# RECURRENT 3D CONVOLUTIONAL NETWORK FOR RODENT BEHAVIOR RECOGNITION

*Van Anh Le, Kartikeya Murari*

Electrical and Computer Engineering, Schulich of Engineering, University of Calgary, AB, Canada

## ABSTRACT

Animal, specially rodent, studies are critical in understanding human health, disease and treatments. Behavior is an important observed outcome in many such studies. Thus, quantifying rodent behaviors is key. This is typically done by trained human observers, making the process very slow and subjective. This has has led to a growing interest in developing automated assessment tools. Existing approaches commonly rely on hand-crafted features which are often obtained through a tracking process. Motivated by state of the art results in image and video analysis using deep learning, we propose a deep architecture which is a combination of recurrent network and 3D convolutional network to learn long and short-term video representations. We test the proposed solution with the dataset collected by [1] and demonstrate that our framework can obtain accuracy on par with human assessment.

***Index Terms*—** Rodent behavior recognition, 3D convnet, LSTM network

## 1. INTRODUCTION

Life scientists often use rodent models to study diseases and treatments, and commonly measure changes in behavior as a result of the progression of disease or the recovery process after treatment. At present, behavioral assessment is mainly conducted by human annotators, making the process subjective, time consuming, and labor-intensive. Therefore, an automated analysis tool for animal behavior can ameliorate those inherent limitations and allow high throughput analysis of experiments (multiple animals at one time). Behavior analysis is one of the most daunting tasks confronting machine vision researchers. Over the last two decades, a myriad of behavior recognition methods have been proposed to extract the information about activities presented in videos. However, research on behaviour recognition mainly focuses on human activities. Numerous methods have been proposed to recognize activities such as 'walking', 'waving', or 'punching' [2] whereas rodents' limbs are small in comparison with their bodies and their limbs' movements are restricted, which make behaviours non-distinctive. Besides, intra- and inter-class similarities make the problem amply challenging. This stems from the fact that activities within the same class may be expressed by different subjects with different body movements. As a result, there is often not a clear difference in posture or movement intensity between certain activities such as 'grooming' and 'eating', or 'sniffing' and 'rearing'. In the literature, a few systems have been developed to automate the phenotyping of animals in their home-cage, typically involving extracting hand-crafted features. The most common approach is to track the animal in videos by tracking their bodies [1, 3, 4, 5, 6] and other specific parts such as nose and tail [7]. From tracking analysis, features such as velocity, acceleration, zone information (distance to feeder, water tube and cage wall) and posture are generated. Efforts to build an automated system for identifying behaviors more complex than locomotion and posture have achieved impressive results recently. For instance, Dollar et al. [8] recognized mouse behavior from the classification of sparse spatio-temporal features, reaching an accuracy of 72%. In [4], a system for the recognition of social behavior of mice using top and side cameras is proposed, in which a large pool of spatio-temporal and trajectory features are generated and followed by a temporal context model. The average accuracy over 13 behavior classes was 61 percent. Similarly, [9] presents an integrated hardware and software system that uses a depth camera along with top and side cameras to extract the body pose and supervised learning algorithms to classify social behaviors. By imitating the organization of the dorsal stream of the visual cortex, which has been shown to play a role in motion processing in biological vision, Jhang et al. [1] created motion features and trajectory features to train a classifier using a Hidden Markov Model Support Vector Machine (SVMHMM). Inspired by breakthroughs in image recognition using deep learning techniques [10], several frameworks have been proposed to recognize animal behavior. Examples include using features at a frame level in a two-class problem [11, 12], and extracting frame-level features via a 2DCNN and learning temporal features with an RNN [13, 14]. Our question is whether abstract spatio-temporal features obtained from deep networks [15, 16], which have been applied successfully to recognize activities in short videos [17], are suitable for recognizing multiple behaviors in long videos, and whether algorithms can work without hand-crafted features that are dominant in most existing frameworks. We

Thanks to Alberta Innovates Technology Futures (AITF) for funding and NVIDIA corporation for donating the GPU used in this research.

propose a framework that uses a 3D Convolutional network (ConvNet) to extract short-term spatio-temporal features from overlapped short clips. Then those local features are fed to a Long Short Term Memory network to learn long-term features which are used for classification. We call our framework LSTM-3DCNN, and show how to learn local spatio-temporal behavioral features using a 3D ConvNet and recognize behaviors in long videos with an LSTM network. The proposed framework was evaluated with the dataset collected in [1].

