Human-Robot Interactions for Single Robots and Multi-Robot Teams

by

Alexander Hong

A thesis submitted in conformity with the requirements
for the degree of Masters of Applied Science

Mechanical and Industrial Engineering
University of Toronto

Human-Robot Interactions for Single Robots and Multi-Robot Teams

Alexander Hong

Master of Applied Science

MECHANICAL AND INDUSTRIAL ENGINEERING
UNIVERSITY OF TORONTO
2016

Abstract

For robots to successfully take part in bi-directional social interactions with people, they must be capable of recognizing and responding to human affect. Such robots would promote effective and engaging interactions with users. In this thesis, a multimodal bi-directional affect architecture is proposed that determines user affect using a unique combination of user body language and vocal intonation. In addition to one-on-one human-robot interactions, multi-robot teams can provide valuable assistance in Urban Search and Rescue (USAR) missions by exploring dangerous environments, while searching for victims. One of the main challenges an operator may face in controlling a multi-robot team is the simultaneous control of multiple robots while juggling between tasks and keeping situational awareness. This thesis also proposes a multi-robot collaboration architecture using a learning-based semi-autonomous controller that provides the ability to effectively allocate sub-tasks to robots to complete USAR missions.

Acknowledgements

I am very grateful for the supervision of Prof. Benhabib, and Prof. Nejat. Their guidance has been invaluable in completing my research and getting it published in international journals and conferences. I am thankful for all the great opportunities they gave me during my research.

I would like to give my thanks out to Onome Igharoro for his contribution of the graphical user interface that was used in the urban search and rescue experiments, Yuma Tsuboi for his implementation of the vocal intonation affect recognition module, and his assisitance in training the multimodal affect recognition system, Yugang Liu for his guidance and contribution of the MAXQ hierarchical reinforcement learning algorithm that was used in the multi-robot control architecture, Derek McColl for his contribution to the affect recognition literature review and his guidance in body language affect recognition, Nolan Lunscher for developing a robot emotional model and questionnaire that was used for affect recognition experiments, and Tianhao Hu for his implementation of robot body language displays that was used in the affect recognition experiments. I could not have finished my thesis without their contributions.

Thank you to all graduate students of the Autonomous Systems and Biomechatronics Laboratory and the Computer Integrated Manufacturing Laboratory for their invaluable assistance and keeping the lab an enjoyable place to work. Their support and encouragement is what made this thesis possible.

Most importantly, I would like to thank my family and friends for their continuous support and their impeccable tenacity for putting up with me.

Table of Contents

Acronyms

AAM	Active Appearance Model
AU	Action Units
DBN	Dynamic Bayesian Network
DCOP	Dynamic Distributed Constraint Optimization Problem
ECG	Electrocardiogram
EDR	Electrodermal Response
EMG	Electromyograms
ESS	Exit Sub-scene
FACS	Facial Action Coding System
GMM	Gaussian Mixture Model
HC	Human Control
HCI	Human-Computer Interaction
HMM	Hidden Markov Model
HRI	Human-Robot Interaction
HRL	Hierarchical Reinforcement Learning
IAS	Interaction Activity Sub-system
IE	Interaction Effort
INS	Inertial Navigation System
KNN	*k*-nearest neighbors
MARS	Multimodal Affect Recognition System
MDP	Markov Decision Process
MFCC	Mel-frequency cepstral coefficient
MIMDP	Mixed Markov Decision Process
MROT	Multi-Robot Operator Team
MLP	Multilayer Perceptron
NG	Navigate
NN	Neural Network
NUR	Navigate to Unvisited Regions
POMDP	Partially Observable MDP
REM	Robot Emotion Model
SA	Situational Awareness
SLAM	Simultaneous Localization and Mapping
SSS	Search Sub-scene
SVM	Support Vector Machine
SVR	Support Vector Regression
TP	Task Performance
UAV	Unmanned Aerial Vehicle
UI	User Interface
USAR	Urban Search and Rescue
USARSim	Unified System for Automation and Robot Simulation
UDK	Unreal Development Kit
VI	Victim Identification

Nomenclature

$\mathbf{c}_m$	Multimodal affect vector
v_b	Body language valence value
v_v	Vocal intonation valence value
v_m	Multimodal valence value
a_b	Body language arousal value
a_v	Vocal intonation arousal value
a_m	Multimodal arousal value
m_i	Recognized multimodal affect for participant i
u_i	User assessed multimodal affect for participant i
n_p	Number of participants
n_a	Total number of affect levels
M_p	Subtask p of MDP
π_p	Policy of subtask p
S_t	State function of task t
SSS_i	Search Sub-scene subtask, where i represents the index of the sub-scene
V	Presence of potential victims
S_S	Sub-scene status (i.e., unexplored, being explored, or explored)
M_G	Collection of 2D occupancy maps of USAR sub-scenes
L_R	Robots' locations within the same sub-scene with respect to the starting location of the first robot
$A_{O,ti}$	Other robots' actions/subtasks, where i represents the index of the sub-scene
$L_{V/R}^{j}$	Potential victim's location with respect to robot j location
C_l^j	Surrounding cells $l = 1$ to 8 of robot j
D_{xy}^{j}	Depth profile information of robot j
s	Robot state
γ	Discount factor
N	Number of transition steps from state s to net state s'.

Chapter 1
Introduction & Motivation

Recent research has promoted the use of robots to expand beyond factories and manufacturing. Robots are now being used in domains such as space exploration, search and rescue, and healthcare. These domains require a close interaction with users and human operators. Hence, the design and evaluation of robotic systems is an important study area in order to better design robots to suit the functional, ergonomic, and emotional requirements of different users. Human-robot interaction (HRI) encompasses both the development of robots that engage humans in specific activities and tasks, as well as the study of how humans and robots interact with each other during such scenarios [1]. HRI can take many forms, from an operator teleoperating mobile robots during search and rescue operations [2], to robotic mail delivery in an office building [3], to a socially engaging nurse robot delivering medicine to elderly long-term care residents [4].

Past HRI research has mainly concentrated on developing efficient ways for humans to control/supervise robot behavior [1]. More recently, research has also focused on developing robots that can detect common human communication cues for more natural interactions [5]-[7]. This thesis focuses on two important HRI applications – to develop robots that can effectively detect common human communication cues during HRI, and to investigate performance in human-robot teams during urban search and rescue (USAR) missions.

1.1 Socially Assistive Robots

Social HRI is a subset of HRI that encompasses robots which interact using natural human communication modalities, including speech, body language and facial expressions [8]. This allows humans to interact with robots without any extensive prior training, permitting desired tasks to be completed more quickly and requiring less work to be performed by the user [8]. However, in order to accurately interpret and appropriately respond to these natural human communication cues, robots must be able to determine the social context of such cues. One application of recognizing cues in socially assistive robots is, for example, elderly care.

Older adults (> 75 years old) may suffer from social isolation, social inactivity, or loneliness

due to physical and cognitive disabilities as well as lifestyle adjustments resulting from old-age [9], [10]. Socially assistive robots can be used as an effective technology for the elderly to provide social interaction and cognitive assistance with activities of daily living. For example, they can support older adults with self-maintenance tasks (e.g., eating, grooming, dressing), and recreational activities (e.g., playing music, games).

In order to promote natural and social HRI, and provide the elderly with suitable assistance, robots would need to be equipped with emotional intelligence. Namely, they would need to have the ability to consider and respond to the emotions, moods, or affect of the person with whom they are interacting [10]. Older adults, including those with dementia, communicate their affective states using facial expressions, body language, and vocal intonation [11].

For robots to successfully take part in bi-directional social interactions with people, they must be capable of recognizing, interpreting and responding effectively to social cues displayed by a human. In particular, during social interactions, a person's affective displays can communicate his/her thoughts and social intent [12]. A person's affect is a complex combination of emotions, moods, interpersonal stances, attitudes, and personality traits that influence his/her own behavior [13]. A robot which is capable to interpret affect will have an enhanced capacity to make decisions and assist humans due to its ability to respond to their affect [14]. Such robots would promote more effective and engaging interactions with users that would lead to better acceptance of these robots by their intended users [15]. The challenge lies in developing social interactive robots with the abilities to recognize and identify complex human behaviors, such as affect and intent, and, in turn, be able to respond appropriately using natural communication modes such as facial expression, vocal intonation, and body language.

