Human Activity Recognition with Pose-driven Attention to RGB

Fabien Baradel, Christian Wolf, Julien Mille

Human Activity Recognition with Pose-driven Attention to RGB

Fabien Baradel
fabien.baradel@liris.cnrs.fr
Université Lyon, INSA Lyon, CNRS, LIRIS, F-69621, Villeurbanne, France

Christian Wolf
christian.wolf@liris.cnrs.fr
Université Lyon, INSA Lyon, LIRIS CITI Laboratory, INRIA, CNRS, Villeurbanne, France

Julien Mille
julien.mille@insa-cvl.fr
Laboratoire d'Informatique de l'Univ. de Tours, INSA Centre Val de Loire, 41034 Blois, France

Abstract

We address human action recognition from multi-modal video data involving articulated pose and RGB frames and propose a two-stream approach. The pose stream is processed with a convolutional model taking as input a 3D tensor holding data from a sub-sequence. A specific joint ordering, which respects the topology of the human body, ensures that different convolutional layers correspond to meaningful levels of abstraction. The raw RGB stream is handled by a spatio-temporal soft-attention mechanism conditioned on features from the pose network. An LSTM network receives input from a set of image locations at each instant. A trainable glimpse sensor extracts features on a set of predefined locations specified by the pose stream, namely the 4 hands of the two people involved in the activity. Appearance features give important cues on hand motion and on objects held in each hand. We show that it is of high interest to shift the attention to different hands at different time steps depending on the activity itself. Finally a temporal attention mechanism learns how to fuse LSTM features over time. State-of-the-art results are achieved on the largest dataset for human activity recognition, namely NTU-RGB+D.

1 Introduction

We address human activity recognition in settings where articulated pose is available, for instance when input is captured from consumer depth cameras. As complementary information we also use the RGB stream, which provides rich contextual cues on human activities, for instance on the objects held or interacted with. Recognizing human actions accurately remains a challenging task, compared to other problems in computer vision and machine learning. We argue that this is in part due to the lack of large datasets. While large scale datasets have been available for a while for object recognition (ILSVRC [30]) and for general video classification (Sports-1M [15] and lately Youtube8M [1]), the more time-consuming acquisition process for videos showing close range human activities limited datasets of this type to several hundreds or a few thousand videos. As a consequence, the best performing methods on this kind of datasets are either based on handcrafted features or suspected to overfit

on the small datasets after years the community spent on tuning methods. The recent introduction of datasets like NTU-RGB-D [31] ($\sim$ 57 000 videos) will hopefully lead to better automatically learned representations.

One of the challenges is the high amount of information in videos. Downsampling is an obvious choice, but using the full resolution at certain positions may help extracting important cues on small or far away objects (or people). In this regard, models of visual attention [7, 27, 33] (see section 2 for a full discussion) have drawn considerable interest recently. Capable of focusing their attention to specific important points, parameters are not wasted on input which is considered of low relevance to the task at hand.

We propose a method for human activity recognition, which addresses this problem by using articulated pose and raw RGB input in a novel way: our method attends to some parts of the RGB stream given information from the pose stream. In our approach, pose has three complementary roles: i) it is used as an input stream in its own right, providing important cues for the discrimination of activity classes; ii) raw pose (joints) serves as an input for the model handling the RGB stream, selecting positions where glimpses are taken in the image; iii) features learned on pose serve as an input to the soft-attention mechanism, which weights each glimpse output according to an estimated importance w.r.t. the task at hand, in contrast to unconstrained soft-attention on the RGB video [33].

The RGB stream model is recurrent (an LSTM), whereas our pose representation is learned using a convolutional neural network taking as input a sub-sequence of the video. The benefits are twofold: a pose representation over a large temporal range allows the attention model to assign an estimated importance for each glimpse point and each time instant taking into account knowledge of this temporal range. As an example, the pose stream might indicate that the person's hand moves into the direction of a different person, which still leaves several possible choices for the activity class. These choices might require attention to be moved to this hand at a specific instant to verify what kind of object is held, which itself may help to discriminate activities.

The contributions of our work are as follows:

- We propose a spatial attention mechanism on RGB videos which is conditioned on deep pose features from the full sub-sequence.
- We propose a temporal attention mechanism which learns how to pool features output from the recurrent (LSTM) network over time in an adaptive way.
- As an additional contribution, we experimentally show that knowledge transfer from a large activity dataset like NTU (57'000 activities) to smaller datasets like SBU Interaction Dataset 3D (300 videos) or MSR Daily Activity (300 videos) is possible.

