Looking deeper into Time for Activities of Daily Living Recognition

Srijan Das, Monique Thonnat, Francois Bremond

INRIA Université Nice Côte d'Azur, France

name.surname@inria.fr

Abstract

In this paper, we introduce a new approach for Activities of Daily Living (ADL) recognition. In order to discriminate between activities with similar appearance and motion, we focus on their temporal structure. Actions with subtle and similar motion are hard to disambiguate since long-range temporal information is hard to encode. So, we propose an end-to-end Temporal Model to incorporate long-range temporal information without losing subtle details. The temporal structure is represented globally by different temporal granularities and locally by temporal segments. We also propose a two-level pose driven attention mechanism to take into account the relative importance of the segments and granularities. We validate our approach on 2 public datasets: a 3D human activity dataset (NTU-RGB+D) and a human-object interaction dataset (Northwestern-UCLA Multiview Action 3D). Our Temporal Model can also be incorporated with any existing 3D CNN (including attention based) as a backbone which reveals its robustness.

1. Introduction

Action recognition is an important problem in the vision community both for its application domains (security, robotics, healthcare) and its challenging issues. Action recognition challenges depend on the types of videos. Datasets such as UCF-101 [28], kinetics [5] with videos from internet have high inter-class variance and changing background (i.e. "ride a bike" vs. "sword exercise"). In our case, we are particularly interested in Activities of Daily Living (ADL) videos. These videos have 1) high intra-class variance with different subjects performing the same action in different ways, 2) low inter-class variation leading to similar visual appearance of different action classes (for instance, *a person wearing and taking off shoes* have similar motion but belong to different action classes). These challenges are all the more important than videos are recorded within the same environment (background image), which prevents us from using contextual information. Recently, 3D convolutional Neural Networks (3D CNNs) have been tailored to capture the short-term dynamics of full 2D+T volume of a video and somewhat alleviating the first aforementioned challenge. However, these models fail to capture long-range temporal information of actions. Thus, the second challenge low inter-class variation requires attention for discriminating them correctly. Such low inter-class variation is often caused by either similar motion with subtle variation such as *taking out something from pocket/putting something inside pocket*, or complex long-term relationship such as *taking off glasses/wearing glasses*. Also, actions with similar motion tends to have discriminative spatio-temporal features over a small time scale. For instance, wearing and taking off a shoe can be distinguished by taking into account whether or not the shoe is separated from the human body in the first few frames. In order to solve the aforementioned challenge, we need to process the videos at multiple time scales to capture specific subtle motion. Thus, our objective is to capture spatio-temporal relations at multiple time scales and link them over time to disambiguate such temporally complex actions.

In this paper, we propose a Temporal Model to have a focus of attention on the spatio-temporal features of the relevant time scale. This is effectuated by splitting the videos into uniform temporal segments at different time scale (namely granularity). This is followed by a two-level attention mechanism to manage 1) relative importance of each segment for a given granularity and to manage 2) the various granularities (see fig 1).

The Temporal Model which comprises the classification network and the attention module, is trained end-to-end for recognizing actions. We make two hypotheses: the input video clip contains a single class label, and the articulated poses are available. Inspired from the recent trend of using poses to guide RGB cue [3, 4, 8], we take the articulated poses as input to the attention module. The articulated poses are highly informative, robust to rotation and illumination, and thus provide a strong clue to select the pertinent sub-

sequences in a video. To summarize, our contributions are the following:

- An end-to-end Temporal Model to address the recognition of temporally complex actions. This is done by
  - splitting a video into several temporal segments at different levels of temporal granularity.
  - employing a two-level pose driven attention mechanism. First to manage the relative importance of the temporal segments within a video for a given granularity. Second to manage the relative importance of the various temporal granularities.
- An extensive ablation study to corroborate the effectiveness of our proposed Temporal Model. Besides, we propose a Global Model to have a generic and complete approach for action recognition.
- A validation of our method on two public datasets. We achieve state-of-the-art results with our proposed Global Model on NTU-RGB-D dataset, a human activity dataset and Northwestern-UCLA, an object-interaction human action recognition dataset.

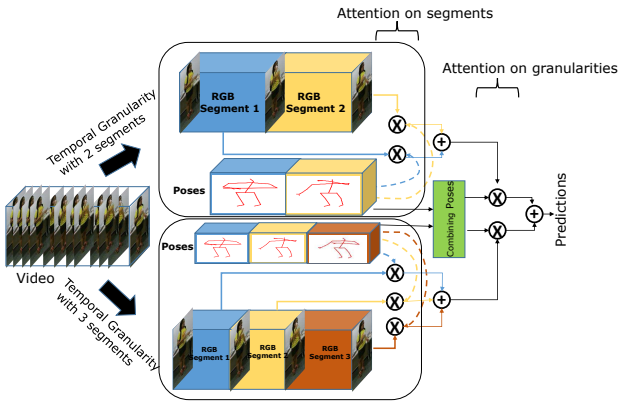

Figure 1. Framework of the proposed approach in a nutshell for two temporal granularities. The articulated poses soft-weight the temporal segments and the temporal granularities using a two-level attention mechanism.