One of the main goals for developing social robots capable of both sensing multimodal inputs and displaying multimodal outputs during HRI is that they can be used to effectively assist humans in a number of everyday tasks. Namely, these robots can intelligently recognize and classify human affective intent and in turn respond appropriately using their own assistive emotional behavior. Human affective cues can be inferred from natural communication modalities, such as facial expressions, vocal intonation, and body language [1], as it has been shown that they correlate strongly to a person's affective state [16]. Recent research in this area has also investigated the use of multimodal affect recognition systems, which combine two or more of the aforementioned modalities for both HCI (human-computer interactions) [17] and

HRI [5] applications.

The use of multimodal inputs, over a unimodal input, provides two key advantages [18]: (*i*) a combination of modalities provides complementary and diverse information, which in turn provide increased robustness and performance, and (*ii*) when one modality is unavailable, due to occlusion and/or noise, a multimodal recognition system can use the remaining modalities to estimate affect. The robot can, then, utilize the user's multimodal affect to determine its own appropriate emotional behavior via a (robot) emotional model.

To date, only a handful of multimodal affect recognition systems have been used for HRI. Multimodal affect recognition for HRI has focused on the primary modes of facial and vocal expressions [19], [20], [21], [22]. Body language has yet to be incorporated into these affect recognition systems. However, body language plays an important role in conveying changes in human emotions during social interactions [23]. Furthermore, vocal intonation has been strongly correlated to body language displays [24]. Vocal intonation plays an important role in conveying changes in emotion through the vocal properties of pitch, tempo, and loudness during social interactions [25].

The proposed multimodal bi-directional affect architecture, in contrast, utilizes both dynamic body language and vocal intonation to determine user affect, and, in turn, determines the robot's appropriate emotions based on a two-layer reactive and deliberative emotional model, which uniquely considers uncertainty in robot emotion expression.

1.2 Urban Search & Rescue Multi-Robot Teams

Robot exploration and victim identification in USAR is a challenging task due to the sensing challenges in USAR environments. Vision-based techniques are difficult to implement due to the inherent challenges of computer vision under unstructured lighting conditions [26]. Low-cost acoustics and CO_2 sensors may experience unexpected behavior due to potential existence of noise and other gases in USAR environments [26]. Thermal sensors prove to be unreliable for victim identification as collapsed structures can give similar heat readings to humans [26]. Furthermore, robots flipping over are one of the biggest challenges for rescue robots [26].

In USAR, rescue robots need to explore USAR environments while detecting hazards and searching for victims. There is minimum presumed knowledge about the environment. This

includes having no knowledge of the victims' locations, existence of hazardous areas (i.e. leaked gas, fire, holes), and scene information (i.e. cluttered, climbable rubble, open space). Furthermore, there is no indication of how long exploration will take. USAR exploration techniques must be robust and applicable to any unstructured environment.

Multi-robot teams can provide valuable assistance in USAR by exploring cluttered and dangerous environments, while searching for potential victims. Mobile robots can effectively explore disaster environments with minimum *a priori* knowledge about the location of victims and scene layout [27], [28]. The majority of past robotic USAR missions, however, have been the utilization of teleoperated single robot teams [28]-[30]. Operators of such robot teams have, typically, experienced perceptual difficulties in trying to understand the 3D cluttered environment via remote visual feedback [28]. Furthermore, these single rescue robot teams have experienced task-handling limitation.

Recently, researchers have considered multi-robot teams for a number of applications, including material transportation [31], reconnaissance and surveillance [32]-[38], inspection and manipulation [39], and USAR missions [29], [40], [41]. It has been claimed that increased efficiency and system robustness can be achieved through robot redundancy. However, one of the main challenges that an operator may face in controlling a multi-robot team in teleoperation mode is the simultaneous control of multiple robots while juggling between their respective tasks and keeping situational awareness of all the robots. Thus, it has been recommended that such multi-robot teams be given some level of autonomy [29].

USAR multi-robot teams face many challenges in searching for potential victims and exploring unknown cluttered scenes in a time-limited situation. To help alleviate some of those challenges, autonomous and semi-autonomous controllers for controlling multi-robot teams have been explored. Using any level of autonomy, however, is a double-edged sword where it may help alleviate some task at hand, but creates new problem for the operators, such as compliance, reliance, and trust [42]. Hence, the challenge lies in developing a control architecture that aims to alleviate operator workload and increase team performance, but also maintains operator situational awareness.

Past controllers have mainly been shown to control operator-to-robot teams with 1:12 ratio, and have been limited in their capabilities to avoid obstacles and identify victims automatically [43]-[46]. The proposed architecture, in contrast, uses hierarchical reinforcement learning (HRL)

[47] to learn from the USAR environment in order to increase performance in autonomous exploration and victim identification.

The learning-based semi-autonomous controller has been developed for single robots [48], as well for non-cooperating [49] and cooperating [47] multi-robots teams. These controllers allow operators and robots to share the important tasks of exploring unknown cluttered USAR scenes and searching for victims. However, as noted by others, in order to effectively implement such semi-autonomous controllers for cooperative multi-robot teams, the impact of increased numbers of robots on system performance must be investigated [40], [43], [50]-[53], as well as the effect of task automation on robot team performance [44], [45], [54]-[62].

The proposed architecture provides the ability to effectively allocate sub-tasks to robots in order to complete the overall mission. The approach provides two primary advantages, when compared to methods currently in use: (*i*) the operator can handle a greater number of robots without significant performance loss due to the controller's ability to only request human assistance when a robot is stuck or when there is uncertainty in human identification; and, (*ii*) the workload of the operator is reduced significantly.

1.3 Problem Statement and Thesis Objectives

This research focuses on enabling bi-directional communication between users and social robots to promote more effective and engaging interactions during social HRI, and effective collaboration between single human operators and multi-robot teams in order to increase overall task performance in urban search and rescue missions.

The two primary HRI objectives of this thesis are:

1. To develop a multimodal recognition architecture for multimodal bi-directional affective communication using dynamic body language and vocal intonation modalities, and
2. To investigate the influence of operator-to-robot ratio in multi-robot USAR teams utilizing a learning-based semi-autonomous controller.

1.4 Proposed Methodology and Thesis Organization

This thesis describes the development, implementation, and evaluation of sensory systems that allow socially assistive robots to identify affective state from both body language and vocal intonation while engaged in HRI, and investigates the influence of operator-to-robot ratio using a learning-based semi-autonomous controller for USAR missions.

Chapter 2 reviews the literature on the development of multimodal affect recognition systems for assistive robots to be able to identify and respond to human affect during HRI including: (*i*) affect categorization models used in HRI, (*ii*) facial affect recognition during HRI, (*iii*) body language affect recognition during HRI, (*iv*) voice-based affect recognition during HRI, (*v*) affective physiological signals recognition during HRI, (*vi*) affect classification techniques, (*vii*) HRI scenarios used, and (*viii*) affect training databases.

Chapter 3 presents a novel multimodal affect recognition system for recognizing and classifying a person's affect from natural displays of dynamic body language and vocal intonation. The system utilizes a 3D sensor and a microphone to capture 3D image data, and sound of the person engaged in one-on-one social HRI. The proposed system was implemented and its performance was tested on a Nao humanoid robot platform in order to engage users in multimodal bi-directional HRI for the purpose of diet and fitness planning.

Chapter 4 reviews the literature on the development and investigation of learning-based semi-autonomous controller architectures for controlling multi-robot teams including: (*i*) robot exploration techniques used in USAR, (*ii*) influence of operator-to-robot ratio on team performance, (*iii*) task automation techniques in multi-robot teams, (*iv*) AI-based approaches for controlling multi-robot teams, (*v*) user interfaces for multi-robot control, and (*vi*) simulation environments used for USAR experiments.

Chapter 5 presents the implementation and evaluation of a learning-based semi-autonomous controller for multi-robot control in USAR missions. Learning-based semi-autonomous controllers allow robots to learn from their surroundings and previous experiences in order to optimize their own behaviors. The objective of the Chapter is to promote the use of a novel controller with MAXQ hierarchical reinforcement learning (HRL) deliberation for improved supervision and control of multi-robot teams in USAR environments. It is conjectured that the proposed controller's efficiency improvement over teleoperation is a direct function of the robotic team's size.

