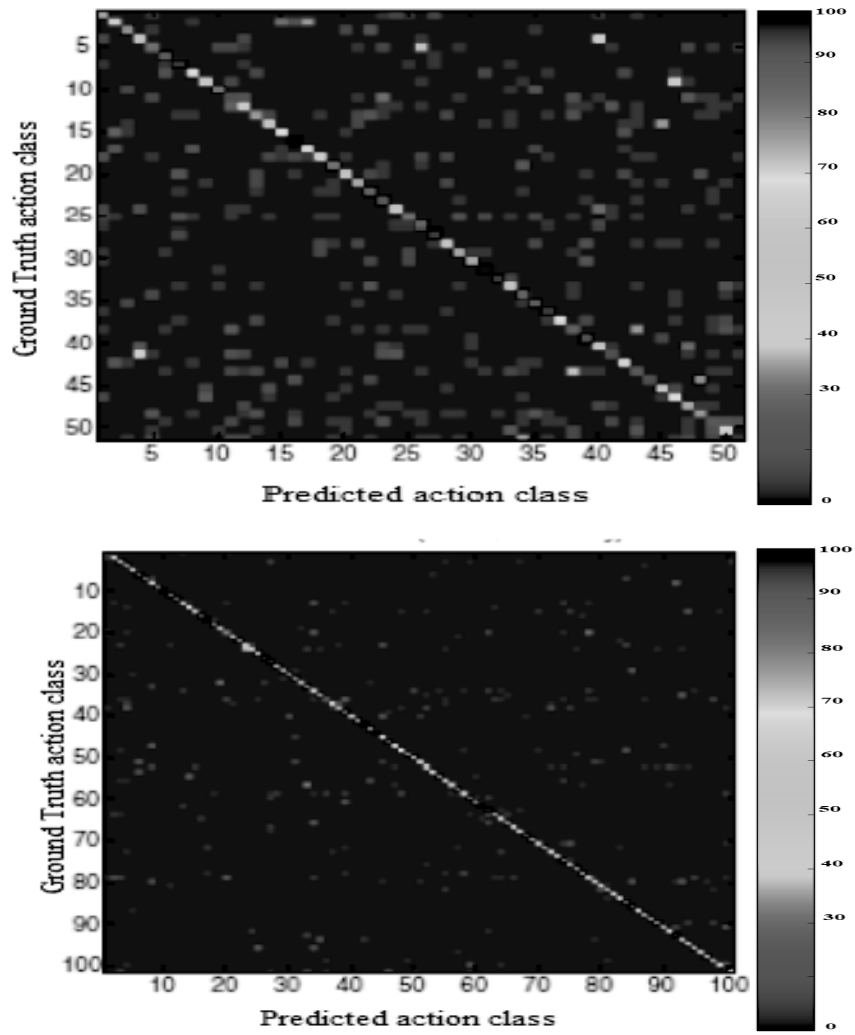

Figure 6. Confusion matrices of the proposed model on HMDB51 and UCF-101 datasets.

Confusion matrixes and recognition visualization

The confusion matrixes for our introduced system on both HMDB51 and UCF-101 datasets are shown in Figure 6. For UCF-101 and HMDB51, we consider all 101 and 51 action categories respectively. Due to the large collection of action classes in these datasets and limited page space, it is not possible to present the confusion matrixes in tabular form. The accuracies in the diagonal cells are indicated by different colors and red cells show the 100% accuracy achieved for the particular action class. The confusion matrix shows the relationship between the classified action class and the ground-truth class. The confusion matrix for the UCF-101 dataset is well diagonalized, where the diagonal portion gives high intensity as recognition accuracy for each action class, and extremely few categories are mixed up when classifying. However, it can be observed from the confusion matrix that some action classes are interfering and misclassifying each other and reporting low scores. The possible reasons for interfering and misclassification are the motion similarity in actions or the same background, objects and scene which shows a similar appearance and motion-based features. For example, action categories throwing and swinging baseball have a similar type of motions as object position is over the head and throwing it away. Motion in both categories produce similar motion features, so the correct action classification is extremely confusing and difficult. Also, two other categories shooting the ball

and dribbling performed in the basketball court and have similar objects and backgrounds so there is a possibility of generation of similar appearance-based features. However, from both given figures, we can observe that the diagonal accuracies dominate in most of the columns, which confirms the good performance of our model.

