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

Giota Stratou and 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

- *G. Stratou is with the Institute for Creative Technologies, University of Southern California, Playa Vista, CA 90094. E-mail: stratou@ict.usc.edu.*
- *L.-P. Morency is with Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: morency@cs.cmu.edu.*

Manuscript received 28 Aug. 2015; revised 30 Aug. 2016; accepted 11 Sept. 2016. Date of publication 27 Sept. 2016; date of current version 6 June 2017. Recommended for acceptance by D. Novak, G. Chanel, A. Koenig, and P. Guillotel. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TAFFC.2016.2614300

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].

Evidently, people learn through experience to naturally integrate multimodal signals, context -by reverse appraisals-, baselines and population filters into a coherent interpretation when perceiving nonverbal behaviors. It follows that automatic behavioral analysis should leverage those same concepts to improve the behavioral understanding elements in applications of affective computing. As a matter of fact, in recent reviews of affective multimodal human computer interaction systems [12], [13], [14], [15], [16] researchers unanimously acknowledge that an ideal automatic system for affect analysis, among others, should be: multimodal, population independent, context-sensitive and dynamics-sensitive. Furthermore, much discussion has arisen on the methods of multimodal fusion and different levels of analysis [5], [15], [17], [18]. Researchers have investigated benefits in fusing modalities in different abstraction levels (e.g., data-, feature-, decision-level) without clear or application independent consensus on which one is better, and very often hybrid fusion techniques emerge from this process [15].

Regardless of this unanimous demand and research activity in this direction, there are still some critical challenges to be dealt with. These challenges have scientific aspects such as fusing different modalities in a flexible way, integrating context and dynamics, and also technical aspects such as desire for speed and robustness of the affective system [14]. Sometimes, some of the technical requirements are deprioritized in research prototypes, but they gain importance when transitioning into a real application that will interact with population in the wild. Moreover, the transition of research products (e.g., complex behavioral models) from the lab to a real-time application is not always straightforward because it is a multi-component process that involves the use of different tools, representations or methods.

We propose a unified approach to behavioral analysis for real world applications by encompassing all domain related functionality (e.g., sensing, tracking, signal interpretation, behavioral model implementation etc.) into one framework. Such a framework will provide a flexible platform that will not only allow exploration of the scientific challenges in the domain of automatic human behavior inference, but also facilitate the transition of the research products into real-world affective computing applications by combining solutions for the technical challenges in the same time. One of the fundamental elements behind this is to support multiple solutions for each of its component (e.g., different types of sensors as plugins, different methodologies for fusion for the data interpretation etc.) while maintaining a unified frame of reference and representation. The product of our approach is a framework for multimodal behavioral analysis and quantification, called MultiSense. This framework and all of its components was first developed during of a long term project for the detection and computational analysis of nonverbal behaviors in healthcare applications and with the aim to support a larger variety of applications beyond that scope. With this work we introduce the MultiSense framework with its components and we demonstrate specific capabilities related to affective computing challenges through its initial use in the mental health domain, where it is used as a 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 infers psychological distress indicators from nonverbal behaviors utilizing context information from the interaction and 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. We will discuss the specific usage as well as how this usage extends to other applications.

In the next section, we review relevant work on i) multimodal behavioral analysis frameworks and ii) automatic behavior analysis in the scope of psychological distress. In Section 3 we briefly describe the framework and how its different components address the functional needs of real-time multimodal analysis in a human computer interaction setting. In Section 4 we present the development process using the framework in the case of the psychological distress analysis application and the results of the automatic distress assessment. We go over the results and discuss the framework use in Section 5 and finally, we deliver our conclusions in Section 6.

2 Related Work

In this section we present the most relevant prior work. Specifically, we will review related work on frameworks for behavioral analysis in human computer interaction settings and on automatic analysis of behaviors in the healthcare domain which is the scope of our real-world application.

2.1 Frameworks for Behavioral Analysis

In this work, we use the term framework to refer to a layered structure of software that addresses a specific domain. As frameworks for behavioral analysis we will review specific software that can be single-component or multiple-component and that address functionalities needed for human sensing and behavioral analysis (e.g., either capturing data of, interpreting, representing, evaluating computational models for human behavior or combinations of those functionalities). Different application needs have prompted different architecture design and focus for each system. From traditional multi-agent distributed architectures [16], [19] recent frameworks are developing to unified platforms that provide templates to build the modules themselves [20], [21].

In the recent years, there has been a substantial interest in developing flexible platforms for multimodal sensing and interaction and certain aspects of these functionalities have been investigated and implemented by existing software. For example, focusing on capturing/recording multimodal streams in a synchronized way, or receiving new modalities as input (like EEG) seem to be very popular among the

research community. Several commercial software are now providing options for multimodal platforms following a modular architecture, such as the iMotions Attention tool,[1] RealSense SDK,[2] and specializing in the synchronized recording aspect such as NOLDUS,[3] These platforms are designed with a user friendly interface and are suited for collecting training corpora. Other frameworks, such as OpenInterface [22], EyesWeb [23], Mudra [24], MINT [25], SoFMIS [26], support considerable set of inputs (at different modalities) and use rule-based, graphical programming environment for constructing multimodal pipelines, but are not specifically equipped for machine learning pipelines. ELAN [27] is another example which supports multimodal input but it's main focus is annotation rather than real-time interaction. World Wide Web Consortium (w3c) is also creating a standard for multimodal architecture and interfaces for the web.[4] Also, on the communication side for collaboration and exchange of information between modules there has been work on defining a protocol fit for the different types of exchange in a unified framework [28], [29] Some of these tools mentioned above are very specialized, but are not flexible for all modalities (especially in terms of real-time synchronization), usages and complex pipelines.

One very well rounded multimodal framework is HCI2 [19] which is based on a Publisher/Subscriber model on top of a message system like ActiveMQ. Another broadly used framework is the Social Signal Interpretation framework (SSI) [30] based on pipeline architecture. Both of these have managed to overcome some of the limitations like being well equipped with real-time multimodal interaction as well as analysis (with feature extraction) and have some advantages like generalizability in the architecture to handle different kinds of signals (images, audio streams etc). No type of architecture choice is superior. Both pipeline as well as publish-subscribe have pros and cons. While on one hand P-S facilitates complex systems, on the other hand, pipeline architecture facilitates synchronization. In all cases, various modes of synchronization, data fusion and framerate control remain challenging issues. Although the existing frameworks solve a lot of the infrastructure and performance issues, they often lack the high level functionality to deal with the different types of information fusion (perhaps with the exception of Wagner et al. [30]). In general, flexibility on fusion techniques is needed to integrate input from multimodal and multicue sources [16].

Promising work on that direction has been reported by the TARDIS project [3], the scope of which includes the development of virtual humans that are able to sense and react to the nonverbal input of the user. Their architecture includes a sensing framework (based on SSI [30]) and automatic analysis of behaviors implemented by the Nova Framework [31] which combines work on annotation tools with technologies to automatically analyze human behavior, offering the novelty that segmentation and labeling of the data happens completely automatically. The work reported in the scope of TARDIS framework deserves special mention because it is one of the few other efforts that take on multiple aspects of the multimodal behavioral analysis challenges, focusing in the same time on logging of data, offline annotation, training methods and real time sensing capabilities towards a real application [32]. Another notable mention is the SEMAINE API [21], which is a framework for building emotion-oriented systems. This is a human-machine interaction framework with special focus on affective computing applications and deserves special mention because it is one of the first efforts towards simplifying and generalizing the building of pipelines among applications by encouraging standard representations and building blocks that can be used as plugins.

We find our motivations aligned with the aims of these last pieces of work in the sense that we want a flexible platform and components that can be customizable to leverage domain specific knowledge in applications but general enough that can have larger usability and appeal. Our approach is to encompass a lot of those functionalities related to the behavioral capture and analysis domain in one common framework, where we can specifically address challenges in behavioral analysis towards real applications. Some of the aspects we focus on, in terms of framework functionality is the ability to support multimodal or multi-cue behavioral analysis in real-time applications and integrate context. The result is a multimodal framework called MultiSense, which includes components for real-time analysis and multi-sensor input build on top of a framework base by SSI [20], a standard logging scheme and a unit specifically designed for computational models for behavioral analysis. This framework includes API that formalizes building of pipelines, internal communication between modules and communication towards external components, thus maintaining a common/standard representation among functionalities, and facilitates the development of complex real-time applications in human computer interaction.