Finally, Chapter 6 presents concluding remarks on the developed multimodal affect recognition system and the learning-based multi-robot single-operator system architecture for USAR missions.

Chapter 2
Review on Affect Recognition during HRI

In HRI, robots should be socially intelligent. They should be able to respond appropriately to human affective and social cues in order to effectively engage in bi-directional communications. Social intelligence would allow a robot to relate to, understand, and interact and share information with people in real-world human-centered environments. One component of social intelligence is affect recognition, where robots can intelligently recognize and classify human affective intent. Human affective cues can be inferred from natural communication modalities, such as facial expression, body language, voice, and physiological signals [5].

This Chapter presents a literature review of affect recognition techniques during HRI. This literature review was published in D. McColl, A. Hong, N. Hatakeyama, G. Nejat, B.Benhabib "*A Survey of Autonomous Human Affect Detection Methods for Social Robots Engaged in Natural HRI*," Journal of Intelligent Robotic Systems, vol. 82, pp. 101-133, 2016. The review was completed in collaboration with Derek McColl and Naoki Hatakeyama. In Section 2.1, affect categorization models used in HRI are discussed. In Section 2.2, an overview of facial affect recognition during HRI is presented. Section 2.3 presents an overview of body language affect recognition techniques during HRI. In Section 2.4, voice-based affect recognition during HRI are discussed. In Section 2.5, affective physiological signals recognition during HRI are presented. Section 2.6 presents a review of multimodal affect recognition systems for HRI. Affect classification techniques are reviewed in Section 2.7. Furthermore, various HRI scenarios discussed in the literature are reviewed in Section 2.8. Section 2.9 provides an overview of affect training databases used in literature. Finally, Section 2.10 presents a summary of the Chapter.

2.1 Affect Categorization Models

The neuroscience and cognitive science fields have identified two leading types of models that describe how people perceive and classify affect: categorical and dimensional [63]. Fig. 2-1 illustrates an example of (*a*) a categorical model, and (*b*) a dimensional affect model. Both models are represented using self-assessment manikins [64]. In general, categorical models consist of a finite number of discrete states, each representing a specific affect category [63],

[65]. For example, a form of affective expression such as a facial expression, body language display, vocal intonation or a combination of these can be classified into one of these states: happy, sad, angry, etc. Dimensional models, on the other hand, also known as continuous models, use feature vectors to represent affective expressions in multi-dimensional spaces, allowing for the representation of varying intensities of affect [63], [66]. Changes in affect are considered in the continuous spectrum for these models. This section discusses existing variations of these two leading models, and how they have been used by robots in HRI scenarios.

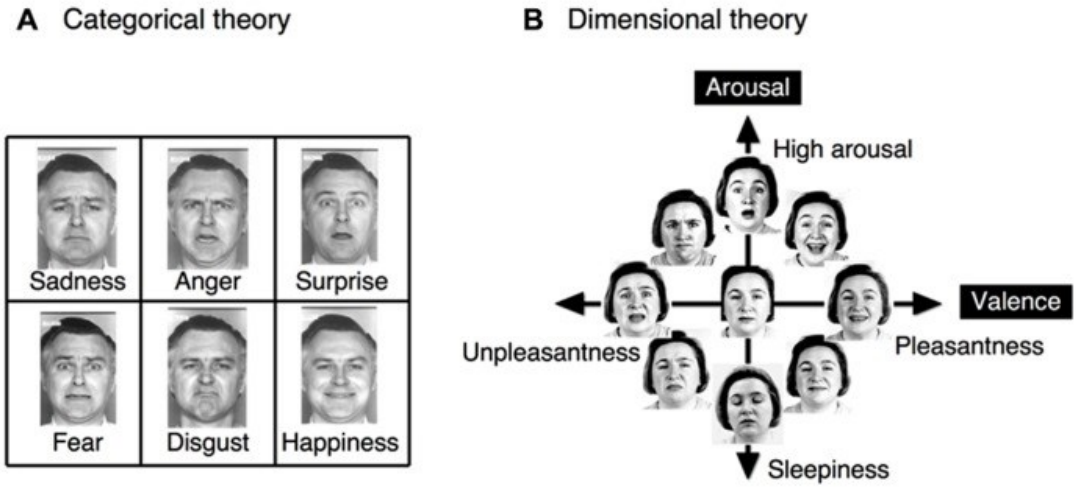


Fig. 2-1. (a) Categorical affect model and (b) the circumplex dimensional affect model [67].

2.1.1 Categorical Model

In 1872, Darwin first classified six universal human affective states that were cross-culturally innate by studying expressions and movements in both humans and animals [68]. These six states were defined to be happiness, surprise, fear, disgust, anger, and sadness. In 1962, Tomkins also proposed that there were a limited number of human affective states by observing human communication movements such as nods and smiles [69]. He introduced the Affect Theory, which included nine discrete affect categories: joy, interest, surprise, anger, fear, shame, disgust, dis-smell, and anguish [70], [71]. Similar to Darwin, Ekman also conducted human cross-cultural studies and determined that human facial affect was a collection of the same six recognizable basic categories [72]. Ekman codified the corresponding facial expressions into these affective states using the Facial Action Coding System (FACS), where positions of facial action units

(AU) are classified as distinct facial expressions of emotions [73]. This categorical model is debatably the most often used model today for classifying affect [74].

The strength in categorical models lies in their ability to clearly distinguish one affect from the others. However, categorical models lack the ability to classify any affect that is not included as a category in the model.

2.1.2 Dimensional Model

In order to address the challenge of classifying affect across a continuous spectrum, dimensional models have been developed. Russell [75] and Barrett [76] argued that affect cannot be separately classified as gradations, and blends of affective responses could not be categorized. Furthermore, they argued that there were no correspondences between discrete affect and brain activity. The majority of dimensional models that have been developed use either two or three dimensions for affect categorization.

In 1897, Wundt first suggested that affect could be described by three dimensions: pleasure *vs.* displeasure, arousal *vs*. calm, and relaxation *vs.* tension [77]. The pleasure-displeasure dimension described the positivity or negativity of mental state, the arousal-calm dimension referred to excitation level, and the relaxation-tension dimension referred to the frequency of affect change [77]. In 1954, Schlosberg proposed a similar model also consisting of three dimensions of affect: pleasure *vs.* displeasure, attention *vs*. rejection, and level of activation (also known as relaxation *vs.* tension) [78]. The attention-rejection dimension described the levels of openness of a person [79].

In 1980, Plutchik designed a three-dimensional model, which included valence, arousal, and power. The model combined both categorical and dimensional model theories as it incorporated blends of eight basic affective states: curiosity, fear, disgust, sadness, acceptance, joy, anger, and surprise [80]. The valence dimension described how positive or negative an affect was, similar to that of the abovementioned pleasure-displeasure scale, and the power dimension described sense of control over the affect.

In [81], Mehrabian suggested the pleasure, arousal, and dominance (PAD) emotional state to incorporate all emotions. Dominance described the dominant or submissive nature of the affect. Russell also proposed a two-dimensional circumplex model containing valence and arousal dimensions [82]. The model illustrated that affect was arranged in a circular fashion around the

origin of the two-dimensional scale, e.g. Fig. 2-1(*b*). The circumplex model has been widely used to evaluate facial expressions [83].

Watson proposed a two-dimensional PANA (Positive Activation-Negative Activation) model to better account for affect consisting of high arousal and neutral valence [84]. The PANA model suggests that positive affect and negative affect are two separate systems, where states of very high arousal are defined by their valence and states of very low arousal tend to be more neutral in valence [85].

Dimensional models have the ability to encompass all possible affective states and their variations [86]. However, they lack the strength to distinguish one prominent affect from another and can often be confused with each other.

2.1.3 Affect Models Used in HRI

A number of HRI studies have incorporated the use of affect classification using categorical emotional models for facial expressions [87]-[90], body language [91]-[94], voice [69], [95], [96]-[100], physiological signals [101]-[103], and multi-modal systems [104]-[105]. These models allow robots to interpret affective states in a similar manner as humans [106]. The most common discrete affective categories used by robots in HRI settings are disgust, sad, surprise, anger, fear, happy, as well as neutral. These affective categories encompass Ekman's six basic affect and have been used to infer the appropriate social robot response in various HRI scenarios.

