Doctoral School of Information and Biomedical Technologies Polish Academy of Sciences

Domain: IT

SUBJECT: Object and Event Identification in Video Guided with Natural Language Processing of Video Content Description

Supervisors, contact: dr hab.inż. Szymon Łukasik (slukasik@ibspan.waw.pl)

Assistant supervisors, contact: dr inż. Anna Wróblewska (anna.wróblewska1@mini.pw.edu.pl)

Place of research: Systems Research Institute Polish Academy of Sciences, Newelska 6, Warszawa

Recruitment & Selection: Interview

Number of positions: 1

Project Description

(220 words)

Automatic video processing and understanding still poses a considerable challenge, caused not only by computational problems but also by complexity of the related tasks (e.g. object tracking). While many predictive models based purely on the image data have been developed, additional information in various forms may have a significant impact on improving their performance. Some research on the potential use of multimodal fusion has been already carried out [4, 5] and it has been demonstrated that multimodal models outperform their unimodal equivalents.

This research project aims to implement hybrid object and event detection models trained on a dataset enriched with textual descriptions and compare results with standard image-based models such as Faster-RCNN [1], YOLO [2], or DETR [3]. Multiple data sources in object detection have already been used in self-driving cars [6], surveillance systems [7], or industrial applications [8]. The authors of these papers use different sensor data to train robust models. Although researchers have recognized the natural connection between text and video data during information processing [10, 11], to the best of our knowledge, very little is known about the combination of textual descriptions and video data in object and event identification [9].

The research will be conducted in collaboration with the NASK National Research Institute which will provide methodological support, data and computational resources.

References

(180 words)

1. *S. Ren et al. Faster R-CNN: towards real-time object detection with region proposal networks. NIPS 2015, 91–99.*
2. *J. Redmon et al. You Only Look Once: Unified, Real-Time Object Detection. CVPR 2016, 779-788.*
3. *N. Carion et al. End-to-End Object Detection with Transformers. ECCV 2020, 213–229.*
4. *Y. Liu et al. Contrastive Multimodal Fusion with TupleInfoNCE. ICCV 2021, 734-743.*
5. *A. Nagrani et al. Attention Bottlenecks for Multimodal Fusion. NeurIPS 2021, 14200-14213.*

6. *D. Feng et al. Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. ITSC 2021, 1341–1360.*
7. *A. Czyżewski et al. Moving object detection and tracking for the purpose of multimodal surveillance system in urban areas. New Directions in Intelligent Interactive Multimedia 2008, 75-84.*
8. *B. Drost et al. 3D Object Detection and Localization Using Multimodal Point Pair Features. 3DIMPVT 2012, 9-16.*
9. *R. Valverde et al. There is More than Meets the Eye: Self-Supervised Multi-Object Detection and Tracking with Sound by Distilling Multimodal Knowledge. CVPR 2021.*
10. *R. Arandjelovic et al. Look, listen and learn. CVPR 2017.*
11. *D. Hu et al. Discriminative sounding objects localization via self-supervised audiovisual matching. NeurIPS 2020.*

Date: 19.05.2022