2.2 Automatic Analysis of Behaviors in Healthcare

Recent advances in sensing technologies in combination with wide spread accessibility and adoption of new sensors (like the Microsoft Kinect[5]) has spawned an interest in developing applications that can measure the state of the user for real world applications. Therefore, domains such as healthcare have fostered sensor-based applications that target, among others, patient treatment (e.g., rehabilitation simulation scenarios [33]) or symptom measurements (e.g., measuring movement in Parkinson disease [34]). Moreover, as affective computing methods come in to the mix, they extend the role of these sensing technologies from pure body capture (physical state) to something more, that is, understanding the affect and the underlying psychological state of the user. The medical community is increasingly embracing the importance of nonverbal communication in clinical settings [35], [36]. Experienced clinicians have learned to observe and read a patient's nonverbal behaviors as part of their diagnosis [37]. This is particularly important for psychological conditions where sometimes trauma mainly manifests in the altered behaviors and affect of the patients.

1. http://imotionsglobal.com/software/attention-tool-core/
2. https://software.intel.com/en-us/intel-realsense-sdk
3. http://www.noldus.com/human-behavior-research/products/media-recorder-0
4. http://www.w3.org/TR/mmi-framework/
5. https://www.microsoft.com/en-us/kinectforwindows/

In the last decade, there has been particular interest in using automatic analysis in medicine, to aid diagnosis, monitor and even treat psychological conditions [38], [39]. Furthermore, technological advances aid in the aspects of delivery and interface of clinical practices. Virtual health agents [40], [41] have been developed for screening of patients and information dispersal to people in need. Virtual humans have of course their limitations but have been shown to bring certain advantages in such interactions, such as a relative feel of anonymity and less judgment that can increase disclosure [42].

In this cross-section of the fields of clinical psychology and automatic methods for interfacing and assessing a patient, contextualization becomes even more important. On the clinical side, alterations in appraisal mechanisms and reactions to stimuli have been particularly diagnostic to certain conditions. For example, recent cognitive research suggests that depressed individuals may appraise emotional stimuli differently than non depressed persons, suggesting a diminished ability among affected persons to experience positive emotion [43]. Also in a healthcare scenario, investigations show that facial responses to positive stimuli (e.g., a smile) could be regulated by the strength of the stimuli [43]. Previous work on the automatic analysis of participant interviews explores contextualization based on the intimacy and polarity levels of the questions asked to the participants: [44], [45] finding improvements in a context-dependent model for assessment of psychological conditions. Therefore, it is important for an automatic system to log and integrate context into the analysis of patient behaviors.

With this work we present our framework for nonverbal behavior analysis, called MultiSense. We demonstrate the capabilities of this framework and its different components through an application in the healthcare domain, for which it was initially developed. We employed this framework for real-time behavioral analysis, where it not only utilizes contextual and interactual cues to improve the inference of behavioral indicators but also takes part in the interaction by broadcasting current state of the user to influence the interaction. With this framework, we leverage the work that has been done in the automatic assessment of psychological distress, and integrated a behavioral inference module into a real system that converses with the user and reports an assessment of his/her distress level at the end of the interaction, as a diagnostic aid.

3 MultiSense Framework

By the term framework we refer to a layered structure of software that includes core components, API that formalizes building application pipelines, communication between internal modules and communication towards external components. Our goal is to unify a lot of the functionalities that relate to human behavior analysis in a human-computer interaction setting under a common framework, in order to provide a flexible platform for research exploration and simplify the development process of applications. Such a platform will lead to technological advances that are generalizable to more than one applications, which is a desirable feature for the community [5], [14]. In Fig. 1 we show a representation of the basic human - machine interaction loop (as discussed in [15]) and show the main functionality that we would want such a framework to cover in different parts of this loop. In this work we focus on the aspects of the loop that are being supported by MultiSense framework (i.e., the domain of human behavior sensing and inference), but a more global view of how the different components of the architecture work together towards a fully automatic agent system can be found in the description of the VHToolkit framework [46].

In this section, we will briefly describe the framework and its components (Section 3.1), we will discuss how to use the framework in a typical development process (Section 3.2), highlight certain technical innovations that make this framework suitable for real world applications in affective computing (Section 3.3) and finally, we will go over integrated technologies and current utilization of the framework in various projects (Sections 3.4 and 3.5).

3.1 Framework Overview

The MultiSense framework specifically, includes the following components:

i) a core API, based on SSI library, for building pipelines that facilitates extensions to support various sensors and plugin technologies.
ii) formal representation of behavioral human state and communication protocol based on PML that allows communication between framework components and external applications.
iii) a database, *MultiSense DB*, with extendable schema to log behavioral signals and interaction context.
iv) a message to database transformer, *MDx*, that logs messages in the database. This component works both live (real-time) and offline for batch processing of existing interaction logs.
v) the *MBU* (Multimodal Behavioral Understanding) unit, for behavioral indicator computation. This unit includes an easy API to query the database and can easily summarize, contextualize, fuse multimodal behavioral signals and infer higher level behavioral indicators (by evaluating machine trained models). This unit can also work both offline and real-time and can broadcast inferred information back to the network using the same communication scheme.

The various components of MultiSense framework cover different functional elements related to human behavior analysis in the human computer interaction loop as seen in Fig. 1.

3.2 Functional Use: Large Scale Development Process with MultiSense

A typical development process for affective agent interaction systems involves the following stages: i) study of real human interaction; this either includes existing behavioral data from relevant domains or ideally a F2F (face-to-face) data collection with a protocol similar to the final application target, ii) data analysis and development of behavioral models targeting the specific application (e.g., inferring agitation of a patient in a healthcare scenario, or inferring agreement from nonverbal in a negotiation setting) and iii) implementation of those models in a human computer interaction application. Sometimes there is a WoZ phase

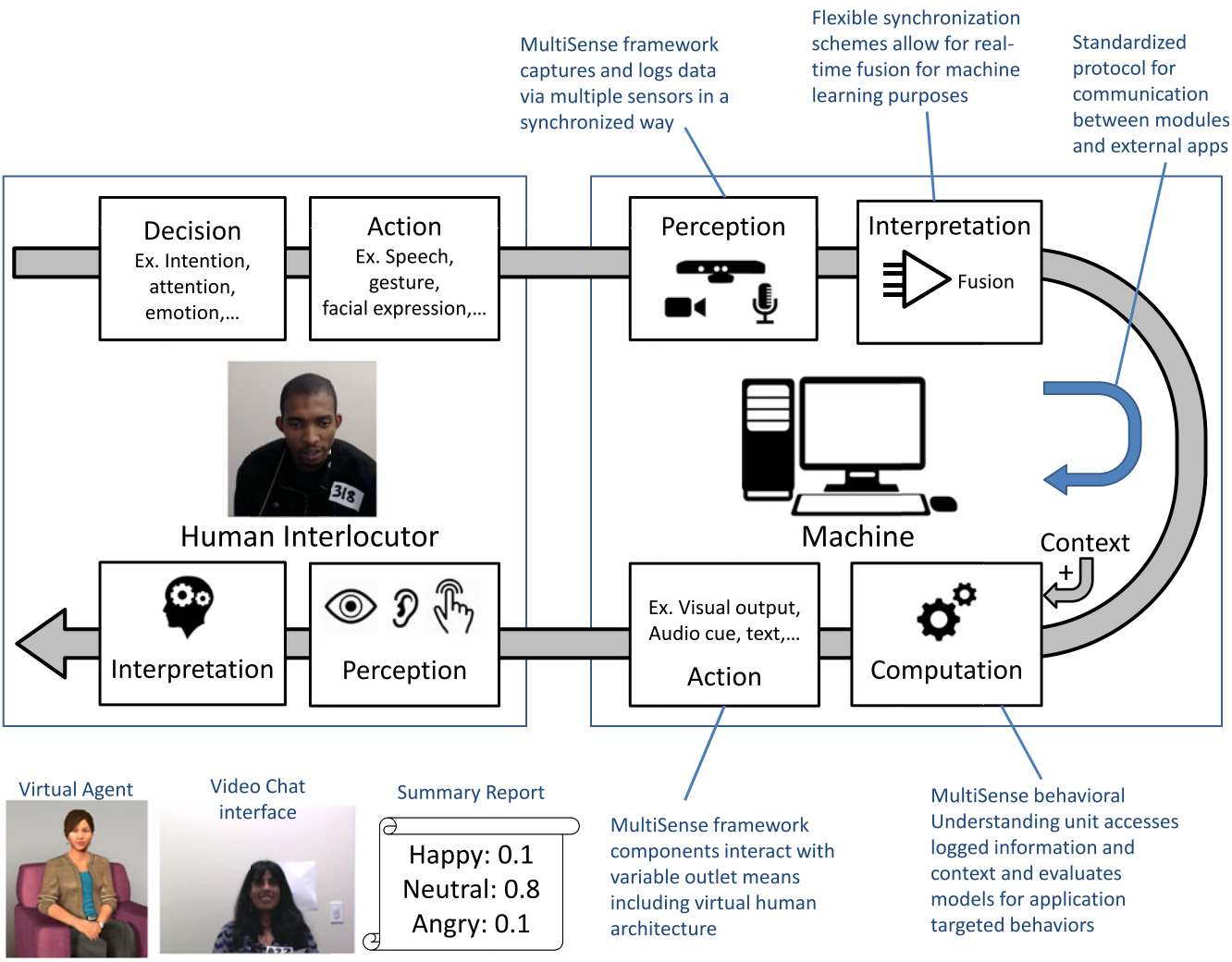

Fig. 1. Overview of MultiSense framework desired capabilities with respect to the human-machine interaction loop (inspired by [15]). *Perception* happens via different sensors, integrated as individual modules to be enabled on demand, which provide multimodal streams that come into the pipeline in a synchronized way. *Interpretation* happens in the middle layers of the pipeline where information from the sensors and trackers is fused in a meaningful way to form higher constructs (either by machine learned models or by rule based methods). This information is communicated via a standardized protocol and logged together with context (e.g., interaction events). *Computation* for behavioral inference happens in different levels, both in intermediate modules of the pipeline for online behaviors, and in the MBU unit which is specifically designed as a parallel process to the pipeline, to query information from the database and implement context based analysis. All the computation modules merge their results back to the system via the same communication protocol and the system then forms an assessment about the state of the user that can be used for *Action*, that is response, back to the user, via selected interface (e.g., text report, virtual agent, or visualization).