The supplementary material contains additional information about our method as well as an explainer video with visualization of our approach at test time.

2 Related Work

Activities, gestures and multimodal data — Recent gesture/action recognition methods dealing with several modalities typically process 2D+T RGB and/or depth data as 3D. Sequences of frames are stacked into volumes and fed into convolutional layers at first stages [9, 14, 28, 29, 41]. When additional pose data is available, the 3D joint positions are typically fed into a separate network. Preprocessing pose is reported to improve performance in some situations, e.g. augmenting coordinates with velocities and accelerations [47]. Pose normalization (bone lengths and view point normalization) has been reported to help in cer-

tain situations [29]. Fusing pose and raw video modalities is traditionally done as late fusion [28], or early through fusion layers [41]. We believe that information extracted from different modalities are complementary but at the same time redundant. Our approach addresses this issue by using learned features from one modality (pose) to attend to some part of another modality (RGB). Hence it can attend to some parts of the RGB stream which are giving discriminative features that can be detected from the pose data.

Architectures for pose data — Recent fined-grained activity recognition methods from pose data are based on recurrent neural networks and/or convolutional neural networks. Part-aware LSTMs [31] separate the memory cells of LSTM networks [11] into part-based sub-cells and let the network learn long-term representations individually for each part, fusing the parts for output. Similarly, Du *et al* [8] use bi-directional LSTM layers which fit anatomical hierarchy. Skeletons are split into anatomically-relevant parts (legs, arms, torso, *etc*), so that each subnetwork in the first layers gets specialized on one part. Features are progressively merged as they pass through layers. Multi-dimensional LSTMs [10] are models with multiple recurrences from different dimensions. Originally introduced for images, they also have been applied to activity recognition from pose sequences [22]. One dimension is time, the second is a topological traversal of the joints in a bidirectional depth-first search, which preserves the neighbourhood relationships in the graph. On the other side convolutional architectures are used from pose data. Convolutional neural networks *et al*. [12, 16, 39] are also used to handle pose sequences. Such approaches require a 3D tensor as input. To satisfy this condition they encode the sequence of pose as a trajectory [16] or into a RGB-like image for benefiting of a Imagenet initialization of the weights [12]. Our solution is close to [39], which stacks the 3D coordinates into a Tensor. However, we follow a topological ordering to extract a better representation of the pose sequence.

Attention mechanisms — Human perception focuses selectively on parts of the scene to acquire information at specific places and times. In machine learning, this kind of processes is referred to as attention mechanism [13], and has drawn increasing interest when dealing with languages [5, 17], images [20] and other data. Integrating attention can potentially lead to improved overall accuracy, as the system can focus on parts of the data, which are most relevant to the task. Attention mechanisms were gradually categorized into two classes. *Hard attention* takes hard decisions when choosing parts of the input data. This leads to stochastic algorithms, which cannot be easily learned through gradient descent and back-propagation. In a seminal paper, Mnih *et al* [27] proposed visual hard-attention for image classification built around a recurrent network, which implements the policy of a virtual agent. A reinforcement learning problem is thus solved during learning [40]. The model selects the next location to focus on, based on past information. Similar approaches have been applied for tackling multiple object recognition [2], generating saliency maps [19] and action detection [43]. On the other hand, *soft attention* takes the entire input into account, weighting each part of the observations dynamically. The objective function is usually differentiable, making gradient-based optimization possible. Soft attention was used for various applications such as neural machine translation [5, 17] or image captioning [42]. Recently, soft attention was proposed for image [7] and video understanding [6, 33, 34, 44], with spatial, temporal and spatio-temporal variants. Sharma *et al* [33] proposed a recurrent mechanism for action recognition from RGB data, which integrates convolutional features from different parts of a space-time volume. Song *et al* [34] propose separate spatial and temporal attention networks for action recognition from pose. At each frame, the spatial attention model gives more importance to the joints most relevant to the current action, whereas the temporal model selects frames. Baradel *et al*. [6] propose to attend to most relevant hands

