Long-term behaviour recognition in videos with actor-focused region attention

Ballan, Luca; Strafforello, Ombretta; Schutte, Klamer

DOI
10.5220/0010215803620369

Publication date
2021

Document Version
Final published version

Published in
VISAPP

Long-term Behaviour Recognition in Videos with Actor-focused Region Attention

Luca Ballan[1,2], Ombretta Strafforello[2,3] and Klamer Schutte[2]
[1]*Department of Math, University of Padova, Italy*
[2]*Intelligent Imaging, TNO, Oude Waalsdorperweg 63, The Hague, The Netherlands*
[3]*Delft University of Technology, The Netherlands*

Keywords: Action Recognition, Region Attention, 3D Convolutional Neural Networks, Video Classification.

Abstract: Long-Term activities involve humans performing complex, minutes-long actions. Differently than in traditional action recognition, complex activities are normally composed of a set of sub-actions, that can appear in different order, duration, and quantity. These aspects introduce a large intra-class variability, that can be hard to model. Our approach aims to adaptively capture and learn the importance of spatial and temporal video regions for minutes-long activity classification. Inspired by previous work on Region Attention, our architecture embeds the spatio-temporal features from multiple video regions into a compact fixed-length representation. These features are extracted with a 3D convolutional backbone specially fine-tuned. Additionally, driven by the prior assumption that the most discriminative locations in the videos are centered around the human that is carrying out the activity, we introduce an Actor Focus mechanism to enhance the feature extraction both in training and inference phase. Our experiments show that the Multi-Regional fine-tuned 3D-CNN, topped with Actor Focus and Region Attention, largely improves the performance of baseline 3D architectures, achieving state-of-the-art results on Breakfast, a well known long-term activity recognition benchmark.

1 INTRODUCTION

Long-term activity recognition is getting increasing attention in the Computer Vision community as it allows for important applications related to video surveillance and sport video analysis. However, this task is intrinsically complex because of the long duration of the videos, the variability in the activities composition and the visual complexity of video frames from real world scenarios. Inspired by previous work on Region Attention (Yang et al., 2017), we introduce a model that can adaptively select and focus on the video regions that are most discriminative for the complex activity classification.

Our method is driven by two assumptions. Firstly, not all the locations and the moments in the videos are equally important. The activity "preparing cereal bowl", for example, has a precise location in the video frames. Other locations belong to the background, namely regions where the activity does not happen. Background locations might show "distracting" elements that might induce to misclassify the activity. Similarly, a correct classification of a cooking activity might be possible just by looking at the last seconds of the videos, that are likely to show the ready dish. On the contrary, some less informative moments might occur elsewhere, for instance when the cook is looking for the ingredients. Following this assumption, we introduce a Region Attention module, that can explicitly choose among multiple spatial and temporal input regions. This setting acts as a natural data augmentation strategy, and allows to retain only the information that is relevant for the classification.

The second assumption is that the most discriminative spatial regions in the videos are the ones placed around the actor that is accomplishing the activity. For example, for cooking activities, the ingredients and the utensils that are characteristic of the actions, are those that the cook interacts with. Therefore, focusing on the cook should give sufficient information to understand what dish is being made. Hence, we introduce an Actor Focus mechanism that allows the model to explicitly center the attention on the actor.

Due to the large intra-class variability, modelling long-term activities can be difficult. The recent solutions in the literature involve 3D-CNNs as effective spatio-temporal feature extractors (Carreira and Zisserman, 2017), combined with additional mod-

Ballan, L., Strafforello, O. and Schutte, K.
Long-term Behaviour Recognition in Videos with Actor-focused Region Attention.
DOI: 10.5220/0010215803620369
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 362-369
ISBN: 978-989-758-488-6

ules that further process the features in the temporal dimension, including temporal convolution (Hussein et al., 2019a) and self-attention (Hussein et al., 2019b; Hussein et al., 2020a; Wang et al., 2018). Even though these works reached competitive results in common long-term activities benchmarks, we argue that the performance of these models is heavily influenced by the quality of the 3D-CNNs backbone training. Despite their potential, 3D-CNN architectures are characterized by the downside of having a large amount of parameters that makes the learning process extremely data hungry. Since the datasets for long-term activities are limited in size (Kuehne et al., 2014; Sigurdsson et al., 2016; Yeung et al., 2018) learning general video representations with these models without overfitting on the training set is unfeasible. That is why our approach based on multiple regions is crucial to reach better generalization. We show that the combination of an optimal backbone fine-tuning, augmented with the multiple regions, with the Region Attention method and the Actor Focus mechanism achieves state-of-the-art results on the Breakfast Actions Dataset benchmark (Kuehne et al., 2014).