HRI studies have also been conducted with robots using dimensional models for affect classification using facial expressions [107]-[109], body language [110]-[112], voice [113], [114], physiological signals [4], [115], [101], [116]-[120], and multimodal systems [121]-[123]. The most common model used in HRI is the two-dimensional circumplex (valence-arousal) model. This model captures a wide range of positive and negative affect encountered in common HRI scenarios.

HRI researchers have also developed their own dimensional affective models. For example, in [124], a four-dimensional model was developed for multimodal HRI to determine affect from voice, body movements, and music. The dimensions of the model were speed, intensity, regularity, and extent. The use of both categorical or dimensional models has allowed social robots to effectively determine a person's affect and in turn respond to it by changing its own behavior which has also included the display of the robot's own affect.

Categorical models have also been used as the main choice of affect classification from facial expressions [87]-[90], [106], [125]-[133], body language [92]-[94], voice [69], [96], [98], [100], [134], [135], and multi-modal inputs [104], [123], [136]-[139]. The affect categories noted range from using Ekman's six basic emotions [72] along with a neutral emotion [88], [125], [126], [128], [132], or using smaller subsets of these emotions ranging from three [69], [96], [123], [128]-[130], [132] to five [98], [106], [127], [128], [131]-[133], [136], [138], [139] affective states. Several of the affect-detection systems designed for social robots used dimensional models, mainly consisting of valence and arousal scales for recognition from facial expressions [107], [108], [155], body language [110], voice [113], [114], physiological signals [4], [101], [119], [140], and multi-modal inputs [140], [141]. A small number of systems have utilized alternative affect classification scales, such as accessibility [111], engagement [91], predictability [113], stance [123], speed regularity and extent [141], stress [95], [118], anxiety [117], [142], and aversion and affinity [143].

2.2 Facial Affect Recognition during HRI

Facial expressions involve the motions and positions of a combination of facial features, which together provide an immediate display of affect [144]. Darwin was one of the first scientists to recognize that facial expressions are an immediate means for humans to communicate their opinions, intentions, and emotions [68]. Schiano *et al.* also identified facial expressions to be the primary mode of communication of affective information due to the inherently natural face-to-face communication in human-human interactions [145]. Fridlund emphasized the cross-cultural universality of emotions displayed by the face [146], and that facial expressions are used for social motives [147], and may directly influence the behaviors of others in social settings [148].

A key feature to social robots is their ability to interpret and recognize facial affect in order to have empathetic interactions with humans [149]. With the advancement of computer-vision technologies, there has been significant research effort towards the development of facial-affect-recognition systems to promote robot social intelligence and empathy during HRI [149]. In order to emulate empathy, robots should be capable of recognizing human affective states, displaying their own affect, and expressing their perspective intentions to others [150]. There are several challenges in recognizing human facial affect during HRI including: lack of control over environmental lighting conditions and working distances [88], [150], real-time computational

requirements [88], [151], processing spontaneous dynamic affective expressions and states [152], and physical and technology constraints [153].

The majority of facial affect recognition systems use 2D onboard cameras to detect facial features. Facial-affect classification is, then, used to estimate affect utilizing binary decision trees [88], 2) AdaBoost [125], multilayer perceptrons (MLPs) [90], support vector machines (SVMs) [126], support vector regression (SVRs) [90], [107], neural networks (NNs) [90], [127]-[130], or dynamic Bayesian networks (DBNs) [106].

The most popular input mode used to infer affect during HRI has been facial expressions [5]. The most common method of capturing facial information has been using a single 2D camera [87]-[90], [106], [109], [121], [128], [131]-[133], [154], [155]. Facial features have mainly been extracted from the eyes, eyebrows, lips, and nose region according to FACS [156]. However, recognition approaches and facial feature extraction techniques usually only perform well when the human frontal face is positioned directly in front of the camera. This is not always the case during HRI. A few research groups have also used stereo vision [125], [127] for better facial affect estimation, but no group, to-date, has used 3D information from 3D sensors in order to infer affect.

2.3 Body Language Affect Recognition during HRI

This section was compiled by Derek McColl from the aforementioned survey paper [5] – it is presented here for completeness. Body language conveys important information about a person's affect during social interaction [157]. Human behavioral research has identified a number of body postures and movements that can communicate affect during social interactions [158]-[160]. For example, in [158], Mehrabian showed that body orientation, trunk relaxation, and arm positions can communicate a person's attitude towards another person during a conversation. In [159], Montepare *et al.* found that the jerkiness, stiffness and smoothness of body language displays can be utilized to distinguish between happy, sad, neutral and angry affective states displayed during social interaction. Furthermore in [160], Wallbott identified specific combinations of trunk, arm and head postures and movements that corresponded to 14 affective states including the social emotions of guilt, pride, and shame. Body movements and postures have also been linked to dimensional models of affect [161], [162]. For example, it has been identified that there exists a strong link between a person's level of arousal and the speed of

movements during a body language display [161] and that head position is important in distinguishing both valence and arousal [162].

To-date, the majority of research on identifying body language during HRI has concentrated on recognizing hand and arm gestures as input commands for a robot to complete specific tasks including navigating an environment [163]-[171], identifying object locations [172]-[174], and controlling the position and movements of a robot arm [175],[176]. A fewer number of automated affect detection systems have been developed to specifically identify body language displays from individuals engaged in HRI [91]-[94], [110]-[112], [177]. These affect from body language systems have been designed for various HRI scenarios, including collaborative HRI [91], assistive HRI [93], [110], [111], mimicry [92], [112], [177], and multi-purpose HRI scenarios [94].

The Kinect™ sensor, which provides a system with 2D and/or 3D information, has been the most commonly used sensor for identifying affect from body language [93], [94], [110], [111]. A wide range of body language features have been investigated, but the Kinect™ skeleton joint positions have been utilized more than any other set of features [93], [94].

2.4 Voice-Based Affect Recognition during HRI

This section was compiled by Derek McColl from the aforementioned survey paper [5] – it is presented here for completeness. During social interactions, people can communicate affect with their voices [114]. Changes in a person's affect can result in changes to the position of his/her larynx, vocal fold tension and position and breathing rate as well as positions and shapes of the lips and mouth muscles, which all influence the tone and quality of the voice [116]. Human vocal displays of affective states are determined by internal physiological changes as well as partially determined by social display rules [114]. For example, in formal social situations a person may produce a pleasant voice quality even when internally feeling rage [114]. Research into human voice has identified features of vocal intonation that directly correspond to specific affective states [117], [118]. In [117], it was identified that the mean and range of the fundamental frequency of a verbal utterance can be utilized to distinguish between a specific set of emotions including cold and hot anger, happiness, sadness, anxiety, elation, panic fear, and despair. In [118], a review of empirical studies that have examined the vocal expression of affect during social interactions identified that the intensity (total energy), fundamental frequency, utterance

contour, high frequency energy, and word articulation rate can be used to distinguish between the affective states of stress, anger, fear, sadness, joy and boredom. More recent research has also found that the fundamental frequency, spectrum shape, speech rate, and intensity of a voice signal reflect changes in arousal, valence and potency/control during social interactions [119].

For social robots to engage in effective bi-directional communication with humans it is important that they have the ability to perceive and interpret human vocal affect. To-date, a number of affect-recognition systems have been developed for social robots to determine affective states from a person's voice [69], [95] ,[96], [98], [100], [113], [114], [134], [135]. These systems, similar to the previous systems, can also be classified based on their use by robots in the following HRI scenarios: collaborative [95], [96], [113], assistive [100], [114], [134], mimicry [69], and multi-purpose HRI [98], [135].

For detecting affect from voice, the most common sensor has been a microphone worn by a user [69], [113], [114], [135]. The most common features detected have included the traditional voice features of pitch and fundamental frequency, which human behavior researchers have directly linked to affect [178], as well as features that have been identified more recently, such as MFCCs.

