AI4media

ARTIFICIAL INTELLIGENCE FOR
THE MEDIA AND SOCIETY

Theme Development Workshop: Trusted AI - The future of creating ethical and responsible AI systems

Breakout Session: **Human-Aligned Video AI**

Invited expert: Vasileios Mezaris (CERTH)

13 Sept. 2023

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 951911

Human-Aligned AI & Video

- Human-Aligned AI: AI that serves humans’ goals and values
- In video understanding tasks, this translates to usefulness of results (accuracy, completeness, timeliness,…); lack of biases; (given that we deal with inevitably non-perfect AI methods,) support to the human user to understand what may have gone wrong and why.
- AI explainability is a means towards the latter goals

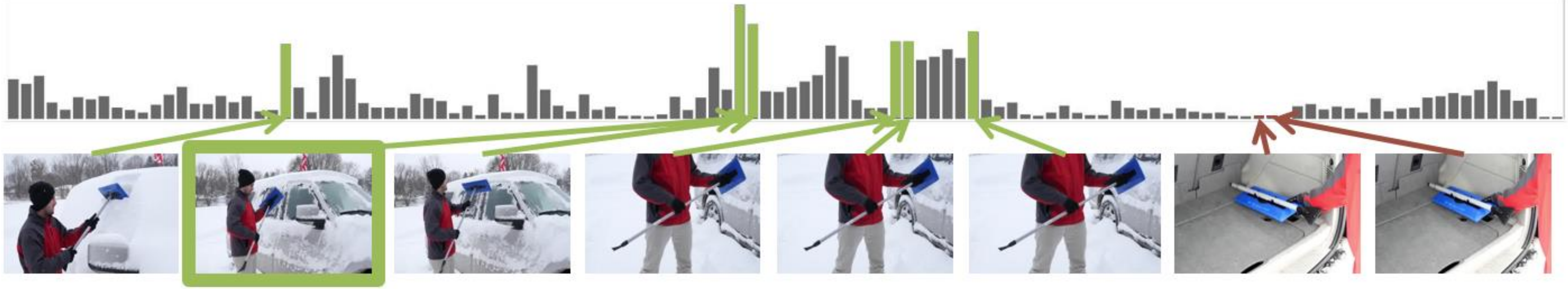
Video event recognition

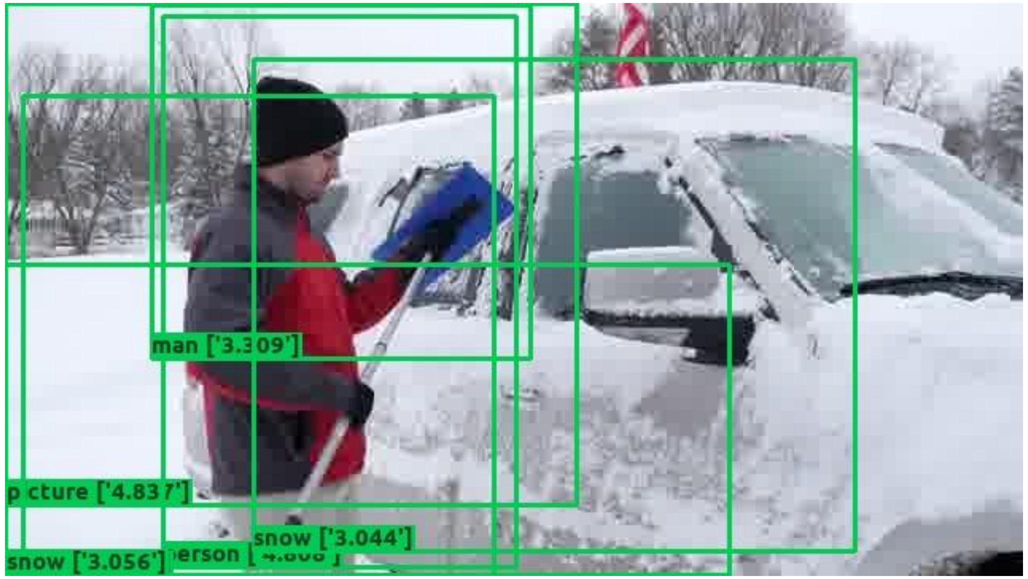
- Goal: in-depth understanding of visual information; recognize real-world events depicted in videos
- Intended use: search for videos showing events of interest
- General approach (ViGAT, Gated-ViGAT):
  - Sample the video to select frames; **detect a multitude of object regions** within each frame
  - Represent both objects and entire frames in a high-dimensional feature space using a pre-trained Transformer network
  - Use a learning architecture based on Graph Attention blocks (“ViGAT head”) to lean **to recognize the target events**, and also to **provide explanations for these recognition decisions**
  - Improve the scalability of the above approach by using part of the explanation signal to guide the network to process in depth (extract & represent objects) only the few frames that are most informative and needed for making a confident event recognition decision for a given video