2 RELATED WORK

Although a wide range of solutions for short-range action recognition have been proposed (Carreira and Zisserman, 2017; Kalfaoglu et al., 2020; Qiu et al., 2019), these are not necessarily transferable to long-term activity recognition, as the two data types are fundamentally different. Short actions (or *unit-actions*), such as "cutting" or "pouring" are limited in duration and consist of a single, possibly periodic, movement. Because of this, they are easily recognizable by looking at a small number of frames, sometimes even one (Schindler and Gool, 2008). On the contrary, long-term activities are composed by a collection of unit-actions, where some of them might be shared among different classes. For example, the action "pouring" belongs both to the classes "making tea" and "making coffee". Because of this, it is not possible to classify a complex activity by looking at a specific moment, but the whole time span should be considered. Therefore, more sophisticated architectures are required.

2.1 Long-term Modelling

The majority of the recently proposed works on long-term modelling enhance the exploitation of the temporal dimension. Timeception (Hussein et al., 2019a), for example, achieves this with multi-scale temporal convolutions which learn flexibly long-term temporal dependencies. Similarly, (Varol et al., 2017) consider different temporal extents of video representations at the cost of decreased spatial resolution. (Wu et al., 2019a) propose a long-term feature bank of information extracted over the entire span of videos as context information in support to 3D-CNNs. (Burghouts and Schutte, 2013) rely on STIP (Spatio-Temporal Interest Points) features weighted by their spatio-temporal probability. Another example of temporal reasoning is provided by the TRN (Temporal Relation Network) (Zhou et al., 2018), that learns dependencies between video frames, at both short-term and long-term timescales. Conditional Gating adopted in TimeGate (Hussein et al., 2020b) enables a differentiable sampling of video segments, to discard redundant information and achieve computational efficiency. According to another recent thread, supported in VideoGraph (Hussein et al., 2019b) and (Wang and Gupta, 2018), a thorough representation of complex activities can be achieved by explicitly modelling the human-object and object-object interactions across time. The VideoGraph method learns this type of information through a fixed set of latent concepts depicting the activity evolution, whereas (Herzig et al., 2019) address directly the object-object interactions, embedding them in a graph structure.

Among the most performing work that utilizes the Breakfast dataset, (Hussein et al., 2020a) propose a new kind of convolutional operation which is invariant to the temporal permutations within a local window. Their proposed model is better suited to handling the weak temporal structure and variable order of the unit-actions that compose the long-term activities. On the other hand, ActionVlad (Girdhar et al., 2017) develops a system that pools jointly across spatio-temporal features provided by a two-stream network. Finally, Non-local Nets (Wang et al., 2018) provide a building block for many deep architectures: computing the response at a position as a weighted sum of the features at all positions, they capture long-term dependencies in a way that is not feasible with standard convolutional or recurrent operations.

2.2 Region Attention

The best attempt of weighted averaging approach that could go under the name of Region Attention, to the best of our knowledge, has been done by (Yang et al., 2017), who believe that a good pooling or aggregation strategy should adaptively weigh and combine the information across all parts of multimedia content. Their Neural Aggregation Networks (NAN) served as a general framework for learning content-adaptive

pooling, emphasizing or suppressing input elements via weighted averaging. The concept of Regional Attention as developed in Section 3 is a direct evolution of what has been applied on Face Expression Recognition in (Wang et al., 2020). The authors built a so-called Region Attention Network (RAN), capable of extracting features from several spatial regions of the original images, and combining them from a weighted perspective. This method is more robust to occlusion and can better attend to the specific face parts that characterize the human expressions.

3 METHOD

In our approach, we use the Inflated 3D ConvNet (I3D) (Carreira and Zisserman, 2017), optimally fine-tuned for the classification task at hand, as a feature extractor for multiple video regions and timesteps. These representations are fed to a novel attention module, that summarizes them into a compact feature vector. We experiment with two variants of the module: (spatial) Region Attention (RA) and Temporal Attention (TA), used both individually and jointly.