2. Related Work

In this section, we mainly study how models in the literature aggregate the frame-level features along time scale for action classification.

In the past, Wang et al. [31, 32] computed video level descriptor from frame based local features by fisher vector encoding [19]. Later, with the emergence of deep networks, the video level descriptors were computed by simple max-min pooling [6]. All such temporal aggregation methods ignore modeling the temporal structure of the actions.

Then, the Recurrent Neural Networks (RNNs) were exploited to model the temporal evolution of the spatial features fed to them [11, 9]. However, these RNNs operate over simple feature vectors extracted from images. Consequently, they do not capture how the state of an object or a human changes over time in a video. To better capture fine temporal relationships of the frame level features, the authors in [27, 2, 3] soft-weight the key frames. By soft-weighting the key frames, we mean soft-weighting the latent variables output from the RNN classification network. In Temporal Segment Network (TSN) [34], a video is divided into a fixed number of segments, and a frame is randomly sampled from each segment. Then a consensus function aggregates the information from the sampled frames. A similar segment based method has been proposed in [26] with self-attention to adaptively pool the frame-level softmax scores for each segment to obtain the video-level prediction. All these methods including the segment based methods and the formerly discussed temporal attention mechanisms can encode the temporal evolution of the image features sparsely sampled from the whole videos. However, these sparsely sampled frames are disconnected which prevents the extraction of local motion patterns. So, these methods perform well on internet videos (videos with strong motion w.r.t human posture and background) and videos with distinctively high human motion (for [27, 2, 3]), whereas they do not model the smooth local temporal structure for ADL. We also argue that the use of optical flow in TSN and other recognition models can only address instantaneous motion but does not model long-term relations of these motion patterns. Zhou et al. [37] proposed a Temporal Reasoning Network (TRN) by learning the temporal relations among the sparsely sampled frames at multiple time scales. Along with missing subtle motion patterns due to the selection of sparse frames, their method also introduces noise by averaging the features from multiple time scale.

Recently, the introduction of spatio-temporal convolutional operations [29] (in C3D network) addresses the aforementioned drawback of the RNNs yielding rich discriminative features for subtle motion patterns. The C3D network has been enhanced to I3D network [5] which takes up to 64 frames of a video clip for classification. The I3D network is effective for action classification on internet videos, but it is not as successful for ADL recognition. This is because it cannot process long-term temporal relationships to disambiguate actions with similar motion occurring in the same environment.

In order to improve the spatio-temporal features extracted from these 3D CNN (like I3D), Wang et al. [35] have proposed a non-local self-attention block. This non-local

block computes the relative distance (using Gaussian embedding) among all its pixels in the spatio-temporal cube. However, this operation computing the affinity between the features does not go beyond the spatio-temporal cube, thus does not account for long-term temporal relations. For ADL recognition, Das et al. [8] proposed a spatial attention mechanism on the spatio-temporal features extracted from I3D network. The spatial attention provides soft-weights to the pertinent human body parts relevant to the action. However, this spatial attention is applied globally over the spatio-temporal features from I3D network, so it fails to capture long-term temporal relations. To fully address ADL challenges pertaining to long-term temporal relationships, we claim that two types of temporal attention are required even in the presence of temporal convolutional operations. We argue that temporal convolutions are mostly designed to extract motion patterns. A first temporal attention is needed to highlight which motion patterns are important, especially when they are subtle. A second temporal attention is also required to model long-term relations between the motion patterns to disambiguate actions with complex temporal relationship which are very common in ADL.

So, instead of extracting frame-level features from the temporal segments as performed in the state-of-the-art [34, 37], we compute spatio-temporal features from the densely sampled frames within the temporal segments to capture subtle motion. Then we propose a focus of attention on the pertinent temporal segment in a video and the pertinent temporal granularity. Furthermore, the capability of the Temporal Model to be combined with the existing 3D CNNs [8, 35] stands it out from the existing approaches [34, 37] for temporally complex actions.

3. Proposed Temporal Model

In this section, we present the Temporal Model for learning and recognizing actions that exhibit complex temporal relationships. This approach involves three stages (see fig. 2) to classify the actions. *Stage A* consists in splitting the video into several temporal segments at different levels of temporal granularity (see section 3.1). *Stage B* classifies the temporal segments of each granularity. It has a Recurrent 3D Convolutional Neural Network ($R-3DCNN$) and an attention mechanism ($TS-att$) so that the different temporal segments are tightly coupled in an optimized manner (see section 3.2). *Stage C* performs the fusion of the different temporal granularities to classify the action videos (see section 3.3).