2.5 Affective Physiological Signals Recognition during HRI

This section was compiled by Derek McColl from the aforementioned survey paper [5] – it is presented here for completeness. Human affect influences the body in many ways, for example by changing a person's heart rate, skin conductance and other ElectroDermal Response (EDR), tension in specific muscles, breathing rate, etc., [142], [179]. These changes in the body can be monitored as physiological signals and utilized to estimate a person's affective state [179]-[181]. For example, a decrease in heart rate, measured with an electrocardiogram (ECG) can signify that a person is feeling disgust or sadness while these two affective states can be distinguished by sadness resulting in a decrease in skin conductance, measured with electrodes on the skin, while disgust results in an increase in skin conductance [179]. It has also been found that increased tension in the corrugator muscle (above the eyebrow) and masseter muscle (upper jaw), measured with electromyograms (EMGs), relate to increases in anxiety and mental stress [142]. Additionally, with respect to dimensional models of affect, it has been found that an increase in skin conductance and heart rate relate to an increase in arousal [181]. Physiological signals are

well suited for HRI due to data being easily obtained from wireless wearable sensors and analyzed in real-time to detect a user's affective state [140].

Several researchers have developed automated affect-detection systems using physiological signals to allow robots to interpret human affective states during HRI [4], [101], [103], [117]-[119], [140], [142], [182]. These systems can also be classified based on the proposed HRI scenarios: collaborative [101], [140], [142] , assistive [4], [117]-[119], [182], and multi-purpose [103].

The majority of systems that detect affect from physiological signals have utilized ECG, EMG and/or skin conductance sensors in order to identify a wide variety of heart rate, facial muscle activity, and skin conductance level measures [4], [101], [117], [118], [140], [142], [182]. One physiological system has investigated blood pressure [119] and another have investigated prefrontal brain activity measured with a near-infrared spectroscopy [103].

2.6 Multimodal Affect Recognition during HRI

The systems and techniques discussed above focus on the recognition of one single input mode in order to determine human affect. The use of multimodal inputs over a single input provides two main advantages: when one modality is not available due to disturbances such as occlusion or noise, a multimodal recognition system can estimate affective state using the remaining modalities, and when multiple modalities are available, the complementarity and diversity of information provide increased robustness and performance [105]. Several researchers have considered the combination of two or more input modes in order to effectively determine human affect during the various HRI applications.

Social robots can interact with people using a number of communication channels in order to establish a social relationship. Human affective states can be inferred from a combination of these communication channels. Multimodal data, however, is more difficult to acquire and process due to the need of many more sensors in order to acquire data from multiple channels, the inherent additional dimensionalities in learning algorithms, and multimodal data fusion [106].

Multimodal systems contain the most sensors compared to all other modalities. Sensors utilized in multimodal systems have included ECG and EMG [113], [124] to record

physiological signals, microphones [124], [136], [138], [183]-[186] to record voice intonations, and 2D cameras to capture facial information [136], [183]-[186], and body language information [138]. These sensors have been integrated together in order to extract many features, including heart rate [113], [124], voice pitch [124], [136], [138], [183]-[186], gait features [138] and facial features [136], [183]-[186]. Future research should continue to investigate a wide range of features for all modes in order to determine which combinations of features result in the highest recognition rates during real-world interactions.

2.6.1 Multimodal Fusion

A combination of modalities could lead to better affect recognition [187]. Multiple modalities can be integrated during decision-level, where individual modal results are combined and decided upon, or during feature-level, where features from different modalities are combined and used for affect recognition. Feature-level fusion is performed by combining the features of multiple modalities into a single feature vector [188]. Feature-level fusion involves feature selection of individual modalities either before or after combining them. The disadvantage of combining two different kinds of modalities is that they may have different time scales and metric levels [188]. Another problem is feature-level fusion results in a feature vector of high dimensionality, which can degrade the performance of the emotion recognition system [188]. Decision-level fusion treats modalities independently, and the single-modal recognition results are combined at decision level. Decision-level fusion overcomes the problem of different time scales and metric levels of audio and visual data, and avoids high dimensionality of feature vector. The disadvantage of this method is that the assumption that modalities are independent may not be true. The assumption of independence results in loss of mutual correlation between modalities [188].

The majority of multimodal affect recognition systems have used decision-level integration for human-robot interaction applications [5]. The most popular way to ultimately determine affective state during decision-level fusion have been to use decision trees that incorporate results from individual affect recognizers [104], [136], [137], [139], [189]. Weighted sums of results from each modality for multimodal affect recognition have also been used [123], [190], [191].

2.6.2 Multimodal Affect Recognition Systems

In [104], the humanoid robot, HIPOP, was used to assist adolescents and young adults with autism spectrum disorder. The robot used a combination of physiological signals, eye gaze direction, and hand gestures to determine user valence and arousal levels. Physiological signals were obtained through electrocardiogram (ECG), accelerometer and respiration sensors embedded on a t-shirt worn by the user, a sensorized glove with electrodes located on the fingers, and a multi-axial accelerometer around the wrist. A high-speed wearable camera was used to capture eye-gaze direction, and a wearable microphone was used to extract high-level features correlated to the user's psychophysiological state. Valence and arousal levels were classified using a Bayesian Decision Tree decision-level fusion approach, providing confidence levels for each affect level. Recognition rates were reported to be greater than 90% for both arousal and valence using 40-fold-cross-validation [184].

In [137], the doll-like robot Maggie, equipped with an onboard 2D camera and a microphone, used user facial and vocal expression information to recognize the user's affective state. Facial expression recognition was achieved using 2D images, the sophisticated high-speed object recognition engine (SHORE) for face detection, and the computer expression recognition toolbox (CERT). The microphone was used to obtain audio signal from the user to determine the vocal features of pitch, flux, energy, signal centroid, and signal-to-noise ratio. The vocal features were classified as affective states using the J48 decision tree classifier and the JRIP decision rules on voice examples from TV shows, audiobooks, interviews, and databases with tagged voice corpus. Affective states from both modalities were classified as happy, neutral, sad, and surprise. The two modalities were combined using a Bayes Theorem based decision-level fusion approach. In an HRI experiment with the robot, participants were asked by the robot to act out each affective state. Results showed that multimodal fusion achieved a classification rate of 77%.

In [141], the humanoid robot, Nao, was used to simulate adult-child interactions. The robot was equipped with four microphones, and a 2D camera to recognize affect from both human voice and gait. Voice features of speech rate, volume range, high-frequency energy ratio, and pitch range, and gait features of walking speed, maximum foot acceleration, step variance, and maximum step length are extracted. Features were mapped onto a 4D affect Gaussian Mixture Model known as SIRE (speed, intensity, irregularity, and extent) by calculating the Z-score of data points relative to the mean and variance over the dataset. Experiments investigated whether

the robot can recognize affective gait based on only training from affective voice. Results showed the classification rates for happiness, sadness, anger, and fear as 62%, 90%, 43%, and 55%, respectively.

Other automated multimodal affect recognition systems have also been proposed [105], [190]-[192], but not yet incorporated in robots engaged in HRI. For instance, in [189], a multimodal affect-recognition system using facial expression and voice information was proposed for multipurpose HRI scenarios. The system recognized seven affective states: anger, happiness, neutral, sadness, surprise. For visual facial expression recognition, an AAM was used to describe and generate both the shape and texture of face from facial features. Face pose and facial features (nose, mouth, and eyes) were recognized by a face-detection module [193], where the module generated a feature-based model for every face found and searched the next frame for similar face model information. The facial features were used to initialize the iterative AAM fitting algorithm. A one-against-all SVM was, then, applied to classify the AAM fitting into the seven affective states.

For speech-affect recognition, EmoVoice [194], a framework that contained analysis on emotional speech databases and emotional speech application, was used. Features extracted were based on the prosodic and acoustic properties of speech signals such as pitch, energy, linear regression and range of frequency spectrum of short-term signal segments, length of voiced and unvoiced parts in an utterance, and number of glottal pulses. Furthermore, SVM with a linear kernel was used for classification of the seven affects. For fusion of both emotion recognition modalities, Bayesian Networks were used where results from the individual facial and voice classifiers were fed into the network and the posteriori probabilities were determined. Performance evaluation of the multimodal recognition system was conducted on the DaFEx database [195]. The overall average recognition rate of the system was 78.17%, compared to 74.46% for facial expression recognition, and 61.90% for voice-affect recognition. Another experiment was conducted on four subjects, where the subjects were asked to display facial expressions of five emotion classes: anger, happiness, neutral, sadness, and surprise with and without speaking. Results showed that the multimodal approach again outperformed the single modal approaches: average emotion recognition rate of 58.15% for multimodal, 55.39% for facial expression modality, and 23.63% for voice modality, respectively.