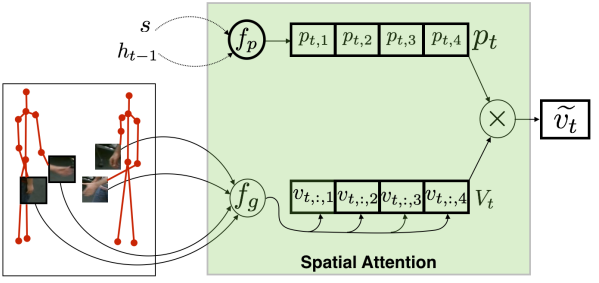

Figure 1: The spatial attention mechanism

in the RGB space given the articulated pose. Our approach is closely related to [6], however our attention mechanism on the RGB space is conditioned on *end-to-end learned deep features* from the pose modality and not only handcrafted pose features. Our pose features need to be discriminative enough for classifying the action and at the same time carry enough information for attending relevant hands.

3 Proposed Model

A single or multi-person activity is described by a sequence of two modalities: the set of RGB input images $I = \{I_t\}$, the set of articulated human poses $x=\{x_t\}$ and we wish to predict the activity class y. We do not use raw depth data in our method, although the extension would be straightforward. Both signals are indexed by time t. Poses x_t are defined by 3D coordinates of K joints per subject, for instance delivered by the middleware of a depth camera. In our case we restrict our application to activities involving one or two people and their interactions. We propose a two-stream model, which classifies activity sequences by extracting features from articulated human poses and RGB frames jointly. Our main contribution comes from the fact that we use features learned from the pose stream to attend to some parts of the RGB stream where all the features are end-to-end learnable.

3.1 Spatial Attention on RGB videos

The sequence of full-HD RGB input images $\{I_t\}$ is arguably not compact enough to easily extract an efficient global representation with a feed-forward neural network. We opt for a recurrent solution, where, at each time instant, glimpses on the seen input is selected using an attention mechanism.

In some aspects similar to [27], we define a trainable bandwidth limited sensor. However, in contrast to [27], and in the lines of [6], our attention process is conditional to the pose input x_t, thus limited to a set of N discrete attention points. In our experiments, we selected N=4 attention points, which are the 4 hand joints of the two people involved in the interaction. We choose hands as attention points because humans use their hands for performing most of their daily actions. Our method can be extended to more attention points. The goal is to extract additional information about hand shape and about manipulated objects. Many activities such as *Reading*, *Writing*, *Eating*, *Drinking* are similar in motion but can be highly correlated to manipulated objects. As the glimpse location is not output by the network, this results in a differentiable soft-attention mechanism, which can be trained by gradient descent.

The glimpse representation for a given attention point i is a convolutional network f_g with parameters θ_g, taking as inputs a crop taken from image I_t at the position of joint i from the set x_t:

$$v_{t,:,i} = f_g(\text{crop}(I_t, x_t, i); \theta_g) \qquad i \in \{1, \ldots, N\} \tag{1}$$

Here and in the rest of the paper, subscripts of mappings f_g and their parameters θ_g choose a specific mapping, they are not indices. Subscripts of variables and tensors are indices. $v_{t,:,i}$ is a (column) feature vector for time t and hand i. For a given time t, we stack the vectors into a matrix $V_t = [v_{t,j,i}]_{i,j}$, where i is the index over hand joints and j is the index over features. V_t is a 2D tensor, since t is fixed for a given instant.

A recurrent model receives inputs from the glimpse sensor sequentially and models the information from the seen sequence with a componential hidden state h_t:

$$h_t = f_h(h_{t-1}, \tilde{v}_t; \theta_h) \tag{2}$$

We chose a fully gated LSTM model including input, forget and output gates and a cell state. To keep the notation simple, we omitted the gates and the cell state from the equations. The input to the LSTM network is the context vector $\tilde{v}_t$, defined further below, which corresponds to an integration of the different attention points (hands) in V_t.

An obvious choice of integration are simple functions like sum and concatenation. While the former tends to squash feature dynamics by pooling strong feature activations in one hand with average or low activations in other hands, the latter leads to high capacity models with low generalization. The soft-attention mechanism dynamically weights the integration process through a distribution p_t, determining how much attention hand i needs with a calculated weight $p_{t,i}$. In contrast to unconstrained soft-attention mechanisms on RGB video [33], our attention distributions not only depend on the LSTM state h, but also on the pose features s (explained in section 3.3) extracted from the sub-sequence, through a learned mapping with parameters θ_p:

$$p_t = f_p(h_{t-1}, s; \theta_p) \tag{3}$$

Attention distribution p_t and features V_t are integrated through a linear combination as

$$\tilde{v}_t = V_t p_t \ , \tag{4}$$

which is input to the LSTM network at time t (see eq. (2)). The conditioning on the pose features in eq. (3) is important, as it provides valuable context derived from motion. Note that the recurrent model itself (eq. (2)) is not conditional [26], this would significantly increase the amount of parameters.

