A Multi-modal approach to Human Action Classifications

Tan Ren Jie*, Kelvin Yeo Ngan Chong †, and Khaing Mon Kyaw‡

Institute of Systems Science, National University of Singapore
Email: *e0267395@u.nus.edu, †e0267573@u.nus.edu, ‡e0146800@u.nus.edu,

***Abstract*—**

I. Introduction

Single Stream [1].
Two Stream [2].
Long-term Recurrent Convolutional Networks (LRCN) [3]
3-D Convolutional Networks [4]
Conv3D with Attention [5]
Two-Stream Network Fusion [6]
Temporal Segment Networks (TSN) [7]
Action Video-Level Aggregation (ActionVLAD) [8]
Hidden Two-Stream Convolutional Networks [9]
Two-Stream Inflated 3D ConvNet (Two-StreamI3D) [10]
Temporal 3D ConvNets (T3D) [11]

- Open with background infomration of the problem case
- Present literature review of other works and their limitations
- Summary of how your work will tackle the limitations of others/solve the problem

II. Methods and Modelling

- Present the derivation of your mdoel and explanations

III. Computational Simulation

- Present the data source and programme/simulation to be conducted

IV. Results

- Present the results of your model simulation

V. Dicussion

- Compare your results with other models and highlight its effectiveness
- Dicuss implications, limitations and future work

VI. Conclusion

References

[1] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*.

[2] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *CoRR*, vol. abs/1406.2199, 2014. [Online]. Available: http://arxiv.org/abs/1406.2199

[3] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *CoRR*, vol. abs/1411.4389, 2014. [Online]. Available: http://arxiv.org/abs/1411.4389

[4] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D: generic features for video analysis," *CoRR*, vol. abs/1412.0767, 2014. [Online]. Available: http://arxiv.org/abs/1412.0767

[5] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing Videos by Exploiting Temporal Structure," *ArXiv e-prints*, Feb. 2015.

[6] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," *CoRR*, vol. abs/1604.06573, 2016. [Online]. Available: http://arxiv.org/abs/1604.06573

[7] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," *CoRR*, vol. abs/1608.00859, 2016. [Online]. Available: http://arxiv.org/abs/1608.00859

[8] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. C. Russell, "Actionvlad: Learning spatio-temporal aggregation for action classification," *CoRR*, vol. abs/1704.02895, 2017. [Online]. Available: http://arxiv.org/abs/1704.02895

[9] Y. Zhu, Z. Lan, S. D. Newsam, and A. G. Hauptmann, "Hidden two-stream convolutional networks for action recognition," *CoRR*, vol. abs/1704.00389, 2017. [Online]. Available: http://arxiv.org/abs/1704.00389

[10] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," *CoRR*, vol. abs/1705.07750, 2017. [Online]. Available: http://arxiv.org/abs/1705.07750

[11] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. V. Gool, "Temporal 3d convnets: New architecture and transfer learning for video classification," *CoRR*, vol. abs/1711.08200, 2017. [Online]. Available: http://arxiv.org/abs/1711.08200