(Wizard-of-Oz) before moving to the fully automatic system, where the interface of the agent is established, but the AI is still controlled by humans (wizards).

MultiSense framework can facilitate this development process by offering key functionalities and applications for the different phases of behavioral capture, analysis and model implementation. In Fig. 2 we highlight some of those key applications, built with MultiSense components. Specifically, the *MultiSense AVRecorder* (Fig. 2, top row) is a framework application that features synchronous multimodal stream logging. This simple application is customizable for different sensory input (for example, we offer the options of audio, video, Kinect sensor and biopack plugins working in parallel) and is very useful for synchronous collection of multimodal data. *MultiSense Offline Replayer* is processing video and audio channels offline through the behavioral analysis pipeline. This pipeline can be seen in Fig. 2, middle row, where again, different tracking technologies can be plugged-in as independent or cooperating components in the tracking block; automatic behavior feature extraction happens in the MBU unit and learned models can be developed and tested there. These two applications are usually employed during the first phases of a project involving behavioral analytics, where systematic collection of data and offline analysis is prominent.

For the development of real-time human-computer interaction applications, one can use a pipeline like the bottom one, in Fig. 2, where tracking and behavioral model assessments are happening live and can be broadcasted to the AI components of the system. Specifically, *MultiSense LIVE* is an application featuring live tracking with most modules already connected to each other in a functional way, including fusion of information, encoding and broadcasting at the end of the pipeline as an output. The benefit of MultiSense LIVE is that it includes an easy configuration scheme to enable/disable modules and functionalities such that any high level user can customize it and run it on demand. At this stage, MultiSense MBU can evaluate live

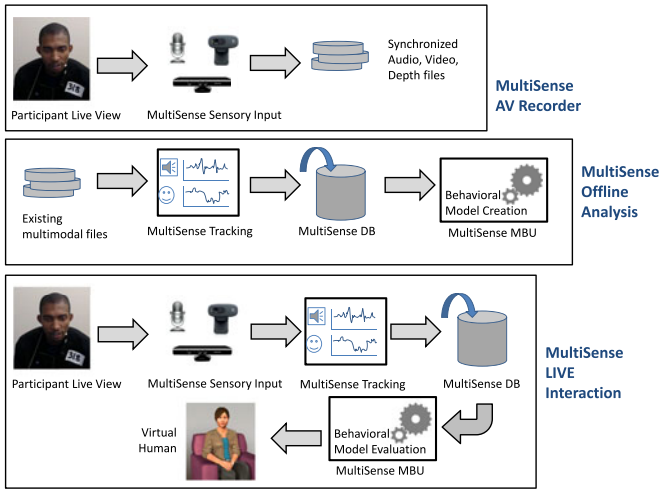

Fig. 2. Basic pipeline view for example usage of MultiSense components in real applications. One should note that the same basic components are re-used both for offline analysis, where we learn a model of behavior, and live implementation. Common blocks and representation facilitate the transition from data driven learning to a live application.

the learned models of behavior and broadcast those back in the live application.

The usage of the same components allows for easy transition from research to implementation and real applications can benefit directly from research output. The main point that we would like to emphasize in the scope of this work is that *a well designed framework that is flexible to serve different functional roles during research, development and final deployment, but maintains a common representation throughout the process, can facilitate the transition between research and real world implementations*. In other words, it is important that behaviors are extracted, logged and analyzed under a common framework and common representation throughout this process. Moreover, the common representation and communication scheme between those components allows for other interesting applications, where, for example, one can re-play directly behaviors logged in the database into a live system, effectively simulating the recorded affect of the user.

3.3 Technical Innovations

Our approach is to develop the core components in a way that they can support customization to fit application specialized needs and help explore different scientific challenges in a global level. However, maintaining the flexibility to address all those functionalities with the same tools is challenging and required the design and implementation of specific features.

3.3.1 Flexible Synchronization and Framework Control

For example, handling input serially during recording or offline analysis is not always the best mode for real-time interaction scenarios where responsivity and short latency are important factors [47]. Providing constant framerate is not crucial for visualization, but becomes important when evaluating a sequential behavioral model that has been trained with controlled input.

The live component of our framework implements effective parallelism to address some of these demands. One of the most functional features in MultiSense is the choice for different synchronization modes, made available via the framework API, which allows for different capabilities. We achieve that by simulating a publish-subscribe functionality in the way modules grab information from input modules using thread notifications in a multithreaded pipeline architecture. Specifically, event based notifications from threads allow us to design different modes and synchronization schemes: *Wait for all*, *Wait for one*, *wait for any*.

Including such capabilities makes possible to implement different control of information flow which is customizable in different parts of the system and allows for effective parallelism in a pipeline architecture. This enables applications with multimodal and multicue (e.g., 3 different facetrackers with different confidence levels) information streams running in parallel, which is useful for building different constructs or diverse emotion models [5]. Additionally, this allows to employ input-driven performance in selected modules that interface with the user, where fast or real-time flow of information is preferable and helps avoid perceptible discontinuities, or *framing*, that disrupt the flow of the application and are potentially distractive to the user. In contrast, in other parts of the system (such as the sensing level) where temporal grabbing consistency is preferred, the framework provides control over a constant framerate. This is especially important for the effective evaluation of machine learned models in the MBU component. Other hybrid framerate-controlled pipeline examples, such as a slow tracker with higher accuracy informing a faster one without compromising the return framerate of the system, have been implemented with the MultiSense framework.

3.3.2 MBU API

Another important feature of the framework is the design of the MBU unit. The MBU's goal is to simplify computation and inference of behavioral indicators by integrating multimodal information on a temporal and contextual level. Specifically, MBU offers an interface to pre-design queries that implement common feature encodings from the database, so it presents the user with an easy API to request features that summarize behaviors. Table 1 has an overview of the signals and possible encodings that have been implemented in MBU and can be extracted in real-time. Most of these signals are based on integrated technologies described in Section 3.4. The flexibility in the design of features and capability of loading different behavioral models makes this one of the main components that can specialize framework functionality in terms of behavioral analysis for specific application needs.

The MBU unit can be configured to work in both offline and real-time modes in combination with the MDx and DB components. The formalization of feature encoding between those two modes allows for a straightforward transition from research prototypes to real application models.

Moreover, the MBU functionality can be linked to the overlaying VH architecture. Specifically, throughout the framework, we are utilizing the Perception Markup Language (PML) [29] protocol to organize the system output into a formal representation and a messaging system to broadcast tracking and inferred results to other components

TABLE 1
Overview of Feature Set That Can Be Extracted Automatically by MultiSense Framework

		Category	Details	Description
SIGNALS	Affect	Smile		plugin: OKAO
		Emotion	ANGER, SADNESS, CONTEMPT, SURPRISE, DISGUST, POSITIVE, JOY, NEGATIVE, FEAR, NEUTRAL	Basic emotion labels plugin: FACET
		AUs	1, 4, 6, 7, 9, 12	Facial action units plugin: FACET
	Gaze	EyeGaze	VERT(UP/DOWN) HORIZ(SIDE) DIR(AWAY)	Eyegaze signals plugin: OKAO
	Body	HeadPose	Px, Py, Pz Rx, Ry, Rz	Head position and orientation plugin: CLNF
		Skeleton	HEAD, NECK SHOULDER (L/R) COLLAR (L/R)	Joint Positions and orientations plugin: MS Kinect
	Speech	Speech	RATE, VAR, ENERGY, FRACTION, FUNDFREQ	Speech dynamics plugin: Emovoice+
		Prosody	NNET0-3	Prosody features plugin: COVAREP
	Dialog	Dialog	RESPONSERATE	Latency to response (from database)
ENCODING		Simple	AVG	Average
			STD	Standard Dev
			RAT	Ratio (activation fraction)
		Composite	DIFF	Difference
CONTEXT		Phase	on conversation phase	application dependent
			(none)	no phase (all interaction)
		Questions	on interaction cue	application dependent