3.1. Temporal Segment Representation

In the first stage (*stage A*), our goal is to split the video into several partitions. However, determining the number of such partitions is a difficult task and depends on the content of the action. Thus, for a coarse-to-fine video analysis, a

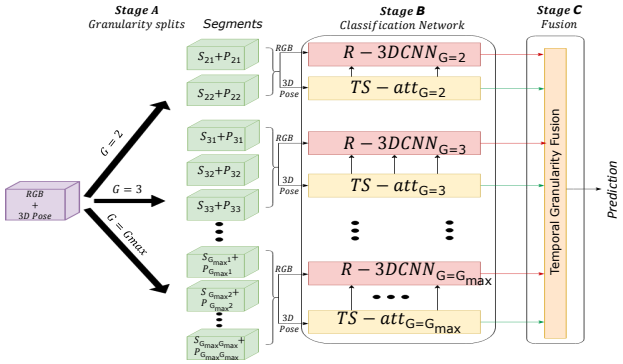

Figure 2. Proposed approach with three *stages*. *Stage A* splits the video into different segments at different granularities. *stage B* is the classification network composed of Recurrent 3D CNN ($R-3DCNN$) and an attention mechanism. *Stage C* performs a fusion of the temporal granularities for predicting the action scores.

hierarchy of temporal segments is built. For a given level in the hierarchy (or granularity), the video is divided into non-overlapping segments of equal length.

Formally, given a video V (RGB+Pose) at granularity G, we divide it into G temporal segments. The video with N frames is processed at different levels of granularity $G = \{2, 3, ..., G_{max} \mid G_{max} \leq N\}$. Thus at granularity G, each temporal segment $S_{Gi} \mid i = \{1, 2, .., G\}$ is a stack of RGB images and $P_{Gi} \mid i = \{1, 2, .., G\}$ is a stack of 3D poses. See an example with a drinking video from NTU-RGB+D [23] in fig. 3.

Note that $G = 1$ represents the whole video and is not input to the proposed Temporal Model. Further discussion can be found in section 4.

3.2. Classification Network

Stage B follows several steps to process the temporal segments for each granularity as described below (see fig. 4).

3.2.1 Recurrent 3D Convolutional Neural Network

A. Processing the Temporal Segments - The first step (step 1) computes the local features for each temporal segment S_{Gi}. These features are computed by a 3D CNN, called $f(.)$. The spatio-temporal representation $ST(V, G)$ is given by:

$$\begin{aligned} ST(V,G) &= ST(\{S_{G1}, S_{G2}, ...S_{GG}\}) \\ &= [f(S_{G1};\theta_w), f(S_{G2};\theta_w), .., f(S_{GG};\theta_w)] \end{aligned}$$

The output of the 3D CNN $f(.)$ with parameters θ_w is a 4-dimensional convolutional feature map. This ST representation is obtained at each level of temporal granularities. In step 2, these convolutional features for each segment S_{Gi}