## 2. RELATED WORK

Action recognition has been studied by the machine vision community for decades. The approaches can be sorted into two groups. The first group involves hand-designed features which typically include spatio-temporal interest points (STIPs) [18] which are obtained by extending Harris corner detectors to 3D, HOG3D [19] and SIFT-3D [20] which are the extensions of HOG [21] and SIFT [22]. The most common classifiers for such features are SVMs. The second group uses spatio-temporal features which are learned directly from the datasets. For instance, Le et al. used independent subspace analysis to learn spatio-temporal features from unannotated videos [23] and Taylor et al. proposed convolutional gated Restricted Boltzmann Machines (RBMs) [24], which can be considered as an extension of convolutional RBMs to 3D, to learn features. Learning visual representations with convolutional neural networks has shown great success on various computer vision tasks [25, 26] and outperformed hand-designed features in large-scale datasets. Extending ConvNet for spatiotemporal features has been proposed and applied to action recognition in recent works [15]. Feichtenhofer et al. constructed ST-ResNet [17] which is the combination of two-stream ConvNets [27] and Residual Networks (ResNets) [25] and Varol et al. has proposed long-term temporal convolutions (LTC) [28]. Both ST-ResNet and LCT achieved the best performance on UCF101 and HMDB51, the two popular action recognition datasets. However, those widely-used benchmark action recognition datasets are segmented short clips, as opposed to continuous videos. Donahue et al. introduced Long-term Recurrent Convolutional Networks (LRCN) [29] which allows encoding long-term temporal information because it is difficult to increase temporal extents in 3D ConvNets due to memory limitations. Our work aims to consolidate the advantages of both 3D ConvNets and RNNs in a framework to work with continuous long videos.

## 3. SINGLE MOUSE BEHAVIOR DATASET

Collected by Serre et al. [1], the dataset is a large database of videos of singly housed home-cage mice which was recorded from a side camera under varying light conditions. Eight behaviors were considered: drinking, eating, grooming, hanging, rearing, walking, resting and micro movements of the head. Examples of video frames are shown in Fig. 1A.

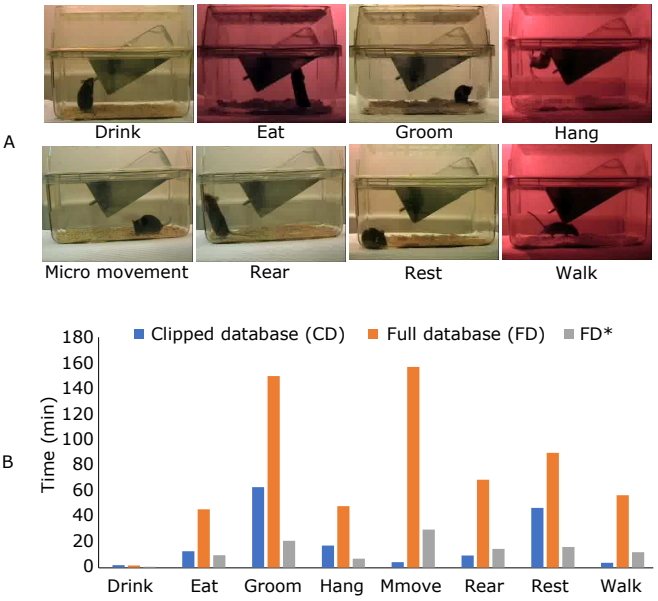


**Fig. 1**: (A) Snapshots taken from exemplary videos under varying light conditions for the eight behaviors of interest. (B) Distribution of behavior annotations for 'clipped database' (CD) and 'full database' (FD) over total time. FD* is a subset of FD annotate by both human groups.

