MultiSense - Context-aware Nonverbal Behavior Analysis Framework: A Psychological Distress Use Case

Giota Stratou, Louis-Philippe Morency

Abstract—During face-to-face interactions, people naturally integrate nonverbal behaviors such as facial expressions and body postures as part of the conversation to infer the communicative intent or emotional state of their interlocutor. The interpretation of these nonverbal behaviors will often be contextualized by interactional cues such as the previous spoken question, the general discussion topic or the physical environment. A critical step in creating computers able to understand or participate in this type of social face-to-face interactions is to develop a computational platform to synchronously recognize nonverbal behaviors as part of the interactional context. In this platform, information for the acoustic and visual modalities should be carefully synchronized and rapidly processed. At the same time, contextual and interactional cues should be remembered and integrated to better interpret nonverbal (and verbal) behaviors. In this article, we introduce a real-time computational framework, MultiSense, which offers flexible and efficient synchronization approaches for context-based nonverbal behavior analysis. MultiSense is designed to utilize interactional cues from both interlocutors (e.g., from the computer and the human participant) and integrate this contextual information when interpreting nonverbal behaviors. MultiSense can also assimilate behaviors over a full interaction and summarize the observed affective states of the user. We demonstrate the capabilities of the new framework with a concrete use case from the mental health domain where MultiSense is used as part of a decision support tool to assess indicators of psychological distress such as depression and post-traumatic stress disorder (PTSD). In this scenario, MultiSense not only infers psychological distress indicators from nonverbal behaviors but also broadcasts the user state in real-time to a virtual agent (i.e., a digital interviewer) designed to conduct semi-structured interviews with human participants. Our experiments show the added value of our multimodal synchronization approaches and also demonstrate the importance of MultiSense contextual interpretation when inferring distress indicators.

Index Terms—MultiSense; System for affective computing; Behavior quantification; Automatic distress assessment; Framework for multimodal behavioral understanding

1 Introduction

Advances in affective computing have made real world applications that perceive and react to the affect of the user a reality. Some promising steps have been taken and there already exist a few systems targeting real applications that not only detect the affect of a human interlocutor but also respond to the sensed affect, therefore closing the *affective loop* [1]. For example, the Affective AutoTutor [2] is an intelligent system that senses states of affect as they relate to the learning experience and combines that with the cognitive states of a user to promote learning and engagement. The TARDIS project [3] aims to build a platform that will help young unemployed population train for job interviews. They have proposed virtual humans that are able to sense and react to the nonverbal input of the user as an interface for job interview simulation scenarios. The Affective Music Player [4] is an application that can provide entertainment and affect the user's mood by choosing songs to induce either calm or energized mood on demand. A validation experiment demonstrated in a real-world office setting that the use of physiological responses (skin conductivity in that case) can be used in real-life affective computing applications. Other examples are mentioned in recent reviews of affective computing systems [5], [6].

As we can see, these emerging technologies span across different domains and use cases and demonstrate that real-world applications that sense and influence the affect of the user are possible. However, there are still a lot of obstacles to overcome; understanding and interpreting human behavior during natural interactions remains a challenging problem. Human communication is by nature multimodal, including but not limited to expressions by facial, vocal, postural and gesture activity. These signals serve important intrapersonal and interpersonal functions and they are attributed causality or meaning based on other information at hand such as the context in which they were expressed [7]. As an example, smile is generally affiliated with positive affect but showing a smile in a negative situation can be perceived as cold and unemotional [8]. A grounding theory behind this is that people naturally form links about the typical relationships between the traits of a situation and emotions expressed, and when observing a new individual's emotional reaction to a situation they rely on that knowledge to infer aspects of the individual's inner state and goals by reverse engineering the underlying appraisals [8], [9].

Even in specialized scenarios such as the healthcare domain, expert clinicians integrate contextualization into their diagnostic methods by looking at discriminative reactions of a patient [10] (anecdotally, clinicians often observe patients in the waiting room to form a reference or *baseline* and sometimes apply different triggers to assert emotional reactions). The reason behind this is that certain psychological illnesses have been found to alter emotional reactivity patterns, for example positive attenuation and negative potentiation have been observed in depression [11].