3.1 I3D and Region Attention

The Region Attention module produces fixed-length representations that highlight the most informative regions received as input. To achieve this, frames are partitioned with an overlapping regular $N \times N$ grid, with $N = 3$, to extract crops. The attention mechanism is built on top of I3D, which processes the raw videos and outputs respective feature representations. The full model can be trained in two steps. To provide coherent features, I3D is fine-tuned on the multiple video regions that will be considered by the attention module. Each video is handled in a fixed mode: *i.* the video frames are converted to RGB and normalized within the range `[-1.0, 1.0]`; *ii.* $T = 64$ timesteps, of 8 consecutive frames each, are uniformly selected from the full clip; *iii.* through a grid-like scheme, R squared spatial regions are cropped from the fixed-length sample, and resized to I3D input's spatial size 224×224. The resulting region crops are partially overlapped, since the cropping portion is 5/8 of a frame. $R = 10$ because the full frame is considered together with the 9 grid regions to preserve global information. $R = 11$ when Actor Focus is applied.
During each I3D training epoch, for each video in the training and validation splits one of the spatial regions is randomly selected. First, this provides data augmentation. Second, I3D extracts features according to the region given as input, instead of always seeing a full frame, thus learning the importance of details in different locations and scale. This behaviour is consistent with the following Region Attention module, that learns to weight the region features, thus making I3D a suitable backbone. Within the Region Attention module a weight in [0.0, 1.0] is assigned to each region feature, through a shared fully-connected layer + Sigmoid activation. The values are used to compute a weighted average of the features, unweighted on the temporal dimension, which is fed into a classification layer. The full process is shown in Figure 1.

3.2 Temporal Attention

A similar scoring mechanism can be applied to the timesteps. The idea of using attention in the temporal dimension derives, for example, from the fact that initial frames generally have a relatively lower relevance compared to the last frames, which show the result of the activity. Also, in some timesteps the activity does not happen at all. However, extended ablation studies showed that Temporal Attention loses its effectiveness when I3D is fine-tuned, as it appears that the I3D model collects already sufficient information from the sequence of the timesteps. Finally, assuming independence between region importance and timesteps importance, we explored the integration of Region Attention and Temporal Attention by using concatenation, as shown in Figure 3.

3.3 Actor-Focus

A further improvement is driven by the consideration that in a high number of cases a single person is performing the activity, generally in a static spatial region of the video. Person detection finds its utility here for the action classification task, due to the following: *i.* detecting the people in the scene allows the focus to be on the subject performing the activity and on the closest involved entities; *ii.* I3D fine-tuning can be carried out exploiting spatial crops centered on the actor, additionally boosting the ability of the framework to prioritize and highlight the activity globe against clutter and irrelevant background.

For each video, FacebookAI's Detectron2 (Wu et al., 2019b) is used to get the person bounding box from each frame. As shown in Figure 2, the coordinates are averaged, images are cropped accordingly and then resized. Specifically, a square with the same center of the average bounding box and dimensions equal to the biggest between height and width of the bounding box is taken. Since the person box has almost always a higher value for the height than for the width, this means that despite the process of having

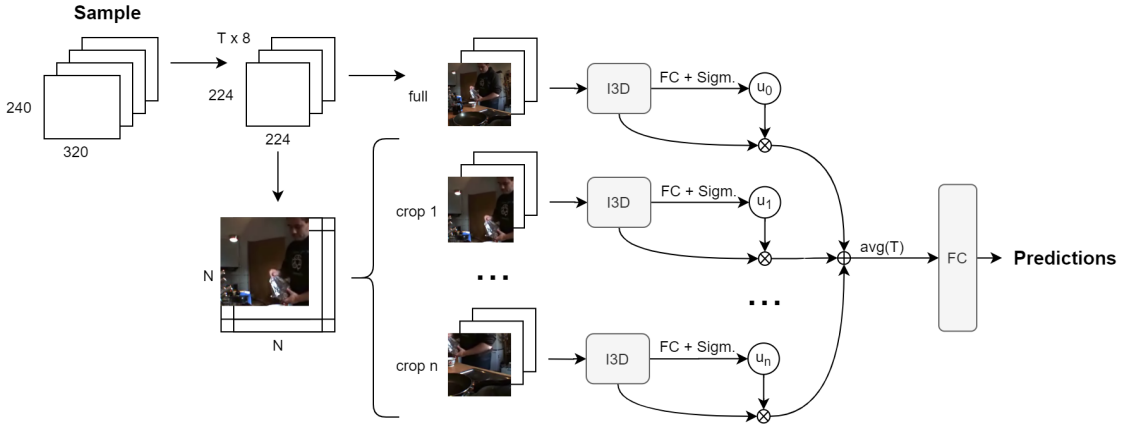

Figure 1: The Region Attention module. From every sample in the dataset a 3 x 3 grid is used, and the extracted crops are placed next to the full frames for I3D feature extraction. A fully-connected (FC) layer and a Sigmoid function attribute to each region a score, through which the features are averaged in a weighted manner and feed the final classification layer.