2.6.3 Multimodal HRI Systems

A number of social robots with automated multimodal affect recognition systems have been reported in [104], [137], [141] as well as with their own emotional response models in [196], [198]-[201]. However, only a few of these robots can recognize multi-modal affect of humans and use this as input to determine their own emotional behavior during HRI [19]-[22].

Robots that engage in bi-directional emotional HRI can be classified into those that mimic the emotions of users [19], and those that determine their own emotions based on a user's affect [20]-[22]. In [19], the humanoid robotic head, Muecas, was used to mimic affect using facial and vocal expressions as inputs from a user. Facial information was obtained using an onboard camera, and facial expression recognition was achieved through Gabor filtering and dynamic Bayesian network classification. Vocal signals were obtained using two microphones, and vocal features of speech rate, pitch, and energy were extracted. Another dynamic Bayesian network classifier was used to determine vocal expressions using these features. A multimodal decision-level fusion, also based on a final dynamic Bayesian network, combined the two modalities to recognize the following affective states: happiness, sadness, anger, neutral, and fear. Experiments showed Muecas was able to recognize and mimic the aforementioned affective states through its own facial expression display using a facial action unit reconstruction model.

In [20], a stuffed-animal-like robot, CuDDler, was developed to respond to user affect during HRI. The affective states recognized were happy, neutral, sad, angry, and surprise. An onboard webcam and two onboard microphones were used to extract frontal facial images, and interaction sounds from the user. Interaction sounds included crying, laughing, or non-voiced occurrences such as patting, stroking, and punching. Local binary pattern features were extracted from facial information and input into a linear support vector machine (SVM) classifier to determine facial affect. Acoustic sound signatures were used to classify sound events. The facial expressions and emotional sounds were synchronized and used to determine an appropriate affective response (happy, sad, and angry) for the robot using a state machine. Experiments with participants interacting with the robot by touching it and displaying facial expressions showed the robot was able to recognize the affective acts of pat, hit, and stroke, and responded appropriately to these situations by blinking its eyes, and displaying gestures and sounds.

In [21], the humanoid robot, AMI, was developed to communicate with a human. An affect recognition system was used to classify multimodal data from facial and vocal expressions into

the affective states of happy, sad, and angry. Facial action unit features of lips, eyebrows, and the forehead, and vocal features of phonetic and prosodic properties were extracted and classified as affective states using neural networks. Multimodal affect recognition was accomplished using decision-level fusion, where the final affect was determined by a weighted sum of the results of the two modalities. The robot used this information to produce its own affect through a synthesizer. The synthesizer used an emotional status model, which considered emotional drives, human emotional status, and a decaying term that restored the robot's emotional status to neutral. It was shown that AMI could recognize user affect and display its own appropriate affect through dialogue, facial, and gesture expressions.

In [22], a mobile robot was utilized to determine a user's neutral, happy, sad, fear, or anger affective states using facial and vocal expressions via an onboard 2D camera and a microphone. Facial features of action units from the eyes, eye brows, noise, and mouth were extracted using principal component analysis. Voice features of sampling frequency, pitch, and volume level were extracted using the Praat vocal toolkit. Both feature vectors, in addition to robot's social profile (sympathetic, anti-sympathetic, and humorous) were inputs to a Bayesian network that determined the robot's own multimodal affect, which it displayed using facial and vocal expressions. Training of the network using live 2D images of facial expressions and a database of vocal expressions was complete when a classification rate of 80% was achieved. Experiments consisted of participants rating the robot's response as funny, neutral, or aggressive based on its social profile.

2.7 Affect Classification Techniques

The majority of affect-classification techniques have successfully incorporated learning algorithms in order to identify distinct affective states or levels. For facial-affect detection this has included the use of Binary Decision Trees [88], AdaBoost [125], Multilayer Perceptrons [90], SVMs [126] and SVRs [90], [107], NNs [90], [127]-[130], and Dynamic Bayesian Networks [106].

Affect-classification techniques using body language have included the use of SVM or KNN [94], nearest neighbor [93], as well as the testing of a variety of learning algorithms [91], [110]. Only a few systems have been developed that incorporate body language features that have been

linked to affect in psychology or human behavior studies [110], [202]. Although identifying affective states from body language is still an open research area in psychology and human behavior research [203], a number of findings in these research areas can be applied to automated affective body language detection during social HRI. For example, it has been found that culture and context are important when attempting to identify affect from body language displays [203]. Utilizing such additional information will allow robotics researchers to develop more effective human affective body language detection systems that can recognize natural (non-acted) body language during HRI.

GMMs [98], [100] and SVMs [114], [135] have been the most common learning techniques used for classifying affective voice features during HRI. Naïve Bayes [69] and HMMs [96] have also been investigated. Future research in this area should focus on expanding the types of learning algorithms investigated for affect detection. The affect-detection systems using physiological signals have utilized fuzzy models [101], [142], HMMs [140], SVM [117], [182], and NN [4] learning techniques for classifying affect from identified features. The majority of these systems reported positive correlations between recognized affect and baseline affect, with only three studies reporting recognition rates. Future research on systems for detecting affect from physiological signals should also focus on providing more quantitative results with respect to the recognition of specific affective states/levels as is the standard with the other modes. Classification techniques utilized by multi-modal systems included HMMs [116], Bayesian network [104], [136], SVMs [137], [138], NNs [123], and statistical methods [141].

2.8 HRI Scenarios

Affective interactions between a human and a robot may take several forms, including: collaborative HRI [204], assistive HRI [205], mimicry HRI [177], and general HRI (e.g., multi-purpose) [206]. Collaborative HRI involves a robot and a person working together to complete a common task [207], [208]. A robot needs to be able to identify a person's affective state to avoid misinterpreting social cues during collaboration and to improve team performance [95]. Assistive HRI involves robots that are designed to aid people through physical, social and/or cognitive assistance [183], [209]. Automated affect detection would promote effective engagement in interactions aimed at improving a person's health and wellbeing, e.g. interventions for children

with autism [115] and for the elderly [210]. Mimicry HRI consists of a robot or person imitating the verbal and/or nonverbal behaviors of the other [211]. Mimicry of affect can also be used to build social coordination between interacting partners and encourage cooperation [212]. Lastly, general or multi-purpose HRI involve robots designed to engage people using bi-directional communication for various applications.

With respect to the specific HRI scenarios, human-affect recognition has been used in a number of game scenarios with robots. A number of them have been specifically with children, ranging from playing chess [91], [109], [121], playing a quiz game [108], playing a basketball game with children with autism [182], simulating a child to engage in adult-child interactions [141], or playing a movement imitation game [92]. Game scenarios with adults have also included playing 20 questions [96] as well as movement imitation [119]. In these scenarios, the affect-aware robots were mainly providing a collaborative role by playing the game with the user, or they were involved in behavioral mimicry. In [182], the robotic arm moving the basketball hoop was used to investigate the use of robotic therapeutic interventions for children with autism. Other assistive applications of robots, in particular for the elderly, have also been considered including delivering medicine [4] and food [125], as well as providing assistance with activities of daily living including meal eating [110]. The remaining HRI scenarios have been more general and not application dependent, namely focusing on the ability to detect affect [94], [98], [103], [107], [123], [126], [129], [130], [154], [132]-[133], [135], [137], or learning to mimic human affect [90], [92], [106], [119], [127], [128], [131], [136], [155].

2.9 Affect Databases

A number of databases have also been used to evaluate affect-recognition methods. Popular databases include the Cohn-Kanade database [213], CMU database [214], JAFFE database [215], DaFEx database [195], EmoVoice [194], and EmoDB database [216]. The majority of existing affect detection systems for social HRI have been tested on acted affect displays from facial expressions [88], [107], [125], [126], [131], [132], body language [92]-[94], voice [69], [98], [100], [134], [135], and multimodal [106], [136], [137], [141] inputs. In general, systems that rely on physiological signals have mainly concentrated on utilizing non-acted data from real-world interactions [4], [101], [103], [117], [119], [142], [140], [182]. A smaller number of single

mode affect-detection systems using facial expressions, voice or body language have utilized non-acted data from real-world HRI scenarios [4], [91], [95], [96], [104], [108]-[111], [113], [114], [121], [125]. Databases and acted evaluations can provide an initial approach to testing the performance of these systems. However, they do not provide the sensory information from the real-world scenarios needed for long-term training or testing of systems being used for affect detection during natural real-world interactions. Experimental studies in intended settings with robots and a large number of participants including the targeted users such as children and the elderly will need to be performed to investigate the performance capabilities of affect-detection systems. Furthermore, investigations into the affect displays of different cultures will need to be conducted to ensure the wide use of such automated affect-detection systems during HRI. Developing systems that are robust to age and cultural backgrounds will allow social robots to engage a larger number of users in natural social HRI.

