Artificial social intelligence is a step towards human-like human-computer interaction. One important step towards building socially intelligent systems, is enabling computers with the ability to process and interpret the social signals of humans in the real world. Social signals include a wide range of emotional responses from a simple smile to expression of complex affects.

This dissertation revolves around computational models for social signal processing in the wild, using multimodal signals with an emphasis on the visual modality. We primarily focus on complex affect recognition with a strong interest in curiosity. In this dissertation, we first present our collected dataset, EmoReact. We provide detailed multimodal behavior analysis across audio-visual signals and present unimodal and multimodal classification models for affect recognition. Second, we present a deep multimodal fusion algorithm to fuse information from visual, acoustic and verbal channels to achieve a unified classification result. Third, we present a novel system to synthesize, recognize and localize facial occlusions. The proposed framework is based on a three-stage process: 1) Synthesis of naturalistic occluded faces, which include hand over face occlusions as well as other common occlusions such as hair bangs, scarf, hat, etc. 2) Recognition of occluded faces and differentiating between hand over face and other types of facial occlusions. 3) Localization of facial occlusions and identifying the occluded facial regions. Region of facial occlusion, plays an important role in recognizing affect and a shift in location can result in a very different interpretation, e.g., hand over chin can indicate contemplation, while hand over eyes may show frustration or sadness. Finally, we show the importance of considering facial occlusion type and region in affect recognition through achieving promising results in our experiments.