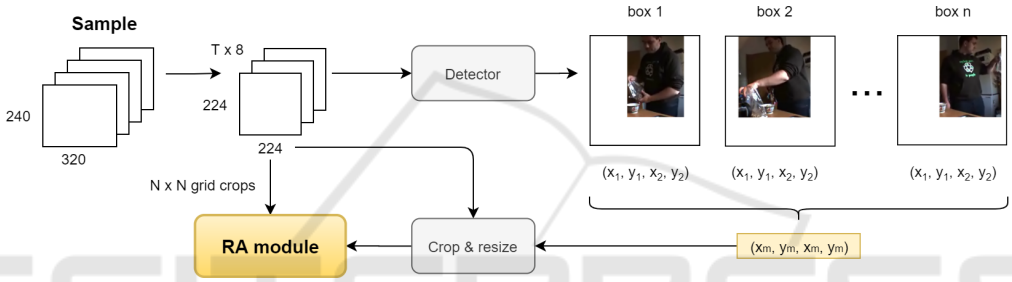

Figure 2: The Actor-Focused crop selection through person detection in video frames. Bounding box coordinates for the actor detected in each frame are averaged and used to crop the original video around the person performing the activity. The selected region is added to the others to feed the Region Attention module.

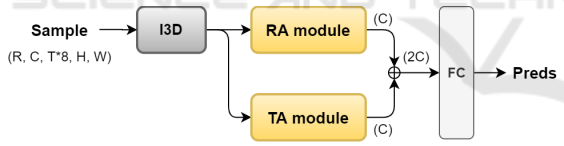

Figure 3: Concatenation of regional and temporal features of a video for classification. The two feature vectors computed separately from the two modules are concatenated along the channels and feed the final classification layer.

a fixed averaged bounding box across the video, the actor is likely not to be cut out of the scene when performing small movements. These actor-centered videos are fed to I3D and Region Attention together with the other regions coming from the fixed-grid selection.

4 EXPERIMENTS

4.1 Dataset

The Breakfast Actions Dataset (Kuehne et al., 2014), on which we achieve state-of-the-art results, comprises 10 classes of long-term activities performed by 52 actors. Videos of the first 44 actors are used for training, the remaining for testing. We keep 5 actors from the training split for validation. This gives, respectively, 1322, 411 and 256 videos. The up-to-10-minutes long videos (2 minutes on average) are handled to be of fixed length and size as explained in Section 3. To obtain equal width and height the horizontal central crop of each original frame is resized and considered as the selected frame. The resulting frames feed both the grid-like region selection and the person detection mechanism.

4.2 Actor-Focused I3D + RA

The full Actor-Focused I3D + RA model, unless otherwise specified, considers 11 regions in total. These include the full frame, kept in order to preserve information about the global spatial context from which the regions are extracted, and the actor-centered region. The original I3D implementation remains unchanged except for the very last layer, which is newly initialized considering a 10-fold output due to the number of Breakfast classes. This allows for the uti-