Work supported by EU projects H2020-832921 MIRROR, H2020-101021866 CRiTERIA, National project T2EDK-00306 QuaLiSID

Video event recognition

- Example of correctly-categorized “Removing ice from car” video, and explanations



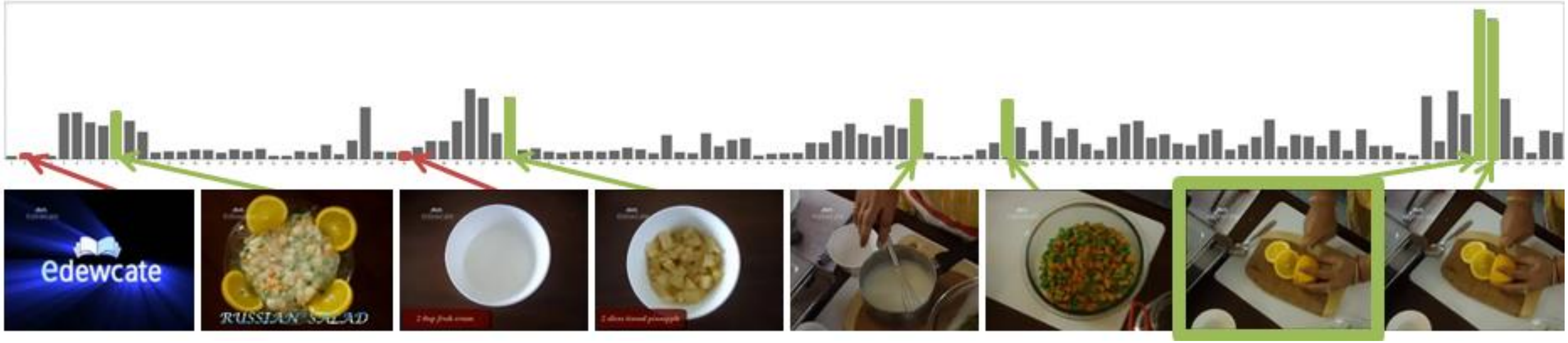


N. Gkalelis, D. Daskalakis, V. Mezaris, "ViGAT: Bottom-up event recognition and explanation in video using factorized graph attention network", IEEE Access, vol. 10, pp. 108797-108816, 2022. DOI:10.1109/ACCESS.2022.3213652.

N. Gkalelis, D. Daskalakis, V. Mezaris, "Gated-ViGAT: Efficient bottom-up event recognition and explanation using a new frame selection policy and gating mechanism", Proc. IEEE Int. Symposium on Multimedia (ISM), Naples, Italy, pp. 113-120, Dec. 2022. DOI:10.1109/ISM55400.2022.00024.

