Human Action Recognition:
Pose-based Attention draws focus to Hands

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

Christian Wolf
Univ Lyon, INSA-Lyon, CNRS, LIRIS
F-69621, Villeurbanne, France
christian.wolf@liris.cnrs.fr

Julien Mille
Laboratoire d'Informatique de l'Université de Tours (EA 6300), INSA Centre Val de Loire
41034 Blois, France
julien.mille@insa-cvl.fr

Abstract

We propose a new spatio-temporal attention based mechanism for human action recognition able to automatically attend to most important human hands and detect the most discriminative moments in an action. Attention is handled in a recurrent manner employing Recurrent Neural Network (RNN) and is fully-differentiable. In contrast to standard soft-attention based mechanisms, our approach does not use the hidden RNN state as input to the attention model. Instead, attention distributions are drawn using external information: human articulated pose. We performed an extensive ablation study to show the strengths of this approach and we particularly studied the conditioning aspect of the attention mechanism. We evaluate the method on the largest currently available human action recognition dataset, NTU-RGB+D, and report state-of-the-art results. Another advantage of our model are certains aspects of explanability, as the spatial and temporal attention distributions at test time allow to study and verify on which parts of the input data the method focuses.

1. Introduction

Human action recognition is an active field in computer vision with a range of industrial applications, for instance video surveillance, robotics, automated driving and others. Consumer depth cameras made a huge impact in research and applications since they allow to estimate human articulated poses easily. Depth input is helpful for solving computer vision problems considered as hard when dealing with RGB inputs only [11]. In this work we address human actiion recognition in settings where human pose is available
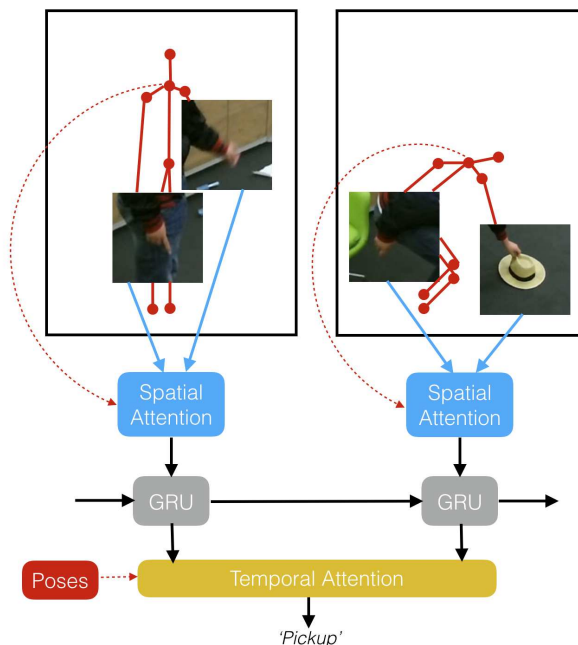

Figure 1: We design a new spatio-temporal mechanism conditioned on pose only able to attend to the most important hands and hidden states.

in addition to RGB inputs. The RGB stream provides additional rich contextual cues on human activities, for instance on the objects held or interacted with.

Understanding human behavior remains a unsolved problem compared to other tasks in computer vision and machine learning in general. This is mainly due to the lack of large datasets. Large datasets, such as Imagenet [29] for

object detection has allowed powerful deep learning methods to reach super-human performances. In the field of human action recognition most of the datasets have several hundreds or few thousand videos. As a consequence, state-of-the-art approaches on this datasets either use handcrafted features or are suspected to overfit on the small datasets after years the community spent on tuning methods. The recent release of large scale datasets like NTU-RGB-D [30] ($\sim$ 57'000 videos) will hopefully lead to better automatically learned representations.

Video understanding is by definition challenging due to its high dimensional, rich and complex input space. Most of the time only a limited area of a video is necessary for getting a fined-grained understanding of the action which occurs. Inspired by neuroscience perspectives, models of visual attention [26, 7, 32] (see section 2 for a full discussion) have drawn considerable interest recently. By attending only to specific areas, parameters are not wasted on input considered as noise for the final task.

We propose a method for human action recognition, which addresses this problem by handling raw RGB input in a novel way. Instead of taking as input the full RGB frame, we take into account image areas cropped around hands only, whose positions are extracted from full body pose estimated by a middleware.

Our model uses two input streams: (i) an RGB stream called *Spatio-Temporal Attention over Hands (STA-Hands)*, and (ii) a pose stream. Both are recurrent over time. A key feature of our method is its ability to automatically draw attention to the most important hands at each time step. Additionally, our approach can also automatically detect the most discriminative hidden RNN states, i.e. most discriminative time instants.