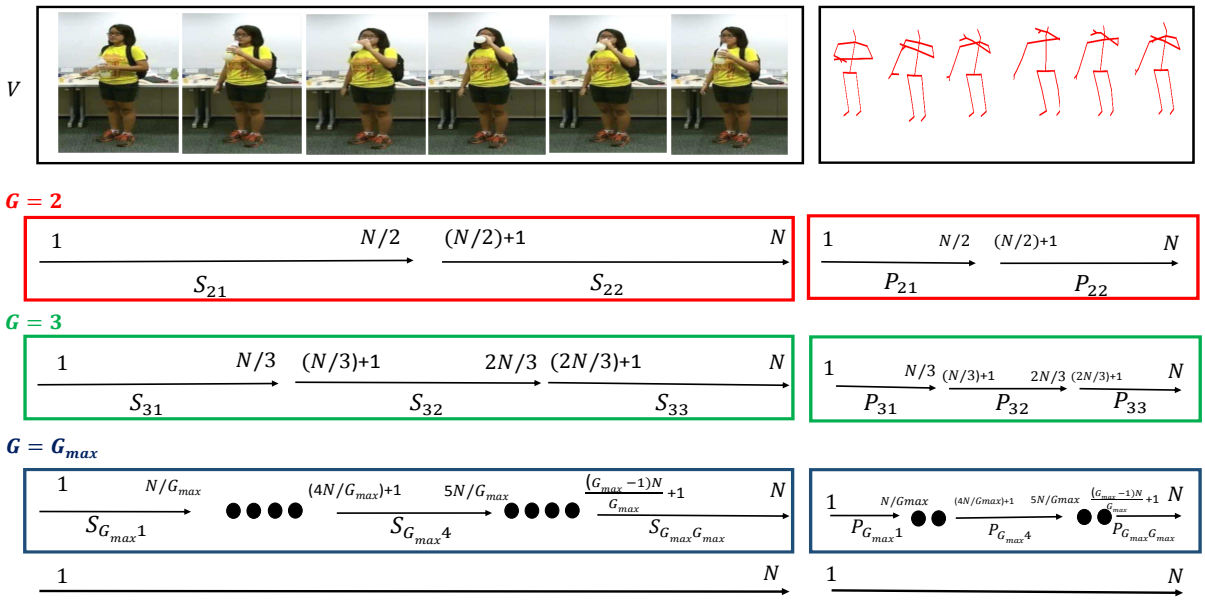

Figure 3. A *drinking* video (from NTU-RGB+D [23]) with RGB frames (at left) and 3D poses (at right) is represented with coarse to fine granularities. G representing granularity ranges from 2 to $G_{max}(\leq N)$.

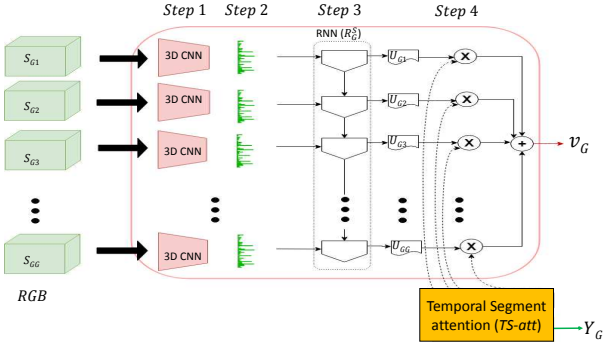

Figure 4. A zoom of the **classification network (stage B)** for a given granularity G. The inputs to the RNN R_G^S are the $flattened$ 3D convolutional features of the temporal segments S_{Gi}. Temporal segment attention soft-weighs the temporal segments.

are resized to a single dimensional tensor by a $flatten(.)$ operation.

B. Combining the Temporal Segments - Step 3 is the global sequential processing of the video at a granularity (G) by the combination of all its temporal segments S_{Gi}. For each granularity G, the aforementioned combination is performed by a recurrent network R_G^S which models the long-term dependencies among the dense temporal segments. Thus, $R{-}3DCNN$ in fig. 4 is the recurrent network R_G^S with 3D CNN $f(.)$ as a backbone. The input of R_G^S with parameters θ_G^S, is the succession of flattened feature maps $f(S_{Gi})$. The output U_{Gi} at each time step i of the recurrent network R_G^S is given by:

$$U_{Gi} = R_G^S(flatten(f(S_{Gi})); \theta_G^S) \quad (1)$$

Step 4 of the classification network combines the output of step 3 with soft-weights provided by a temporal attention mechanism, which is described below.

3.2.2 Attention on Temporal Segments

For a video, some of the segments may contain discriminative information while the others provide contextual information. We argue that poses (3D joint coordinates) are clear indicators to select the prominent sub-sequences in a video as proposed in [27, 2]. This is because of their capability to understand the human body dynamics which is an important aspect in daily living actions.

For a granularity G, the temporal segment attention ($TS-att$) includes two parts (see fig. 5). First, the 3D poses of the temporal segments P_{Gi} are processed by an RNN $R_{G,i}^p$ (with parameters $\theta_{G,i}^p$). Then, the output set of the first RNNs are processed by another RNN R_G^p (with parameters θ_G^p) to combine all the temporal segments into G weights corresponding to the importance of the temporal segments. The soft attention $\alpha_{G,j}$ for j^{th} segment of a given granularity G is predicted by learning the mapping:

$$\alpha_{G,j} = \frac{exp(R_G^p(R_{G,j}^p(P_{Gj}; \theta_{G,j}^p); \theta_G^p))}{\sum_{i=1}^{G} exp(R_G^p(R_{G,i}^p(P_{Gi}; \theta_{G,i}^p); \theta_G^p))} \quad (2)$$

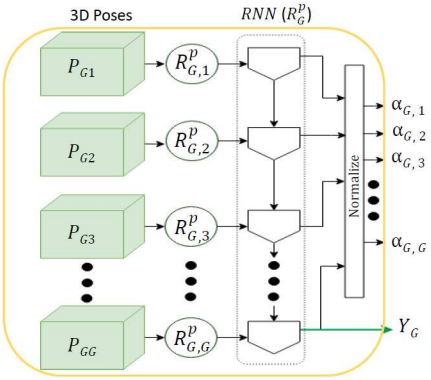

Figure 5. Temporal Segment attention ($TS-att$) from 3D poses for a given granularity G. P_{Gi} being input to the RNN $R^p_{G,i}$ followed by their combination using RNN g^p_G to assign soft-weights $\alpha_{G,j}$.