Video event recognition

- Example of “Preparing salad” video, mis-recognized as “Making Lemonade”, and explanations

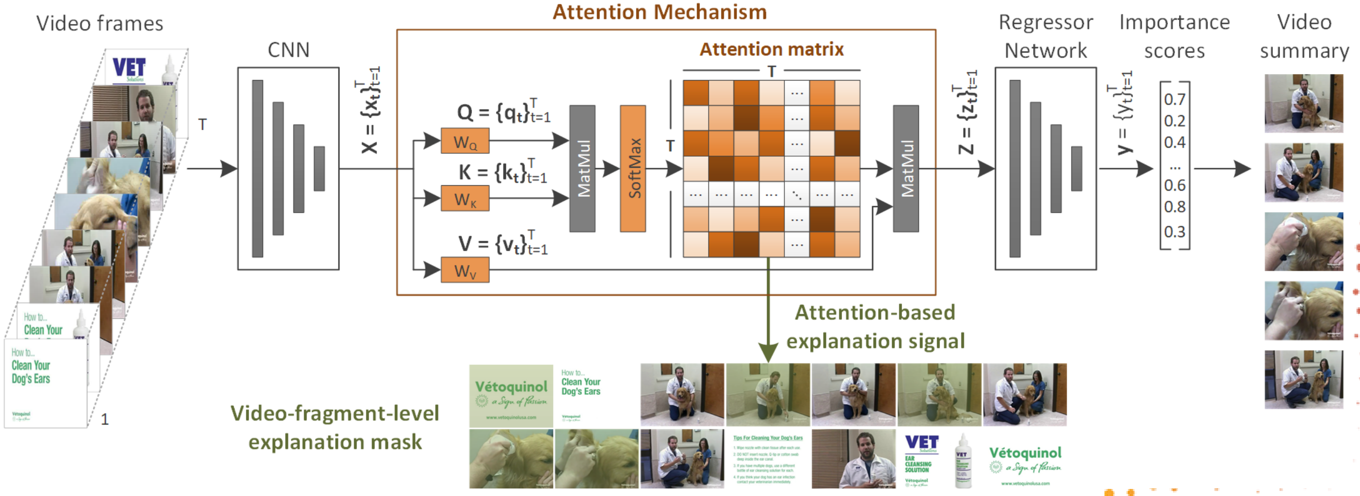


N. Gkalelis, D. Daskalakis, V. Mezaris, "ViGAT: Bottom-up event recognition and explanation in video using factorized graph attention network", IEEE Access, vol. 10, pp. 108797-108816, 2022. DOI:10.1109/ACCESS.2022.3213652.

N. Gkalelis, D. Daskalakis, V. Mezaris, "Gated-ViGAT: Efficient bottom-up event recognition and explanation using a new frame selection policy and gating mechanism", Proc. IEEE Int. Symposium on Multimedia (ISM), Naples, Italy, pp. 113-120, Dec. 2022. DOI:10.1109/ISM55400.2022.00024.

Video summarization

- Goal: generate video summaries out of an original, longer video
- Intended use: create videos suitable for various channels (e.g. social platforms)
- General approach (CA-SUM, XAI-SUM):
    - Video frames are represented using pre-trained CNNs (e.g., GoogleNet)
    - Video summarization networks estimate the frames’ importance
    - Given a video fragmentation and a time budget, the video summary is formed by selecting fragments that maximize its importance (Knapsack problem)
    - We examine possible explanation signals can be formed using the Attention matrix that is in the core of the summarization network