In the three sections of the Table we present the SIGNALS that were captured automatically based on MultiSense LIVE plugins, the possible ENCODINGS that are implemented in the MBU unit and CONTEXT examples implemented for behavior contextualization.

utilizing VHmsg.[6] The latter, also allows incoming messages from external components to interact with MultiSense modules. Specifically, in terms of behavioral analysis, we allow external VH architecture components (e.g., the dialog system) to log interaction events in the same database and this enables the MBU to access interaction information for contextualization of behavior (see context-based encoding of features in Table 1, based on interaction cues or specific timings). In the same time, inferred behaviors can be broadcasted by the MBU to the external VH architecture through the same messaging system. The framerate of the output packages is configurable by the API, and can be easily changed depending on application needs. For example, in a virtual human avatar scenario, MultiSense can broadcast low level tracking information such as position of the head and facial features in 30 Hz, whereas in a virtual human interaction scenario, where the virtual human needs to know if the person is paying attention, MultiSense can broadcast higher level inferences computed by the MBU, such as attention, in a lower framerate to inform the interaction. These two elements effectively facilitate the usage of MultiSense as a part of a larger distributed system and in terms of behavioral analysis, allow context-based model design in the MBU.

6. https://confluence.ict.usc.edu/display/VHTK/VHMsg

3.3.3 Context Base Modeling

This being an important part of the design specifications of our framework, we would like to clarify the usage of the term context in the scope of our presented work. In general, context represents the set of circumstances or facts that surround a particular event or situation and as such is often a vague term by nature. For example, when we refer to the context of an interaction influencing the perceived behaviors, this could mean either the global scope of the interaction (e.g., the goal of the interaction: think of a date interaction, versus a clinical interview), or the local scope (what is the conversation state right now, or what was the previous question asked), and sometimes it can even be loosely used to describe other intangible elements that surround an interaction such as population bias (e.g., under what culture, racial profile or gender is a certain behavior being observed). It is important to observe that all of those types of context are crucial towards interpreting a natural interaction and by framework design we should ideally be able to handle all of those cases, if we wish to research the scientific importance of context in behavior understanding. Context as the global scope is usually attached to an application and often exists inherently in the behavioral models that were trained in a specific application. Context in the local scope sense, can be utilized by logging the specific events of an interaction. Finally population context could be given as a parameter to the behavioral analysis module to target specific participants.

To give an example, in Fig. 3 one can observe different couplings of expressions (or facial reactions) with questions to which they respond to. Most literature and empirical data suggest that by having knowledge of the questions when interpreting the facial expressions of a target person, the assessment of the state of said person changes.

To address context-based modeling, our framework is directly integrate-able in the virtual human architecture, as described in the previous subsection. This allows of systematic event and interactional context logging. The MBU component references this information and also has the capability of storing reference population statistics for modeling and decision making in specific domains. We will address contextualization in the scope of the healthcare

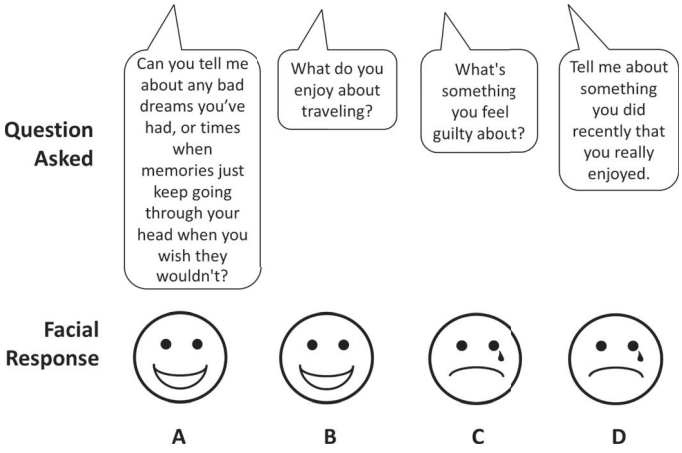

Fig. 3. An oversimplified example where the context of the conversation (in this case question asked) can influence the interpretation of a facial expression. For example, for both cases A and B the facial response is exactly the same, but after reading the questions asked most people would find reaction A to be atypical. Similarly, comparing cases C and D, perhaps most people would find reaction D to be atypical in general population. Such atypical patterns of behavior could sometimes be indicators of a different psychological state.

application in the next section, since it is an important feature of our experiments.

3.4 Integrated Technologies

MultiSense is currently being used in different applications where various plugin software have been integrated and tested. Those include licensed technologies such as OKAO vision[7] face tracker, Emotient SDK, with FACET[8] emotion and facial action unit (AU) recognizer, Microsoft Kinect skeleton tracking,[9] Cogito Audio Compute Library,[10] and Biopack SDK, plus the open-source modules that are publicly shared with the SSI base such as Emovoice,[11] and SHORE facetracker [48]. Also the list includes open source shared software such as GAVAM headtracker [49], CLNF facetracker [50], real-time hCRF library for machine learning,[12] FAAST[13] action coder, CLM-CSIRO facetracker[14] and COVAREP audio toolbox [51]. Improvements and synchronization schemes in the core made possible to have all these technologies running in parallel in a single 6 core computer with a 25 Hz framerate output of broadcasted information, and real-time visualization of user state and tracked behaviors (an example of which can be seen in Fig. 4).

3.5 Utilization of MultiSense Framework

Flexibility in application design and tracking technology plugins has allowed for different functional usage of the framework components. In the context of detection, computation and analysis of psychological signals MultiSense framework has been used in a novel application that targets Telemedicine and assists the clinician directly as an end

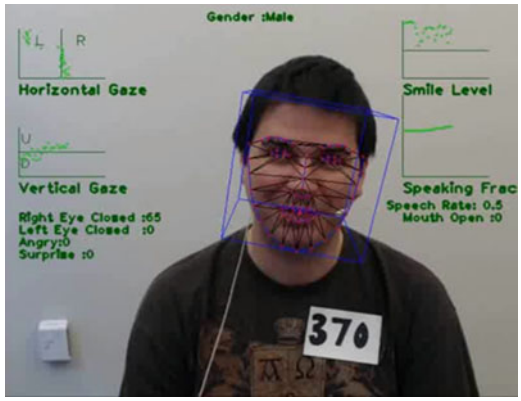

Fig. 4. MultiSense LIVE tracking visualization. Here we see an overlay of the live tracking by MultiSense, as used in SimSensei, on the participant camera view. The participant's tracked state is logged for analysis and broadcasted to the rest of the architecture for interactional purposes.

user. With this application, *Telecoach* [52], tracked behaviors are quantified and presented to the clinician on an interface over a tele-mediated interaction (in this implementation it was a Skype call) and can inform him/her of diagnostic nonverbal events from the side of the patient and potentially improve his/her assessment [53].

Also, the multimodal framework and architecture is used in other domains like for public speaking training [54], [55] or tutoring systems. In these cases as well, it offers general user state behavior summaries and context depended assessments. For example, in the case of the public speaking training, MultiSense LIVE can track the behaviors of the presenter and give live feedback to a virtual audience which will react accordingly (e.g., presenter looking down a lot will decrease the audience attention) [56]. Also, it quantifies important behaviors that are needed for user feedback (ex. the presenter was smiling only 2 percent of the time). MultiSense is also integrated in a prototype tutoring system, PAL3, providing affective feedback to the system components, making the agents reactive to the student, and influencing the flow of learning tasks via AutoTutor.[15]

An existing integration of MultiSense Live application in a virtual human architecture is also available and openly shared with the VHToolkit,[16] adding perception messages which enable user behavior feedback to the virtual agents. Finally, specific applications such as the *AVRecorder* for synchronized recording are actively being shared with the research community and used for collection of corpora in different domains such as clinical [57], training systems for negotiation or public speaking [54], [58] etc. In most of those cases, the AVrecorder and offline analysis MultiSense applications are used for collection of multimodal data and automatic annotation of behaviors.

4 Use Case in Healthcare

In this section we will focus on the utilization of the MultiSense framework in a real application in the healthcare domain, for which our framework MultiSense plays an instrumental role. We will give a brief overview of the

7. http://www.omron.com/r_d/coretech/vision/okao.html
8. http://www.emotient.com/products#FACETVision
9. http://www.microsoft.com/en-us/kinectforwindows/
10. http://www.cogitocorp.com/
11. http://www.informatik.uni-augsburg.de/lehrstuehle/hcm/projects/tools/emovoice/
12. http://sourceforge.net/projects/hcrf/
13. http://projects.ict.usc.edu/mxr/faast/
14. http://face.ci2cv.net/
15. http://ict.usc.edu/prototypes/personal-assistant-for-life-long-learning-pal3/
16. https://vhtoolkit.ict.usc.edu/

Fig. 5. SimSensei interface. A side by side view of the participant (left) as captured by the system's camera and Ellie (right) the virtual human who conducts the interview.

application, *SimSensei* (Section 4.1), and we will focus on the part that MultiSense played in the development process (Section 4.2) as well as present behavioral analysis results in this domain (Section 4.3). For the behavioral analysis we will emphasize in the importance of a context-based approach achieved via the MultiSense framework.