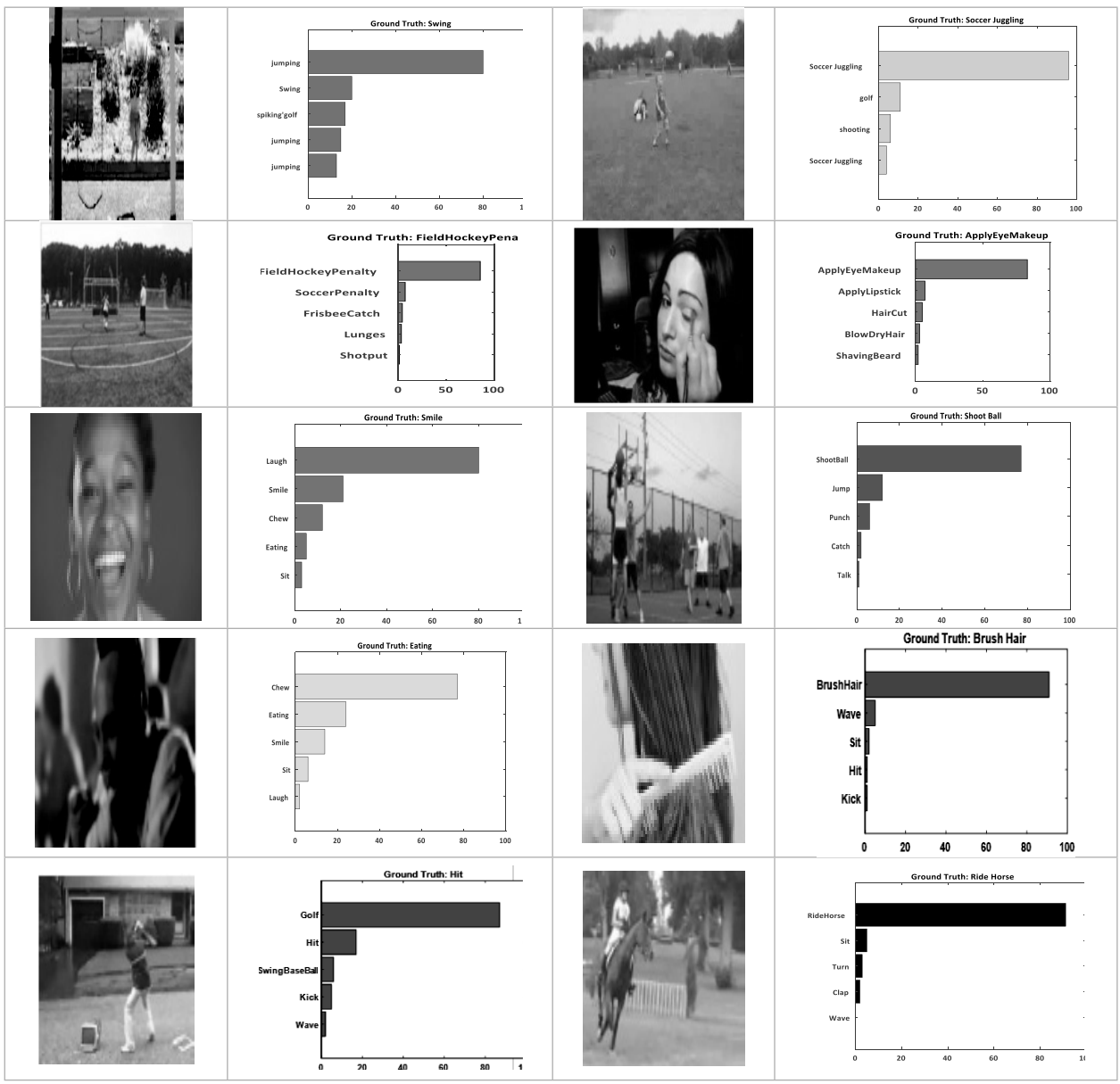

Figure 7. Failed and successful predictions on HMDB51 and UCF-101 datasets.

