Learning to Represent Spatio-Temporal Features for Fine Grained Action Recognition

Kaustubh Sakhalkar
STARS, INRIA
Sophia Antipolis, 2004 Rte des Lucioles, 06902,
Valbonne, France
https://sakhalkark.github.io/

François Brémond
STARS, INRIA
Sophia Antipolis, 2004 Rte des Lucioles, 06902,
Valbonne, France
francois.bremond@inria.fr

***Abstract*—Convolutional neural networks have pushed the boundaries of action recognition in videos, especially with the introduction of 3D convolutions. But it is an open ended question on how efficiently a 3D CNN can model temporal information? which we try to investigate and introduce a new optical flow representation to improve the motion stream. We use the baseline inflated 3D CNN networks and separate the convolutional filters into spatial and temporal, which reduces the number of parameters with minimal loss of accuracy. We evaluate our approach on NTU RGBD dataset which is the largest human action dataset and outperform the state-of-the-art by a large margin.**

***Index Terms*—action recognition, 3D convolutions, optical flow.**

I. Introduction

Research in video action recognition has gained significant traction in the last couple of years with applications in surveillance, robotics, HCI, healthcare and autonomous driving. Particularly human action recognition in daily living actions is challenging for the following reasons, (1) Similarity in background (2) Some actions are similar in pose but differ only in the context of the object used (3) Temporally "opposite" actions are harder to distinguish eg. stacking/unstacking pairs (wearing/removing shoes) (4) Discernable motion is less in certain activities.

Human actions can be broadly divided into daily living actions and outdoor actions. Daily living actions are essential for our activities and we perform them frequently throughout the day e.g. brushing, wearing clothes, reading, writing etc. Outdoor actions which we do while we are outside are relatively easier to distinguish because the scene is dynamic. For example cricket, tennis, abseiling etc. the background of the scene can give us information on the activity even from a single image. Example datasets for Outdoor actions are Kinetics [1], UCF101 [2], HMDB [3], etc

In our work we focus on daily living actions which usually happen indoors. Daily living actions are the most frequently performed and have applications in assisted living for the elderly. They are harder to distinguish than outdoor actions. Example datasets for daily living actions are NTU RGBD [4], MSRDailyActivity3D [5] and STAIRS [6]. Another reason why daily living actions are harder to distinguish is because of similar poses can represent different actions e.g. eating and drinking. The only subtle hint separating these two action classes is whether the person has a tumbler or not. Other challenges in daily living action recognition are temporally opposite actions. For example stacking and unstacking pairs of actions like wearing/removing shoes or sitting down and standing up. Classifiers often confuse such action pairs.

Fig. 1. Sample images from the NTU RGBD dataset

Convolutional neural networks have become the de facto method for visual recognition problems. But they did not give significant improvement in action recognition over hand crafted techniques until a few years back. With the introduction of 2D CNN for images, it was natural to extend it to 3D CNN [7] for videos. It is still an open ended question on how efficiently 3D convolutional kernels can represent temporal information. Until now we represented temporal information using optical flow [8], [9].

We argue that raw optical flow representing x and y magnitudes of the motion is not the best representation of motion information and introduce a better optical flow representation by changing the colour space. Other attempts have also been made at representing temporal information by using colour coded channels [10] (see [11] [12] for similar ideas).

II. State of the Art

Handcrafted features were popular earlier and the features were manually engineered (see [13] for a comprehensive review). But with the rise of deep learning we moved towards learned feature representations for images as well as videos. One obvious trick was framewise pooling of features from the last or second last layer of a 2D CNN [14] [8]. Another attempt was made by using LSTM's after 2D CNN [15] [16] instead of framewise pooling and used a categorical cross-entropy loss.

Then came two stream networks proposed by Simonyan & Zisserman [8] which represented temporal information with ten frames of optical flow and used a class fusion to combine it with RGB. In [9] authors extend the two stream network by fusing the RGB and optical flow streams. Pose based CNN [17] modelled information from five different patches of the image taking cues from different body part in the image and use pooling for aggregation.