4.1 Automatic Indicators for Distress

The SimSensei system is a fully automated agent designed to create an engaging face-to-face interaction with a user over a semi-structured interview related to psychological distress. The system was developed over a two year iterative process [59], in the scope of a larger project investigating the detection, computation and analysis of psychological signals.

The SimSensei interface can be seen in Fig. 5 side by side with one of the participants. For the purposes of this work we will mention that the system was designed with two main goals: 1) to create an engaging environment where people can express themselves in a natural way and feel free to open up about their problems and 2) to create an environment for natural interaction rich in expressed behaviors that can indicate psychological distress such as depression or post-traumatic stress disorder (PTSD). The intent of such a system is to give additional access to clinical resources to people that seek it and do not have access to a doctor, or that are hesitant to contact one because of the stigma associated with seeking therapy. Besides opening the door to the clinical process, the automatic system can also serve as a diagnostic aid to a clinician, by automatically assessing behaviors during the interview and detecting nonverbal indicators of psychological distress. The system in its final form provides a behavioral report and distress assessment of the user right after the interaction is completed.

As a part of the SimSensei project, a large database was collected of people giving interviews about distress in the different phases of the system development [57]. Participants were recruited from two distinct populations living in the Greater Los Angeles metropolitan area, veterans of the U.S. armed forces and from the general public, and are coded for depression, PTSD and anxiety based on accepted psychiatric questionnaires. The system had a good reception and participants felt safe to share information and express themselves. A study on user feedback shows that the anonymity of a fully automatic system, developed as a final stage of SimSensei caused participants to feel less fear to disclose and express more facial displays of sadness among others [42].

4.2 Development Process

SimSensei, followed the typical development process mentioned in Section 3.2, and described in detail in relevant SimSensei publication [59]. We can summarize the roles of MultiSense framework as follows:

First, MultiSense AVRecorder was the main means of multimodal data collection in all phases of data collection [57], by employing synchronized recording of video, audio, skeleton and depth data.

Second, MultiSense was the framework for automatic analysis and behavior quantification, in all stages of analysis. In the initial phases of the systems, the audio and video of the participant would be processed offline by MultiSense to extract nonverbal behaviors. The first phases were very important in the design of the behavioral cue inference aspects of the final system because they provided information on which behaviors are most informative for assessing psychological distress. For example, automatic extracted behaviors referring to audio quality [44], facial expression and head movement [60] or verbal channels [61] and gestures from automatic and manual annotation [62] were investigated in relation with psychological conditions such as depression and PTSD in the first phases of the system (face-to-face and Wizard-Of-Oz). Then followed research exploration on how to best integrate those behaviors into models of multimodal behavioral indicators for distress [45], [63], [64].

Based on this analysis phase, MultiSense framework components were customized to support a live implementation of such models for the final system. Specifically, the MultiSense LIVE tracking component was customized to include the necessary tracking technologies that were needed for analysis running live, and the MBU component was tuned to encode the features in the necessary format and context. This phase was important in defining the feasibility and fine tuning of certain feature extraction and model evaluation methods in a real system.

Finally, and moving towards the fully automatic phase of the system, MultiSense MBU was linked to the rest of the architecture in the way we described in the earlier section. MultiSense helped enhance the interaction by providing live feedback of the user tracked behaviors to the SimSensei agent. The virtual human on receiving this information by different parts of its AI, could both branch the conversation in a different dialog path or generate appropriate nonverbal behavior. For example, this enabled Ellie to smile back at the user and headnod when the user was pausing, among other behaviors, thus closing an affective interaction loop with the user based on nonverbal behaviors. Symmetrically, SimSensei components were informing the MBU about the context of the interaction by logging relevant conversation events, allowing for live evaluation of context-based models.

An overview of the behavioral analysis and the other functional roles (such as interaction enhancement and recording) that MultiSense played in the SimSensei system can be seen in Fig. 6. There, we present the human computer interaction loop instantiated for the case of the SimSensei application and specifically as it relates to the behavioral analysis feedback that the user receives at the end (the distress report).

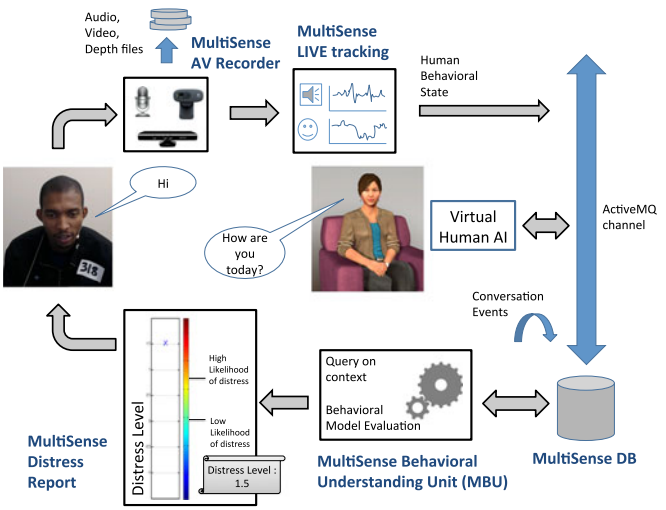

Fig. 6. Behavioral analysis overview on the SimSensei system. Instantiation of the affective computing loop utilizing the MultiSense framework in the SimSensei architecture and particularly as it relates to the distress analysis.

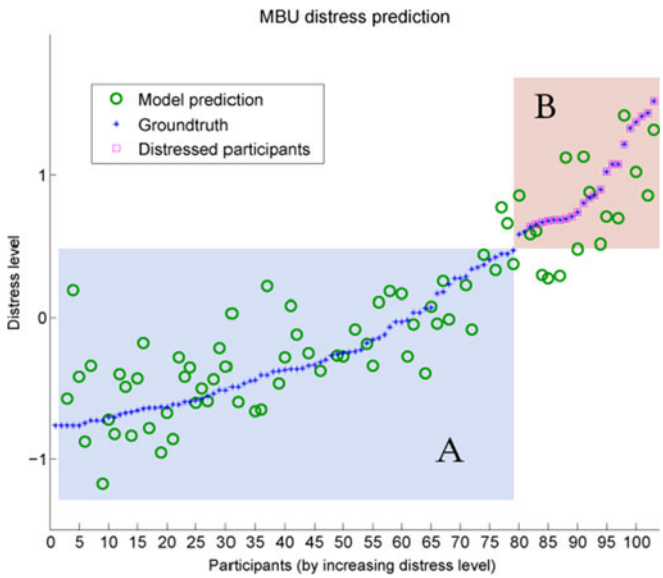

Fig. 7. Distress level output. This is a visualization of the performance of the final model that utilizes context and gender information. The results are plotted by increasing level of groundtruth distress. Correlation between groundtruth and prediction is r = 0.8822. The model returns distress level estimates that distinguish the non-distressed and distressed population (highlighted in areas A and B respectively). This is a good indication that the distress level predictor can also give a distress label assessment.

4.3 Distress Analysis

4.3.1 Automatic Features and Context Integration

Based on the integrated technologies described in Section 3.4 and the MBU feature encoding functionality we were able to design a rich collection of features (Table 1).

The basic encodings that were used can be seen in Table 1. Besides the standard average (AVG), standard deviation (STD) and ratio of activation (RAT) over timeslot, we also provide interface to request linear combination of signals (ex. difference of two signals). Based on the MBU interface, besides holistic features (features that refer to the whole interaction, independent of the conversation state), we were able to contextualize behaviors based on different interaction events, specific to the SimSensei application.

In the scope of this work, we accept that the global context of a clinical interview with a virtual human is common among all our participants and is inherent in the distress model we created based on those interactions. Also, population bias is usually difficult to handle automatically but one aspect of it, gender, that is possible to assess automatically, has been investigated in previous work [60] and being utilized in our current model as well. In our current investigation, we will focus mostly on the local context of an interaction, that is the conversation or dialog phase. Most literature and empirical data suggest that by having knowledge of the questions when interpreting the facial expressions of a target person, the assessment of the state of said person changes. Specifically, in the case of clinical interviews, atypical reactions to certain questions can be indicators of certain psychological state.