Beyond of giving state-of-the-art results on the NTU dataset, our spatio-temporal mechanism also features certain aspects of explainablity. In particular, it gives insights into key choices made by the model at test time in the form of two different attention distributions: a spatial one (which hands are most important at which time instant?) and a temporal one (which time instants are most important?)

The contributions of our work are as follows:

- We propose a spatial attention mechanism on human hands on RGB videos which is conditioned on the estimated pose at each time step.

- We propose a temporal attention mechanism which learns how to pool features output from the RNN over time in an adaptive way conditioned on the poses over the full sequence.

- We show by an extensive ablation study that soft-attention mechanisms (both spatial and temporal) can be done using external variables in contrast to usual approaches which condition the attention mechanism on the hidden RNN state.

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 RGB frames are stacked into volumes and fed into convolutional layers at first stages [3, 15, 27, 28, 38]. 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 acceleration [42]. Pose normalization (bone lengths and view point normalization) has been reported to help in certain situations [28]. Fusing pose and raw video modalities is traditionally done as late fusion [27], or early through fusion layers [38]. In [22], fusion strategies are learned together with model parameters with by stochastic regularization.

Recurrent architectures for action recognition — Most recent human action recognition methods are based on recurrent neural networks in some form. In the variant Long Short-Term Memory (LSTM) [12], a gating mechanism over an internal memory cell learns long-term and short-term dependencies in the sequential input data. Part-aware LSTMs [30] separate the memory cell 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 [24]. One dimension is time, the second is a topological traversal of the joints in a bidirectional depth-first search, which preserves the neighborhood relationships in the graph.

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, and has drawn increasing interest when dealing with languages, images 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.

In computer vision, visual attention mechanisms date as far back as the work of Itti *et al* for object detection [14] and has been inspired by works from the neuroscience community [16]. Early models were highly related to saliency maps, i.e. pixelwise weighting of image parts that lo-

cally stand out, no learning was involved. Larochelle and Hinton [21] pioneered the incorporation of attention into a learning architecture by coupling Restricted Boltzmann Machines with a foveal representation.

More recently, 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* [26] 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 [37]. The model selects the next location to focus on, based on past information. Ba et al [2] improved the approach to tackle multiple object recognition. In [20], a hard attention model generates saliency maps. Yeung *et al* [41] use hard-attention for action detection with a model, which decides both which frame to observe next as well as when to emit an action prediction.

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, 18] or image captioning [39]. Recently, soft attention was proposed for image [7] and video understanding [32, 33, 40], with spatial, temporal and spatio-temporal variants. Sharma *et al* [32] proposed a recurrent mechanism for action recognition from RGB data, which integrates convolutional features from different parts of a space-time volume. Yeung et al. report a temporal recurrent attention model for dense labeling of videos [40]. At each time step, multiple input frames are integrated and soft predictions are generated for multiple frames. An extended version of this work has been proposed [23] by also taking into account the optical flow. Bazzani *et al* [6] learn spatial saliency maps represented by mixtures of Gaussians, whose parameters are included into the internal state of a LSTM network. Saliency maps are then used to smoothly select areas with relevant human motion. Song *et al* [33] 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.

Up to our knowledge, no attention model has yet taken advantage of articulated pose for attention over RGB sequences.

Our method has slight similarities with [26] in that crops are done on locations in each frame. However, these operations are not learned, they depend on pose. On the other hand, we learn a soft-attention mechanism, which dynam-
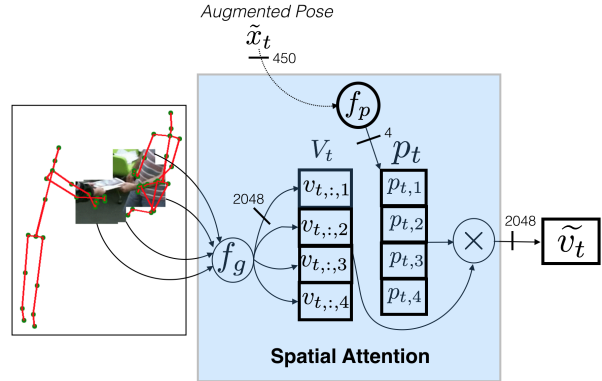

Figure 2: The spatial attention mechanism: *SA-Hands*.

ically weights features from several locations. The mechanism is conditional on pose, which allows it to steer its focus depending on motion.

