–4497, Washington, DC, USA, 2015. IEEE Computer Society.

[39] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. *arXiv preprint arXiv:1704.06904*, 2017.

[40] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011.

[41] J. Wang, X. Nie, Y. Xia, Y. Wu, and S. Zhu. Cross-view action modeling, learning, and recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2649–2656, June 2014.

[42] P. Wang, W. Li, C. Li, and Y. Hou. Action recognition based on joint trajectory maps with convolutional neural networks. *Knowledge-Based Systems*, 158:43 – 53, 2018.

[43] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. *arXiv preprint arXiv:1511.06984*, 2015.

[44] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[45] S. Zhang, X. Liu, and J. Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 148–157, March 2017.

[46] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, and C. Shi. Cross-view action recognition via a continuous virtual path. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2690–2697, June 2013.

[47] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2923–2932. IEEE, 2017.