In our analysis, we will utilize the different conversational phases that were defined in the interaction by system design: *Intro* (for rapport building, where simple get to know you questions were asked to get the participant comfortable with the system), followed by *Hot* phase (where the intimate and potentially more provocative questions were asked that referred to traumatic experiences, relationship with the family, sleeping patterns etc.) and ending with the *Cool* or Cooldown phase (where participants were asked some positive hopefully uplifting questions to end the interaction in a good spirit). Also, since we are dealign with a structured interview we can also consider particular questions as specific events. In this analysis we added the *LHappy* question (*"When was the last time you felt really happy?"*) as a single question phase, based on a preliminary study that showed that this question was eliciting diagnostic behaviors.

4.3.2 Results

To demonstrate the effectiveness of our framework design specifications in the healthcare scenario, we will present the final model on automatic distress assessment in the SimSensei application, implemented via the MultiSense MBU unit. This model exists in the real system and a report of the results can be produced with the push of a button after the interaction. With our experiments we aim to show the value of the context based design in the model performance.

These experiments refer to one of the latest stages of development of SimSensei, where 100 participants from general and U.S veteran population interacted with the fully automatic version of the system. The groundtruth for the *distress level* is a construct created by depression and PTSD scores combined together, as discussed in previous literature [62].

The method used is a linear regression model[17] with 25 terms, that were chosen automatically from the pool of all possible multimodal features by a greedy forward selection method. The prediction was tested for 100 participants in a Leave-One-Out testing manner. The results of the Leave-One-Out testing are demonstrated in Fig. 7, reporting correlation 0.8822 with the groundtruth distress label and a total

17. http://www.mathworks.com/help/matlab/data_analysis/linear-regression.html

TABLE 2
Comparison of Context-Based and No-Context Models, Measured by Model RMSE and Correlation

	RMSE	Correlation	R-squared
No-context	0.370	0.7448	0.738
Context-based	**0.263**	**0.8822**	**0.868**

We show that context-based modeling improves performance.

RMSE of 0.265. Moreover, our model manages to separate the non-distressed and distressed population (see Fig. 7, areas A and B).

To further showcase how context-based methods improve the performance of our model we compare with a model that only uses holistic features, namely features encoded over the whole interaction. In this case the method was exactly the same, but the feature pool to choose from was restricted to only holistic (no phase) features, whereas in the first case all types of features (both holistic and context-based) were included. The results can be seen in Table 2, where we show both in terms of total RMSE and in terms of prediction correlation that utilizing context information in the interaction and adding phase encoded features to our model improves performance.

Further analysis into the feature terms that were selected in the linear model allows us to make some interesting observations about the effect of selected terms and the way they were combined to contribute to the model. First, behaviors encoded on the phase level were preferred. Second, we observed pairwise matching of certain behaviors on different phases, *usually with opposite effects*. As an example we present the effect of model terms STD_POSITIVE_ACTV and STD_AU7_EVID both in COOL_PHASE, and HOT_PHASE in Fig. 8). This demonstrates that the linear model often benefits from the difference of certain behaviors between phases of the interaction, finding discriminative information in the differential of observed behaviors rather than on holistic representation of behaviors.

In Table 3 we show the first five features that were selected by the greedy forward method for both cases of the full (context-based) model and the no-context model. Phase encoded features seem to be preferred when available.

5 Discussion

In this section we would like to discuss the different aspects of framework design and utilization in a real world application. We will also mention some additional information about the real system.

On the Framework and Real-World Applications. We demonstrated how the specific components and functionalities of MultiSense framework support aspects of multimodal behavioral sensing and analysis in a real world application. We would like to make a note of our empirical observations that transitioning research prototypes to real-world utilization is not always a straightforward process and we believe there is still a long way to go in terms of both technology and scientific challenges, to transition most research prototypes towards a robust real world application.

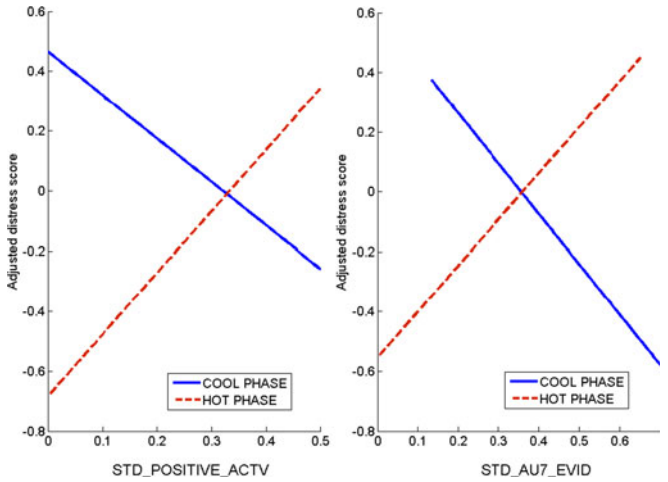

Fig. 8. Effects of selected model terms, highlighting the utilization of different phases in the final model. In the adjusted response plots from the distress model certain features have been selected from more than one PHASE. In those cases the linear model learned from the difference of those features, assigning opposite coefficients to the PHASE terms.

A lot of these challenges are obviously hampered from the desire that a framework should be generalizable for different usage and applications. Focusing the framework usage into a specific application type would resolve some of these issues, however, we strongly believe that creating a common platform to serve different functional aspects in different applications will help explore different scientific challenges and will lead to technological and scientific advances that are more generalizable to other applications (e.g., behavioral models that can be implemented and applied to other scenarios). The same point stands even when referring to the development phase of one application, transitioning from a research prototype where data are annotated and models are trained, to a real-world automatic system. We believe that maintaining a flexible framework for this process facilitates and speeds up the launching of new applications. There is a long way to go, but if we start by technically unifying the way we collect data and the way we analyze data and we implement models for real-time systems then this is a step for our systems to be more generalizable, easier to extend to data and cases in the wild as opposed to only laboratory settings.

Moreover, we would like to mention that this work focused on the importance of contextualization in the model and the framework features that allow such analysis, but to bring this model in a real-world scenario there were additional considerations in feature robustness and in the design. Specifically, we evaluated the model with incomplete data

TABLE 3
Five First Features Selected by the Greedy Forward Method for Each of the Two Models Described in the Experiment, in Order of Addition to the Model

Context-based model	No-context model
AVG_SURPRISE_INTS_LHAPPY_PHASE	AVG_SURPRISE_EVID
STD_AU7_EVID_HOT_PHASE	STD_NEGATIVE_INTS
STD_POSITIVE_ACTV_HOT_PHASE	AVG_AU7_EVID
STD_SURPRISE_ACTV	STD_NNET_FEAT1
STD_POSITIVE_ACTV_COOL_PHASE	STD_HEADR_Rz

and made sure we always get a response (minimize lost cases) based on automatic tracker confidence. These are important elements to consider when implementing a model in a real application with natural interactions.

Limitations. MultiSense framework was mainly developed under the scope of the healthcare interview application. Although, its components have been designed to be customizable and it supports a span of applications already, there are certainly still improvements to be made to reach a point where it can support any general application related to behavioral analysis. Its original development spawned with live interactions that require a complex behavioral analysis network in mind. We believe that it serves that role well, but the multithreaded architecture is probably not optimal for simple, one tracker pipelines and should adjust appropriately. Similarly, the core architecture is not optimal for mobile use (with limited cores) and would require adjustments to serve a mobile application that requires live behavioral analysis. One of our current goals is to expand this to mobile usage through a server architecture, so a lot of the core components are currently being re-evaluated for this purpose. As another note, while high level programming (e.g., to enable/disable modules or to interpret outgoing messages) is easy and user friendly with MultiSense, lower level customizations (e.g., creating new functionality or integrating a new tracker) rely on understanding the core API which is more complex, so it would need developer effort. Finally, manual annotation has not been addressed fully in this framework yet. In an ideal behavioral analysis framework we would like to allow for manual annotations in the same multimodal streams of information that is being recorded and processed. A UI that will allow this, in our opinion, would be a great addition to the current framework functionalities.

On the Behavior Analysis. We covered in the introduction how reverse appraisal theories support the usefulness of context for a better interpretation of observed behaviors. Moreover, we would claim that our results hint on something more, and that is, that analyzing observations (behaviors) from the same person under multiple situations can be more informative. This makes sense in an applied behavioral analysis level because of the notion of *behavioral baselining*, meaning that behaviors may make more sense when observed as differences from an initial state rather than looking at them at one instance in time, in a similar manner to how physiological signals are treated [65]. Humans are naturally integrating discriminative information (for example, we judge someone's personality better when observed under different situations, rather than in one context) and in the healthcare domain, expert clinicians sometimes integrate this into their diagnostic methods by looking at discriminative reactions of a patient [10], [11].

Another point to be made, since we are using automatic methods, is in an algorithmic level. Analysis of patterns often relies on discriminative models, and given people's idiosyncratic features we can imagine that the automatic techniques perform better in picking up differences rather than absolute measures (some facial recognition softwares, e.g., FACET [66], even support baseline input of a neutral face for improved performance). This may be another reason why it is important to contextualize features by phases of an interaction for automatic behavioral analysis.