3.2 Temporal Attention

Recurrent models can provide predictions for each time step t. Most current work in sequence classification proceeds by temporal pooling of these predictions, e.g. through a sum or average [33]. We show that it can be important to perform this pooling in an adaptive way. In recent work on dense activity labelling, temporal attention for dynamical pooling of LSTM logits has been proposed [44]. In contrast, we perform temporal pooling directly at feature level. In particular, at each instant t, features are calculated by a learned mapping given the current hidden state:

$$u_{:,t} = f_u(h_t; \theta_u) \tag{5}$$

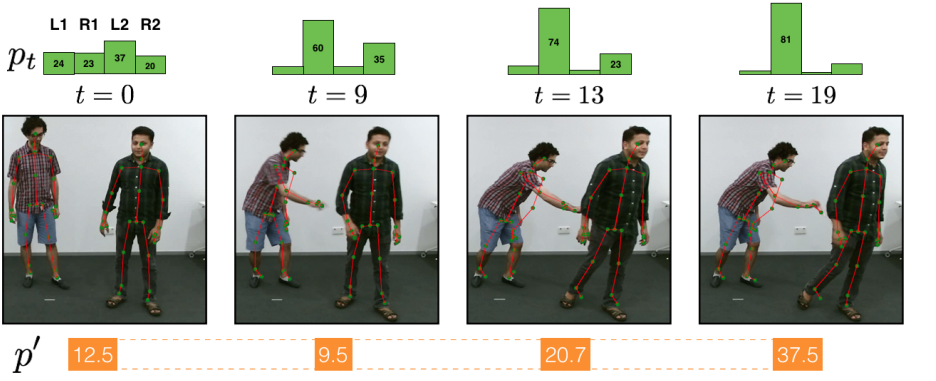

Figure 2: Spatial attention over time: putting an object into the pocket of someone will make the attention shift to this hand.

The features for all instants t of the sub-sequence are stacked into a matrix $U=\{u_{j,t}\}$, where j is the index over the feature dimension. A temporal attention distribution p' is predicted through a learned mapping. To be efficient, this mapping should have seen the full sub-sequence before giving a prediction for an instant t, as giving a low weight to features at the beginning of a sequence might be caused by the need to give higher weights to features at the end. In the context of sequence-to-sequence alignment, this has been addressed with bi-directional recurrent networks [4]. To keep the model simple, we benefit from the fact that (sub) sequences are of fixed length and that spatial attention information is already available. We conjecture that (combined with pose) the spatial attention distributions p_t over time t are a good indicator for temporal attention, and stack them into a single vector P, input into the network predicting temporal attention:

$$p' = f_{p'}(P, s; \theta'_p) \tag{6}$$

This attention is used as weight for adaptive temporal pooling of the features U, i.e. $\tilde{u} = U \times p'$.

3.3 Convolutional pose features

Given the K body joints, we wish to extract features which model i) the temporal behaviour of the pose(s) and ii) correlations between different joints. An attention mechanism on poses could have been an option, similar to [34]. We argue that the available pose information is sufficiently compact to learn a global representation and show that this is efficient. In our case, attention is performed on RGB conditioned on pose instead, as described earlier. We also argue for the need to find a hierarchical representation which respects the spatio-temporal relationships of the data. In the particular case of pose data, joints also have strong neighbourhood relationships in the human body.

In the lines of [22], we define a topological ordering of the joints in a human body as a connected cyclic path over joints. The path itself is not Hamiltonian as each node can be visited multiple times: once during a forward pass over a limb, and once during a backward pass over the limb back to the joint it is attached to. The double entries in the path are important, since they ensure that the path preserves neighbourhood relationships.

In [22], a similar path is used to define an order in a multi-dimensional LSTM network. In contrast, we propose a convolutional model which takes three-dimensional inputs (tensors)

calculated by concatenating pose vectors over time. In particular, input tensors X are defined as $X=\{X_{t,j,k}\}$, where t is the time index, j is the joint & coordinate index, and k is a feature index: each line corresponds to a time instant; the first three columns correspond to the x, y and z coordinates of the first joint, followed by the x, y and z coordinate of the second joint, which is a neighbour of the first etc. The first channel corresponds to raw coordinates, the second channel corresponds to first derivates of coordinates (velocities), the third channel to second derivates (accelerations). Poses of two people are stacked into a single tensor along the second dimension.

