Seminar "Video Understanding"

November 2020

(Initial Papers and) Topics

Wu, Chao-Yuan, Manzil Zaheer, Hexiang Hu, R. Manmatha, Alexander J. Smola, and Philipp Krahenbuhl. "Compressed Video Action Recognition" [in en]. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6026–6035. Salt Lake City, UT: IEEE, June 2018. ISBN: 978-1-5386-6420-9. https://doi.org/10.1109/CVPR.2018.00631.

1 Compressed-Video Action Recognition.

Carreira, Joao, and Andrew Zisserman. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset" [in en]. *arXiv:1705.07750 [cs]*, February 2018. arXiv: 1705.07750 [cs].

2 Deep Net Architectures for Action Recognition.

Butepage, Judith, Michael J. Black, Danica Kragic, and Hedvig Kjellstrom. "Deep Representation Learning for Human Motion Prediction and Classification" [in en]. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1591–1599. Honolulu, HI: IEEE, July 2017. ISBN: 978-1-5386-0457-1. https://doi.org/10.1109/CVPR.2017.173.

3 Representation Learning for Human Motion Prediction.

Tran, Du, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. "A Closer Look at Spatiotemporal Convolutions for Action Recognition" [in en]. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6450–6459. Salt Lake City, UT: IEEE, June 2018. ISBN: 978-1-5386-6420-9. https://doi.org/10.1109/CVPR.2018.00675.

4 Spatio-Temporal Convolutions for Action Recognition.

Ji, Shuiwang, Wei Xu, Ming Yang, and Kai Yu. "3D Convolutional Neural Networks for Human Action Recognition" [in en]. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, no. 1 (January 2013): 221–231. ISSN: 0162-8828, 2160-9292. https://doi.org/10.1109/TPAMI.2012.59.

5 Spatio-Temporal CNN for Action Recognition.

Lee, Namhoon, Wongun Choi, Paul Vernaza, Christopher B. Choy, Philip H. S. Torr, and Manmohan Chandraker. "DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents" [in en]. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2165–2174. Honolulu, HI: IEEE, July 2017. ISBN: 978-1-5386-0457-1. https://doi.org/10.1109/CVPR.2017.233.

6 Future Prediction for Interacting Agents.

Pathak, Deepak, Ross Girshick, Piotr Dollar, Trevor Darrell, and Bharath Hariharan. "Learning Features by Watching Objects Move" [in en]. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6024–6033. Honolulu, HI: IEEE, July 2017. ISBN: 978-1-5386-0457-1. https://doi.org/10.1109/CVPR.2017.638.

7 Self-Supervision: Learning Features from Moving Objects.

Gordon, Ariel, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. "Depth From Videos in the Wild: Unsupervised Monocular Depth Learning From Unknown Cameras" [in en]:10.

8 Self-Supervision: Unsupervised Depth Learning from Video.

Godard, Clement, Oisin Mac Aodha, and Gabriel J. Brostow. "Unsupervised Monocular Depth Estimation with Left-Right Consistency" [in en]. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6602–6611. Honolulu, HI: IEEE, July 2017. ISBN: 978-1-5386-0457-1. https://doi.org/10.1109/CVPR.2017.699.

8* Self-Supervision: Unsupervised Depth Learning from Video.

Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "Unsupervised Learning of Video Representations Using LSTMs" [in en]:10.

9 Self-Supervision: Unsupervised Learning of Video Representations.

Luc, Pauline, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun. "Predicting Deeper into the Future of Semantic Segmentation" [in en]. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 648–657. Venice: IEEE, October 2017. ISBN: 978-1-5386-1032-9. https://doi.org/10.1109/ICCV.2017.77.

10 Predicting the Future in Video.

Mathieu, Michael, Camille Couprie, and Yann LeCun. "Deep Multi-Scale Video Prediction beyond Mean Square Error" [in en]. *arXiv:1511.05440 [cs, stat]*, February 2016. arXiv: 1511.05440 [cs, stat].

10* Predicting the Future in Video.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need" [in en]. *arXiv:1706.03762 [cs]*, December 2017. arXiv: 1706.03762 [cs].

11 Transformers in Language Processing.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale" [in en]. *arXiv:2010.11929 [cs]*, October 2020. arXiv: 2010.11929 [cs].

12 Visual Transformers.

Wu, Bichen, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. "Visual Transformers: Token-Based Image Representation and Processing for Computer Vision" [in en]. *arXiv:2006.03677 [cs, eess]*, July 2020. arXiv: 2006.03677 [cs, eess].