Given the trade-off between generalizability and context utilization we tried in this work to stay on an interpretable set of subconstructs, utility and meaning of which can be applied to other domains. This is facilitated by the framework that offers common representation of behaviors.

On the Distress Analysis. To the best of our knowledge, this work represents one of the first fully automatic systems that conducts a clinical interview and delivers behavioral assessment reports. As discussed in the section about context, the implemented models refer to this application and inherent the global context of the interaction with the virtual human, Ellie, by design, so it is fair to only think of them in this scope. However, we presented the thought behind the design and implementation of a flexible framework that will hopefully enable in the future to unify those results with results from other analyses in the same domain, towards cohesive observations about nonverbal behaviors in distress.

6 Conclusion

We presented a framework that is designed to support multimodal affective computing applications and computer interactions. We designed the framework with the aim to support context based analysis and flexible design of applications as a platform for researchers to investigate and advance on the challenges that confront multimodal systems. We demonstrated through the SimSensei application from the healthcare domain that such a framework can be used in a real-world scenario, by supporting the implementation of a fully automatic system that conducts interviews with patients and both helps distressed people open up about their problems and logs information that can be used in a diagnostic way in a model implementation.

Acknowledgments

This work is supported by DARPA under contract (W911NF-04-D-0005) and by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred. The authors would also like to thank: Apar Suri, Edward Fast, Chirag Merchant for helping with the implementation of different parts of the framework, Johannes Wagner for his support and the ICT DCAPS team for making SimSensei.

References

[1] C. Conati, S. Marsella, and A. Paiva, "Affective interactions: The computer in the affective loop," in *Proc. 10th Int. Conf. Intell. User Interfaces*, 2005, pp. 7–7. [Online]. Available: http://doi.acm.org/10.1145/1040830.1040838

[2] S. D'mello and A. Graesser, "Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 4, pp. 23:1–23:39, Jan. 2013. [Online]. Available: http://doi.acm.org/10.1145/2395123.2395128

[3] K. Anderson, et al., "The TARDIS framework: Intelligent virtual agents for social coaching in job interviews," in *Advances in Computer Entertainment*, D. Reidsma, H. Katayose, and A. Nijholt, Eds. Berlin, Germany: Springer, 2013, pp. 476–491. [Online]. Available: http://dx.doi.org/10.1007/978–3-319-03161-3_35

[4] M. van der Zwaag, J. Janssen, and J. Westerink, "Directing physiology and mood through music: Validation of an affective music player," *IEEE Trans. Affective Comput.*, vol. 4, no. 1, pp. 57–68, Jan.-Mar. 2013.

[5] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surveys*, vol. 47, no. 3, 2015, Art. no. 43.

[6] R. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affective Comput.*, vol. 1, no. 1, pp. 18–37, Jan.-Jun. 2010.

[7] Y. Trope, "Identification and inferential processes in dispositional attribution." *Psychological Rev.*, vol. 93, no. 3, 1986, Art. no. 239.

[8] S. Hareli and U. Hess, "What emotional reactions can tell us about the nature of others: An appraisal perspective on person perception," *Cognition Emotion*, vol. 24, no. 1, pp. 128–140, 2010.

[9] C. de Melo, P. Carnevale, S. Read, and J. Gratch, "Reverse appraisal: The importance of appraisals for the effect of emotion displays on people's decision-making in a social dilemma," in *Proc. 34th Annu. Meet. Cognitive Sci. Soc.*, 2012, pp. 270–275.

[10] L. M. Bylsma, B. H. Morris, and J. Rottenberg, "A meta-analysis of emotional reactivity in major depressive disorder," *Clinical Psychology Rev.*, vol. 28, no. 4, pp. 676–691, 2008.

[11] J. Rottenberg, J. J. Gross, and I. H. Gotlib, "Emotion context insensitivity in major depressive disorder." *J. Abnormal Psychology*, vol. 114, no. 4, 2005, Art. no. 627.

[12] M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proc. IEEE*, vol. 91, no. 9, pp. 1370–1390, Sep. 2003.

[13] M. Pantic, N. Sebe, J. F. Cohn, and T. Huang, "Affective multimodal human-computer interaction," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 669–676.

[14] M. Pantic, A. Pentland, A. Nijholt, and T. S. Huang, "Human computing and machine understanding of human behavior: A survey," in *Artifical Intelligence for Human Computing*. Berlin, Germany: Springer, 2007, pp. 47–71.

[15] B. Dumas, D. Lalanne, and S. Oviatt, "Multimodal interfaces: A survey of principles, models and frameworks," in *Human Machine Interaction*, D. Lalanne and J. Kohlas, Eds. Berlin, Germany: Springer, 2009, vol. 5440, pp. 3–26. [Online]. Available: http://dx.doi.org/10.1007/978–3-642-00437-7_1

[16] A. Jaimes and N. Sebe, "Multimodal human-computer interaction: A survey," *Comput. Vis. Image Underst.*, vol. 108, no. 1–2, pp. 116–134, Oct. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2006.10.019

[17] S. Chhabria, R. Dharaskar, and V. Thakare, "Survey of fusion techniques for design of efficient multimodal systems," in *Proc. Int. Conf. Mach. Intell. Res. Advancement*, Dec. 2013, pp. 486–492.

[18] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, 2010. [Online]. Available: http://dx.doi.org/10.1007/s00530–010-0182-0

[19] J. Shen, W. Shi, and M. Pantic, "Hci2 workbench: A development tool for multimodal human-computer interaction systems," in *Proc. IEEE Int. Conf. Automat. Face Gesture Recognit. Workshops*, 2011, pp. 766–773.

[20] J. Wagner, F. Lingenfelser, and E. André, "The social signal interpretation framework (SSI) for real time signal processing and recognition," in *Proc. Interspeech*, 2011, pp. 3245–3248.

[21] M. Schröder, "The SEMAINE API: Towards a standards-based framework for building emotion-oriented systems," *Advances Human-Computer Interaction*, vol. 2010, 2010, Art. no. 2.

[22] M. Serrano, L. Nigay, J.-Y. L. Lawson, A. Ramsay, R. Murray-Smith, and S. Denef, "The openinterface framework: A tool for multimodal interaction." in *Proc. Extended Abstracts Human Factors Comput. Syst.*, 2008, pp. 3501–3506. [Online]. Available: http://doi.acm.org/10.1145/1358628.1358881

[23] A. Camurri, P. Coletta, G. Varni, and S. Ghisio, "Developing multimodal interactive systems with EyesWeb XMI," in *Proc. 7th Int. Conf. New Interfaces Musical Expression*, 2007, pp. 305–308. [Online]. Available: http://doi.acm.org/10.1145/1279740.1279806

[24] L. Hoste, B. Dumas, and B. Signer, "Mudra: A unified multimodal interaction framework," in *Proc. 13th Int. Conf. Multimodal Interfaces*, 2011, pp. 97–104. [Online]. Available: http://doi.acm.org/10.1145/2070481.2070500

[25] S. Feuerstack and E. Pizzolato, "Building multimodal interfaces out of executable, model-based interactors and mappings," in *Proc. 14th Int. Conf. Human-Comput. Interaction Design Development Approaches - Vol. Part I*, Jul. 9–14, 2011, pp. 221–228.

[26] T. Halic, S. A. Venkata, G. Sankaranarayanan, Z. Lu, W. Ahn, and S. De, "A software framework for multimodal interactive simulations (SoFMIS)," in *Medicine Meets Virtual Reality 22*, vol. 163, J. D. Westwood, et al., Eds. Amsterdam, Netherlands: IOS Press, 2011, pp. 213–217. [Online]. Available: http://dblp.uni-trier.de/db/conf/mmvr/mmvr2011.html#HalicVSLAD11

[27] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: A professional framework for multimodality research," in *Proc. Language Resources Evaluation Conf.*, 2006, pp. 1556–1559.

[28] D. Heylen, S. Kopp, S. C. Marsella, C. Pelachaud, and H. Vilhjálmsson, "The next step towards a function markup language," in *Intelligent Virtual Agents*. Berlin, Germany: Springer, 2008, pp. 270–280.

[29] S. Scherer, et al., "Perception markup language: Towards a standardized representation of perceived nonverbal behaviors," in *Proc. 12th Int. Conf. Intell. Virtual Agents*, Sep. 2012, pp. 455–463. [Online]. Available: http://ict.usc.edu/pubs/Perception

[30] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, "The social signal interpretation (SSI) framework: Multimodal signal processing and recognition in real-time," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 831–834. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502223

[31] T. Baur, I. Damian, F. Lingenfelser, J. Wagner, and E. André, "Nova: Automated analysis of nonverbal signals in social interactions," in *Human Behavior Understanding*, vol. 8212, A. Salah, H. Hung, O. Aran, and H. Gunes, Eds. Berlin, Germany: Springer, 2013, pp. 160–171. [Online]. Available: http://dx.doi.org/10.1007/978–3-319-02714-2_14

[32] T. Baur, et al., "Context-aware automated analysis and annotation of social human–agent interactions," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 2, pp. 11:1–11:33, Jun. 2015. [Online]. Available: http://doi.acm.org/10.1145/2764921

[33] C.-Y. Chang, et al., "Towards pervasive physical rehabilitation using Microsoft Kinect," in *Proc. 6th Int. Conf. Pervasive Comput. Technol. for Healthcare*, May2012, pp. 159–162.