We learn a pose network f_{sk} with parameters θ_{sk} on this input, resulting in the pose feature representation s:

$$s = f_{sk}(X;\theta_{sk}) \tag{7}$$

f_{sk} is implemented as a convolutional neural network alternating convolutions and max-pooling.

3.4 Stream fusion

Each stream, pose and RGB, leads to its own set of features, with the particularity that pose features s are input to the attention mechanism for the RGB stream. We first train the pose stream and then the RGB stream. The final model fuse both streams on logit level. More sophisticated techniques, which learn fusion [29], do not seem to be necessary.

Our model is similar to [6] in some respects, and here we would like to highlight the main differences. Baradel *et al*. [6] draw the spatial attention distribution p_t from the *augmented pose* which corresponds to handcrafted features of x_t. In our approach the spatial attention is conditioned on h_{t-1} and s, which can be trained end-to-end, and makes our approach more data driven. The temporal attention distribution differs in a similar way. We calculate temporal attention from learned pose features s whereas Baradel *et al*. [6] use handcrafted features of the pose. In the experiments, we show that these differences are key design choices. Finally, we would like to highlight about the nature of the pose features s. They are extracted and learned by a convolutional architecture while preserving the topological order of the joints. This is a crucial point which makes s being able to keep information about the most discriminative joints and timesteps of the full sequence. Since s is an input of our attention mechanisms, our model can decide to focus on some parts (i.e. hands, hidden states) of the sequence which can lead to improve the understanding of the full-sequence and hence enrich the final representation of the video.

Details about the full architecture structure and the training steps can be found in the supplementary material.

4 Network architectures and Training

Architectures — The pose network f_{sk} consists of 3 convolutional layers of respective sizes 8×3, 8×3, 5×75. Inputs are of size $20\times300\times3$ and feature maps are, respectively, 10×150, 5×75 and $1\times1\times1024$. Max pooling is employed after each convolutional layer, activations are ReLU. The glimpse sensor f_g is implemented as an Inception V3 network [35]. Each vector $v_{t,:,i}$ corresponds to the last layer before output and is of size 2048. The LSTM network f_h has a single recurrent layer with 1024 units. The spatial attention network f_p is an MLP with a single hidden layer of 256 units and sigmoid activation. The temporal attention network f'_p is an MLP with a single hidden layer of 512 units and sigmoid activation.

The feature extractor f_u is a single linear layer with ReLU activation. The output layers of both stream representations are linear layers followed by softmax activation. The full model (without glimpse sensor f_g) has 38 millions trainable parameters.

Training — All classification outputs are softmax activated and trained with cross-entropy loss. The glimpse sensor f_g is trained on the ILSVRC 2012 data [30]. The pose learner is trained discriminatively with an additional linear+softmax layer to predict action classes. The RGB stream model is trained with pose parameters θ_{sk} and glimpse parameters θ_g frozen.

Implementation details — Following [31], we cut videos into sub sequences of 20 frames and sample sub-sequences. During training a single sub-sequence is sampled, during testing we sample 10 sub-sequences and average the logits. We apply a normalization step on the joint coordinates by translating them to a body centered coordinate system with the "middle of the spine" joint as the origin. If only one subject is present in a frame, we set the coordinates of the second subject to zero. We crop sub images of static size on the positions of the hand joints (50×50 for NTU, 100×100 for SBU and MSR). Cropped images are then resized to 299×299 and fed into the Inception model.

Training is done using the Adam Optimizer [18] with an initial learning rate of 0.0001. We use minibatches of size 64 and dropout with a probability of 0.5. Following [31], we sample 5% of the initial training set as a validation set, which is used for hyper-parameter optimization and for early stopping. All hyper-parameters have been optimized on the validation sets of the respective datasets. When transferring knowledge from NTU to SBU, the target networks were initialized with models pre-trained on NTU. Skeleton definitions are different and were adapted. All layers were finetuned on the smaller datasets with an initial learning rate 10 times smaller then the learning rate for pre-training.