The dataset contains two subsets, denoted 'clipped database' (CD) and 'full database' (FD). The first set includes short clips that have been human annotated with very high confidence. These are the most exemplary samples of each behavior. The second set has 12 long videos that are labeled frame-by-frame by two human annotator groups. Group 1 annotated FD in entirety, and group 2 annotated a subset of it, denoted FD*, in order to evaluate agreement between the two annotator groups. Fig. 1B shows the distribution of behavior labels for CD, FD and FD*.

## 4. DESCRIPTION OF LSTM-3DCNN

The overall architecture of our proposed framework is shown in Fig. 2. The input to the framework is a sliding sequence of frames from the long video. The deep architecture consists of three parts: a 3D ConvNet for extracting local spatio-temporal features, an LSTM network for learning long-term temporal features, and a softmax classifier to recognize behavioral classes. In the following subsections, we elaborate on how short and long-term features are learned.

### 4.1. 3D ConvNet for learning local spatiotemporal features

We simplify notations by removing the channels. Thus, input, kernel, and output are considered $L \times H \times W$ 3D tensors, where $L$ is temporal length, $H$ is height and $W$ is width. The detailed architecture of our 3D ConvNet is shown in Fig. 2. Our structure is quite similar to the C3D model [15]. How-

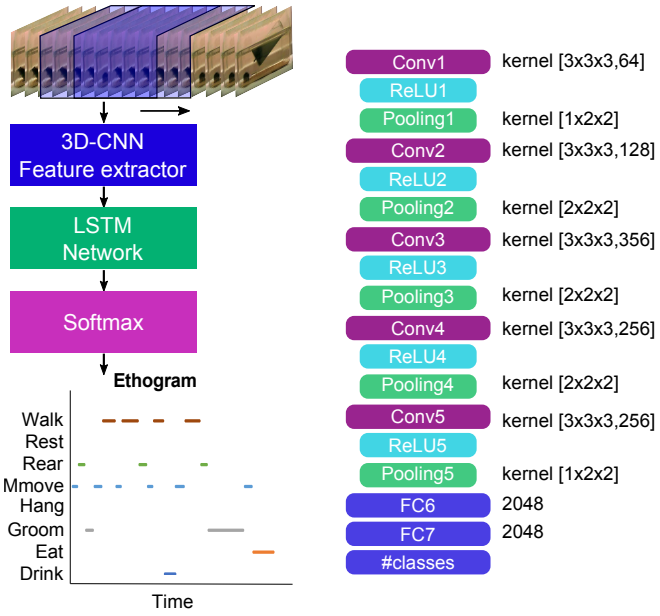


**Fig. 2**: Illustration of pipeline of proposed architecture with details of 3D Convnet architecture for learning short term spatiotemporal representations with 5 convolutional layers. Each convolutional layer is followed by a rectified linear unit (ReLU) and a max pooling layer.

ever, due to GPU memory limits, our 3D ConvNet is modified to have 5 convolutional layers, 5 pooling layers, followed by two fully connected layers and a softmax output layer. The network uses an $8 \times 128 \times 128$ input and all convolutional kernels are $3 \times 3 \times 3$ with stride $1 \times 1 \times 1$. Pooling1 and Pooling5 have kernel size of $1 \times 2 \times 2$ with stride $1 \times 2 \times 2$ while other pooling layers use $2 \times 2 \times 2$ kernels with stride $2 \times 2 \times 2$. Both fully connected layers have 2048 units.