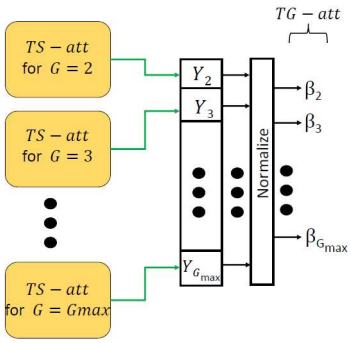

Figure 6. Attention for temporal granularity ($TG-att$) to globally focus on the video representation v_G for a given granularity. The model extended from fig 5 soft-weights the video representations for G_{max} granularities of the video.

Thus the final output v_G of the classification network is a result of adaptive pooling of U_{Gi}, given by:

$$v_G = \sum_{j=1}^{G} \alpha_{G,j} \cdot U_{Gj} \tag{3}$$

For each granularity G, $(G+1)$ recurrent networks are required, which may look expensive but at the same time they operate on **lightweight 3D pose information**. So, they are computationally very efficient.

3.3. Fusion of different temporal granularities

Clipping videos into shorter segments may not be an optimal solution for capturing the subtle motion of an action. So, we propose a temporal granularity attention ($TG-att$) to find the extent of fine temporal segments required to recognize an action. In *stage C* of fig. 2, the temporal segment attention ($TS-att$) described above is extended to soft-weight the output features of the classification network ($R-3DCNN$) for each granularity (see fig. 6). The last timestep output features of the pose based RNN R^p_G are concatenated to form a feature vector Y. So, $Y = [Y_2, Y_3, .., Y_{G_{max}}]$ where $Y_G = R^p_G(R^p_{G,j}(P_{Gj}; \theta^p_{G,j}); \theta^p_G)$ for $j \in [1, G_{max}]$ and $G \in [2, G_{max}]$. The attention weight β_G for G^{th} granularity is computed by

$$\beta_G = \frac{exp(Y_G)}{\sum_{i=2}^{G_{max}} exp(Y_i)} \tag{4}$$

This attention weight is used for focusing on the pertinent temporal granularities. Finally, the prediction for C classes is the weighted summation of the scores at all the granularities followed by a softmax operation:

$$prediction = softmax(\sum_{G=2}^{G_{max}} \beta_G \cdot v_G) \tag{5}$$

4. Global Model for Action Recognition

In the above subsection, we have described our **Temporal Model** where temporal segments with different granularities ($G_{max} \geq 2$) are adaptively fused to classify actions. We call a 3D CNN used for the whole video sequence without any temporal decomposition (i.e. $G = 1$), as the **Basic Model**. The Basic Model is simply the backbone of the Temporal Model to classify the actions. As stated in [30], temporally segmenting videos can destroy the local structure of some short actions. So, we define a **Global Model** for action classification by performing a late fusion of the proposed Temporal Model and the Basic Model. This is done by performing dot product operation of the model scores at logit level. We do not perform soft weighting of the temporal segment with $G = 1$ (i.e., the Basic Model) to classify the actions. The reason is the presence of asymmetric operations in the subnetworks (RNN and 3D CNN) with $G = 1$ and $G > 1$ makes the proposed attention model difficult to train.

5. Network architectures

For the 3D CNN $f(.)$, we use I3D [5] pre-trained on ImageNet [14] and kinetics [5]. Shareable parameters are used to extract spatio-temporal representations of each temporal segment. Spatio-temporal features are extracted from the *Global Average Pooling layer* of I3D. Recurrent networks R^S_G modeling global temporal structure are Gated Recurrent Networks (GRUs) with single hidden layer of size 512. All the recurrent networks for $R^p_{G,j}$ and R^p_G are also GRUs with a hidden state of size 150. We use 3D pose information from depth based middleware [25].

6. Experiments

6.1. Dataset description

We performed our experiments on the following two public human action recognition datasets: NTU RGB+D Dataset [23] and Northwestern-UCLA Multiview Action 3D Dataset [33].

NTU RGB+D Dataset (NTU) - The NTU dataset is currently one of the largest action recognition dataset containing samples with varied subjects and camera views. It was acquired with a Kinect v2 sensor. It contains 56880 video samples with 4 million frames and 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 [23]. We followed the Cross-Subject (CS) and Cross-View (CV) split protocol from [23].

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 contain interaction between human and object which is difficult to model making this dataset challenging as described in [4]. We performed our experiments by following Cross-View protocol from [33], 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.

6.2. Implementation details

For training, first the 3D CNN (I3D) backbone is pre-trained separately on the full human body crops. Then the classification network is trained using the Adam Optimizer [13] with an initial learning rate of 0.0005. We use minibatches of size 32 on 4 GPUs. Straightforward categorical cross-entropy with no regularization constraints on the attention weights has been used to train the network end-to-end. For training the pose driven attention network, similar to [23], we uniformly sample the pose segments into sub-sequences of respectively 5 and 4 frames for NTU and N-UCLA. We use the 3D CNN (I3D) trained on NTU as a pre-trained model and fine-tuned it on N-UCLA.

