BMVA

The British Machine Vision Association and Society for Pattern Recognition

BMVA Symposium on Video Understanding

One Day BMVA symposium in London Wednesday 25th September 2019

Chairs: Hilde Kuehne, Dima Damen, Juergen Gall & Ivan Laptev

www.bmva.org

On Wednesday 25th of September, at the new offices of the British Computer Society in London, the largest BMVA one-day meeting was held with around 100 attendees and a line-up of 18 speakers. As new models for video understanding remain a bottleneck for research despite the increase in the number of large-scale video datasets, this symposium wanted to question the current challenges and opportunities in video understanding. Speakers included leading researchers who have contributed significantly to video understanding, and collectively participated in the collection of datasets such as: UCF, HMDB, THUMOS, Hollywood, Hollywood2, Kinetics, Charades,ActivityNet, YouTube8M, YouCook, Breakfast, EPIC-Kitchens, DALY and AVA, amongst others! With a star-studded list of speakers, we aim to ask the question, what is missing in video understanding to catch up with the success witnessed in object detection or language translation?

The morning session started by a keynote from Prof Jeff Zacks (Washington n St. Louis). His talk questioned the general consensus on events, and long-term monitoring of daily life. The talk triggered a number of questions regarding how computer vision researchers currently address some of the questions when compared to cognitive scientists. Prof Rahul Sukthankar (Google) presented the distinction between object recognition and action recognition as research problems, arguing for video understanding as a better-defined problem. He questioned the term ‘Video’ and proposed a distinction between ‘Created Media’ – that is the set of footage curated by individuals for a personal purpose, e.g. sharing using social media, and ‘Situated Video’ which have not been curated and represent both tedious and interesting activities, e.g. surveillance. Prof Cees Snoek (Amsterdam) delved into his latest works on action understanding using a two-in-one stream network, using optical flow only during training to moderate attention in the spatial stream. Prof Mubarak Shah (Central Florida) focused on spatio-temporal localisation of actions, and the need for accurate localisation for better video understanding, suggesting the upcoming video challenge is that of video object segmentation.

Following a coffee break, Prof Cordelia Schmid (INRIA) presented her group's latest work on video-bert, sequential visual-textual attention model trained from weakly supervised instructional videos with narrations. Dr Dima Damen (Bristol) presented a collection of her group's work on fine-grained action understanding, particularly for egocentric videos of object interactions including the collection of the EPIC-Kitchens dataset. Dr Juan Carlos Niebles (Stanford– Toyota) argued for graph-based convolutional approaches, as a mid-level representation for spatio-temporal information in video, including his latest works on retrieval of complex activities. Dr Du Tran (Facebook) presented a collection of the latest architectures for video understanding, developed by his group. He presented practical guidance on training multi-modal fusion video models. Dr Jason Corso (Michigan) addressed the problem of passively collecting narrations for all examples, replacing this by alternative hybrid approaches where access to a human 'expert' is available during inference.

The lunch break featured 27 posters, with engaging discussions about ongoing works as well as recently published methods. A few groups presented their coming ICCV papers on multi-modal retrieval, holistic as well as fine-grained video understanding, self-supervised learning and actor-object localisation in video.

Prof Andrew Zisserman (Oxford) initiated the afternoon session with a keynote on self-supervised learning in videos – the need and the potential. Prof Lorenzo Torresani (Dartmouth College, Facebook) discussed self-supervision multi-modal training as a pre-training approach to action recognition. He also presented efficient sampling architectures for action recognition. Dr Hilde Kuehne (MIT-IBM Watson Lab) focused on action localisation in untrimmed videos, including the series of successful datasets she's collected for this purpose from HMDB to Breakfast. Prof Ivan Laptev (INRIA) argued for the need to link video understanding to robotics focusing on embodied recognition, replacing unhelpful intermediate representations by the success in accomplishing a task. He presented their latest effort to collect HowTo100M dataset from instructional youtube videos. Dr Nazli Ikizler-Cinbis (Hacettepe, Ankara) presented a collection of her group's latest papers on multi-actor activities, including collaborative activities like sports, as well as face and gesture recognition.

After a concluding coffee break, Dr Jan van Gemert (Delft) presented an overview of datasets and approaches currently used in video understanding. He argued for better uniformity in how datasets are collected and annotated, as well as the tasks they are evaluated on. Prof Juergen Gall (Bonn) focused on action anticipation as a primary goal for intelligent video understanding. His group has produced a number of potential solutions to solve anticipation and identify uncertainty in this prediction. Dr Angela Yao (Singapore) questioned the focus on cooking-related videos, and presented their latest 'Tasty' dataset of instructional videos. Their work showed how sequences could be primarily learnt from steps of written recipes. Finally, Dr Efstratios Gavves (Amsterdam) presented his group's work novel Timeception and time-aligned neural architectures.

Following a dense day, during the conclusions, the audience were asked to select two topics to 'invest' in, out of a number of topics highlighted as important during the day. The audience selected from: Multi-modal learning, self-supervision, fine-grained understanding,

from instructional videos to robotics, spatio-temporal architectures, video-object segmentation, larger datasets, collaborative activities, video anticipation and pose estimation. The highest number of votes were devoted to the first two research topics: multi-modal learning and self-supervised learning.

While presenting this summary, the reader is very much encouraged to watch the recordings of the day, as these represent a more accurate account of the keynotes. The organisers (Hilde, Dima, Ivan and Juergen) would like to take this opportunity to thank all the speakers for very engaging discussions, and thank all the researchers (postdocs and PhD students) who enriched the day with their posters.

Dima Damen