To learn local spatio-temporal features, we train the network with the videos in 'clipped database' which consists of 4,200 clips (262,360 frames equivalent to ~2.5 h.) [1]. Each video belongs to 1 of 8 behavioral categories. As can be seen in Fig. 1B, the 'clipped database' is highly imbalanced. To address this, we augment the behaviors having fewer examples by various forms of data augmentation. We randomly adjust image brightness and contrast, and split videos of the under-represented behaviors into more densely overlapped clips. Therefore, the final training set becomes balanced. For training, we resize input clips to $8 \times 128 \times 128$. The 3D Convnet was trained from scratch with 'clipped database' with a batch size of 32 examples and an initial learning rate of 0.002. The learning rate was divided by 10 after every 3 epochs. The training was terminated after 15 epochs and yields ~95% accuracy on the test split.

After training, the 3D ConvNet is used as a feature extractor. We split 12 long videos in the 'full database' into overlapped clips. These clips are passed to the network to

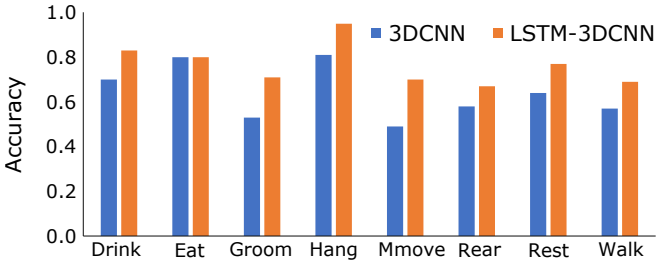


**Fig. 3**: Improved classification with long-term temporal features

extract `fc6` activations which are then L2-normalized to become 2048-dim descriptors. If we consider each clip as a volume $\mathbb{R}^{L \times H \times W}$, the C3D network transforms it into a feature vector $\mathbb{R}^{2048}$.

### 4.2. LSTM network for learning long-term features

In order to capture long-term sequential information, we build a recurrent network with LSTM cells [30]. Our architecture consists of three stacked LSTM layers, each with 512 memory cells, and then a softmax classifier which outputs the probability of the behavior. The LSTM network was trained with long continuous videos from the 'full database' to learn long-term spatiotemporal features. Each video is split into a sequence of 8-frame clips with 6 frames overlapped between two consecutive clips. Those clips are then passed to the 3D Convnet which was previously trained successfully with the 'clipped database' to be transformed into a sequence of features before being input to the LSTM model. Again, an imbalanced dataset is a challenging issue in this stage (Fig. 1B). To cope with this problem, we duplicate training samples of poorly represented behavioral categories.

## 5. EXPERIMENTAL RESULTS

The system was implemented with Tensorflow. The proposed networks were trained with one NVIDIA TITAN X GPU. To evaluate the accuracy of the system, we trained and tested it in the same way as the original paper [1]. We use a leave-one-video-out procedure, in which we trained the system with 11 videos and evaluate on the remaining video. The procedure was repeated 12 times for all videos and the results averaged. The system was trained and tested based on the annotations done by Annotator group 1. To investigate the effects of learning long-term features, we removed the LSTM network from our system and used only the 3D ConvNet to classify each indivisual clip segmented from the 'full database'. We compare the accuracy of classification between 3DCNN and LSTM-3DCNN in Fig. 3. Classification accuracy of all behaviors, except eating, was significantly improved.

Fig. 4 shows the confusion matrices to measure agreement between our framework and Annotator group 1, between the system proposed in [1] and Annotator group 1,