3D CNN was an extension to 2D CNN's and were expected to model spatial and temporal information since they had filters in 3D [7] [18]. C3D had less number of layers but more number of parameters than comparable 2D CNN of equivalent depth. Significant improvement was made with I3D [19] which introduced the idea of using 2D kernels for 3D action recognition. [20] aggregate activations of 3D CNN into descriptors based on joints which are more effective than simple descriptors.

More recent approaches were focused towards attention mechanism in actions. [21], [22] have proposed a spatio-temporal attention mechanism where they have used RNNs. This moves away from the trend of traditional soft attention mechanism. Glimpse clouds [23] uses a gated recurrent unit find the next glimpse in the sequence by using three workers. Chained multistream [24] networks make use of markov chaining from C3D networks with multimodal input of RGB, optical flow and pose. In [25] a 3D CNN with skeleton is proposed that encodes the 3D positions of the joints in space and time. Junnan Li et al [26] propose an unsupervised learning framework by extrapolating cross-view motions. Dividing aggregating net [27] propose a two level fusion for different views using a conditional random field. This strategy will be effective in cross-view evaluations of datasets. Fine to coarse net [28] simultaneously extracts spatial and temporal features of skeletons for 3D action recognition. Varol et al in [29] propose long term convolutions for action recognition and increase the temporal extents of the video. Wang [30] have extended [8] by 3D convolutions and use videos of undefined length to accommodate the variance in length of video clips. Luvizon in [31] have modeled an architecture for both image and video action recognition. [32] claim that the correct temporal order is not necessary for action recognition for datasets like kinetics and UCF101 but it is important in fine grained action recognition. Their work proposes a CNN with gated recurrent units for modelling temporal information. A more comprehensive understanding of temporal representation can be found in [33]. Authors in [34] propose spatio temporal pyramids for video action recognition which is similar to [35].

Non local neural networks [36] introduce a new block which takes the weighted sum of all features. This block can be plugged anywhere in the network.

III. Limitations of Current Approaches

Handcrafted features failed to scale up to large datasets and were prone to errors due to changes in illumination and contrast. Pooling of framewise features from 2D CNN output ignores temporal information and may not represent all the frames. The idea of LSTM from CNN features was introduced but LSTMs are harder to regularize. 3D CNNs have a very large number of parameters which makes them require significant hardware and time for training. The accuracy of recognition from number of multiple visual cues like RGB, depth, pose etc. relies heavily on the feature fusion technique(see [24] ablation studies for variation in accuracy). An interesting way of representing temporal information was shown in [10] where they use a colour scheme from starting to the ending frame.

Skeleton data is usually inaccurate and is not always available due to occlusions. Moreover two or more actions can represent the same skeletal poses.

IV. Proposed Method

A. Separating the 3D convolutions

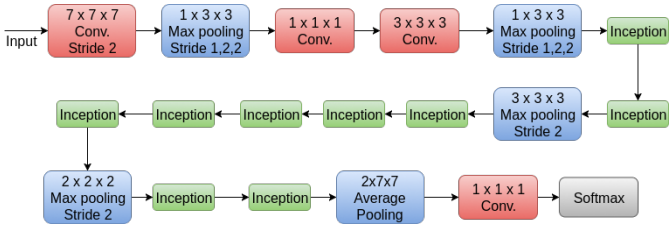

Fig. 2. The i3d network [19]

We use I3D [19] network as our baseline approach which inflates the 2D convolutions to 3D by replicating the kernels. The network was pretrained on Kinetics [1] dataset. The network uses the original 2D inception weights from ImageNet and converts the filters to 3D using the same weights. Then the network was trained with the kinetics dataset [1]. This pretraining has shown to yield a significant improvement in accuracy.The I3D network starts with downsampling the input dimension for the first few layers. Then we have inception_v3 blocks with pooling at regular intervals. The inception backbone allows us to go deeper with the network and batch normalization [37] is used after every convolutional layer. We also do an average pool to bring the 3D convolutions back

to 2D before the softmax layer and use ReLu activations throughout the network. The last layer is a fully connected softmax function to obtain the prediction logits.