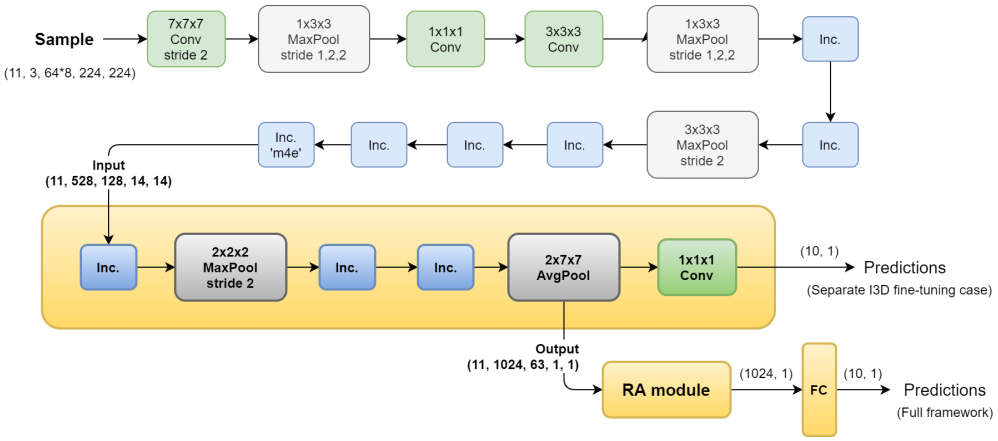

Figure 4: I3D + RA architecture. The fine-tuned section (last 3 Inception blocks), together with the RA module and the classification layer, composes the trainable part of the framework, highlighted in yellow. Note that the 1x1x1 Convolution, used as a fully-connected layer in the original I3D architecture, is not used when extracting the features from fine-tuned I3D.

lization of pre-trained I3D checkpoints obtained from Kinetics 400 (Carreira and Zisserman, 2017).

Experiments where run on Nvidia GeForce GTX 1080 and Tesla V100 GPUs. Due to the large size of the input and the huge number of parameters of I3D (tens of millions), the devices capacity enabled a maximum batch size of 4 for the backbone fine-tuning. In addition, to make the computation feasible, we restrict the fine-tuning only to the the last three Inception blocks and freeze the bottom layers. The features processed by I3D are extracted from the $2 \times 7 \times 7$ *AvgPool* layer, and feed the conclusive RA step. Again, RA calculates importance scores for each input region and uses them in a weighted average, to aggregate the multi-regional input in a compact representation. The output is a 1024-dimensional vector (2048 in the Region + Temporal Attention setting) and is used for the final classification step. The full architecture, detailed on input and output shapes, is shown in Figure 4.

The developed framework is implemented using PyTorch and trained on single GPU for 100 epochs, using Adam optimizer with learning rate 10^{-3}, ε value 10^{-8}, weight decay coefficient 10^{-5}, and CrossEntropy loss function calculated on the 10-fold logits of the last fully-connected layer. Results are calculated on the test set, while our best models are chosen based on the best validation accuracy obtained in 100 epochs.

4.3 Ablation Studies

4.3.1 Temporal Dimension

First, we show that the amount of timesteps considered has a remarkable impact on classification. Consequently, we confute the assumption that only a few specific moments in time are sufficient for the classification of complex activities. Previous work (Hussein et al., 2019b) shows that a uniform selection works generally better than sampling timesteps randomly. Therefore, we keep this setup, and vary instead the quantity of input timesteps, from 4 to 128. Each timestep is composed of 8 consecutive frames.

Table 1: Full framework results varying the timestep number. Best accuracy on the test set has been reached with T = 64.

T	4	16	64	128
Acc. %	68.13	83.94	**89.84**	86.13

The results, shown in Table 1, indicate that, for an accurate classification, a sufficiently but not exceedingly high number of timesteps from the videos should be considered. This finding is coherent with the complex and variable nature of long-term activities, that are characterized by the presence of several unit-actions. The unit-actions should be represented by the selected video timesteps. Also, sampling a large amount of timesteps helps reduce the noise in input signals, leading to a more robust modelling of the underlying features. However, the results show that an excessively long input might not be optimal. In fact, the highest accuracy obtained with our full model (89.84%) is achieved with T = 64, while the accuracy drops when using 128 timesteps. This unexpected outcome can be motivated by considering that many videos in Breakfast are shorter than 128×8 frames = 1024 frames. In this short videos, the 128 selected segments significantly overlap, thus introducing high redundancy and altering the temporal dynamics.

Following the analysis on the number of video timesteps, we demonstrate that the overall temporal order of the timesteps carries valuable information. First, we shuffle the timesteps during the I3D fine-

Table 2: Comparison between frame filtering methods on validation and test sets. Despite a lower accuracy in validation, selecting uniformly 64 timesteps from each video gives better results on the test set. Here, I3D is fine-tuned according to the Multi-Regional with Actor-Focus setting.

	I3D		I3D + RA	
	val. acc. %	test acc. %	val. acc. %	test acc. %
512 equally spaced frames	83.59	80.05	87.89	86.86
T = 64 (8 frames each)	82.03	82.97	87.50	**89.84**

tuning. As convolution is not a permutation invariant operator, the shuffling has a negative impact on the backbone, and consequently on the Region Attention. With this setup, we obtain an accuracy of 79.81%. We report the results in Table 3, under "Sh. timesteps".