6.3. Hyper-parameter settings

The hyperparameter G_{max} is the most sensitive choice in our Temporal Model. We have tested different values of G_{max}: 2,3, and 4. In the ablation study, we show that the choice of taking up to 4 granularities is meaningful for the short actions described above.

6.4. Ablation study for Temporal Attention

In this section, we show the effectiveness of our proposed two-level attention mechanism on NTU (CS and CV) and NUCLA datasets. We provide two ablation studies to evaluate the benefit of the **(A) temporal segment attention** ($TS-att$)**, (B) temporal granularity attention** ($TG-att$) compared to baseline I3D.

(A) Fig. 7 is a plot of action classification accuracy w.r.t. the number of granularities. The dotted and solid lines represent the classification results without/ with ($TS-att$) respectively. The relatively higher accuracy scores of the solid line for $G > 1$ as compared to the dashed line indicates the effectiveness of the proposed first level $TS-att$ attention. Fig. 7 also shows that, as we introduce the Temporal Model for $G > 1$, the classification performance improves as compared to the performance of baseline I3D network ($G = 1$) for NTU. This implies that the temporal decomposition in the Temporal Model improves the classification of temporally complex action videos (examples provided at the end of this section). As we go for finer granularities from $G = 2$ to 4, the action classification accuracy goes down, say from 89.7% to 87.4% for NTU-CS with $TS-att$. This is due to the short duration of actions present in the database mentioned above such as *clapping* (-7.2%), *taking out something from pocket* (-6.4%) which lacks temporal structure. It is interesting to note that the classification performance degrades for N-UCLA, when processed in segments. However, we observe that actions like *pick up something with one hand or two hands* are now classified correctly when processed with Temporal Model rather than the Basic Model. Thus, the visual features learned in the Temporal Model are complementary to that of the Basic model.

(B) Table 1 shows the improvement of the classification score with the combination of granularities ($G = 2, 3, 4$). For instance, the accuracy of the Temporal Model from the Basic model improves by 4.5% on NTU, even without employing temporal granularity attention ($TG-att$). $TG-att$ attention further improves the action classification score by 0.8% on NTU dataset. Table 1 also shows the importance of fusing together the Basic and Temporal Model into a Global Model. There are some actions which are correctly recognized by the Basic Model but mis-classified by the Temporal Model such as *punching* (-13.4%) and *throwing* (-8.4%). Temporal decomposition of these actions with very few key frames, does not improve their recognition. So, thanks to the late fusion of the aforementioned Models, we manage to recover the correct recognition of these actions. Thus, the Global Model improves the action classification performance by approximately 2% as compared to the Temporal Model over all the datasets.

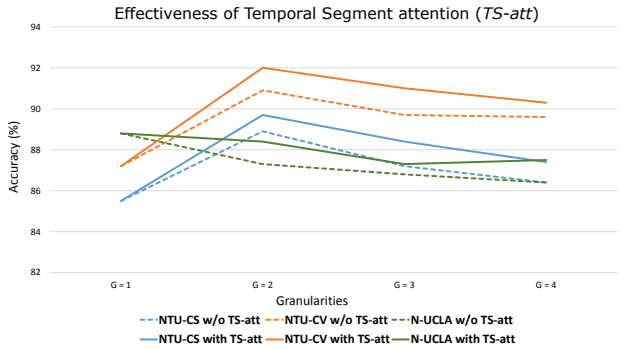

Figure 7. A plot of Accuracy in % (vertical axis) vs number of granularities G (horizontal axis) to show the effectiveness of the temporal segment attention ($TS-att$) on NTU-RGB (CS & CV) and N-UCLA($V^3_{1,2}$). Note that the accuracy for $G = 1$ is on the I3D base network.

Table 1. Ablation study to show the effectiveness of the temporal granularity attention ($TG-att$) and the Global Model compared to the Basic and Temporal Models on NTU-RGB (CS & CV) and N-UCLA($V^3_{1,2}$). Acc. denotes action classification accuracy.

Model	G	$TG-att$ Acc. (%)	NTU-CS Acc. (%)	NTU-CV Acc. (%)	N-UCLA
Basic	1	×	85.5	87.2	88.8
Temporal	2,3,4	×	89.9	91.9	88.2
Temporal	2,3,4	✓	90.6	92.8	89.5
Global	1,2,3,4	✓	**92.5**	**94.0**	**91.0**

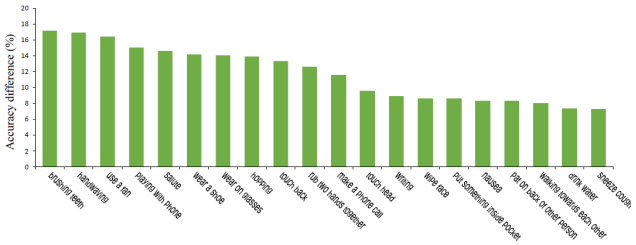

Figure 8. Accuracy difference per action label of the 20 best classes for NTU dataset between the Temporal Model and the Global Model. The base network is I3D and the results are averaged over the CS and CV protocols.