The network was inflated from 2D filters to 3D filters which have been pretrained on ImageNet. Such inflation has shown to improve performance of 3D CNN's. In multimodal data the fusion used plays a great role in determining the performance of the model.

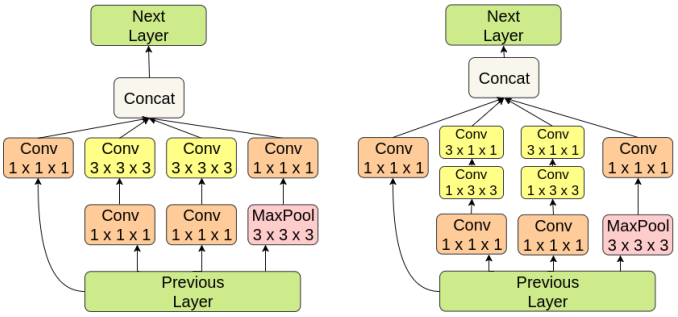

Fig. 3. a. (left) The original Inception v3 module. b. (Right) The inception module after separating the blocks into spatial and temporal

3D convolutions have more parameters than 2D convolutions. Hence we separate the convolutional blocks into 2D and 1D convolutions to mimic spatial and temporal convolutions. Separating the $3 \times 3 \times 3$ block into $3 \times 3 \times 1$ and $1 \times 1 \times 3$ reduces the number of parameters [38].

B. Optical flow with magnitude and orientation

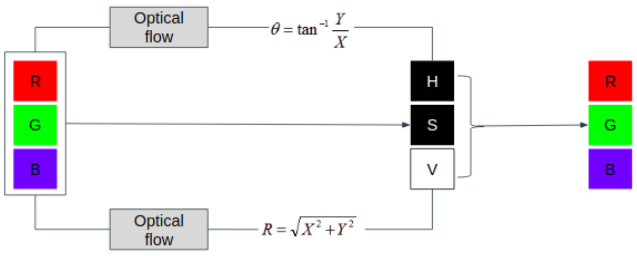

Fig. 4. Optical flow representation framework

Optical flow represents short term temporal information and was introduced in two-stream network [8]. We revisit this optical flow representation and use the traditional TVL optical flow. We then convert the x and y flows into magnitude and orientation which gives us more information about the pixel motion rather than plain optical flow[2]. Then we move the RGB image to HSV color space and use the saturation value along with the magnitude and orientation which is converted to RGB again. We compute magnitude as $M = \sqrt{x^2 + y^2}$ and orientation as $O = tan^{-1}(y/x)$ where x and y represent flow directions (See Fig). The hue and value is replaced by orientation and magnitude respectively while saturation is kept the same using a 3 channeled representation instead of a two channel.

This optical flow representation for action recognition is inspired from how the human vision system percieves color. Instead of directly using X and Y directions of optical flow, we find the magnitude and orientation for the motion (optical flow). Then we use them as hue(orientation) and value(magnitude) which is similar to 3D HSV representation (HSV is represented in 3D by radius, angle and height). Finally we convert it back to RGB to mimic the human vision system. This allows us to preserve the motion properties and as well as have a similar representation for temporal and spatial features.