2.10 Summary

This Chapter reviewed affect models used in HRI, and affect recognition during HRI. The two types of affect model used in HRI are categorical and dimensional. The most common discrete affective categories in categorical models are Ekman's six basic affect – disgust, sad, surprise, anger, fear, and surprise. On the other hand, the most common dimensional model used by robots in HRI has been the two-dimensional circumplex (valence-arousal) model. Affect recognition has been accomplished using 4 different modalities – facial expression, body language, voice, and physiological signals. There have been systems that also combine one or more modalities.

So far, multimodal affect recognition for HRI has focused on the primary modes of facial and vocal expressions [19], [20], [21], [22]. Body language has yet to be incorporated into these affect recognition systems. However, body language plays an important role in conveying changes in human emotions during social interactions [23]. Furthermore, vocal intonation has been strongly correlated to body language displays [24]. Vocal intonation plays an important role in conveying changes in emotion through the vocal properties of pitch, tempo, and loudness during social interactions [25]. In the next Chapter, a multimodal HRI architecture is presented utilizing the unique inputs of body language and vocal intonation.

Chapter 3
Multimodal Affect Recognition for Socially Assistive Robots

This Chapter presents the development of a novel multimodal emotional HRI architecture to promote natural and engaging bi-directional communications between a social robot and a human user. The proposed architecture allows the robot to recognize and classify user affect via a unique combination of body language and vocal intonation. The architecture has been implemented via a humanoid robot to perform diet and fitness counselling during HRI. Herein, specifically, the robot's ability to detect user affect and adapt its emotional response accordingly is demonstrated. Extensive experimental results have shown that the robot can effectively detect user valence and arousal levels during real-time HRI. In Section 3.1, the proposed multimodal affect recognition system is presented. In Section 3.2, an HRI experiment using the architecture and its results are presented. Finally, Section 3.3 summarizes the Chapter.

3.1 Proposed Multimodal Affect Recognition System

The multimodal emotional HRI system architecture, Fig. 3-1, comprises three main subsystems: (1) the multimodal affect recognition sub-system (MARS), (2) the robot emotion model (REM), and (3) the interaction activity sub-system (IAS). The inputs are from both the user and the robot itself. Sensory information from a 3D depth sensor and microphone are used by MARS to determine both body language and vocal intonation features and, then, to classify a user's overall affect based on valence and arousal levels using both modalities.

Both the user's affect and verbal responses during an interaction are used by the robot to determine its own emotional behavior. The user's affect is used by REM along with inputs from the robot's onboard sensors (e.g., touch sensors and 2D camera) and the robot's own internal state to determine appropriate emotions for the robot to express. The IAS uses these emotional expressions as well as the utterances of the user to determine the robot's emotional behaviors during HRI.

The focus of this Chapter is the MARS component of the multimodal HRI system architecture. The REM, and IAS components of the architecture were completed by Nolan Lunscher, and Tianhao Hu for a complete multimodal HRI architecture.

The proposed multimodal affect recognition system is used to classify a person's affect based on a 2D valence-arousal scale in real-time. Valence is used to define a user's level of pleasure, and arousal is used to define his/her excitation level [217]. The valence-arousal scale was chosen as it encompasses all possible affective states and their variations [218]. In addition, valence and arousal better represent experimental and clinical findings compared to categorical emotional models (e.g., happy, angry, sad) [219].

A decision-level fusion is utilized to effectively estimate a user's multimodal affect during HRI based on both body language and vocal intonation. Namely, affect from each of the two modalities is determined first and, then, combined to determine overall affect.

3.1.1 Body Language

Postures and body movements have been shown to be directly correlated to a person's affect [220], [221] and be used effectively to communicate affect during social interactions [5]. Body language has been defined as an interaction of at least two seconds long [220].

Previous work in this area by others in our research team has focused on identifying body language features and validating these features for automated recognition and classification by a robot [222]. Herein, these features are utilized for the body language mode for multimodal affect classification. Namely, the body language descriptors of bowing/stretching trunk, opening/closing of the arms, vertical head position and motion of the body, forward/backwards head position and motion of the body, expansiveness of the body, and speed of the body are utilized, Appendix A. Real-time identification and tracking of these features is achieved herein by using 3D information of the user's body provided by a Kinect™ 3D sensor. Namely, 20 position coordinate points of the body are identified and tracked using the Kinect Skeleton [223], including on the head, shoulder center, spine, hip center, both left and right hands, wrists, elbows, shoulders, hips, knees, ankles, and feet, Fig. 3-2. Dynamic body language features are, then, calculated using the tracked points, forming a feature vector at a sampling rate of 30 fps.

Once the feature vector is obtained, affect classification takes place. Random forest decision trees are used herein to classify both valence and arousal.

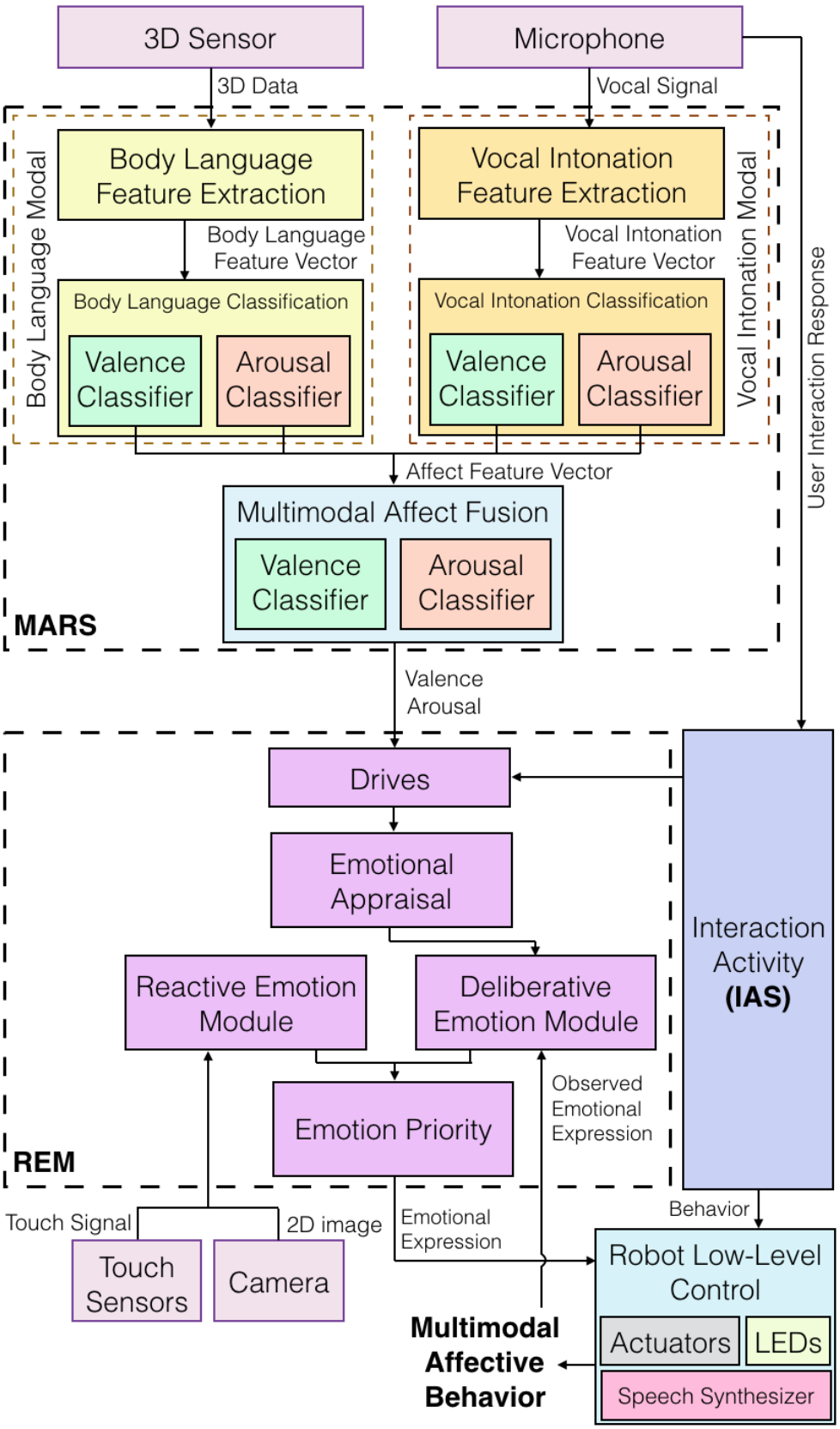


Fig. 3-1. Proposed Multimodal Emotional HRI Architecture.

