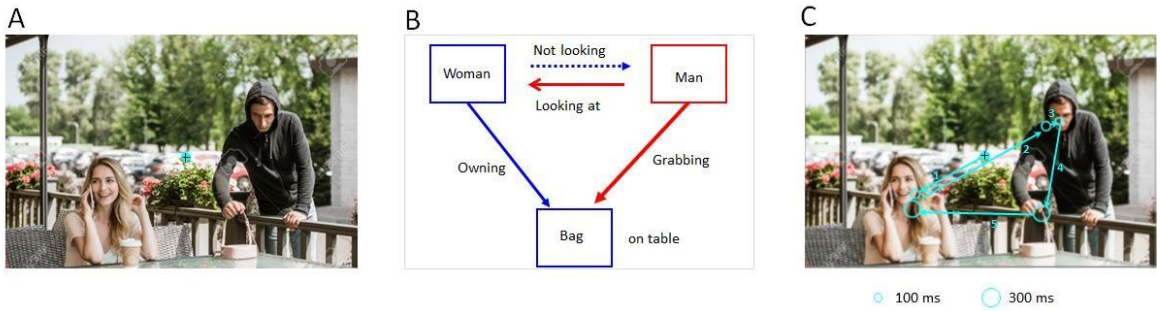
A natural visual scene typically enormous amounts of information, and only a fraction of it can be processed at any given moment, due to several factors (e.g. limited channel capacity, poor peripheral resolution, etc.). Yet, despite these severe constraints, our perception and comprehension of complex natural scenes is remarkably efficient. We can extract crucial elements of a scene (e.g. the physical layout and the individuals in the environment), from a mere glimpse. However, understanding *complex social interactions*, typically requires further analysis to put together all the scene details. Consider, for example, the *theft scene*, depicted below (in **A**). If the picture is presented for 100 ms and followed by a mask, observers typically report that the scene included two adults, a woman and man, in an outdoor scene. However, they often fail to understand the *scene narrative*. With further scrutiny (provided that the image is still present) they notice that the woman, who is the bag owner, is focused on her conversation, and thus cannot see that the man is grabbing her bag. By sequentially gathering all these clues (see **B)**, typically using fixations at critical elements in the scene (indicated in **C**), observers make sense of the scene



Such seamless scene understanding is the product of a complex visual image analysis involving multiple features (motion, color, texture, disparity, shading, etc) and a series of processing stages. It begins with extraction of contours and boundaries in *low-level* vision, followed by identification of 3D contours at *mid-level* vision, and *reaches full* visual cognition capacity (e.g. recognition of specific objects and agents, understanding of social settings, action preparation and outcome prediction) at the final stages of visual analysis.

To complicate matters the information flow is bi-directional, allowing for prior visual experience and context to affect the analysis of incoming visual information.