Runtime — For a sub-squence of 20 frames, we get the following runtimes for a single Titan-X (Maxwell) GPU and an i7-5930 CPU: A full prediction from features takes 1.4ms including pose feature extraction. This does not include RGB pre-processing, which takes additional 1sec (loading Full-HD video, cropping sub-windows and extracting Inception features). Classification can thus be done close to real-time. Fully training one model (w/o Inception) takes $\sim$4h on a Titan-X GPU. Hyper-parameters have been optimized on a computing cluster with 12 Titan-X GPUs. The proposed model has been implemented in Tensorflow.

5 Experiments

The proposed method has been evaluated on three datasets: NTU RGB+D (NTU), SBU Kinect Interaction (SBU) and MSR Daily Activity (MSR). NTU [31] is the largest dataset for human activity recognition with 56K videos and 60 different activities. We follow the cross-subject and cross-view split protocol from [31]. We extensively tested on NTU and we shows two transfer experiments on smaller datasets SBU and MSR. SBU is an interaction dataset features with two people with a total of 282 sequences and 8 activities while MSR is an daily action dataset features with one people with a total of 320 videos and 16 actions. We follow the standard experimental protocols of [45] and [38] respectively for SBU and MSR.

Details about the implementations can be found in the supplementary material.

Comparisons to the state-of-the-art — We show comparisons of our model against the state-of-the-art methods in table 1, 5 and table 4 respectively. We achieve state of the art performance on the NTU dataset with the full model fusing both streams. We also show a

Methods	Pose	RGB	CS	CV	Avg
Part-aware LSTM [51]	X	-	62.9	70.3	66.6
ST-LSTM + TrustG. [52]	X	-	69.2	77.7	73.5
STA-LSTM [53]	X	-	73.4	81.2	77.2
GCA-LSTM [54]	X	-	74.4	82.8	78.6
JTM [55]	X	-	76.3	81.1	78.7
MTLN [56]	X	-	79.6	84.8	82.2
VA-LSTM [63]	X	-	79.4	87.6	83.5
View-invariant [24]	X	-	80.0	87.2	83.6
DSSCA - SSLM [58]	X	X	74.9	-	-
C3D† [57]	-	X	63.5	70.3	66.9
Resnet50+LSTM†	-	X	71.3	80.2	75.8
STA-Hands [8]	○	X	73.5	80.2	76.9
STA-Hands + DeepGRU [8]	X	X	82.5	88.6	85.6
Ours (pose only)	X	-	**77.1**	**84.5**	**80.8**
Ours (RGB only)	○	X	**75.6**	**80.5**	**78.1**
Ours (pose +RGB)	X	X	**84.8**	**90.6**	**87.7**

Table 1: Results on the NTU RGB+D dataset with Cross-Subject (CS) and Cross-View (CV) settings (accuracies in %); († indicates method has been re-implemented).

Methods	Attention Conditional to pose	CS	CV	Avg
RGB only	-	66.5	72.0	69.3
RGB only	X	75.6	80.5	78.1
Multi-modal	-	83.9	90.0	87.0
Multi-modal	X	84.8	90.6	87.7

Table 2: Results on NTU: conditioning the attention mechanism on pose (RGB only, accuracies in %).

Methods	CS	CV	Avg
Random joint order	75.5	83.2	79.4
Topological order w/o double entries	76.2	83.9	80.0
Topological order	77.1	84.5	80.8

Table 3: Results on NTU: pose only, effect of joint ordering.

Methods	Pose	RGB	Depth	Acc.
Raw skeleton [15]	X	-	-	49.7
Joint feature [15]	X	-	-	80.3
Raw skeleton [16]	X	-	-	79.4
Joint feature [16]	X	-	-	86.9
Co-occurence RNN [50]	X	-	-	90.4
STA-LSTM [53]	X	-	-	91.5
ST-LSTM + Trust Gate [52]	X	-	-	93.3
DSPM [21]	-	X	X	93.4
VA-LSTM [63]	-	X	X	97.5
Ours (Pose only)	X	-	-	90.5
Ours (RGB only)	○	X	-	72.0
Ours (Pose + RGB)	X	X	-	94.1

Table 4: Results on SBU Kinect Interaction dataset (accuracies in %)

Methods	Pose	RGB	Depth	Acc.
Action Ensemble [38]	X	-	-	68.0
Efficient Pose-Based [8]	X	-	-	73.1
Moving Pose [40]	X	-	-	73.8
Moving Poselets [42]	X	-	-	74.5
MP [43]	X	-	-	79.4
Depth Fusion [45]	-	-	X	88.8
MMMP [43]	X	-	X	91.3
DL-GSGC [43]	X	-	X	95.0
DSSCA - SSLM [58]	-	X	X	97.5
Ours (Pose only, no finetuning)	X	-	-	72.2
Ours (Pose only)	X	-	-	74.6
Ours (RGB only)	○	X	-	75.3
Ours (Pose + RGB)	X	X	-	90.0

Table 5: Results on MSR Daily Action dataset (accuracies in %)

good generalization of our model by showing competitive results on SBU and MSR.