Furthermore, we also give recognition visualizations on video frames from two standard datasets for a better understanding of our proposed model. The proposed approach is tested on 30% videos of both UCF-101 and HMDB51 datasets. Some of the examples of intermediate frames of action categories along with correct and failed visual recognition results are given in Figure 7. In this figure, two initial rows show some successful and failed results from UCF 101 dataset and the last three rows indicate the successful and failed results from the HMDB51 dataset. Our models receive test video stream as input and C3D captures spatial and motion features which are given

input to the bidirectional LSTM in small segments for time interval *T*. The bidirectional LSTM identify the temporal dynamics and generates output for each segment and then the video action is predicted for the highest frequency class for obtained outputs. In Figure 7, we can see that some action classes are interfering each other, where "Hit" is classified as "Golf", "Eating" is classified as "Chew" "Smile" is predicted as "Laugh". These actions are taken from the HMDB51 dataset, at which our method cannot discriminate fined-grained examples. The possible reason is that the HMDB-51 dataset is not that large in terms of both diversity and scale and these failed predictions are because of the similarity in the camera motion, visual content, appearances, and changes in subject body parts in both confused action classes. While in the case of UCF-101 "Swing" is classified as "Jumping" but most of the action categories are correctly classified. UCF-101 dataset is a comparatively large dataset than HMDB51 for training which can recognize fine-grained examples. Thus, from the qualitative examples, we conclude that our proposed framework can obtain impressive performance in practice.

Table 4. Comparison with the state-of-the-art methods on UCF-101 and HMDB51 datasets.

Features	State of the art approach	Year	UCF101	HMDB51	Combined
Traditional	iDT-FV	2013	84.7%	57.2%	70.9%
	Ordered Trajectories	2013	72.8%	47.3%	60.0%
	MPR	2015	-	65.5%	65.5%
	MoFAP	2016	88.3%	61.7%	75.0%
	Trajectory Rejection	2017	85.7%	58.1%	71.9%
DCNN	Two-stream ConvNets	2016	88.0%	59.4%	73.7%
	FSTCN	2015	88.1%	59.1%	73.6%
	C3D	2015	85.2%	-	-
	EMV-CNN	2016	86.4%	-	-
	DANN	2016	89.2%	63.3%	76.2%
	Dynamic Images	2016	89.1%	65.2%	77.1%
	3D Convolution	2013	91.8%	64.6%	78.2
	FCNs-16	2017	90.5%	63.4%	76.9%
	LTC-CNN	2018	92.7%	67.2%	79.9%
RNN	LSTM	2015	88.6%	-	-
	LRCN	2017	82.9%	-	-
	VideoLSTM	2018	89.2%	56.4%	72.8%
	STPP-LSTM	2017	91.6%	69.0%	80.3%
	Hidden-Two-Stream	2018	90.3%	58.9%	
	RMDN	2017	82.8%	-	-
	L2STM	2017	93.6%	66.2%	79.9%
	TS-LSTM	2019	94.1%	69.0%	81.5%
	Multi-LSTM	2018	90.8%	-	-
	RSTAN [54]	2018	**94.6**%	70.5%	82.55%
Hybrid Model	TDD-iDT+FV	2015	91.5%	65.9%	78.7%
	C3D-iDT	2015	90.4%	-	-
	TSN	2016	94.2%	69.4%	81.8%
	3D Convolution + iDT	2013	93.5%	69.2%	81.3%
	SCLSTM	2017	84.0%	55.1%	69.5%
	FCNs-16 + iDT	2017	93.0%	70.2%	81.6%
	LTC-iDT	2018	92.7%	67.2%	79.9%
Ours	SC-BDLSTM		94.2%	73.9%	84.0%

Comparison to the existing state-of-the-art methods

In our previous subsections, we already explore our proposed model in different aspects. This subsection further verifies the effectiveness and feasibility of our model. The recognition accuracy of our proposed approach is compared with various existing successful and prominent Human Action Recognition approaches on UCF101 and HMDB51 video datasets. The comparison performance in terms of accuracy is listed in Table 4. We categorize these baseline models concerning the type of extracted features and network architecture being used, including traditional (handcrafted), deep convolutional neural networks (DCNNs), recurrent neural networks (RNNs) and hybrid features. Because of the non-availability of recognition results of some methods on particular datasets, some cells are left blanks in the table.