12* Visual Transformers.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" [in en]. *arXiv:1502.03044 [cs]*, April 2016. arXiv: 1502.03044 [cs].

13 Image Captioning with Visual Attention.

Lei, Jie, Licheng Yu, Mohit Bansal, and Tamara L. Berg. "TVQA: Localized, Compositional Video Question Answering" [in en]. *arXiv:1809.01696 [cs]*, May 2019. arXiv: 1809.01696 [cs].

14 Video Question Answering.

Pramanik, Subhojeet, Priyanka Agrawal, and Aman Hussain. "OmniNet: A Unified Architecture for Multi-Modal Multi-Task Learning" [in en]. *arXiv:1907.07804 [cs, stat]*, July 2020. arXiv: 1907.07804 [cs, stat].

15 Multi-Task Learning.

Gao, Lianli, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. "Video Captioning With Attention-Based LSTM and Semantic Consistency" [in en]. *IEEE Transactions on Multimedia* 19, no. 9 (September 2017): 2045–2055. ISSN: 1520-9210, 1941-0077. https://doi.org/10.1109/TMM.2017.2729019.

16 Video Captioning with Attention Modelling.

Shen, Zhiqiang, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. "Weakly Supervised Dense Video Captioning" [in en]. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5159–5167. Honolulu, HI: IEEE, July 2017. ISBN: 978-1-5386-0457-1. https://doi.org/10.1109/CVPR.2017.548.

17 Weakly-Supervised Dense Video Captioning.

Rahman, Tanzila, Bicheng Xu, and Leonid Sigal. "Watch, Listen and Tell: Multi-Modal Weakly Supervised Dense Event Captioning" [in en]. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 8907–8916. Seoul, Korea (South): IEEE, October 2019. ISBN: 978-1-72814-803-8. https://doi.org/10.1109/ICCV.2019.00900.

18 Visual-Lingual Video Captioning.

Tsai, Yao-Hung Hubert, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. "Multimodal Transformer for Unaligned Multimodal Language Sequences" [in en]. *arXiv:1906.00295 [cs]*, June 2019. arXiv: 1906.00295 [cs].

19 Analyzing Multi-Modal Time Series.

Dai, Angela, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. "BundleFusion: Real-Time Globally Consistent 3D Reconstruction Using On-the-Fly Surface Reintegration" [in en]. *ACM Transactions on Graphics* 36, no. 3 (July 2017): 1–18. ISSN: 0730-0301, 1557-7368. https://doi.org/10.1145/3054739.

20 Real-Time Globally Consistent 3D Reconstruction.

Dou, Mingsong, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. "Motion2fusion: Real-Time Volumetric Performance Capture" [in en]. *ACM Transactions on Graphics* 36, no. 6 (November 2017): 1–16. ISSN: 0730-0301, 1557-7368. https://doi.org/10.1145/3130800.3130801.

21 Volumetric Motion Capture.

Tulsiani, Shubham, Abhishek Kar, Joao Carreira, and Jitendra Malik. "Learning Category-Specific Deformable 3D Models for Object Reconstruction" [in en]. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, no. 4 (April 2017): 719–731. ISSN: 0162-8828, 2160-9292. https://doi.org/10.1109/TPAMI.2016.2574713.

22 Learning Deformable 3D Models.

Feng, Andrew, Dan Casas, and Ari Shapiro. "Avatar Reshaping and Automatic Rigging Using a Deformable Model." In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, 57–64. MIG '15. New York, NY, USA: Association for Computing Machinery, November 2015. ISBN: 978-1-4503-3991-9. https://doi.org/10.1145/2822013.2822017.

23 Automatic Rigging Using a Deformable Model.

Le, Binh Huy, and Zhigang Deng. "Robust and Accurate Skeletal Rigging from Mesh Sequences" [in en]. *ACM Transactions on Graphics* 33, no. 4 (July 2014): 1–10. ISSN: 0730-0301, 1557-7368. https://doi.org/10.1145/2601097.2601161.

23* Skeletal Rigging from Mesh Sequences.

Tagliasacchi, Andrea, Matthias Schröder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. "Robust Articulated-ICP for Real-Time Hand Tracking" [in en]. *Computer Graphics Forum* 34, no. 5 (August 2015): 101–114. ISSN: 01677055. https://doi.org/10.1111/cgf.12700.

24 Real-Time Hand Tracking.