Guest Editorial: Video Recognition

Ivan Laptev¹ · Deva Ramanan² · Josef Sivic¹

Published online: 22 June 2016
© Springer Science+Business Media New York 2016

“Video Recognition” is a fundamental research area in visual recognition, required for true perceptual understanding in any practical scenario where image streams are processed. Compared to static image analysis, videos introduce several additional challenges and opportunities, including activity parsing, spatiotemporal representation learning, and data-driven temporal learning. Video recognition has seen exciting progress over the last decade. First, new low-level and mid-level video representations have opened up the possibility of activity recognition “in the wild” as demonstrated on a number of recent benchmarks composed from YouTube, TV and movie videos. Second, new challenging tasks have been addressed, extending the field beyond activity classification towards modeling the temporal and spatial structure of events as well as relations between human actions, objects and scenes; recognizing person-person and person-object interactions; recognizing group activities and social roles of people; understanding activities in the first-person view video.

The goal of this special issue is to present recent papers addressing different aspects of video recognition. The issue contains eight contributions that are briefly discussed below.

- “A robust and efficient video representation for action recognition” (doi:[10.1007/s11263-015-0846-5](https://doi.org/10.1007/s11263-015-0846-5)) by Wang et al. improves dense trajectory video features by explicit

camera motion estimation and demonstrates superior performance of action recognition on a number of video benchmarks.

- “EXMOVES: Mid-level Features for Efficient Action Recognition and Video Analysis” (doi:[10.1007/s11263-016-0905-6](https://doi.org/10.1007/s11263-016-0905-6)) by Tran and Torresani proposes a scalable mid-level representation for video analysis based on exemplar movement classifiers.
- “MoFAP: A Multi-Level Representation for Action Recognition” (doi:[10.1007/s11263-015-0859-0](https://doi.org/10.1007/s11263-015-0859-0)) by Wang et al. introduces a multi-level video representation consisting of local motion, atomic actions, and composites, illustrating results on activity understanding.
- “LIBSVX: A Supervoxel Library and Benchmark for Early Video Processing” (doi:[10.1007/s11263-016-0906-5](https://doi.org/10.1007/s11263-016-0906-5)) by Xu and Corso presents an extensive benchmark evaluation of video-based segmentation algorithms, offering diagnostic analysis of scenarios where particular techniques work well.
- “Circulant temporal encoding for video retrieval and temporal alignment” (doi:[10.1007/s11263-015-0875-0](https://doi.org/10.1007/s11263-015-0875-0)) by Douze et al. describes a method for video event retrieval based on frequency-based descriptors, offering gains in both efficiency and temporal localization.
- “First-Person Activity Recognition: Feature, Temporal Structure, and Prediction” (doi:[10.1007/s11263-015-0847-4](https://doi.org/10.1007/s11263-015-0847-4)) by Ryoo and Matthies examines interaction-level human activities from a first-person viewpoint, considering temporal patterns generated by various types of interactions.
- “Learning Relevance Restricted Boltzmann Machine for Unstructured Group Activity and Event Understanding” (doi:[10.1007/s11263-016-0896-3](https://doi.org/10.1007/s11263-016-0896-3)) by Zhao et al. describes latent-variable models that learn mid-level

✉ Ivan Laptev
ivan.laptev@inria.fr

¹ Département d’Informatique de l’Ecole Normale Supérieure, Inria-Willow Team, Paris, France

² Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA

representations of activities defined by multiple group members.

- “Recognizing Fine-Grained and Composite Activities using Hand-Centric Features and Script Data” (doi:[10.1007/s11263-015-0851-8](https://doi.org/10.1007/s11263-015-0851-8)) by Rohrbach et al. exam-

ines the task of fine-grained activity understanding, demonstrating that human posture and hand pose are an important cues for such scenarios.