In computer vision, context refers to any information that may influence how visual media are understood. Traditionally, researchers have studied the influence of several sources of context in relation to the object detection problem in images. In this dissertation, we present a multifaceted study of the problem of context. Context is analyzed as a source of improvement in the object detection problem, not only in images but also in videos. In the case of images, we also investigate the influence of the semantic context, determined by objects, relationships, locations, and global composition, to achieve a general understanding of the image content as a whole. In our research, we also attempt to solve the related problem of finding the context associated with visual media. Given a set of visual elements (images), we want to extract the context that can be commonly associated with these images in order to remove ambiguity.

The first part of this dissertation concentrates on achieving image understanding using semantic context. In spite of the recent success in tasks such as image classification, object detection, image segmentation, and the progress on scene understanding, researchers still lack clarity about computer comprehension of the content of the image as a whole. Hence, we propose a Top-Down Visual Tree (TDVT) image representation that allows the encoding of the content of the image as a hierarchy of objects capturing their importance, co-occurrences, and type of relations. A novel Top-Down Tree LSTM network is presented to learn about the image composition from the images and their TDVT representation. Given a test image, our network allows to detect objects and determine the hierarchical structure that they form, encoded as a TDVT representation of the image.

A single image could have multiple interpretations that may lead to ambiguity about the intentionality of an image. What if instead of having only a single image to be interpreted, we have multiple images? This question is the focus of the second part of this dissertation. In this part, we extend the TDVT representation to handle multiple images and develop a method that can automatically cluster similar images and generate a summary of the context associated with these clusters.

The public is welcome to attend.