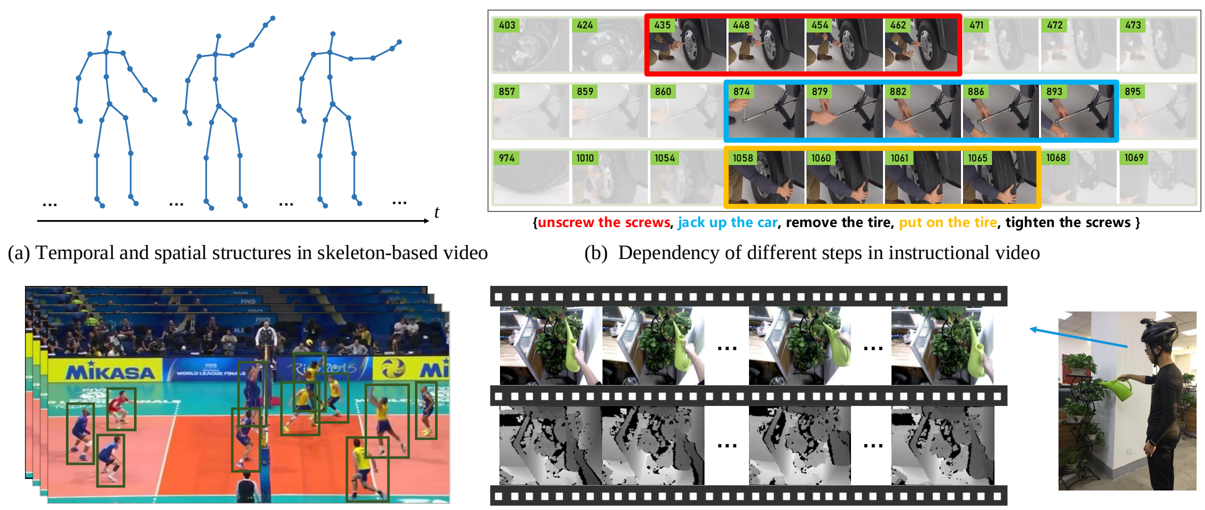
Research Statement – Yansong Tang

yansong.tang@eng.ox.ac.uk - https://andytang15.github.io

As an important and fundamental problem in computer vision, human activity understanding has wide practical applications, such as human-computer interaction, sport video analysis, video retrieval and many others. Compared with the conventional image-based visual understanding tasks, understanding human activity in videos is more challenging because of the various and complex structures in different videos. For example, the temporal and spatial structures of different frames and joints in skeleton-based video (Fig. 1 (a)), the dependency of different steps in long-term instructional video (Fig. 1 (b)), the relation of different people in group activity video (Fig. 1 (c)), and the sharable-distinctive charateristics of multiple modalities in RGBD egocentric video (Fig. 1 (d)). Although great progress has been achieved for learning general video representations in recent years, **there is still plenty of room to leverage these structures for enhancing the understanding capability of different activities. Towards this goal, my research concerntrats on devising specialized algorithms to explore the corresponding structures in videos for better results.** I will introduce my research on the four directions accordingly as follow.



(a) Temporal and spatial structures in skeleton-based video (b) Dependency of different steps in instructional video

(c) Relation of different people in group activity video (d) Sharable-distinctive characteristics of multiple modalities in RGBD egocentric video

Figure 1: There are rich structure information in different types of videos, *e.g.,* (a) the temporal and spatial structures of different frames and joints in skeleton-based video, (b) the dependency of different steps in instructional video. (c) the relation of different people in group activity video, and (d) the sharable and distinctive charateristics of multiple modalities in RGBD egocentric videos. All figures are best viewed in color.

Mining Temporal and Spatial Structure in Skeleton-based Video. The skeleton-based videos are comprised of compact coordinates of human joints at each time-stamp. In order to explore temporal strcuture of different frames, we propose a deep progressive reinforcement learning (DPRL) method which aims to select the most informative frames for action recognition (CVPR 2018 [7]). Moreover, we employ the graph-based convolutional neural network to capture the spatial dependency between the joints for feature learning.

Besides, we pioneer a new unsupervised domain adaptation setting for skeleton-based action recognition, where the action labels are only available on a source dataset, but unavailable on a target dataset. We present two approaches for this problem, the first is building on the widely used adversial learning strategy, but performing transfer at the relation level rather than the vanilla feature level (TCSVT 2020 [10]). The second takes recent advance on self-supervised learning [2]. By segmenting and permuting temporal segments or human body parts, we design two self-supervised learning classification tasks to explore the temporal and spatial dependency of a skeleton-based action and improve the generalization ability of the model.

Exploring Dependency among Different Steps in Instructional Video. There are substantial instructional videos on the Internet, which enables novices to acquire knowledge for completing various tasks. Compared with short clips, instructional videos have longer durations and an interesting task is to localize the different steps in the videos. To tackle this problem, we first collect a dataset called "COIN" for COmprehensive INstructional video analysis (CVPR 2019 [1]). It contains 11,827 videos of 180 tasks from 12 domains. As a by product, we contribute a new toolbox to effectively annotate a series of step descriptions and the corresponding temporal boundaries. Furthermore, we exploit two important characteristics (*i.e.,* task-consistency and ordering-dependency) for localizing important steps in instructional videos (TPAMI 2020 [4]). Accordingly, we propose two simple yet effective methods, which can be easily plugged into conventional proposal-based action detection models.