3. Proposed Model

A single or multi-person action is described by a sequence of two modalities: the set of RGB input images $I=\{I_t\}$, and the set of articulated human poses $x=\{x_t\}$. Both signals are indexed by time t. Poses x_t are defined by 3D coordinates of joints. We propose a hands spatio-temporal attention based mechanism conditioned on pose. This stream processes RGB data I and also uses pose information x (human body joint locations and their dynamics). Our two-stream model comprises the aggregation of the streams presented below.

3.1. SA-Hands: Spatial Attention on Hands

Most of the existing approaches for human action recognition focus on pose data, which provides good high level information of the body motion in an action but somewhat limits feature extraction. A large number of actions such as *Reading*, *Writing*, *Eating*, *Drinking* share the same body motion and can be differentiated only by looking at manipulated objects and hands shapes. Performing fine-grained understanding of human actions can be handled by extracting cues from the RGB streams.

To solve this, we define a glimpse sensor able to crop images around hands at each time step. This is motivated by the fact that humans perform most of their actions using their hands. The cropping operation is done using the pixel coordinates of each hand detected by the middleware (up to 4 hands for human interactions between 2 people). The glimpse operation is fully-differentiable since the exact locations are inputs to the model. The goal is to extract information about hand shapes and about manipulated objects and to draw attention to specific hands.

The glimpse representation for a given hand i is a convolutional network f_g with parameters θ_g (e.g. a pretrained Inception v3), taking as input a crop taken from image I_t at the position of hand i:

$$\boldsymbol{v}_{t,:,i} = f_g(\text{crop}(I_t, \text{hand}_i); \theta_g) \qquad i=\{1,\dots 4\} \qquad (1)$$

Here and in the rest of the paper, subscripts of mappings f and their parameters θ choose a specific mapping, they are not indices. Subscripts of variables and tensors are indices. $\boldsymbol{v}_{t,:,i}$ is a (column) feature vector for time t and hand i. For a given time t, we stack the vectors into a matrix $\boldsymbol{V}_t=\{\boldsymbol{v}_{t,:,i}\}$, where i is the index over hand joints and j the index over the feature dimensions . $\boldsymbol{V}_t$ is a matrix (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 $\boldsymbol{h}_t$:

$$\boldsymbol{h}_t = f_h(\boldsymbol{h}_{t-1}, \tilde{\boldsymbol{v}}_t; \theta_h) \qquad (2)$$

We select the GRU as our recurrent function f_h. To keep the notation simple, we omitted the gates from the equations. The input fed to the recurrent network is the context vector $\tilde{\boldsymbol{v}}_t$, defined further below, which corresponds to an integration of the different features vectors extracted from hands in $\boldsymbol{V}_t$.

An obvious choice of integration are simple functions like sums and concatenations. 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.

We employ a soft-attention mechanism which dynamically weighs the integration process through a distribution $\boldsymbol{p}_t$, determining how much attention hand i needs with a calculated weight $\boldsymbol{p}_{t,i}$. We define the *augmented pose* vector $\tilde{\boldsymbol{x}}_t$ defined by the concatenation of the current pose $\boldsymbol{x}_t$, the acceleration $\dot{\boldsymbol{x}}_t$ and the velocity $\ddot{\boldsymbol{x}}_t$ for each joint over time. At each time step, $\tilde{\boldsymbol{x}}_t$ gives a brief overview of human poses on the scene and their dynamics. In contrast to mainstream soft-attention based mechanisms [32, 1, 23], our attention distribution does not depend on the previous hidden state $\boldsymbol{h}_{t-1}$ of the recurrent network, but exclusively depends on an external information defined just above: the *augmented pose* $\tilde{\boldsymbol{x}}_t$.

Finally the spatial attention weights $\boldsymbol{p}_t$ are given through a learned mapping with parameters θ_p:

$$\boldsymbol{p}_t = f_p(\tilde{\boldsymbol{x}}_t; \theta_p) \qquad (3)$$

Remark that if we replace $\tilde{\boldsymbol{x}}_t$ by $\boldsymbol{h}_{t-1}$ in equation 3 we get the usual soft-attention mechanism by conditioning the attention weights on the hidden state [32]. Attention distribution $\boldsymbol{p}_t$ and features $\boldsymbol{V}_t$ are integrated through a linear
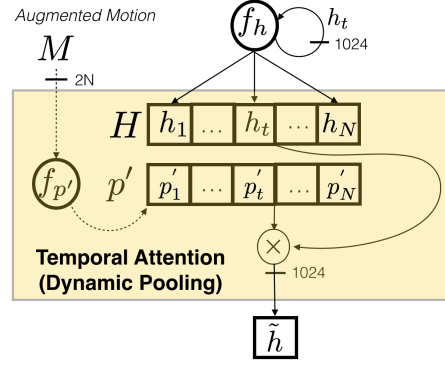