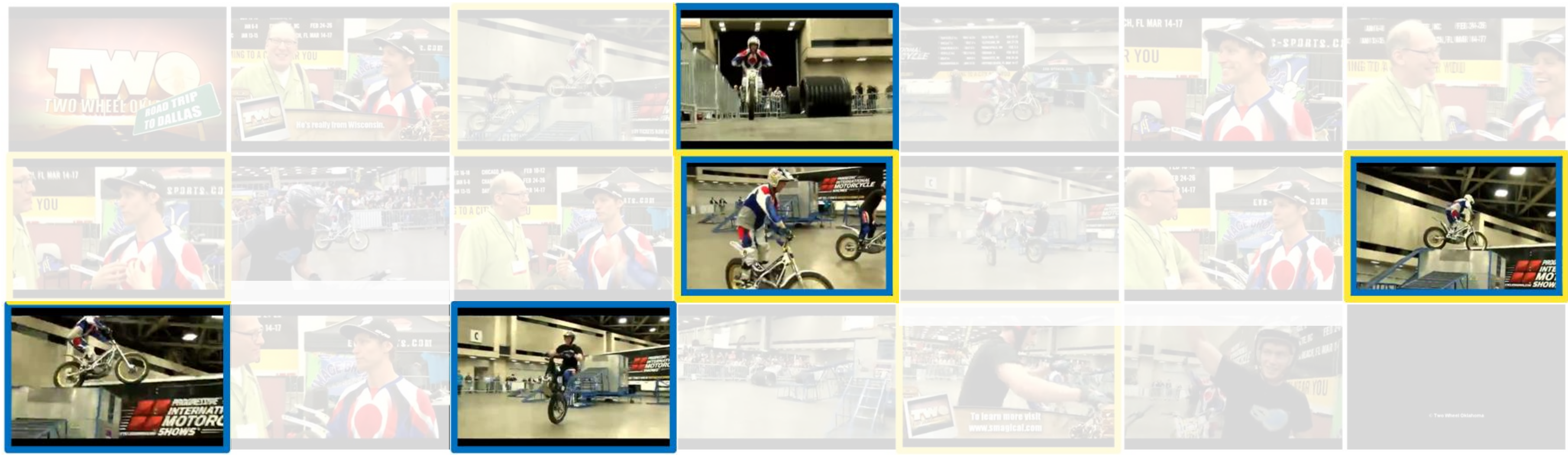


Work supported by EU projects H2020-832921 MIRROR, H2020-951911 AI4Media

Video summarization

- Summary of Motorcycle Stunt Show video, & explanation (good example)

Key-frame-based video overview. Blue boxes: summary (its top-5 fragments). Yellow boxes: explanation (top-5 influential fragments)



Video summary (blue)

- Associated with the motorcycle riders doing tricks (5 / 5 selected fragments)
- Contains visually diverse fragments (all are clearly different)

Attention mechanism (yellow)

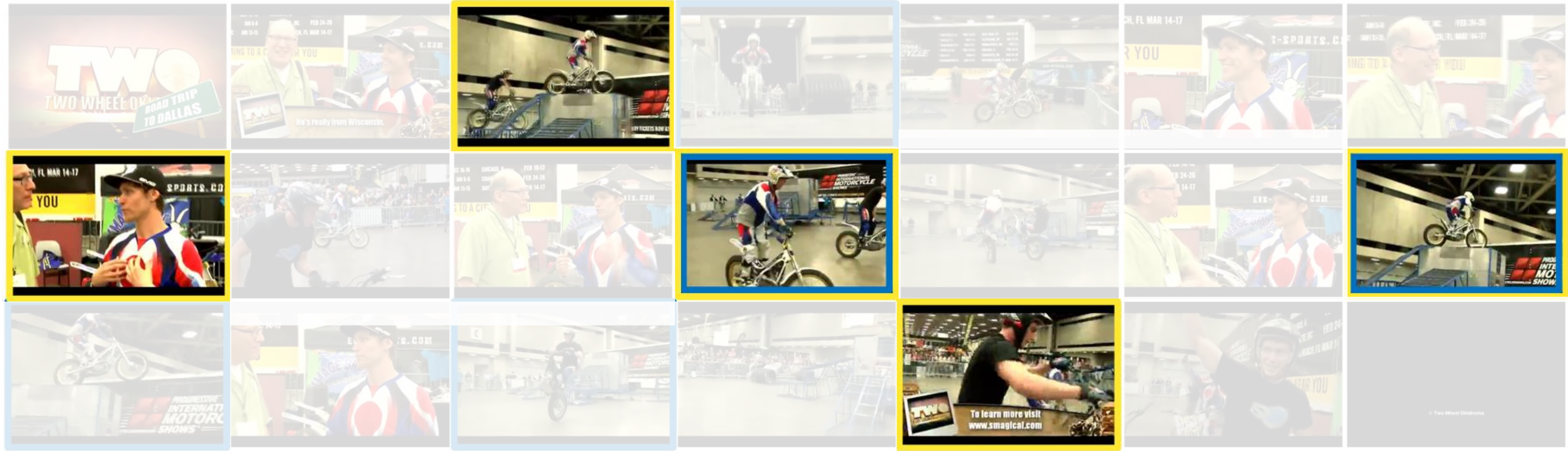
- Pays more attention on parts showing the tricks made by the riders
- Pays less attention to the logo of the TV-show and the interview

E. Apostolidis, G. Balaouras, V. Mezaris, I. Patras, "Explaining video summarization based on the focus of attention", Proc. IEEE Int. Symposium on Multimedia (ISM), Naples, Italy, pp. 146-150, Dec. 2022. DOI:10.1109/ISM55400.2022.00029.

