In this dissertation we address the problem of detecting humans and vehicles, tracking their identities in crowded scenes, and finally determining activities in aerial video. Even though this is a well explored problem in the field of computer vision, many challenges still remain when one is presented with realistic data. In this work, we propose a number of novel methods based on exploiting scene constraints from the imagery itself to aid in the detection and tracking of objects. First we tackle the problem of detecting moving as well as stationary objects in scenes that contain parallax and shadows. For finding moving objects we use median background modelling which requires few frames to obtain a workable model, and is very robust to having a large number of moving objects in the scene while the model is being constructed. We then remove false detections due to parallax and registration errors using gradient information of the background image. We propose a method for constraining the search based on a number of geometric constraints obtained from the metadata. Specifically, we obtain the orientation of ground plane normal, the orientation of shadows cast by out of plane objects in the scene, and the relationship between object heights and the size of their corresponding shadows. We utilize the above information to detect shadow casting out of plane (SCOOP) objects in the scene. These SCOOP candidate locations are then classified as either human or clutter. Additionally we combine a regular SCOOP and an inverted SCOOP candidates to obtain vehicle candidates. We also show that we can extend the SCOOP detection method to automatically estimate the orientation of shadow in the image without relying on metadata.

Next we propose a novel method for tracking a large number of densely moving objects in aerial video. In order to keep complexity of the tracking problem manageable when dealing with a large number of objects, we divide the scene into grid cells, solve the tracking problem optimally within each cell using bipartite graph matching and then link tracks across cells. Besides tractability, grid cells allow us to define a set of local scene constraints such as road orientation and object context. We use these constraints as part of cost function to solve the tracking problem which allows us to track fast-moving objects in low framerate videos.

Finally we propose a method for recognizing human actions in aerial video from as few examples as possible (a single example in the extreme case). We use the bag of words action representation, and use 1vsAll multi-class classification framework. We assume that most of the classes have many examples, and construct Support Vector Machine models for each class. We then use Support Vector Machines that were trained for classes with many examples to improve the decision function of the Support Vector Machine that was trained using few examples via late fusion of weighted decision values.
The public is welcome to attend.