Recognition of complex events captured under real-world settings, has emerged as a challenging area of research across both computer vision and multimedia community. In this dissertation, we present a bottom-up approach towards an end to end framework for complex event recognition in widely available, consumer uploaded web-scale videos. We structure our approach into the following key stages and highlight our novel contributions in each of these stages: (a) Extraction of novel semi-local features, (b) Construction of intermediate representations from these features, (c) Integration of the intermediate representations into mid-level spatio-temporal concepts, and (d) Modeling temporal interactions between the spatio-temporal concepts, to recognize complex events. Promising results achieved through extensive experiments demonstrate the efficacy and importance of each stage in the proposed computational approach to obtain state-of-the-art performance in complex event recognition.