Second, we investigate two methods for the feature extraction, that are allowed by the peculiar architecture of I3D. Specifically, thanks to the cascading layers containing max pooling, I3D shrinks the temporal dimension of the input of a factor. As each timestep is composed of 8 consecutive frames, the output feature representation has the same length as the number of timesteps. Because of this, it is possible to extract the features one timestep at a time (*One-at-a-time*) and concatenate the results on the time dimension dimension, or to feed in input all the segments together (*One-shot* fashion), without changing the output size. The difference between the two settings is given by the fact that in the *One-at-a-time* case, the modelling of one specific timestep is not affected by the neighbouring timesteps. On the other hand, in the *One-shot* way the full I3D temporal receptive field is exploited, combining local with global information.

Experiments show that the *One-shot* setting brings a noticeable improvement over *One-at-a-time* features. Intuitively, considering the context in which timesteps are placed, helps achieve a better feature representation. The results from these two setting, respectively, are 89.84% versus 83.7%, as shown in Table 3.

The variability in length of Breakfast videos, also within the same class, makes it challenging to represent all the videos fairly in a fixed-length vector. To this extent, short videos are well represented by $T = 64$ timesteps, but this amount of timesteps might not be enough to cover all the unit-actions in longer videos. Other than the uniform and random 8-frame timestep selection evaluated in previous work (Hussein et al., 2019b), we experiment with 512 equally spaced frames (*One-shot + 512 f.*) in Table 3. Despite achieving slightly better performance in validation (Table 2), the *One-shot + 512 f.* setup results in lower accuracy on the test set. This is probably due to the fact that sampling equidistant frames introduces variable frame frequency in the I3D input. Opposite to this, when sampling timesteps instead of frames, the frequency within each timestep is fixed, as all the videos have the same frame rate. The variable frame frequency alters the motion dynamics modeled by I3D, making the learning process harder.

The last experiment with regards to the temporal dimension is about Temporal Attention, used as an alternative of spatial Region Attention or in conjunction with it. As shown in Table 3, applying TA and TRA (combined Temporal-Region Attention, as described in Section 3) on top of the convolutional backbone does not result in interesting improvements. Apparently, I3D itself learns sufficiently strong fine-grained and long-term temporal patterns in the fine-tuning phase, thus making Temporal Attention superfluous. On the other side, it is interesting to note that without fine-tuning I3D, the best performances are given by the combination of Temporal and Region Attention. All the above results are summarised in Table 3.

Table 3: Ablation results considering the temporal axis. Table sections from the top: i. Region Attention (RA), Temporal Attention (TA), Temporal-Region Attention (TRA) on top of not fine-tuned I3D; ii. RA/TA/TRA on top of fine-tuned I3D; iii. same of ii. with different input settings.

I3D setting	**T**	**Acc.**	**Top**	**Acc.**
Not fine-tuned	64	58.88	TA	65.94
			RA	69.59
			TRA	71.53
One-shot	64	**82.97**	TA	84.67
			RA	**89.84**
			TRA	86.62
Sh. timesteps	64	73.97	RA	79.81
One-at-a-time	64	77.62	RA	83.70
One-shot	512 f.	80.05	RA	86.86

4.3.2 Spatial Dimension

Having discussed the experiments on the temporal axis, we now analyse the spatial dimension. In the following experiments we compare our full model with two model variations: *i.* a simple *Region Mean* model processes 11 video regions and computes a compact representation by taking the arithmetic mean of the features, neglecting the variable importance of the video regions; *ii.* the multi-regional fine-tuning strategy for I3D is replaced with a single region, that corresponds to the person-centered crop in each training video. To this end, we exploit the Actor-Focus mechanism described in Section 3.