We conducted extensive ablation studies to understand the impact of our design choices.

Joint ordering — The joint ordering in the input tensor X has an effect on performance, as shown in table 3. Following the topological order described in section 3.3 gains >1.6 percentage point on the NTU dataset w.r.t. random joint order, which confirms the interest of a meaningful hierarchical representation. As anticipated, keeping the redundant double joint entries in the tensors gives an advantage, although it increases the amount of trainable parameters. More visualizations can be found in the video.

The effect of the attention mechanism — The attention mechanism on RGB data has a significant impact in term of performance as shown in table 6. We compare it to baseline summing (B) or concatenating (C) features. In these cases, hyper-parametres where optimized for these meta-architectures. The performance margin is particularly high in the case of the single stream RGB model (methods E and G). In the case of the multi-modal (two-stream) models, the advantage of attention is still high but not as high as for RGB alone. A part of the gain of the attention process seems to be complementary to the information in the pose stream, and it cannot be excluded that in the one stream setting a (small) part of the pose information is translated into direct cues for discrimination through an innovative (but

	Methods	Pose	RGB	Attention			CS	CV	Avg
				Spatial	Temporal	Pose			
A	Pose only	X	-	-	-	-	77.1	84.5	80.8
B	RGB only, no attention (sum of features)	-	X	-	-	-	61.5	65.9	63.7
C	RGB only, no attention (concat of features)	-	X	-	-	-	63.2	67.2	65,2
E	RGB only + spatial attention	○	X	X	-	X	67.4	71.2	69.3
G	RGB only + spatio-temporal attention	○	X	X	X	X	75.6	80.5	78.1
H	Multi-modal, no attention (A+B)	X	X	-	-	-	83.0	88.5	85.3
I	Multi-modal, spatial attention (A+E)	X	X	X	-	X	84.1	90.0	87.1
K	Multi-modal, spatio-temporal attention (A+G)	X	X	X	X	X	84.8	90.6	87.7

Table 6: Results on NTU: effect of attention. ○ means that pose is only used for the attention mechanism.

admittedly not originally planned) use of the attention mechanism. However, the gain is still significant, with ∼2.5 percentage points compared to the baseline.

Figure 2 shows an example of the effect of the spatial attention process: during the activity of *Putting an object into the pocket of somebody*, the attention shifts to the "putting" hand at the point where the object is actually put.

Pose-conditioned attention mechanism — Making the spatial attention model conditional to the pose features s is confirmed to be a key design choice, as can be seen in table 2. In the multi-modal setting, a full point is gained, >12 points in the RGB only case.

Pose: end-to-end features vs handcrafted features — Conditioning the attention on end-to-end pose features is shown to be an important component of our model. Compared to [6] which use handcrafted pose features for drawing the attention over the RGB stream we show a gain of ∼2 points (78.1 vs 76.9 for the RGB stream and 87.7 vs 85.6 for the two-stream model) as shown in table 1. We argue that the nature of f_{sk} which output the pose features s is a key design choice. The convolutional architecture of f_{sk} and the topological joint ordering of its input are important for making sure that s keep enough information about the most important joints and time instants of the full-sequence. Hence our approach can focus on the most discriminative attention points on the RGB space and the most important hidden states of f_h to extract a stronger final representation of the video.

Comparison with RGB only methods — There is a clear gap between our approach and standard methods for action recognition on RGB data such C3D and CNN+LSTM (+21.8 for C3D and +12.1 for CNN+LSTM) as shown in table 1. These methods need to downsample the RGB stream to a lower resolution which leads to poor performances for fined-grained action recognition. Using some parts of the high resolution RGB stream such as done by our method is important for extracting discriminative features.

