Video Interpretation and Summarization
Prof. Dr. Jürgen Döllner, Jobin Wattasseril, Willy Scheibel
Computer Graphics Systems

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Video Interpretation and Summarization

Overview

Master's project focusing on design, development and evaluation of video interpretation and summarization technology:

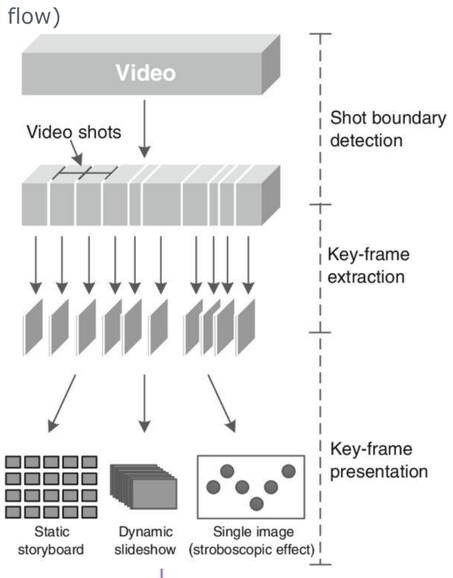
- **Goals**:
  - Develop algorithms for video analysis and content interpretation
  - Develop algorithms for visually and textually summarize video contents
  - Evaluate uses cases such as for content previews, incident reports, how-to videos
- **Project duration**: Winter Term 2024/2025, typically involving 2-4 students
- **Participants**: Masters students in IT-Systems Engineering, Data Engineering, Software Systems Engineering, Computer Science
- **Supervisors**: Prof. Dr. Jürgen Döllner, Jobin Wattasseril, Willy Scheibel

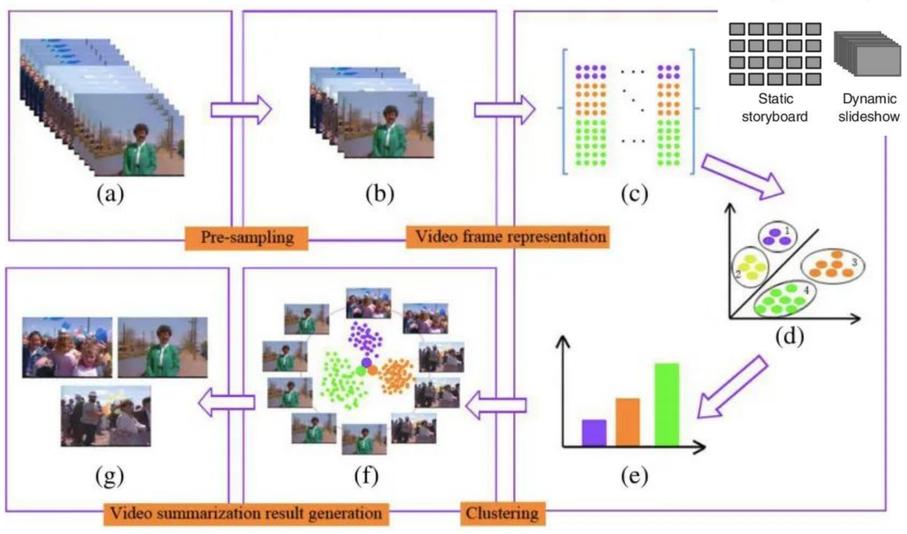
Sample application of video interpretation and summarization: Generating comics from videos.

1) Feature-Based Interpretation

Develop algorithms to extract representative keyframes and dynamic video summaries by merging video segments, based on:

- Low-level measurements (e.g., color histograms, motion/optical flow)
- Semantic measurements (e.g., people, objects, activities)
- Audio cues analysis (e.g., pitch, tone)
- Word frequency analysis and topic modeling
- Shot and scene segmentation
- Maximizing information coverage
- Highlighting important moments

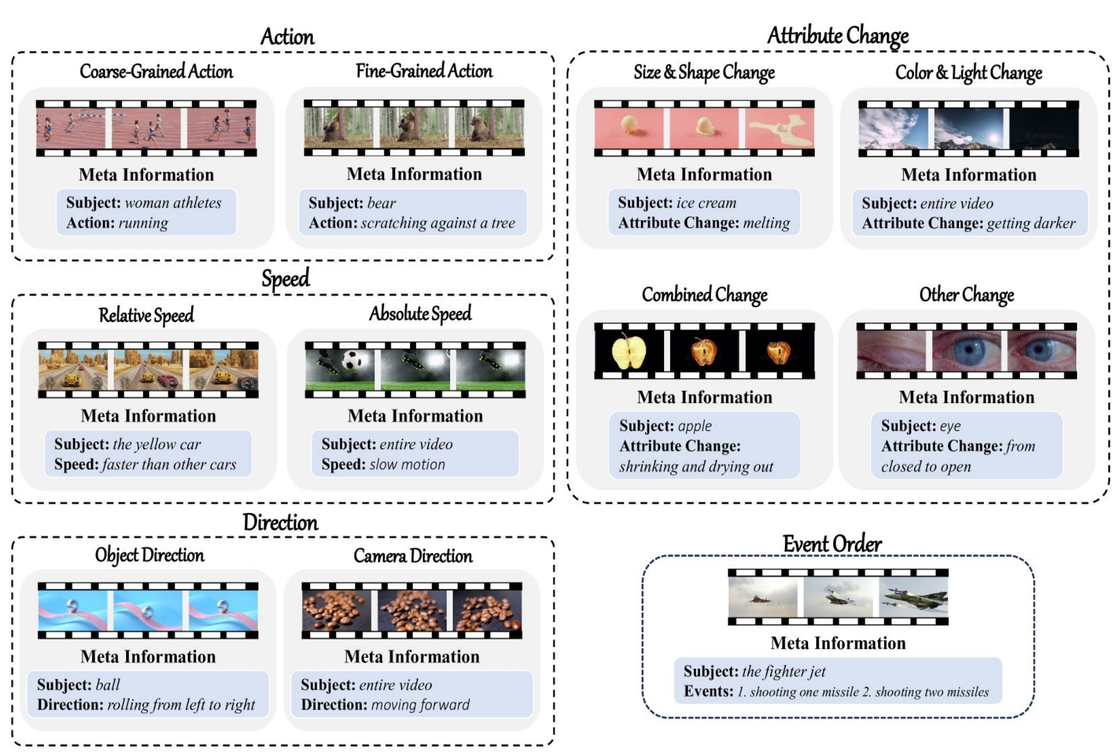




2) Text-Based Interpretation

Generate textual summaries describing video content using:

- Image/video captioning models
- Text summarization models or large language models (LLMs) for speech transcripts
- Multimodal LLMs for keyframe descriptions and visual question answering
- Retrieving and personalizing summaries based on user preferences and queries



3) Visual Video Summarization

Visualize results of video analysis by visual summaries:

- Storyboards with keyframes and captions
- Comic-style summaries with narrative balloons
- Audio narration of summaries using text-to-speech models
- Mosaic-based summaries for panoramic views of video shots

Technologies:

Analysis Component: Python, PyTorch, LLaMA, OpenCV, Hugging Face
Client Component: React, D3, gradio, plotly
Server Component: FastAPI, Flask