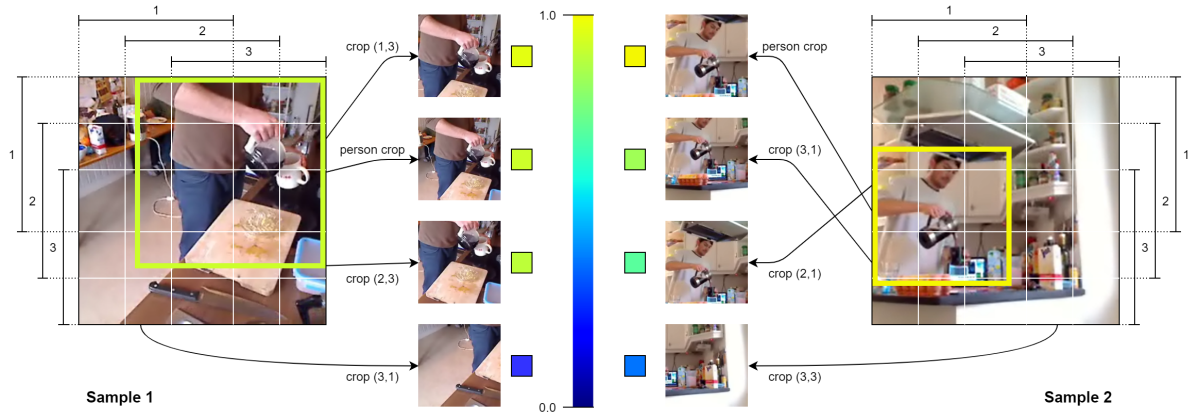

Figure 5: Visualization of the different scores that the Region Attention module attributes to the video regions. The four regions that are visualized correspond, respectively, to the top-3 and last crops, for 2 samples of the activity "preparing coffee". The coloured square in each frame represents the Actor-Focus region. The RA module sets higher scores for the person-centered and grid-central crops.

The first setting aims to show the improvements brought by RA scoring mechanism. Without the weighted average, the drop in accuracy is around 1.76%, as shown in Table 4 (*Region Mean* versus *RA*). Secondly, the comparison with the one region I3D fine-tuning proves the benefit of the multi-regional setup. In fact, training the network with multiple region crops from the same videos acts as a convenient data augmentation strategy. In addition, this learning process produces spatio-temporal features that are more representative of what the following Region Attention module expects as input. When fine-tuning the backbone only with the Actor-Focus crop, the accuracy is 86.62%, with a drop of 3.22% compared to the Multi-Regional setup, as shown in Table 4.

Figure 5 provides a visualization of the variable importance scores attributed to different video regions through the attention mechanism. According to the prior assumption that the regions of interest for activity recognition revolve around the actor performing the action, RA assigns the highest scores to the person-centered and central grid crops. On the contrary, background regions such as lower and "corner" crops score weights that are close to zero.

Finally, we measure the benefit brought by the Actor-Focus mechanism. The model is trained with and without the Actor-Focus crop. The inclusion of the latter region appears to have a huge impact in the action recognition performance, that increases from 86.62% (*MR I3D* setting in Table 4) to the final result of 89.84%.

4.3.3 I3D Fine-Tuning

The extensive experimental comparison between current state-of-the-art methods, is partially limited by the lack of hardware resources. In all the above experiments, I3D is fine-tuned only in the last three convolutional layers and only one region at a time is fed for each video. We leave the end-to-end training of the full Multi-Regional I3D + RA for future work. However, the classification accuracies achieved when fine-tuning the last three layers of I3D or the full model are nearly equal. Respectively, these correspond to 80.05% and 80.64% (Hussein et al., 2020a). As the difference is not significant, we do not expect substantial improvements with a full fine-tuning.

Table 4: Ablation results considering the spatial axis. Table sections from the top: i. different I3D fine-tuning settings and Region Attention (RA); ii. best I3D model with Region Mean or RA; iii. former state-of-the-art results on Breakfast. Note: "MR I3D" indicates Multi-Regional fine-tuning on 10 regions (no person-centered region), while "AF I3D" indicates fine-tuning only on person-centered region. R specifies the number of regions. The RA setting is intended to be placed on top of the respective I3D setting.

Backbone	R	Acc.	RA setting	Acc.
I3D not f.t.	1	58.88	RA	72.02
I3D	1	80.05	RA	83.45
MR I3D	10	81.02	RA	86.62
AF I3D	1	81.75	RA	86.62
AF MR I3D	11	**82.97**	Region mean	88.08
			RA	**89.84**
I3D full f.t.	1	80.64	ActionVlad	82.67
			Nonlocal	83.79
			Timeception	86.93
			PIC	89.84

