Spatial and temporal localization of objects and actions in videos using text and video analysis

Jan Neumann (StreamSage/Comcast)
Jana Kosecka (George Mason University)
Evelyne Tzoukermann (StreamSage/Comcast)

Multimedia content is a growing focus of search and retrieval, personalization, categorization, and information extraction. Video analysis allows us to find both objects and actions in video, but recognition of large amounts of categories is very challenging. Textual information is very good at describing objects and actions at a semantic level and often outlines the salient information present in the video. In this inter-disciplinary proposal we combine natural language processing, computer vision and machine learning to investigate how the semantic information contained in textual sources can be leveraged to improve the detection of complex actions.

This semantic information allows us to constrain the range of objects and actions that we need to search for in a given video to the semantically interesting set. This is especially helpful for activities that involve the interaction of humans and objects. The ability to identify such objects and the activities operating on them is a challenging computational problem. We expect that the use of language will sufficiently limit the range of possible objects in the scene, so that we can successfully detect and localize the objects and actions based on their appearance and dynamics, as well as the spatial and temporal relationships between them that correspond to the language description.

We propose to investigate how we can utilize textual descriptions of a video to tell us what objects and actions are present and need to be localized in a given scene. After parsing textual descriptions of a video segment into its constituents of verb-objects dependencies, we use lexical associations in knowledge databases to identify the set of words describing the objects and actions to localize in the video. We then query external image databases to gather visual exemplars of the objects and actions to train domain-specific classifiers. Finally, we plan to localize the objects and actions in the video using the trained classifiers as well as motion information, and graphical models encoding the relationships between objects and actions using graphical models.

For this workshop we will select one domain (e.g. kitchen or life-style show) and will evaluate our approaches on a hand-annotated ground-truth data set that will be provided to the community as part of the workshop.

StreamSage/Comcast offers resources that are critical to addressing these issues: (a) large amount of audio video content, including news, TV series, movies, music, sports, lifestyle shows, with associated metadata information; a large amount of tagged data at the scene level is also available; (b) topic detection algorithms using either closed captions or speech recognition output where several levels of granularity can be identified; and (c) a term-indexing system based on a word co-occurrence model that labels a segment of relevant content around each occurrence of a term.