6 Conclusion

We propose a general method for dealing with pose and RGB video data for human action recognition. A convolutional network on pose data processes specifically organized input tensors. A soft-attention mechanism crops on hand joints and allows the model to collect relevant features on hand shapes and on manipulated objects. Adaptive temporal pooling further increases performance. Our method shows state-of-the-art results on the NTU RGB+D dataset and competitive performance by performance transfer learning on SBU Interaction dataset and MSR Daily Activity.

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arxiv*, 1609.08675, 2016.

[2] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. In *International Conference on Learning Representations (ICLR)*, 2015.

[3] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *HBU*, 2011.

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[5] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015.

[6] Fabien Baradel, Christian Wolf, and Julien Mille. Human action recognition: Pose-based attention draws focus to hands. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) Workshop*, 2017.

[7] K. Cho, A. Courville, and Y. Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE-T-Multimedia*, 17:1875 – 1886, 2015.

[8] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[9] Abdalrahman Eweiwi, Muhammed S. Cheema, Christian Bauckhage, and Juergen Gall. Efficient pose-based action recognition. In *Asian Conference in Computer Vision (ACCV)*, pages 428–443, 2014.

[10] A. Graves and J. Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In *Neural Information Processing Systems (NIPS)*, 2009.

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

[12] Y. Hou, Z. Li, P. Wang, and W. Li. Skeleton optical spectra based action recognition using convolutional neural networks. In *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[13] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 20(11):1254–1259, 1998.

[14] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(1):221–231, 2013.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[16] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[17] Y. Kim, C. Denton, L. Hoang, and A.M. Rush. Structured attention networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[18] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Machine Learning (ICML)*, 2015.

[19] J. Kuen, Z. Wang, and G. Wang. Recurrent attentional networks for saliency detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3677, 2015.

[20] H. Larochelle and G. Hinton. Learning to combine foveal glimpses with a third-order Boltzmann machine. In *Neural Information Processing Systems (NIPS)*, pages 1243–1251, 2010.

[21] Liang Lin, Keze Wang, Wangmeng Zuo, Meng Wang, Jiebo Luo, and Lei Zhang. A deep structured model with radius-margin bound for 3d human activity recognition. *International Journal of Computer Vision (IJCV)*, 2015.

[22] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition. In *European Conference in Computer Vision (ECCV)*, pages 816–833, 2016.

[23] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C. Kot. Global context-aware attention lstm networks for 3d action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

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

[25] J. Luo, W. Wang, and H. Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, 2013.

[26] T. Mikolov and G. Zweig. Context dependent recurrent neural network language model. In *Spoken Language Technology Workshop*, 2016.

[27] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent models of visual attention. In *Neural Information Processing Systems (NIPS)*, pages 2204–2212, 2014.

[28] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4207–4215, 2016.

[29] N. Neverova, C. Wolf, G.W. Taylor, and F. Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(8):1692–1706, 2016.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[31] 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)*, pages 1010–1019, 2016.

[32] A. Shahroudy, T.-T. Ng, Q. Yang, and G. Wang. Deep multimodal feature analysis for action recognition in rgb+d videos. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016.

[33] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *International Conference on Learning Representations (ICLR) Workshop*, 2016.

[34] 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 *Association for the Advancement of Artificial Intelligence (AAAI)*, 2016.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[36] Lingling Tao and Rene Vidal. Moving poselets: A discriminative and interpretable skeletal motion representation for action recognition. In *ICCV Workshops*, pages 303–311, 2015.

[37] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[38] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1297, 2012.

[39] Pichao Wang, Wanqing Li, Chuankun Li, and Yonghong Hou. Action Recognition Based on Joint Trajectory Maps with Convolutional Neural Networks. In *ACM Conference on Multimedia*, 2016.

[40] R.J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 2012.

[41] D. Wu, L. Pigou, P.-J. Kindermans, N. Do-Hoang Le, L. Shao, J. Dambre, and J.M. Odobez. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(8):1583–1597, 2016.

[42] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference in Machine Learning (ICML)*, pages 2048–2057, 2015.

[43] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[44] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision (IJCV)*, 2015.

[45] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, pages 28–35, 2012.

[46] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. Interactive body part contrast mining for human inter- action recognition. In *ICMEW*, 2014.

[47] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2752–2759, 2013.

[48] P. Zhang, C. Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *International Conference on Computer Vision (ICCV)*, 2017.

[49] W. Zhu, W. Chen, and G. Guo. Fusing multiple features fordepth-based action recognition. In *ACM TIST*, 2015.

[50] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2016.