Most of the hand-crafted feature-based techniques make use of trajectories such as iDT-FV (Wang et al., 2013a), ordered trajectories (OD) (Murthy et al., 2013), trajectory rejection (TR) (Seo et al., 2017), Motion part regularization (MPR) (Ni et al., 2015) and Mofap (Wang et al., 2016b) perform well and have competitive results, however, our approach outperforms them by a fair margin on both datasets. Compared with prominent deep learning models such as 3D Convolution, C3D, FCN-16 (Yu et al., 2017), FSTCN (Sun et al., 2015), DANN (Wang et al., 2016c), (Bilen et al., 2016) and LTC-CNN (Varol et al., 2018), the proposed method reported slightly better results in accuracy on UCF-101 and HMDB51 dataset respectively against the best accuracy of LTC-CNN. It can be also seen that some RNN based methods such STPP-LSTM (Wang et al., 2017a), L2STM, RMDN (Bazzani et al., 2017), Hidden-Two-Stream (Zhu et al., 2018), Multi-LSTM (Yeung et al., 2018) and TS-LSTM obtained extremely competitive results on UCF-101 datasets and specially RSTAN (LSTM based attention model) (Wenbin et al., 2018) shows better performance by a minimal margin on the UCF101 dataset. However, our introduced method outperforms these RNN based methods on the HDMB51 dataset and show a higher recognition rate on the small-scale dataset. In contrast to the manually crafted features and DCNN and RNN models, some frameworks like TDD+ iDT+ FV (Wang et al., 2015), FCNs-16+iDTand LTC-IDT, which integrates deep-learning features and hand-crafted features also produced state-of-the-art results but still, our model outperforms these hybrid models by a fair margin. We can conclude that a combination of bidirectional LSTM with the 3D convolutional network for RGB and Saliency-aware streams achieves better results and obtains the recognition rate of 94.2% and 73.9% on UCF101 and HDMB51 datasets respectively. Moreover, we also demonstrate the combined recognition accuracy by considering both datasets and introduced model achieved better results than all. For combined recognition accuracy, we only consider the accuracy of those models which provides the recognition results for both of the datasets. The symbol "-" means results are not reported for this particular dataset. Thus, our model in the presence of RGB and saliency-aware stream explores more relationships between video clips and salient regions and the introduction of bidirectional LSTM captures the long-term temporal dependencies from video frames.

CONCLUSIONS

This research work introduced an effective framework for human action recognition which combines both modalities i.e., C3D and bidirectional LSTMs. Two streams, RGB and saliency-based streams, are used to obtain strong video representation for action prediction in videos. Firstly, we apply the saliency-based method to capture saliency-aware videos, which are extremely useful in enhancing the significance of foreground objects and regions in the video frame. This method also avoids the computation complexity as we usually find in optical flow data. Frame level features are extracted by C3D and clip level features are processed by BD-LSTM and the time-series pooling layer. Our BD-LSTM comprises two stacking layers in both backward and forward directions. This proposed architecture performs well to discover the hidden and complex sequential patterns in video features. The achieved recognition results demonstrate that the recognition performance of our model outperforms the other existing prominent models on HMDB51 and UCF-101 datasets. Successful validations proved that our approach is suitable for the processing of sequential visual data. In future research work, we can retrain our framework on larger datasets such as Kinetic human action video dataset, ActivityNet, 1M-sports, and Charades for further evaluation and improvement of our proposed framework. Also, attention mechanisms can be incorporated into our proposed method to enhance the learning of video representation and successful recognition of complex human actions such as actions with a series of subactivities.

REFERENCES

Arif, S., Wang, J., Ul Hassan, T. & Fei. Z. 2019. 3D-CNN-Based Fused Feature Maps with LSTM Applied to Action Recognition. Innovative Topologies and Algorithms for Neural Networks, Future Internet (MDPI). **11**(2): 1-17.

Bazzani, L., Larochelle, H. & Torresani, L. 2017. Recurrent mixture density network for spatiotemporal visual attention. 5th International Conference on Learning Representations ICLR. Toulon, France.

Bilen, H., Fernando, B. & Gavves. 2016. Dynamic image networks for action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA.

Dalal, N. & Triggs, B. 2005. Histograms of oriented gradients for human detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). San Diego, CA, USA.

