Images and videos can be naturally represented by graphs, with spatial graphs for images and spatiotemporal graphs for videos. However, for different applications, there are usually different formulations of the graphs, and algorithms for each formulation have different complexities. Therefore, wisely formulating the problem to ensure an accurate and efficient solution is one of the core issues in Computer Vision research. We explore three problems in this domain to demonstrate how to formulate all of these problems in terms of spatiotemporal graphs and obtain good and efficient solutions.

The first problem we explore is video object segmentation. The goal is to segment the primary moving objects in the videos. This problem is important for many applications, such as content based video retrieval, video summarization, activity understanding and targeted content replacement. In our framework, we use object proposals, which are object-like regions obtained by low-level visual cues. Each object proposal has an object-ness score associated with it, which indicates how likely this object proposal corresponds to an object. The problem is formulated as a directed acyclic graph, for which nodes represent the object proposals and edges represent the spatiotemporal relationship between nodes. A dynamic programming solution is employed to select one object proposal from each video frame, while ensuring their consistency throughout the video frames. Gaussian mixture models (GMMs) are used for modeling the background and foreground, and Markov Random Fields (MRFs) are employed to smooth the pixel-level segmentation.

In the above spatiotemporal graph formulation, we consider the object segmentation in only single video. Next, we consider multiple videos and model the video co-segmentation problem as a spatiotemporal graph. The goal here is to simultaneously segment the moving objects from multiple videos and assign common objects the same labels. The problem is formulated as a regulated maximum clique problem using object proposals. The object proposals are tracked in adjacent frames to generate a pool of candidate tracklets. Then an undirected graph is built with the nodes corresponding to the tracklets from all the videos and edges representing the similarities between the tracklets. A modified Bron-Kerbosch Algorithm is applied to the graph in order to select the prominent objects contained in these videos, hence relate the segmentation of each object in different videos.

In online and surveillance videos, the most important object class is the human. In contrast to generic video object segmentation and co-segmentation, specific knowledge about humans, which is defined by a pose (i.e. human skeleton) can be employed to help the segmentation and tracking of people in the videos. We formulate the problem of human pose estimation in videos using the spatiotemporal graph. In the formulation, the nodes represent different body parts in the video frames and edges represent the spatiotemporal relationship between body parts in adjacent frames. The graph is carefully designed to ensure an exact and efficient solution. The overall objective for the new formulation is to remove the simple cycles from the traditional graph-based formulations. Dynamic programming is employed in different stages in the method to select the best tracklets and human pose configurations.
The public is welcome to attend.