Video summarization

- Summary of Motorcycle Stunt Show video, & explanation (good example)

Key-frame-based video overview. Blue boxes: summary (its top-5 fragments). Yellow boxes: explanation (top-5 influential fragments)



Video summary (blue)

- Associated with the motorcycle riders doing tricks (5 / 5 selected fragments)
- Contains visually diverse fragments (all are clearly different)

Attention mechanism (yellow)

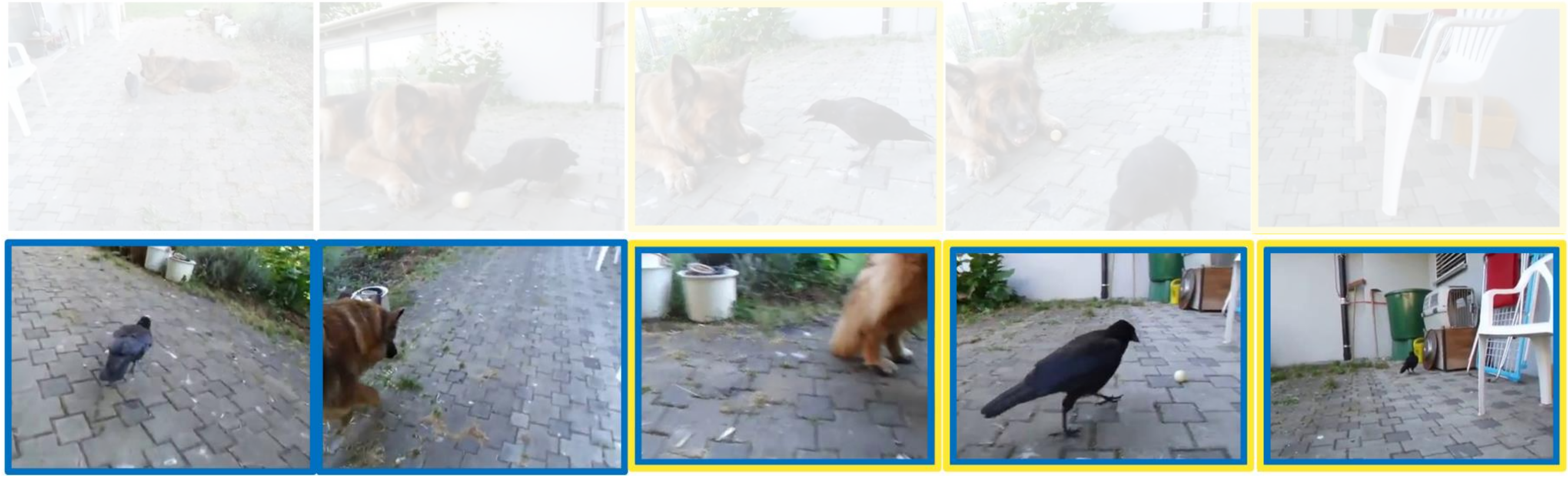
- Pays more attention on parts showing the tricks made by the riders
- Pays less attention to the logo of the TV-show and the interview

E. Apostolidis, G. Balaouras, V. Mezaris, I. Patras, "Explaining video summarization based on the focus of attention", Proc. IEEE Int. Symposium on Multimedia (ISM), Naples, Italy, pp. 146-150, Dec. 2022. DOI:10.1109/ISM55400.2022.00029.