C. Frame selection in videos

The frame selection plays an important role in action recognition heavily influences the accuracy as seen by our results. Although ablation studies are already done with changing the gap between every selected frame and the best results were obtained when frames are skipped with a stride of 2. In our work we perform ablation studies with respect to the starting point with sample size (64, 224, 224, 3) which is the standard with I3D. We first choose 64 frames from the starting frame consecutively, this fixes the starting point to the first frame. Then we choose 64 frames randomly from the video but the order is maintained. In this case the sample spans the whole video. Thirdly we divide the video into 64 equal parts and choose one frame randomly from each part. In the last method we choose the starting point randomly and choose 64 consecutive frames with an interval of 2 frames. The first two methods have starting points very close to the first frame and the third method and fourth method will have random starting points.

V. Experiments

NTU RGB+D Dataset [4] is the largest human action dataset so far with 56,880 sequences. The skeleton data was collected with a Microsoft Kinect. Each skeleton contains 25 human joints. In this dataset, there are 60 action classes of three types viz. daily actions, health-related actions, and interactive actions. All the actions are performed by 40 distinct subjects. The actions are recorded simultaneously by three camera sensors located at different angles at a separation of 45 degrees This dataset is challenging due to the large variations of viewpoints and sequence also the large amount of videos. We use the train and test splits for cross subject evaluation as mentioned in the original paper. We adhere to the standard two stream model using I3D networks [19]. We train on 4 GPUs Nvidia 1080Ti with an initial learning rate of 0.01 instead of 0.1 that was used originally to train on kinetics dataset and use a mini-batch size of 2 per GPU with SGD optimization. We do only cross subject evaluations since cross-view is out of scope of our work.

VI. Results and Discussion

Daily living actions are more challenging to distinguish than outdoor actions due to high intra-class and low inter-class variations and most of the previous approaches have used multimodal data for classification as show in section 2. We

address this problem by introducing spatio-temporal separation in the same network and improving optical flow for improving classification of activities where discernable motion is less.

Frame selection strategy (I3D RGB training)	Accuracy on NTU dataset
Choose the first 64 frames	81.98%
Choose 64 frames randomly	82.07%
Choose 64 frames at equal intervals (randomly in interval)	83.97%
Choose the starting point randomly with stride 2	89.08%

TABLE I
FRAME SELECTION STRATEGIES AND THEIR RESULTS

Randomly selecting the starting point in the video and choosing consecutive frames has proven to be the best strategy. We attribute its success to multiple random points chosen during training and during testing we choose 5 different starting points in the same video and average the logits from the predictions. Choosing the first 64 frames does not see the whole action in the video leading us to the poorest performance. Choosing 64 frames randomly can have irregular distributions but the starting point of the clip is very close to the starting point of the video, hence the network will see nearly the same clip every time.

Method	Accuracy
F2CSkeleton [28]	79.6%
Chained Mutlistream [24]	80.8%
Glimpse clouds [23]	86.6%
Dividing Aggregating Net [27]	88.12%
2D3D [31]	85.5%
TSN [39]	84.93%
Ours	92.67%

TABLE II
COMPARISON WITH THE STATE OF THE ART ON NTU DATASET [4]

Method	Accuracy on NTU dataset
I3D (RGB)	89.08%
I3D (OF)	81.17%
I3D (MOF)	82.97%
I2.5D (RGB)	84.87%
I2.5D (OF)	78.75%
I2.5D (MOF)	80.45%
I2.5D (RGB + OF)	86.41%
I2.5D (RGB + MOF)	86.89%
I3D (RGB + OF)	91.23%
I3D (RGB + MOF)	92.67%

TABLE III
I3D REPRESENTS THE SAME NETWORK DESCRIBED IN [19]. I2.5D IS WHEN THE INCEPTION BLOCK IS REPLACED WITH SEPARABLE CONVOLUTIONS. OF IS THE VANILLA TVL OPTICAL FLOW AND MOF REPRESENTS THE IMPROVED REPRESENTATION AS DESCRIBED.

The best result is obtained by using vanilla I3D with improved optical flow but the training was faster on I2.5D (time reduced by 25%) and the number of parameters reduced from 12M to 8M. The loss in accuracy was less compared to the gain in time and memory efficiency.