To analyze the gain obtained by the Temporal Model, we study the difference in classification accuracy between the Basic Model and the Global Model for the 20 best classes in fig. 8. Our Global Model improves **53 out of 60** action classes. The most significant improvements concern actions with repetitive cycles like *brushing teeth* (+17.1%), *handwaving* (+16.9%), and *use a fan* (+16.4%). These actions have long-term temporal structure (the repetition of actions) which our proposed Temporal Model successfully decipher. The Basic Model fails when it has to distinguish between action pairs with similar poses and subtle motion, such as *wearing and taking off a shoe* and *wearing and taking off glasses*. On the contrary, the temporal decomposition of these actions into segments enables the classifier to discriminate between similar pairs, and thus improves the recognition of *wearing a shoe* (+14.1%) and *wearing glasses* (+14.0%). For these actions, the temporal segments contain very specific and discriminative parts which enables the classifier to discriminate the similar ones. See fig. 9 in which our proposed Global Model outperforms the Basic Model (I3D). For action *taking on a shoe*, the first temporal segment S_{21} for granularity $G = 2$ discriminates it from *taking off a shoe*. Similarly, for action *put something inside pocket*, the second temporal segment S_{32} for temporal granularity $G = 3$ enables the classifier to recognize the action correctly. Actions like *cross hands in front* (-4.0%) and *punching* (-3.1%) are the two major worst classes. The Global Model has difficulties recovering these actions because the Temporal Model may add noise to the recognition score acquired by the Basic Model during their fusion. However, these drops in performance are not as significant as the improvements.

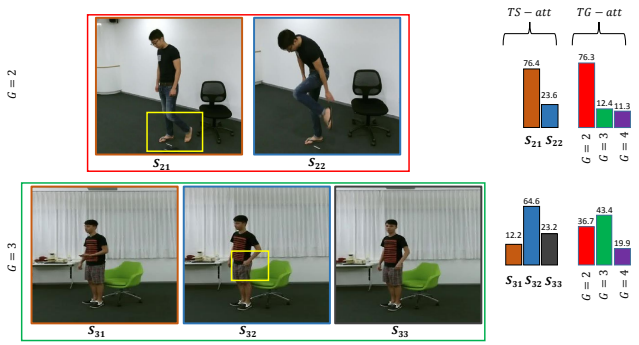

Figure 9. Examples of videos at left (*taking on a shoe* and *put something inside pocket*) and attention weights of temporal segments and granularities at right. Our proposed Global Model classifies these action videos correctly but Basic Model (I3D) does not. The distinctive context or gesture in the pertinent temporal segment is highlighted with yellow box.

6.5. Comparison with the State-of-the-art

In this section, our Global Model is compared with previous methods. We achieve state of the art performance on the NTU and N-UCLA datasets as shown in Table 2 and Table 3. For input modalities, a pose is defined as the 3D body joint information whereas depth is defined as the depth map from RGB-D sensor. Note that in table 2 and 3, Glimpse Cloud [4] uses pose information only for learning but performs significantly worse for cross-subject protocol on NTU dataset. In table 2, the Temporal Model compete closely with PEM [18] which uses evolution of heatmaps of pose estimation. But we argue that, in real world settings this pose estimation can be noisy, especially in case of oc-

clusions. Thus, the reliability of this pose estimation weakens the (PEM) method. In comparison, we only use pose information to learn the attention weights. Consequently, classification is not affected by the wrong poses. Table 2 and Table 3 also show the effectiveness of our Temporal Model when adopted on top of rich discriminative features from existing spatio-temporal attention models [8, 35]. We call them Global Model (I3D-NL base) or (P-I3D base) - the base network in parentheses. The attention mechanism of non-local blocks [35] from convolutional feature maps are not view-invariant and thus perform worse than simple I3D as backbone of the Temporal Model in CV protocols. P-I3D [8] with 42M trainable parameters as compared to simple I3D's 12M trainable parameters outperforms the state-of-the-art results on NTU (95% average over CS and CV) and NUCLA (93.5%) datasets when used as a backbone of the Temporal Model. The Global Model with P-I3D as base network has 80M trainable parameters and improves action, with similar motion like *wearing glasses* (+2.5%) and *taking off glasses* (+2.1%) compared to the Basic Model (P-I3D).

