

“Video, Language, and Contextual Reasoning-StoryLine Model for Video Understanding”

Jianbo Shi
University of Pennsylvania

Abstract:

There has been a significant interest in utilizing visual and contextual models for high level semantic reasoning in video. There are many weakly annotated images and videos available on the internet, along with other rich sources of information such as dictionaries, which can be used to learn visual and contextual models for recognition.

The goal of this work is to analyze videos of human activities not only by recognizing actions (typically based on their appearances), but also by determining the story/plot of the video. The storyline of a video describes causal relationships between actions. Beyond recognition of individual actions, discovering causal relationships helps to better understand the semantic meaning of the activities. We present an approach to learn a visually grounded storyline model of videos directly from weakly labeled data. The storyline model is represented as an AND-OR graph, a structure that can compactly encode storyline variation across videos. The edges in the AND-OR graph correspond to causal relationships which are represented in terms of spatio-temporal constraints. We formulate an Integer Programming framework for action recognition and storyline extraction using the storyline model and visual groundings learned from training data.

This is a joint work with Abhinav Gupta, Praveen Srinivasan, and Larry S. Davis