Figure 3: The temporal attention mechanism: *ST-Hands*
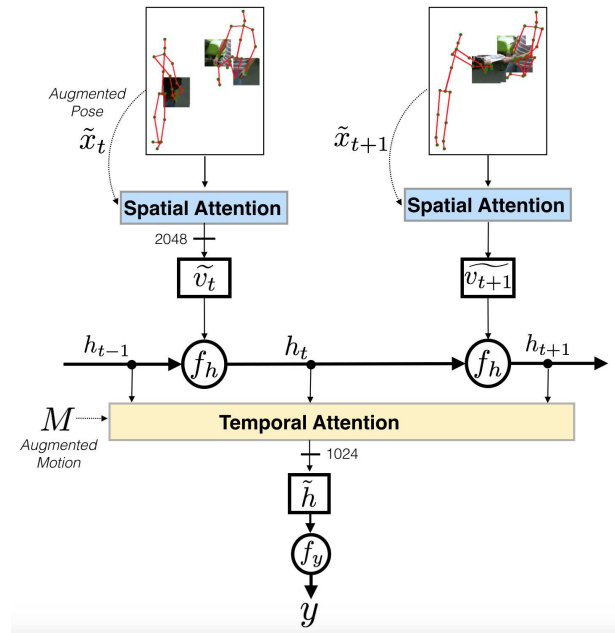.

Figure 4: The spatio-temporal attention mechanism: *STA-Hands*. The spatial mechanism is detailed in figure 2 and the temporal one is details in figure 3

combination as

$$\tilde{\boldsymbol{v}}_t = \boldsymbol{V}_t \boldsymbol{p}_t , \qquad (4)$$

which is input to the GRU network at time t (see eq. (2)). The conditioning on the *augmented pose* in 3 is important, as it provides valuable body motion information at each timestep (see the ablation study in the experimental section).

We refer to this model as *SA-Hands* in our table. For a better understanding of this module, a visualization can be found in Figure 2.
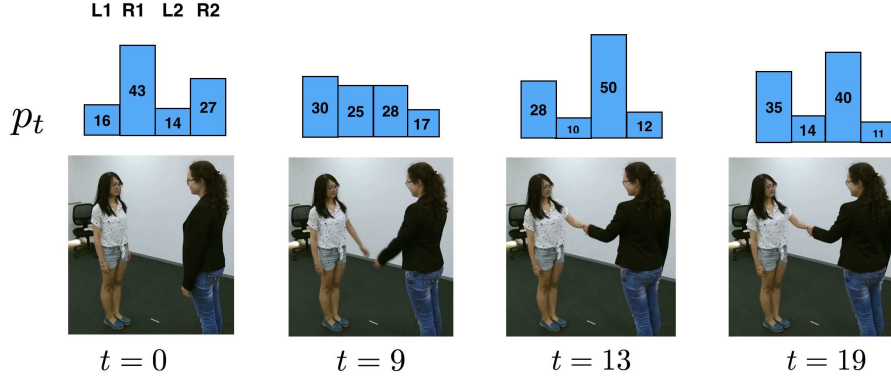
Figure 5: Spatial attention over time: shaking hands will make the attention shift to hands in action.

3.2. ST-Hands: Temporal Attention on Hidden States

Recurrent models can provide predictions for each time step t by performing a mapping directly from the hidden state $\boldsymbol{h}_t$. Some hidden states are more discriminative than other one. Following this idea we perform a temporal pooling on the hidden state level in an adaptive way. At the end of the sequence an attention mechanism automatically gives weights for each hidden states.

The hidden states for all instants t of the sequence are stacked into a 2D matrix $\boldsymbol{H}=\{\boldsymbol{h}_{j,t}\}$, where j is the index over the hidden state dimension. A temporal attention distribution $\boldsymbol{p}'$ is predicted through a learned mapping to automatically identify the most important hidden states (i.e. the most important time instants t). To be efficient, this mapping should have seen the full 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.