5 CONCLUSIONS

We introduce Multi-Regional I3D fine-tuning with Actor-Focused Region Attention, a neural framework

dedicated to the spatio-temporal modelling of long-term activities in videos. We show that the model can learn long-term dependencies across timesteps, resulting in robust representations, and that it is not possible to accurately classify long activities from a few timesteps only. We give insights on the amount of timesteps, their order and the importance of the frame frequency. Next, a Region Attention module supports spatio-temporal data to adaptively learn the importance of the spatial cues in different video regions, which also allow the backbone to learn rich feature representations. Lastly, an Actor-Focus mechanism drives the attention on the truly discriminative video regions where the actor is performing the activity, neglecting background and irrelevant regions. We demonstrate the effectiveness of the architecture, benchmarking our model on the Breakfast Actions Dataset, with a SOTA-matching accuracy of 89.84%. Because of the modularity of our architecture and of related work (Hussein et al., 2019a; Hussein et al., 2019b; Hussein et al., 2020a), our framework could complement other approaches. Due to the fact that the strength of our model relies on the way the backbone is fine-tuned and on the use of attention to account for the spatial dimension, further modelling of the time dimension could improve the results. Both PIC (Hussein et al., 2020a) and Timeception (Hussein et al., 2019a) successfully exploit the time axis and can be juxtaposed on existing backbones, integrated with our RA module. Experiments are left for future work. Finally, future work may include studies on the full I3D fine-tuning and on a I3D + Region Attention end-to-end training.

ACKNOWLEDGEMENTS

This work is supported by the Early Research Program Hybrid AI, and by the research programme Perspectief EDL with project number P16-25 project 3, which is financed by the Dutch Research Council (NWO), Applied and Engineering Sciences (TTW).

REFERENCES

Burghouts, G. J. and Schutte, K. (2013). Spatio-temporal layout of human actions for improved bag-of-words action detection. In *Pattern Recognition Letters*.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

Girdhar, R., Ramanan, D., Gupta, A., et al. (2017). Actionvlad: Learning spatio-temporal aggregation for action classification. In *CVPR*.

Herzig, R., Levi, E., Xu, H., et al. (2019). Spatio-temporal action graph networks. In *ICCV*.

Hussein, N., Gavves, E., and Smeulders, A. W. M. (2019a). Timeception for complex action recognition. In *CVPR*.

Hussein, N., Gavves, E., and Smeulders, A. W. M. (2019b). Videograph: Recognizing minutes-long human activities in videos. In *arXiv*.

Hussein, N., Gavves, E., and Smeulders, A. W. M. (2020a). Pic: Permutation invariant convolution for recognizing long-range activities. In *arXiv*.

Hussein, N., Jain, M., and Bejnordi, B. E. (2020b). Timegate: Conditional gating of segments in long-range activities. In *arXiv*.

Kalfaoglu, M., Kalkan, S., and Alatan, A. (2020). Late temporal modeling in 3d cnn architectures with bert for action recognition. In *arXiv*.

Kuehne, H., Arslan, A., and Serre, T. (2014). The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*.

Qiu, Z., Yao, T., Ngo, C., et al. (2019). Learning spatio-temporal representation with local and global diffusion. In *CVPR*.

Schindler, K. and Gool, L. V. (2008). Action snippets: How many frames does human action recognition require? In *IEEE Conference on Computer Vision and Pattern Recognition*.

Sigurdsson, G. A., Varol, G., Wang, X., et al. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*.

Varol, G., Laptev, I., and Schmid, C. (2017). Long-term temporal convolutions for action recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, K., Peng, X., Yang, J., et al. (2020). Region attention networks for pose and occlusion robust facial expression recognition. In *IEEE Transactions on Image Processing*.

Wang, X., Girshick, R., Gupta, A., et al. (2018). Non-local neural networks. In *CVPR*.

Wang, X. and Gupta, A. (2018). Videos as space-time region graphs. In *ECCV*.

Wu, C. Y., Feichtenhofer, C., Fan, H., et al. (2019a). Long-term feature banks for detailed video understanding. In *CVPR*.

Wu, Y., Kirillov, A., Massa, F., et al. (2019b). Detectron2. https://github.com/facebookresearch/detectron2.

Yang, J., Ren, P., Zhang, D., et al. (2017). Neural aggregation network for video face recognition. In *CVPR*.

Yeung, S., Russakovsky, O., Jin, N., et al. (2018). Every moment counts: Dense detailed labeling of actions in complex videos. In *International Journal of Computer Vision*.

Zhou, B., Andonian, A., Oliva, A., et al. (2018). Temporal relational reasoning in videos. In *ECCV*.