Video summarization

- Summary of Playing ball video, & explanation (summarization failure case)

Key-frame-based video overview. Blue boxes: summary (its top-5 fragments). Yellow boxes: explanation (top-5 influential fragments)



Video summary (blue)

- Contains parts showing the bird and the courtyard (e.g., paving, chair)
- Misses parts showing the dog and the bird playing together

Attention mechanism (yellow)

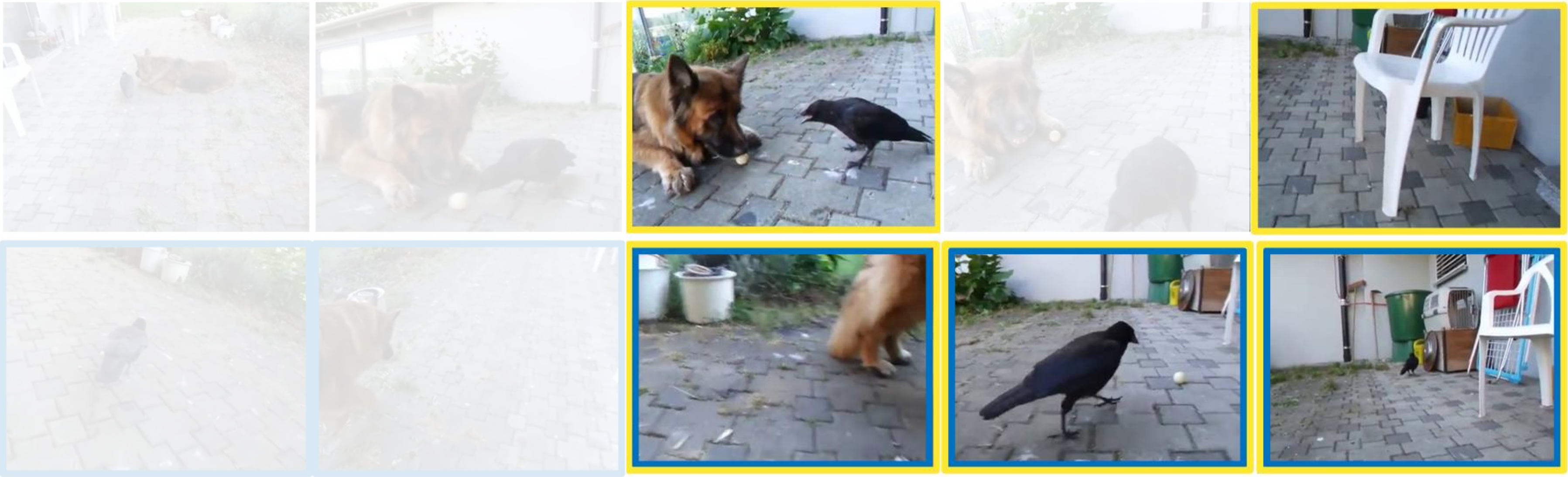
- Pays more attention on parts showing the courtyard (3 / 5 fragments)
- Pays less attention on parts showing the dog and the bird’s playing (1 fragment)

E. Apostolidis, G. Balaouras, V. Mezaris, I. Patras, "Explaining video summarization based on the focus of attention", Proc. IEEE Int. Symposium on Multimedia (ISM), Naples, Italy, pp. 146-150, Dec. 2022. DOI:10.1109/ISM55400.2022.00029.

Video summarization

- Summary of Playing ball video, & explanation (summarization failure case)

Key-frame-based video overview. Blue boxes: summary (its top-5 fragments). Yellow boxes: explanation (top-5 influential fragments)



Video summary (blue)

- Contains parts showing the bird and the courtyard (e.g., paving, chair)
- Misses parts showing the dog and the bird playing together

Attention mechanism (yellow)

- Pays more attention on parts showing the courtyard (3 / 5 fragments)
- Pays less attention on parts showing the dog and the bird's playing (1 fragment)

E. Apostolidis, G. Balaouras, V. Mezaris, I. Patras, "Explaining video summarization based on the focus of attention", Proc. IEEE Int. Symposium on Multimedia (ISM), Naples, Italy, pp. 146-150, Dec. 2022. DOI:10.1109/ISM55400.2022.00029.

AI4media

Follow us

@ai4mediaproject

www.ai4media.eu

info@ai4media.eu

Our Consortium

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911