To keep the model simple, we benefit from the fact that sequences are of fixed length. We define a statistic called *augmented motion* $\boldsymbol{m}_t$ given by the sum of the absolute acceleration and the sum of the absolute velocity of all body joints at each time step t. $\boldsymbol{m}_t$ is a vector of size 2 and we obtain M by stacking all $\boldsymbol{m}_t$. M gives a good overview of when most important moments occur. Our assumption is that higher values of $\boldsymbol{m}_t$ indicate more useful instants t. But of course the network can learn more complex mappings reacting to more complex motion or poses. The temporal attention weights are given by the mapping:

$$\boldsymbol{p}' = f_p'(\boldsymbol{M}; \theta_p') \tag{5}$$

This attention is used as weight for adaptive temporal pooling of the features $\boldsymbol{H}$, i.e.

$$\tilde{\boldsymbol{h}} = \boldsymbol{H}\boldsymbol{p}' \quad .$$

We called this module *ST-Hands*. A visualization of the module can be found in figure 3.

The spatial and temporal attention mechanism are independent of each other. When both are combined we call the model *Spatio-Temporal Attention over Hands (STA-Hands)*. A visualization of the overall RGB stream can be found in figure 4.

Related work — note that most current work in sequence classification proceeds by temporal pooling of individual predictions, e.g. through a sum or average [32] or even by taking predictions of the last time step. We show that it can be important to perform this pooling in an adaptive way. In recent work on dense activity labeling, temporal attention for dynamical pooling of LSTM logits has been proposed [40]. In the context of sequence-to-sequence alignment, temporal pooling has been addressed with bi-directional recurrent networks [4].

3.3. Deep GRU: Gated Recurrent Unit on Poses

Above, the pose information was used as valuable input to the RGB stream. Articulated pose is also used directly for classification in a second stream, the pose stream. We process the sequence of pose, where at each time step t, $\boldsymbol{x}_t$ is a vector which represents the concatenation of 3D coordinates of joints of all subjects. The raw pose vectors are input into a RNN.

In particular, we learn a pose network f_{sk} with parameters θ_{sk} on this input sequence $\boldsymbol{x}$, resulting in a set of hidden state representation $\boldsymbol{h}^{sk}=\{\boldsymbol{h}_t^{sk}\}$:

$$\boldsymbol{h}_t^{sk} = f_p(\boldsymbol{h}_{t-1}^{sk}, \boldsymbol{x}_t; \theta_{sk}) \tag{6}$$

We call this baseline on poses *Deep GRU* in our tables.

3.4. Stream fusion

Each stream, pose and RGB, leads to its own features respectively $\boldsymbol{h}^{sk}$ for the pose stream and $\tilde{\boldsymbol{h}}$ for the RGB stream. Each representation is classified with its own set

p_t

L1 R1 L2 R2

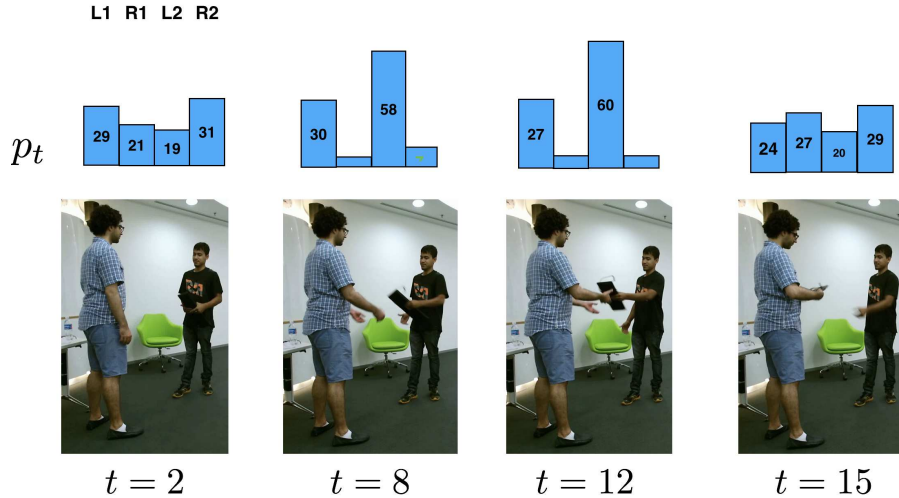| 29 | 21 | 19 | 31 | 30 | 58 | 27 | 60 | 24 | 27 | 20 | 29 |

$t = 2$ $t = 8$ $t = 12$ $t = 15$

Figure 6: Spatial attention over time: giving something to other person will make the attention shift to the active hands in the action.

of parameters using a standard classification approach as defined further below in 4. We fuse both streams on logit level by summing. More sophisticated techniques, such as features concatenation and learned fusion [28] have been evaluated and rejected.

4. Network architectures and Training

Architectures — The pose network f_p consists of a stack of 3 GRU each with an hidden state of size 150.