Dalal, N., Triggs, B. & Schmid. C. 2006. Human detection using oriented histograms of flow and appearance. In European Conference on Computer Vision. Graz, Austria.

Donahue, J., Hendricks, L. & S. Guadarrama, S. 2017. Long-term recurrent convolutional networks for visual recognition and description. IEEE Transactions on Pattern Analysis and Machine Intelligence. **39**(4): 677-691.

Feichtenhofer, C., Pinz, A. & Zisserman, A. 2016. Convolutional two-stream network fusion for video action recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA.

Graves, A., Fernández, S. & Schmidhuber, J. 2005. Bidirectional LSTM networks for improved phoneme classification and recognition. International Conference on Artificial Neural Networks: Formal Models and Their Applications–ICANN. Warsaw, Poland.

Hochreiter, S., & J. Schmidhuber, J. 1997. Long short-term memory. 1997. Neural Computation, **9**(8): 1735-1780.

Ji, S., Xu, W. & Yang, M. 2013. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. **35**(1): 221-231.

Jia, Y., Shelhamer, E. & Donahue, J. 2014. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the ACM International Conference on Multimedia. Pp: 675-678.

Jiang, Y.G., Liu, J. & Roshan, Z.A. 2013. THUMOS challenge: Action recognition with a large number of classes. THUMOS'13 International Workshop on Action Recognition with a Large Number of Classes Program. Sydney, Australia.

Karpathy, A., Toderici, G., Shetty, S. & Leung, T., Sukthankar, R & Li, F.F. 2014. Large-scale video classification with convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA.

Kuehne, H., Jhuang, H. & Garrote, E. 2011. Hmdb: a large video database for human motion recognition. IEEE International Conference on Computer Vision, Barcelona, Spain.

Li, Z., Gavves, E. & Jain, M. 2018. VideoLSTM convolves, attends and flows for action recognition. Computer Vision and Image Understanding. **166**: 41-50.

Lowe, DG. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision. **60**(2): 91-11.

Ma, C.Y., Chen, M.H. & Kira, Z. 2019. Ts-lstm and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. Signal Processing: Image Communication. **71**: 76-87.

Mahasseni, B. & Todorovic, S. 2016. Regularizing long short term memory with 3D human-skeleton sequences for action recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA.

Murthy, O.V. & Goecke, R. 2013. Ordered trajectories for large scale human action recognition. In proceeding IEEE conference on computer vision and pattern recognition. Sydney, NSW, Australia.

Ni, B., Moulin, P. & Yang, X. 2015. Motion part regularization: Improving action recognition via trajectory selection. In Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA.

Poppe, R. 2010. A survey on vision-based human action recognition. Image and Vision Computing. **28**(6): 976-990.

Scovanner, P., Ali, S. & Shah, M. 2007. A 3-dimensional SIFT descriptor and its application to action recognition. Proceedings of the 15th ACM international conference on Multimedia. Pp: 357-360.

Seo, J., Kim, H & Ro, Y.M. 2017. Effective and efficient human action recognition using dynamic frame skipping and trajectory rejection. Journal Image and Vision Computing, **58**:76-85.

Sharma, S., Kiros, R. & Salakhutdinov, R. 2015. Action recognition using visual attention. Mathematics, Computer Science, International Conference on Learning Representations ICLR.

Simonyan, K., & Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. Proceedings of the 27th International Conference on Neural Information Processing Systems, **1**: 568-576.

Soomro, K., Zamir, A.R. & Shah, M. 2012. UCF101: A dataset of 101 human action classes from videos in the wild. Center Res. Comput. Vis. Univ. Florida, Orlando, USA.

Srivastava, N., Mansimov, E & Salakhutdinov, R. 2015. Unsupervised learning of video representations using lstms. Proceedings of the 32nd International Conference on International Conference on Machine Learning, **37**: 843-852.

Sun, L., Jia, K. & Shi, B.E. 2015. Human action recognition using factorized spatio-temporal convolutional networks. In Proceedings of IEEE International Conference on computer vision (ICCV), Santiago, Chile.

Sun, L., Jia, K. & Chen, K. 2017. Lattice Long Short-Term Memory for Human Action Recognition. IEEE International Conference on Computer Vision, Venice, Italy.