VII. CONCLUSIONS

We present a better and faster representations for 3D convolutions and find scope for more improvement in 3D CNN. Separating the blocks into spatial and temporal has proved to be beneficial. Some future directions for this work are to use $1 \times 3 \times 3$ convolutions as well. Other ideas include learning temporal structure without using optical flow. Attention mechanisms have also gained significant traction in the community, and in the future we will propose efficient RNN representation akin to convolutional networks.

REFERENCES

[1] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[2] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[3] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[4] 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.

[5] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1290–1297.

[6] Y. Yoshikawa, J. Lin, and A. Takeuchi, "Stair actions: A video dataset of everyday home actions," *arXiv preprint arXiv:1804.04326*, 2018.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

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

[9] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.

[10] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "Potion: Pose motion representation for action recognition," in *CVPR 2018*, 2018.

[11] L. Sevilla-Lara, Y. Liao, F. Guney, V. Jampani, A. Geiger, and M. J. Black, "On the integration of optical flow and action recognition," *arXiv preprint arXiv:1712.08416*, 2017.

[12] C. A. Caetano, V. H. C. De Melo, J. A. dos Santos, and W. R. Schwartz, "Activity recognition based on a magnitude-orientation stream network," in *Graphics, Patterns and Images (SIBGRAPI), 2017 30th SIBGRAPI Conference on*. IEEE, 2017, pp. 47–54.

[13] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer vision and image understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[15] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[16] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[17] G. Chéron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3218–3226.

[18] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

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

[20] C. Cao, Y. Zhang, C. Zhang, and H. Lu, "Action recognition with joints-pooled 3d deep convolutional descriptors." in *IJCAI*, vol. 1, 2016, p. 3.

[21] F. Baradel, C. Wolf, and J. Mille, "Pose-conditioned spatio-temporal attention for human action recognition," *arXiv preprint arXiv:1703.10106*, 2017.

[22] ——, "Human action recognition: Pose-based attention draws focus to hands," in *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*. IEEE, 2017, pp. 604–613.

[23] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, "Glimpse clouds: Human activity recognition from unstructured feature points," *Computer Vision and Pattern Recognition (CVPR)(To appear)*, vol. 3, 2018.

[24] 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*. IEEE, 2017, pp. 2923–2932.

[25] J. Tu, M. Liu, and H. Liu, "Skeleton-based human action recognition using spatial temporal 3d convolutional neural networks," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[26] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Unsupervised learning of view-invariant action representations," *arXiv preprint arXiv:1809.01844*, 2018.

[27] D. Wang, W. Ouyang, W. Li, and D. Xu, "Dividing and aggregating network for multi-view action recognition," in *European Conference on Computer Vision*. Springer, 2018, pp. 457–473.

[28] T. M. Le, N. Inoue, and K. Shinoda, "A fine-to-coarse convolutional neural network for 3d human action recognition," in *British Machine Vision Conference*, 2018.

[29] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1510–1517, 2018.

[30] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-stream 3d convnet fusion for action recognition in videos with arbitrary size and length," *IEEE Transactions on Multimedia*, 2017.

[31] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2018.

[32] D. Dwibedi, P. Sermanet, and J. Tompson, "Temporal reasoning in videos using convolutional gated recurrent units," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1111–1116.

[33] D. Wei, J. J. Lim, A. Zisserman, and W. T. Freeman, "Learning and using the arrow of time," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8052–8060.

[34] Y. Wang, M. Long, J. Wang, and S. Y. Philip, "Spatiotemporal pyramid network for video action recognition." in *CVPR*, vol. 6, 2017, p. 7.

[35] P. Wang, Y. Cao, C. Shen, L. Liu, and H. T. Shen, "Temporal pyramid pooling-based convolutional neural network for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2613–2622, 2017.

[36] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," *arXiv preprint arXiv:1711.07971*, vol. 10, 2017.

[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[38] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, 2018.

[39] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 20–36.