The glimpse senor f_g is implemented as an Inception V3 network [34]. Each vector $v_{t,:,i}$ corresponds to the last layer before output and is of size 2048. The GRU 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 with ReLu activation. The temporal attention network f_p' is an MLP with a single hidden layer of 32 units with ReLu activation. Output layers of attention networks f_p and f_p' use the softmax activation in order to get the sum of the attention weights equal to 1. The full model (without glimpse sensor f_g) has 10 millions trainable parameters.

Training — All classification are done using a simple fully-connected layer followed by a softmax activation and trained with cross-entropy loss. For the pose stream *Deep GRU* the classification is learned from all the hidden states h_t^{sk}. At test time we average the predictions given by each time step since it gives better results than taking predictions from the last hidden state.

For the RGB stream, classification using *STA-Hands* is learned from the feature vector $\tilde{h}$. When the temporal attention (i.e. *SA-Hands*) is not employed in the RGB stream we follow the same settings as described for the pose stream. The glimpse sensor f_g is pretrained on the ILSVRC 2012

data [29] and is frozen during training. Both spatial p and temporal attention weights p' are initialized to be equal for each input modality. This set up leads to faster convergence and better stability during training.

5. Experiments

The proposed method has been evaluated on the largest human action recognition dataset: NTU RGB+D. We extensively tested all aspects of our model by conducting an ablation study. This leads to get a proper understanding of the choice of our proposed new spatio-temporal mechanism and specially its conditioning aspect.

The NTU RGB+D Dataset (NTU) [30] has been acquired with a Kinect v2 sensor and contains more than 56K videos and 4 millions frames with 60 different activities including individual activities, interactions between 2 people and health related events. The actions have been performed by 40 subjects and with 80 viewpoints. The 3D coordinates of 25 body joints are provided in this dataset. We follow the cross-subject and cross-view split protocol from [30]. Due to the large amount of videos, this dataset is highly suitable for deep learning modeling.

Implementation details — Following [30], we cut videos into sub sequences of 20 frames and sample subsequences. During training a single sub-sequence is sampled, during testing 5 sub-sequences are extracted and logits are averaged. 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 50×50 on the positions of the hand joints (pixel loca-

Methods	Pose	RGB	CS	CV	Avg
Lie Group [35]	X	-	50.1	52.8	51.5
Skeleton Quads [9]	X	-	38.6	41.4	40.0
Dynamic Skeletons [13]	X	-	60.2	65.2	62.7
HBRNN [8]	X	-	59.1	64.0	61.6
Deep LSTM [30]	X	-	60.7	67.3	64.0
Part-aware LSTM [30]	X	-	62.9	70.3	66.6
ST-LSTM + TrustG. [24]	X	-	69.2	77.7	73.5
STA-LSTM [33]	X	-	73.2	81.2	77.2
GCA-LSTM [25]	X	-	74.4	82.8	78.6
JTM [36]	X	-	76.3	81.1	78.7
MTLN [17]	X	-	79.6	84.8	82.2
DSSCA - SSLM [31]	X	X	74.9	-	-
Deep GRU [A]	X	-	**68.0**	**74.2**	**71.1**
STA-Hands [B]	○	X	**73.5**	**80.2**	**76.9**
A+B	X	X	**82.5**	**88.6**	**85.6**

Table 1: Results on the NTU RGB+D dataset with Cross-Subject (CS) and Cross-View (CV) settings (accuracies in %, ○ means that pose is only used for the attention mechanism).

tions of each hands are given by the middleware). Cropped images are then resized to 299×299 and fed into the Inception model.

Training is done using the Adam Optimizer [19] with an initial learning rate of 0.0001. We use minibatches of size 32, dropout with a probability of 0.5 and train our model up to 100 epochs. Following [30], we sample 5% of the initial training set as a validation set, which is used for hyperparameter optimization and for early stopping. All hyperparameters have been optimized on the validation sets.

Comparisons to the state-of-the-art — We show comparisons of our model to the state-of-the-art methods in table 1. We achieve state of the art performance on the NTU dataset with the two-stream model even if we explicitly implemented a weak model, *Deep GRU*, on the pose stream. That shows the strength of our RGB stream called *STA-Hands* at extracting cues. Comparing one by one our two streams (RGB vs pose) demonstrate that *STA-Hands* gets better results than *Deep GRU*.