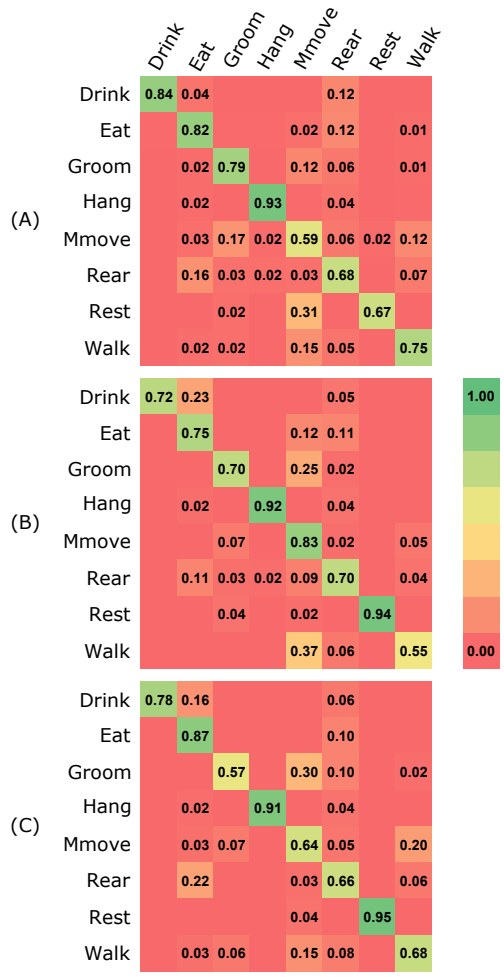


**Fig. 4**: Confusion matrices for comparing the agreement between (A) our proposed framework and human scoring (B) the system proposed by [1] and human scoring and (C) human to human scoring [1]

and between Annotator group 2 and Annotator group 1. All three were evaluated on FD*, annotated by both the human groups. We also compared overall accuracy over all frames and over all behaviors in FD. The data is summarized in Table 1. Compared to the previously described system that used handcrafted features [1], our system without handcrafted features did better with behaviors drink, eat, groom, and walk. However, it performed worse for rest and micro movement. Since these behaviors accounted for ∼40% of FD, though our behavior level performance is similar (75.9% vs. 76.4%), the frame-level performance is worse (71.2% vs. 77.3%). This was also true for a comparison over the entire 'full database'. Compared to the human annotators, our system achieved similar results at both the behavior level (75.9% vs. 75.7%) and the frame level (71.2% vs. 71.6%). In Fig. 5, the sequence of behavior generated automatically by our system is compared with human assessment over 5 minutes. The total time for each behavior is shown in the right columns.

| | This work | Jhuang et al. [1] | Human |
|---|---|---|---|
| FD* over behaviors | 75.9% | 76.4% | 75.7% |
| FD* over frames | 71.2% | 77.3% | 71.6% |
| FD over behaviors | 76.5% | 77.1% | |
| FD over frames | 73.5% | 78.3% | |

**Table 1**: Behavior recognition accuracies across behaviors and frames computed over the 'full database' (FD) and the subset annotated by both human groups (FD*).

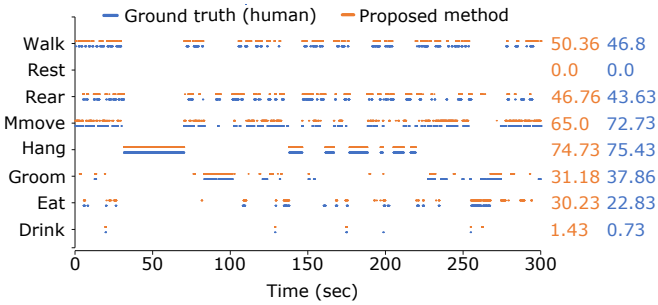


**Fig. 5**: Comparison between ground truth (Annotator group 1) and our system over 5 min of testing video.

## 6. CONCLUSION

We presented a deep framework for recognizing mouse behaviors. We described how to learn local behavioral representations and integrate those features further to learn global features. We tested our system on a large-scale dataset collected by [1] containing multiple rodent behaviors. We showed that long-term temporal features can significantly improve the performance, and our system, without using any hand-crafted features, can achieve a performance comparable to human assessment and state of the art automated assessment. We believe that combining deep features with hand-designed features can further automate rodent behavior recognition.

## 7. REFERENCES

[1] Hueihan Jhuang, Estibaliz Garrote, Xinlin Yu, Vinita Khilnani, Tomaso Poggio, Andrew D Steele, and Thomas Serre, "Automated home-cage behavioural phenotyping of mice," *Nature Communications*, vol. 1, pp. 68, 2010.

[2] Jake K Aggarwal and Michael S Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, pp. 16, 2011.