[34] B. Galna, G. Barry, D. Jackson, D. Mhiripiri, P. Olivier, and L. Rochester, "Accuracy of the Microsoft Kinect sensor for measuring movement in people with Parkinson's disease," *Gait And Posture*, vol. 39, no. 4, pp. 1062–1068, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0966636214000241

[35] J. A. Hall, J. A. Harrigan, and R. Rosenthal, "Nonverbal behavior in clinician patient interaction," *Appl. Preventive Psychology*, vol. 4, no. 1, pp. 21–37, 1996.

[36] P. Philippot, R. S. Feldman, and E. J. Coats, *Nonverbal Behavior in Clinical Settings*. New York, NY, USA: Oxford University Press, 2003.

[37] R. Boice and P. M. Monti, "Specification of nonverbal behaviors for clinical assessment," *J. Nonverbal Behavior*, vol. 7, no. 2, pp. 79–94, 1982.

[38] M. Valstar, "Automatic behaviour understanding in medicine," in *Proc. Workshop Roadmapping Future Multimodal Interaction Res. Including Bus. Opportunities Challenges*, 2014, pp. 57–60. [Online]. Available: http://doi.acm.org/10.1145/2666253.2666260

[39] J. M. Girard and J. F. Cohn, "Automated audiovisual depression analysis," *Current Opinion Psychology*, vol. 4, pp. 75–79, 2015, depression. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352250X14000219

[40] A. Rizzo, et al., "Simcoach: An intelligent virtual human system for providing healthcare information and support," *Int. J. Disability Human Development*, vol. 10, no. 4, pp. 277–281, 2011.

[41] T. Bickmore and L. Pfeifer, "Relational agents for antipsychotic medication adherence," in *Proc. CHI. workshop Technol. Mental Health*, 2008.

[42] G. M. Lucas, J. Gratch, A. King, and L.-P. Morency, "It's only a computer: Virtual humans increase willingness to disclose," *Comput. Human Behavior*, vol. 37, pp. 94–100, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563214002647

[43] L. I. Reed, M. A. Sayette, and J. F. Cohn, "Impact of depression on response to comedy: A dynamic facial coding analysis." *J. Abnormal Psychology*, vol. 116, no. 4, 2007, Art. no. 804.

[44] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and PTSD," in *Proc. Interspeech*, 2013, pp. 847–851.

[45] S. Ghosh, M. Chatterjee, and L.-P. Morency, "A multimodal context-based approach for distress assessment," in *Proc. 16th Int. Conf. Multimodal Interaction*, 2014, pp. 240–246. [Online]. Available: http://doi.acm.org/10.1145/2663204.2663274
[46] A. Hartholt, et al., "All together now—introducing the virtual human toolkit," in *Proc. 13th Int. Conf. Intell. Virtual Agents*, 2013, pp. 368–381.
[47] J. Shen and M. Pantic, "A software framework for multimodal human-computer interaction systems," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2009, pp. 2038–2045.
[48] C. Kublbeck and A. Ernst, "Face detection and tracking in video sequences using the modifiedcensus transformation," *Image Vis. Comput.*, vol. 24, no. 6, pp. 564–572, 2006.
[49] L. Morency, J. Whitehill, and J. Movellan, "Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation," in *Proc. 8th IEEE Int. Conf. Automatic Face Gesture Recognition*, 2008, pp. 1–8.
[50] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2013, pp. 354–361.
[51] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarepa collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 960–964.
[52] J. Gratch, et al., "User-state sensing for virtual health agents and telehealth applications," in *Proc. Med. Meets Virtual Reality 20*, 2013, pp. 151–157.
[53] G. Lucas, J. Gratch, S. Scherer, J. Boberg, and G. Stratou, "Towards an affective interface for assessment of psychological distress," in *Proc. Int. Conf. Affective Comput. and Intell. Interaction*, 2015, pp. 539–545.
[54] C. W. Leong, L. Chen, G. Feng, C. M. Lee, and M. Mulholland, "Utilizing depth sensors for analyzing multimodal presentations: Hardware, software and toolkits," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 547–556.
[55] L. M. Batrinca, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer, "Cicero—towards a multimodal virtual audience platform for public speaking training," in *Proc. 13th Int. Conf. Intell. Virtual Agents*, 2013, pp. 116–128.
[56] M. Chollet, G. Sratou, A. Shapiro, L.-P. Morency, and S. Scherer, "An interactive virtual audience platform for public speaking training," in *Proc. Int. Conf. Autonomous Agents Multi-Agent Syst.*, 2014, pp. 1657–1658.
[57] J. Gratch, et al., "The distress analysis interview corpus of human and computer interviews" in *Proc. 9th Int. Conf. Language Resources Evaluation*, 2014, pp. 3123–3128.
[58] L. Chen, G. Feng, J. Joe, C. W. Leong, C. Kitchen, and C. M. Lee, "Towards automated assessment of public speaking skills using multimodal cues," in *Proc. 16th Int. Conf. Multimodal Interaction*, 2014, pp. 200–203. [Online]. Available: http://doi.acm.org/10.1145/2663204.2663265
[59] D. DeVault, et al., "SimSensei kiosk: A virtual human interviewer for healthcare decision support," in *Proc. Int. Conf. Autonomous Agents Multi-Agent Syst.*, 2014, pp. 1061–1068. [Online]. Available: http://dl.acm.org/citation.cfm?id=2617388.2617415
[60] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and PTSD: The effect of gender," *J. Multimodal User Interfaces*, pp. 1–13, 2014. [Online]. Available: http://dx.doi.org/10.1007/s12193–014-0161-4
[61] D. DeVault, et al., "Verbal indicators of psychological distress in interactive dialogue with a virtual human," in *Proc. SIGDIAL Conf.*, Aug. 2013, pp. 193–202. [Online]. Available: http://www.aclweb.org/anthology/W13–4032
[62] S. Scherer, et al., "Automatic behavior descriptors for psychological disorder analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit Workshops*, 2013, pp. 1–8. [Online]. Available: http://dblp.uni-trier.de/db/conf/fgr/fg2013.html#SchererSMBGRM13
[63] Z. Yu, et al., "Multimodal prediction of psychological disorders: Learning verbal and nonverbal commonalities in adjacency pairs," in *Proc. 17th Workshop Semantics Pragmatics Dialogue*, pp. 160–169, 2013.
[64] S. Scherer, G. Stratou, and L.-P. Morency, "Audiovisual behavior descriptors for depression assessment," in *Proc. 15th ACM Int. Conf. Multimodal Interaction*, 2013, pp. 135–140.
[65] J. J. Blascovich, E. Vanman, W. B. Mendes, and S. Dickerson, *Social Psychophysiology for Social and Personality Psychology*. Thousand Oaks, CA, USA: Sage Publications Ltd, 2011.
[66] G. Littlewort, et al., "The computer expression recognition toolbox (CERT)," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit Workshops, 2011*, 2011, pp. 298–305.

Giota Stratou received the Master's degree in control and dynamical systems from CalTech and the Master's in CS: Game Development from USC. She is currently a research programmer in the Emotions Group, USC's Institute for Creative Technologies (ICT). Since 2010 she has been working closely with affective computing researchers exploring and implementing novel computer vision algorithms and sensing technologies. Her research expertise includes recognition of nonverbal behaviors from vision and understanding human underlying affective state. In the last couple of years, her work has led to several publications in IEEE and ACM peer reviewed conferences. She received a best paper award in IEEE's Face and Gesture conference for her work on automatic detection of distress indicators from videos. She is also the lead developer of a complex framework for sensing and multimodal behavior understanding (MultiSense) which has been used in many projects, including the DARPA funded SimSensei.

Louis-Philippe Morency received the Master and PhD degrees from MIT Computer Science and Artificial Intelligence Laboratory. He is assistant professor in the Language Technology Institute, the Carnegie Mellon University where he leads the Multimodal Communication and Machine Learning Laboratory (MultiComp Lab). In 2008, he was selected as one of "AI's 10 to Watch" by *IEEE Intelligent Systems*. He has received 7 best paper awards in multiple ACM- and IEEE-sponsored conferences for his work on context-based gesture recognition, multimodal probabilistic fusion and computational models of human communication dynamics. He was lead Co-PI for the DARPA-funded multi-institution effort called SimSensei which was recently named one of the years top ten most promising digital initiatives by the NetExplo Forum, in partnership with UNESCO.