We have to keep in mind that the pose is used as external data in our RGB stream but only for the cropping operation around hands and for computing of the attention distributions. Poses are never directly fed as input to the GRU in *STA-Hands* for updating the hidden state. The purpose of *STA-Hands* is to extract cues from hand shapes or manipulated objects. By its design choice *STA-Hands* is not able to extract body motion since pose is only used for drawing an attention distribution over hands. However this stream achieves better performance than the pose one. This shows

that RGB data should not be put aside for human action recognition.

We conducted extensive ablation studies to understand the impact of our design choices on the full model, and in particular on the spatial attention mechanism *STA-Hands*.

Conditioning the spatial attention — Conditioning the spatial attention on the statistics of the pose (*augmented pose*) at each time step is a key design choice, as shown in table 2 (*SA-Hands* rows). Compared to mainstream soft-attention mechanisms, which condition attention on the hidden state, we gain 2 points on average (75.0 vs 73.0). Interestingly, conditioning using both the hidden state and the pose statistics deteriorates the performances (75.0 vs 73.6) showing that different kinds of information are contained in these two latent variables. The recurrent unit is not able to combine those two informations or at least ignore the hidden state. We can conclude that the *augmented pose* is a better latent variable for weighting the spatial attention compared to the internal hidden state of the GRU. Compared to simple baseline like summing the different inputs, our methods improves the average accuracy by 3.5 points (75.0 vs 71.5). This opens new perspectives for creating attention mechanisms conditioned on new latent variables which can be external to the GRU (but highly correlated to the inputs and to the final task).

Effect of the temporal attention — Weighted integration of the hidden states over time seems to be an important design choice as shown in table 2. Compared to classical baselines, like averaging the predictions, we improve performance by 3.3 points in average (74.8 vs 71.0). Taking only the final predictions even leads to worst performance. Again we can see that pose and its statistics, in this case the *augmented motion*, are good latent variables for (though external from the input data but highly correlated) for computing the temporal attention weights.

A powerful spatio-temporal attention mechanism — We show consistent results by combining spatial and temporal attention trained end-to-end. Conditioning the spatial and temporal attention mechanisms on statistics of the pose (respectively *augmented pose* and *augmented motion*) leads to the best results. In average we gain up to 5.4 and 4.9 points compared to baseline without any attention modules like summing or concatenating the inputs (76.9 vs 71.5 and 72.0).

Impact of the attention on the two stream model — Again we get consistent results when going from RGB stream only to two-stream model (pose and RGB streams). Even if both streams are trained separately and fused at the logit level they extract complementary features. Spatial attention seems to be more important than temporal one (85.6 vs 84.2). Compared to baseline like summing inputs on the RGB stream, our full spatio-attention mechanism conditioned on poses beats the baseline by 2.8 points on the

Methods	Spatial Attention		Temporal Attention	CS	CV	Avg
	Hidden state	Augmented Pose	Augmented Pose			
Sum	-	-	-	68.3	74.6	71.5
Concat	-	-	-	68.9	75.2	72.0
	X	-	-	69.8	76.2	73.0
SA-Hands	-	X	-	**71.0**	**78.9**	**75.0**
	X	X	-	70.5	76.6	73.6
ST-Hands	-	-	X	**71.1**	**78.5**	**74.8**
	X	-	X	72.2	77.8	75.0
STA-Hands	-	X	X	**73.5**	**80.2**	**76.9**
	X	X	X	72.8	78.3	75.6

Table 2: Effects of the conditioning on the spatial attention and the temporal attention (RGB stream only, accuracies in %).

RGB stream methods	Spatial Attention		Temporal Attention	CS	CV	Avg
	Hidden state	Augmented Pose	Augmented Motion			
Sum-Hands	-	-	-	79.5	85.9	82.8
	X	-	-	80.5	86.8	83.7
SA-Hands	-	X	-	81.4	87.4	84.4
	X	X	-	81.0	86.9	84.0
ST-Hands	-	-	X	80.8	87.6	84.2
	X	-	X	81.4	87.4	84.4
STA-Hands	-	X	X	**82.5**	**88.6**	**85.6**
	X	X	X	81.6	88.0	84.8

Table 3: Effects of conditioning the spatio-temporal attention on different latent variables in the RGB stream for the two-stream model (accuracies in % on NTU). The pose stream is always the same: (*Deep GRU*) for every row.

two-stream model.

Runtime — For a sequence of 20 frames, we get the following runtimes for a single Titan-X (Maxwell) GPU and an i7-5930 CPU: A full prediction from Inception 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 ∼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.

Pose noise — Crops are performed on hand locations given by the middleware. In case of noise, crops could end up not being on hands. We saw, that the attention model can cope with this problem in many cases.