Table 2. Accuracy results on NTU RGB+D dataset with cross-subject (CS) and cross-view (CV) settings along with indicating the input modalities (accuracies in %); Att indicates attention mechanism, ∘ indicates that the modality has only been used for training. *The code has been reproduced on this dataset.

Methods	Pose	RGB	Att	CS	CV
STA-LSTM [27]	✓	×	✓	73.2	81.2
TS-LSTM [15]	×	✓	×	74.6	81.3
GCA-LSTM [16]	✓	×	✓	74.4	82.8
DSSCA-SSLM [24]	×	✓	×	74.8	-
MTLN [38]	×	✓	×	79.6	84.8
VA-LSTM [36]	✓	×	×	79.4	87.6
STA-Hands [2]	✓	✓	✓	82.5	88.6
altered STA-Hands [3]	✓	✓	✓	84.8	90.6
Glimpse Cloud [4]	∘	✓	✓	86.6	93.2
I3D-NL [35]*	×	✓	✓	88.4	87.1
PEM [18]	✓	✓	×	91.7	95.2
P-I3D [8]	✓	✓	✓	93	95.4
Global Model (I3D base)	✓	✓	✓	92.5	94.0
Global Model (I3D-NL base)	✓	✓	✓	92.6	93.9
Global Model (P-I3D base)	✓	✓	✓	**93.9**	**96.1**

6.6. Runtime

Training the Temporal Model end-to-end takes 3 hours with a single job spread over 4 GTX 1080 Ti GPUs. Pre-training the Basic Model on the NTU dataset takes 15 hours. 3D CNN (I3D) features are extracted in parallel over 16 GPUs for 4 granularities and thus varying the granularity does not affect the run time of the model. At test time, RGB pre-processing takes one second (loading Full-HD video and extracting 3D CNN features). The Temporal Model with granularity G_{max}= 4, takes 1.1 ms including the prediction from the Basic Model on a single GPU. The temporal attention module is very efficient because it works only on the 3D pose joints. Classification can thus be done close to real-time. The proposed model has been implemented in Keras [7] with tensorflow [1] as back-end.

Table 3. Accuracy results on Northwestern-UCLA Multiview Action 3D dataset with cross-view $V^3_{1,2}$ settings along with indicating input data modalities (accuracies in %); Att indicates attention mechanism.

Methods	Data	Att	$V^3_{1,2}$
HPM+TM [21]	Depth	×	91.9
HBRNN [12]	Pose	×	78.5
view-invariant [17]	Pose	×	86.1
Ensemble TS-LSTM [15]	Pose	×	89.2
nCTE [10]	RGB	×	75.8
NKTM [20]	RGB	×	85.6
Glimpse Cloud [4]	RGB+Pose	✓	90.1
P-I3D [8]	RGB+Pose	✓	93.1
Global Model (I3D base)	RGB+Pose	✓	91.0
Global Model (P-I3D base)	RGB+Pose	✓	**93.5**

7. Conclusion

In this paper, we have presented an end-to-end Temporal Model for human action recognition. The Temporal Model includes the notions of granularity and temporal segments for each granularity. A two-level attention mechanism manages the relative importance of each temporal segment for a given granularity and handles the various granularities. The proposed attention model is driven by articulated poses. As now 3D poses can be obtained from RGB frames using [22], our recognition approach is not restricted for RGB-D videos. Our ablation study shows the potential of the proposed Temporal Model to capture complex temporal relationships, finally resulting in better action classification. Existing attention/no-attention based spatio-temporal CNN architectures can be combined with our Temporal Model as its backbone. For example, the Temporal Model when combined with existing spatial attention based 3D CNN, outperforms the state-of-the-art performance on NTU and N-UCLA datasets. Future work will be to adapt the proposed Temporal Model for untrimmed action detection.

Acknowledgement

We are grateful to INRIA 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, and J. Mille. Human activity recognition with pose-driven attention to rgb. In *The British Machine Vision Conference (BMVC)*, September 2018.

[4] 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.

[5] 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.

[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, A. Chaudhary, F. Bremond, and M. Thonnat. Where to focus on for human action recognition? In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 71–80, Jan 2019.

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

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

[12] 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.

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

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

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-net: Localization-classification-regression for human pose. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1216–1224, July 2017.

[23] 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.

[24] 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.

[25] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011.

[26] S. Song, N.-M. Cheung, V. Chandrasekhar, and B. Mandal. Deep adaptive temporal pooling for activity recognition. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, pages 1829–1837, New York, NY, USA, 2018. ACM.

[27] 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.

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

[29] 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.

[30] G. Varol, I. Laptev, and C. Schmid. Long-term temporal convolutions for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1510–1517, 2018.

[31] 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.

[32] H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*, Sydney, Australia, 2013.

[33] 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.

[34] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Val Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

[35] X. Wang, R. B. Girshick, A. Gupta, and K. He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[36] 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.

[37] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, pages 831–846, 2018.

[38] 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.