Tran, D., Bourdev, L. & Fergus, R. 2015. Learning spatiotemporal features with 3d convolutional networks. IEEE International Conference on Computer Vision, Santiago, Chile.

Ullah, A., Ahmad, J. & Muhammad, K. 2017. Action Recognition in Video Sequences using Deep Bi-directional LSTM with CNN Features. Visual Surveillance and Biometrics: Practices, Challenges, and Possibilities, IEEE Access **6**: 1155-1166.

Varol, G., Laptev, I. & Schmid, C. 2018. Long- term temporal convolutions for action recognition. IEEE transactions on pattern analysis and machine intelligence, **40**: 1510-1517.

Veeriah, V., Zhuang, N. & Qi, G.J. 2015. Differential recurrent neural networks for action recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Santiago, Chile.

Wang, H & Schmid, C. 2013a. Action recognition with improved trajectories. IEEE International Conference on Computer Vision, Sydney, NSW, Australia.

Wang, H., Klaser, A. & Schmid, C. 2013b. Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, **103**: 60-79.

Wang, J., Wang, W. & Wang, R. 2016c. Deep alternative neural network: exploring contexts as early as possible for action recognition. Proceedings of 30th Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain.

Wang, L., Qiao, Y. & Tang, X. 2015. Action recognition with trajectory-pooled deep-convolutional descriptors. Proceedings of the IEEE conference on computer vision and pattern recognition. Boston, MA, USA

Wang, L., Xiong, Y. & Z. Wang, Z. 2016a. Temporal segment networks: towards good practices for deep action recognition. European Conference on Computer Vision, Amsterdam, The Netherlands.

Wang, L., Qiao, Y. & Tang, X. 2016b. Mofap: a multi-level representation for action recognition. International Journal of Computer Vision, **119**: 119-254.

Wang, W. & Shen, J. 2015. Saliency-aware geodesic video object segmentation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA.

Wang, X., Gao, L. & Wang, P. 2017a. Two-stream 3D ConvNet Fusion for Action Recognition in Videos with Arbitrary Size and Length. IEEE transaction on multimedia, **20**(3): 634-644.

Wang, X., Gao, L. & Song, J. 2017b. Beyond Frame-level CNN: Saliency-Aware 3-D CNN with LSTM for Video Action Recognition. IEEE signal processing letters, **24**(4): 510-514.

Wang, Y., Wang, S. & Tang, J. 2016d. Hierarchical attention network for action recognition in videos. In ArXiv.

Wenbin, D., Wang, Y. & Qiao, Y. 2018. Recurrent Spatial-Temporal Attention Network for Action Recognition in Videos. IEEE Transactions on image processing, **27**(3): 1-14.

Wu, Z., Wang, X. & Jiang, Y. 2015. Modelling spatial-temporal clues in a hybrid deep learning framework for video classification.

In Proceedings of the 23rd ACM international conference on Multimedia. Pp: 461-470.

Xingjian, S., Chen, Z. & Wang, H. 2015. Convolutional lstm network'. A machine learning approach for precipitation nowcasting. Proceedings of the 28th International Conference on Neural Information Processing Systems. Pp: 802-810.

Yeung, S., Russakovsky, O. & N. Jin. 2018. Every Moment Counts: Dense Detailed Labeling of Actions in Complex Videos. International Journal of computer vision, **126**: 375-389.

Yu, S., Cheng, Y. & Xie, L. 2017. Fully convolutional networks for action recognition. Institution of Engineering and Technology (IET). **11**(8): 744-749.

Yue-Hei, J., Hausknecht, M. & Vijayanarasimhan, S. 2015. Beyond short snippets: Deep networks for video classification. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA.

Zaremba, W., Sutskever, I. & Vinyals. O. 2014. Recurrent neural network regularization. arXiv preprint arXiv:1409.2329.

Zhang, B., Wang, L. & Wang, Z.Y. 2016. Real-time action recognition with enhanced motion vector CNNs. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA.

Zhou, Z., Shi, F. & Wu, W. 2015. Learning spatial and temporal extents of human actions for action detection. IEEE Transaction on Multimedia, **17**(4): 512-525.

Zhu, Y., Zhengzhong, L. & Newsam, S. 2018. Hidden two-stream convolutional networks for action recognition. Asian Conference on Computer Vision (ACCV). Perth, WA, Australia.