6. Conclusion

We propose a new method for dealing with RGB video data for human action recognition given pose. A soft-attention mechanisms crops on hand joints allowing the model to collect relevant features on hand shapes and on manipulated objects from more relevant hands. Adaptive temporal pooling further increases performance. We show that conditioning attention mechanisms on pose leads to better results compared to standard approach which conditioned on the hidden state. Our method on RGB stream can be seen as a plugin which can be added to any powerful pose stream. Our two-stream approach shows state-of-the-art results on the largest human action recognition even by employing a weak pose stream.

7. Acknowledgements

References

[1] Show , Attend and Tell : Neural Image Caption Generation with Visual Attention. In *ICLM*, 2015. 4

[2] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. In *ICLR*, 2015. 3

[3] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *HBU*, 2011. 2

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. 5

[5] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 3

[6] L. Bazzani, H. Larochelle, and L. Torresani. Recurrent mixture density network for spatiotemporal visual attention. In *ICLR*, 2017 (to appear). 3

[7] K. Cho, A. Courville, and Y. Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE-T-Multimedia*, 17:1875 – 1886, 2015. 2, 3

[8] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, June 2015. 2, 7

[9] G. Evangelidis, G. Singh, and R. Horaud. Skeletal quads:human action recognition using joint quadruples. In *ICPR*, pages 4513–4518, 2014. 7

[10] A. Graves and J. Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In *NIPS*, 2009. 2

[11] J. Han, L. Shao, D. Xu, , and J. Shotton. Enhanced computer vision with microsoft kinect sensor: A review. In *IEEE Trans- actions on Cybernetics*, 2013. 1

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 2

[13] J. Hu, W.-S. Zheng, J.-H. Lai, and J. Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *CVPR*, pages 5344–5352, 2015. 7

[14] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998. 2

[15] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE TPAMI*, 35(1):221–231, 2013. 2

[16] J. Jonides. Further toward a model of the mind's eye's movement. *Bulletin of the Psychonomic Society*, 21(4):247–250, Apr 1983. 2

[17] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 7

[18] Y. Kim, C. Denton, L. Hoang, and A. Rush. Structured attention networks. In *ICLR*, 2017 (to appear). 3

[19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICML*, 2015. 7

[20] J. Kuen, Z. Wang, and G. Wang. Recurrent attentional networks for saliency detection. In *CVPR*, pages 3668–3677, 2015. 3

[21] H. Larochelle and G. Hinton. Learning to combine foveal glimpses with a third-order Boltzmann machine. In *NIPS*, pages 1243–1251, 2010. 3

[22] F. Li, N. Neverova, C. Wolf, and G. Taylor. Modout: Learning to Fuse Face and Gesture Modalities with Stochastic Regularization. In *FG*, 2017. 2

[23] Z. Li, E. Gavves, M. Jain, and C. G. M. Snoek. Videolstm convolves, attends and flows for action recognition. In *CVPR*, 2016. 3, 4

[24] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition. In *ECCV*, pages 816–833, 2016. 2, 7

[25] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. 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. 7

[26] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent models of visual attention. In *NIPS*, pages 2204–2212, 2014. 2, 3

[27] 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 *CVPR*, pages 4207–4215, 2016. 2

[28] N. Neverova, C. Wolf, G. Taylor, and F. Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE TPAMI*, 38(8):1692–1706, 2016. 2, 6

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1, 6

[30] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *CVPR*, pages 1010–1019, 2016. 2, 6, 7

[31] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang. Deep multimodal feature analysis for action recognition in rgb+d videos. In *PAMI*, 2016. 7

[32] S. Sharma, R. Kiros, and R. Salakhutdinov. Action recognition using visual attention. *ICLR Workshop*, 2016. 2, 3, 4, 5

[33] 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 Conf. on AI*, 2016. 3, 7

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 6

[35] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, pages 588–595, 2014. 7

[36] P. Wang, W. Li, C. Li, and Y. Hou. Action Recognition Based on Joint Trajectory Maps with Convolutional Neural Networks. In *ACM Conference on Multimedia*, 2016. 7

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

[38] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J. Odobez. Deep dynamic neural networks

for multimodal gesture segmentation and recognition. *IEEE TPAMI*, 38(8):1583–1597, 2016. 2

[39] 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 *ICML*, pages 2048–2057, 2015. 3

[40] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *arXiv preprint arXiv:1507.05738*, 2015. 3, 5

[41] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. In *CVPR*, 2016. 3

[42] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *ICCV*, pages 2752–2759, 2013. 2