Modelling the Relation among Different People in Group Activity Video. Compared with conventional action recognition based on single person, group activity recognition is a more challenging task as it requires further understanding of high-level relationships among different people. Different from the common attention mechanisms, we develope a Teacher Network to leverage the prior knowledge in the semantics domain, and explore the discriminative information of different people by transfering the semantics-preserving attention learned by the Teacher Network to the Student Network in the appearance domain (ACM MM 2018 [8]). To our best knoledge, this is the original effort leveraging attention in both semantics and appearance clues to perform group activity recognition. Furthermore, we utilize the graph convolutional modules to reason about the relationship of different people and extend our method for group activity detection in untrimmed long videos (TIP 2019 [3]).

Learning Sharable-Distinctive Charateristics of Multiple Modalities in RGB-D Egocentric Video. Recent years have witnessed rapid development on egocentric action recognition due to the development of wearable cameras such as GoPro and Google Glass. Generally speaking, most existing works on this topic are mainly based on RGB videos, which contain the spatial appearance and temporal information. However, the primary limitations of RGB videos are the absence of 3D information and the sensitivity to illumination variations, while an exclusive depth modality is capable of covering these shortages. To address this, we introduce a new dataset for RGB-D egocentric action recognition (ICIP 2017 [6]), and develop a multi-stream deep neural networks (MDNN) method (TCSVT 2019 [9]) to exploit the shared properties and distinctive characteristics for different modalities (*i.e.,* RGB frames, optical flows and depth frames). Moreover, we provide over 200M-pixel hand annotation in our dataset and strengthen our MDNN by incorporating with hand cues in the egocentric videos.

In the future, I will continue working in the field of human activity understanding and branch out to explore problems related to action transfer and action enhancement. I outline three topics that I plan to pursue as follow.

Understanding Activity in Long-term Video. While great progress has been achieved for action recognition in trimmed videos, understanding the long untrimmed videos is still a challenging and important problem. Specifically, in our collected COIN dataset [1], the state-of-the-art performance on the step localization and action segmentation tasks are still barely satisfactory. Although there are some encouraging results suggesting that leveraging the dependency of different steps would help [4], it is still a long way to go along this direction.

Egocentric Action Transfer. In our previous works [9, 6], we investigate the egocentric action recognition. Moving towards an interesting direction, I plan to transfer the action to the robot with the technology of hand pose estimation and reinforcement learning. In this way, the robot could imitate the human action to accomplish different tasks.

Action Quality Enhancement. We have made a primary study on the problem of action quality assessment in sport videos (CVPR 2020 [5]). As an extension, I plan to reconstruct the sport action in a virtual enviroment given an real-world video. With the reference of the assessed scores, I hope to enhance the actions to achieve higher scores. In this way, the enhanced action videos could serve as important resources to guide the sporters for their training process.

References

[1] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou. COIN: A large-scale dataset for comprehensive instructional video analysis. In *CVPR*, pages 1207–1216, 2019.

[2] Y. Tang, X. Liu, X. Yu, D. Zhang, J. Lu, and J. Zhou. Learning from temporal spatial cubism: A self-supervised domain adaptation approach for skeleton-based action recognition. In *Submission*, 2019.

[3] Y. Tang, J. Lu, Z. Wang, M. Yang, and J. Zhou. Learning semantics-preserving attention and contextual interaction for group activity recognition. *TIP*, 28(10):4997–5012, 2019.

[4] Y. Tang, J. Lu, and J. Zhou. Comprehensive instructional video analysis: The COIN dataset and performance evaluation. *TPAMI, accepted*, 2020.

[5] Y. Tang*, Z. Ni*, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *CVPR*, 2020.

[6] Y. Tang, Y. Tian, J. Lu, J. Feng, and J. Zhou. Action recognition in RGB-D egocentric videos. In *ICIP*, pages 3410–3414, 2017.

[7] Y. Tang*, Y. Tian*, J. Lu, P. Li, and J. Zhou. Deep progressive reinforcement learning for skeleton-based action recognition. In *CVPR*, pages 5323–5332, 2018.

[8] Y. Tang, Z. Wang, P. Li, J. Lu, M. Yang, and J. Zhou. Mining semantics-preserving attention for group activity recognition. In *ACM MM*, pages 1283–1291, 2018.

[9] Y. Tang, Z. Wang, J. Lu, J. Feng, and J. Zhou. Multi-stream deep neural networks for RGB-D egocentric action recognition. *TCSVT*, 29(10):3001–3015, 2019.

[10] Y. Tang, Y. Wei, X. Yu, J. Lu, and J. Zhou. Graph interaction networks for relation transfer in human activity videos. *TCSVT, accepted*, 2020.