[3] Mayank Kabra, Alice A Robie, Marta Rivera-Alba, Steven Branson, and Kristin Branson, "Jaaba: interactive machine learning for automatic annotation of animal behavior," *Nature Methods*, vol. 10, no. 1, pp. 64, 2013.

[4] Xavier P Burgos-Artizzu, Piotr Dollár, Dayu Lin, David J Anderson, and Pietro Perona, "Social behavior recognition in

continuous video," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1322–1329.

[5] Malte Lorbach, Ronald Poppe, Elsbeth A van Dam, Lucas PJJ Noldus, and Remco C Veltkamp, "Automated recognition of social behavior in rats: The role of feature quality," in *International Conference on Image Analysis and Processing*. Springer, 2015, pp. 565–574.

[6] Elsbeth A van Dam, Johanneke E van der Harst, Cajo JF ter Braak, Ruud AJ Tegelenbosch, Berry M Spruijt, and Lucas PJJ Noldus, "An automated system for the recognition of various specific rat behaviours," *Journal of Neuroscience Methods*, vol. 218, no. 2, pp. 214–224, 2013.

[7] Malte Lorbach, Elisavet I Kyriakou, Ronald Poppe, Elsbeth A van Dam, Lucas PJJ Noldus, and Remco C Veltkamp, "Learning to recognize rat social behavior: Novel dataset and cross-dataset application," *Journal of Neuroscience Methods*, 2017.

[8] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 65–72.

[9] Weizhe Hong, Ann Kennedy, Xavier P Burgos-Artizzu, Moriel Zelikowsky, Santiago G Navonne, Pietro Perona, and David J Anderson, "Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning," *Proceedings of the National Academy of Sciences*, vol. 112, no. 38, pp. E5351–E5360, 2015.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[11] Ulrich Stern, Ruo He, and Chung-Hui Yang, "Analyzing animal behavior via classifying each video frame using convolutional neural networks," *Scientific Reports*, vol. 5, pp. 14351, 2015.

[12] Zhongzheng Ren, Adriana Noronha Annie, Vogel Ciernia, and Yong Jae Lee, "Who moved my cheese? automatic annotation of rodent behaviors with convolutional neural networks," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE, 2017, pp. 1277–1286.

[13] Zachary Nado, "Deep recurrent and convolutional neural networks for automated behavior classification," *Undergraduate Thesis*. Brown University, 2016.

[14] Gregory Kramida, Yiannis Aloimonos, Chethan Parameshwara, Cornelia Fermüller, Nikolas Francis, and Patrick Kanold, "Automated mouse behavior recognition using VGG features and LSTM networks," in *Visual observation and analysis of vertebrate and insect behavior workshop*. Cancun, 2016, pp. 1–3.

[15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4489–4497.

[16] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri, "Convnet architecture search for spatiotemporal feature learning," *arXiv preprint arXiv:1708.05038*, 2017.

[17] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes, "Spatiotemporal residual networks for video action recognition," in *Advances in Neural Information Processing Systems*, 2016, pp. 3468–3476.

[18] Ivan Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[19] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association, 2008, pp. 275–1.

[20] Paul Scovanner, Saad Ali, and Mubarak Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM international conference on Multimedia*. ACM, 2007, pp. 357–360.

[21] Navneet Dalal, Bill Triggs, and Cordelia Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision*. Springer, 2006, pp. 428–441.

[22] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[23] Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3361–3368.

[24] Graham W Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler, "Convolutional learning of spatio-temporal features," in *European Conference on Computer Vision*. Springer, 2010, pp. 140–153.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[26] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[27] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[28] Gul Varol, Ivan Laptev, and Cordelia Schmid, "Long-term temporal convolutions for action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[29] Jeffrey Donahue, Lisa Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor 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.

[30] Felix A Gers, Nicol N Schraudolph, and Jürgen Schmidhuber, "Learning precise timing with LSTM recurrent networks," *Journal of Machine Learning Research*, vol. 3, no. Aug, pp. 115–143, 2002.