References

[1] Sathyanarayanan Aakur, Fillipe DM de Souza, and Sudeep Sarkar. Going deeper with semantics: Exploiting semantic contextualization for interpretation of human activity in videos. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019. 1, 6, 8

[2] Sathyanarayanan N. Aakur, Fillipe DM de Souza, and Sudeep Sarkar. Towards a knowledge-based approach for generating video descriptions. In *Conference on Computer and Robot Vision (CRV)*. Springer, 2017. 1

[3] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4575–4583, 2016. 2, 5, 6, 7, 8

[4] Bharat Lal Bhatnagar, Suriya Singh, Chetan Arora, CV Jawahar, and KCIS CVIT. Unsupervised learning of deep feature representation for clustering egocentric actions. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1447–1453. AAAI Press, 2017. 2, 7

[5] Yi Bin, Yang Yang, Fumin Shen, Xing Xu, and Heng Tao Shen. Bidirectional long-short term memory for video description. In *ACM Conference on Multimedia (ACM MM)*, pages 436–440. ACM, 2016. 1

[6] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision (ECCV)*, pages 628–643. Springer, 2014. 2, 7, 8

[7] Gail A Carpenter and Stephen Grossberg. *Adaptive resonance theory*. Springer, 2016. 2

[8] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1932–1939. IEEE, 2009. 3

[9] Fillipe DM de Souza, Sudeep Sarkar, Anuj Srivastava, and Jingyong Su. Spatially coherent interpretations of videos using pattern theory. *International Journal on Computer Vision (IJCV)*, pages 1–21, 2016. 6, 8

[10] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 7

[11] Joaquin M Fuster. The prefrontal cortex and its relation to behavior. In *Progress in brain research*, volume 87, pages 201–211. Elsevier, 1991. 2

[12] Ana Garcia del Molino, Joo-Hwee Lim, and Ah-Hwee Tan. Predicting visual context for unsupervised event segmentation in continuous photo-streams. In *ACM Conference on Multimedia (ACM MM)*, pages 10–17. ACM, 2018. 2

[13] Zhao Guo, Lianli Gao, Jingkuan Song, Xing Xu, Jie Shao, and Heng Tao Shen. Attention-based lstm with semantic consistency for videos captioning. In *ACM Conference on Multimedia (ACM MM)*, pages 357–361. ACM, 2016. 1

[14] Catherine Hanson and Stephen José Hanson. Development of schemata during event parsing: Neisser's perceptual cycle as a recurrent connectionist network. *Journal of Cognitive Neuroscience*, 8(2):119–134, 1996. 2

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3, 5

[16] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *European Conference on Computer Vision (ECCV)*, pages 137–153. Springer, 2016. 2, 6, 7, 8

[17] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *Neural Information Processing Systems*, pages 667–675, 2016. 4

[18] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, 2014. 1

[19] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 780–787, 2014. 1, 2, 5, 6, 7, 8

[20] Hilde Kuehne, Juergen Gall, and Thomas Serre. An end-to-end generative framework for video segmentation and recognition. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016. 7, 8

[21] Colin Lea, Austin Reiter, René Vidal, and Gregory D Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *European Conference on Computer Vision (ECCV)*, pages 36–52. Springer, 2016. 2, 6, 7

[22] Peng Lei and Sinisa Todorovic. Temporal deformable residual networks for action segmentation in videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6742–6751, 2018. 7

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 2

[24] Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. *arXiv preprint arXiv:1503.01558*, 2015. 2, 7, 8

[25] Katherine Metcalf and David Leake. Modelling unsupervised event segmentation: Learning event boundaries from prediction errors. In *CogSci*, 2017. 2

[26] Ulric Neisser. Cognitive psychology new york: Appleton-century-crofts. *Google Scholar*, 1967. 2

[27] Colin Lea Michael D Flynn René and Vidal Austin Reiter Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 6, 7

[28] Alexander Richard and Juergen Gall. Temporal action detection using a statistical language model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3131–3140, 2016. 7

[29] Alexander Richard, Hilde Kuehne, and Juergen Gall. Weakly supervised action learning with rnn based fine-to-coarse