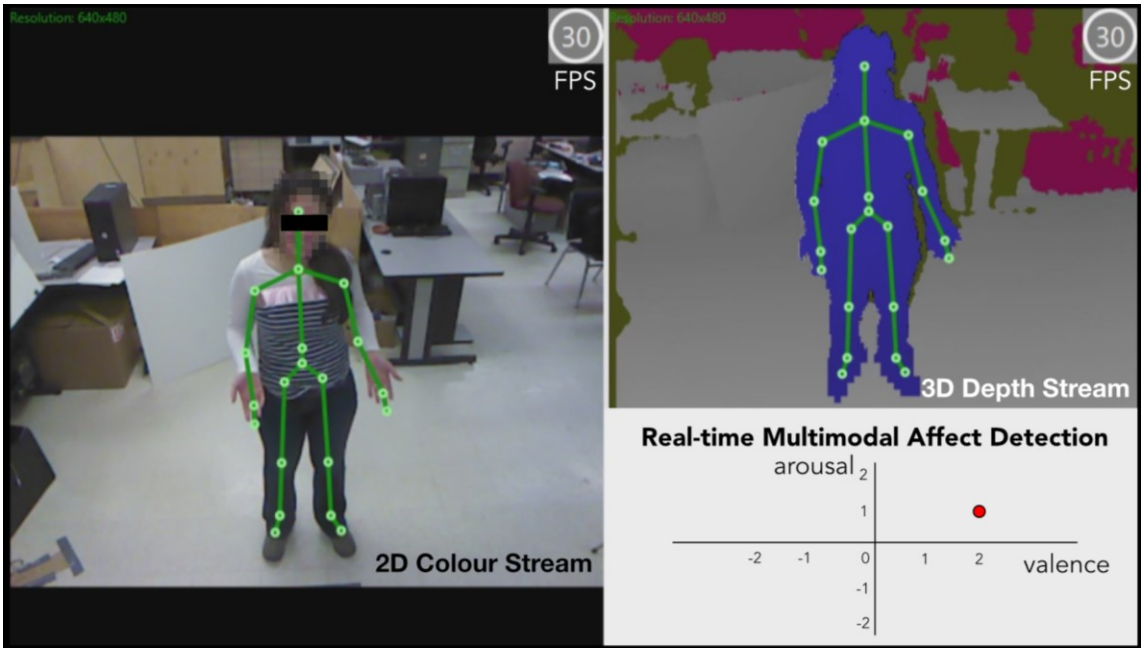


Fig. 3-2. Skeleton Fitting in Real-time.

3.1.2 Vocal Intonation

The vocal intonation affect recognition method was developed by Yuma Tsuboi. Audio signal patterns in vocal intonation, exclusive of speech content, are used herein to estimate vocal intonation of the user. Vocal intonation features are identified based on the local extrema and flatness in amplitude of vocal signals [224]. These are based on the peaks and plateaus of the signal. In order to extract vocal intonation features in real-time HRI scenarios, a noise-cancelling microphone array is needed, such as the Voice Tracker II array microphone used herein. These features are directly extracted from the vocal signals using the Nemesysco QA5 SDK software [225].

The extracted features in the vocal intonation affect recognition are level of content, anger, excitement, upset, energy, hesitation, embarrassment, stress, extreme state, arousal factor, imagination activity, intensive thinking, concentration level, uncertainty, brain power, max amplitude volume, voice energy, and emotion-cognition ratio, Appendix B. Vocal intonation is classified using model trees (decision trees with linear regression). This vocal intonation system has also been applied to recognizing affective states in elderly, Appendix C.

3.1.3 Multimodal Affect Fusion

Decision-level fusion combines the classified valence, v_b, and arousal, a_b, values from body language, and the valence, v_v, and arousal, a_v, values from vocal intonation, forming an affect feature vector for decision-level affect classification. Both body language and vocal intonation affective values have a one-to-one correspondence as they are taken from the same interaction time interval for decision-level fusion.

In order to classify the affect feature vector, a Bayesian network is used for multimodal valence and arousal classification. For the multimodal affect vector, $\mathbf{c}_m = \begin{bmatrix} v_m \\ a_m \end{bmatrix}$, containing multimodal valence value, v_m, and multimodal arousal value, a_m, the joint probability function is defined as:

$$P(\mathbf{c}_m, v_b, a_b, v_v, a_v) = P(v_b \mid \mathbf{c}_m)P(a_b \mid \mathbf{c}_m)P(v_v \mid \mathbf{c}_m)P(a_v \mid \mathbf{c}_m)P(\mathbf{c}_m). \quad (3\text{-}1)$$

From the joint distribution, the posterior probability of $\mathbf{c}_m$ can be obtained by applying Bayes' Theorem:

$$P(\mathbf{c}_m \mid v_b, a_b, v_v, a_v) = \frac{P(v_b \mid \mathbf{c}_m)\, P(a_b \mid \mathbf{c}_m)\, P(v_v \mid \mathbf{c}_m)\, P(a_v \mid \mathbf{c}_m)}{P(v_b, a_b, v_v, a_v)}. \quad (3\text{-}2)$$

The multimodal class, $\mathbf{c}_m$, with the highest probability is chosen as the output of the multimodal affect recognition system, and the corresponding valence and arousal values are passed to REM. The REM, then, utilizes a user's affect levels (from MARS) and interaction responses as inputs in order to generate appropriate emotional states and expressions for the robot.

3.1.4 Implementation with Robot Platform for HRI

The multimodal emotional HRI architecture can be applied to a number of different bi-directional HRI scenarios between social robots and users as defined within the interaction activity sub-system. One such HRI scenario is presented in this Chapter. This Chapter briefly describes the IAS component of the architecture completed by Nolan Lunscher and Tianhao Hu.

The objective of the IAS, considered herein, is to determine the appropriate robot behavior to motivate a user to live a healthy lifestyle through meal and exercise planning, and to assist him/her to do so by offering suggestions of meals and exercises each day. A one-on-one multi-modal HRI scenario is designed where a robot can provide social assistance with this task. Currently, Autom is the only robot that has been designed to provide and monitor meal and exercise plans [226]. However, Autom utilizes user inputs provided through its tablet PC and does not engage in bi-directional affective communication.

For experiments, interactions between the robot and users took place twice a day: one in the morning, where the robot made recommendations for the rest of the day, and one in the evening, where the robot checked-in with the user. The behavior of the robot was designed using a finite-state machine for the morning and evening interactions:

Morning Interaction: During this first interaction, at the start of the day, the robot greets the user and introduces itself, inquires about the weather and the user's dietary restrictions, and provides healthy-lifestyle meal and exercise suggestions.

Evening Interaction: During this interaction, at the end of the day, the robot carries out a social exchange, such as asking about the user's day and, then, inquires whether the suggested meals and exercise activities were complied with. The robot provides positive feedback if the user has eaten the suggested meals or healthy alternatives, and completed the suggested exercise plan and has been active. Otherwise, the robot encourages the user to follow the suggestions in the future.

The HRI architecture was tested through a Nao robot platform, developed by Aldebaran Robotics. The Nao robot has 25 degrees of freedom mobility, 8 RGB LEDs around each eye, which are used herein to display the robot's multimodal emotional behavior, a synthesized voice that can be controlled via the pitch, speed and volume, touch sensors on its hands, feet and top of the head, and two cameras in its head. Each emotion had a set of expressions (high intensity and low intensity) defined by a unique combination of eye color, body language and vocal intonation. Some of these expressions were adapted and extended herein from [227].

3.2 Experiments

The proposed multimodal HRI architecture was evaluated through a HRI experiment with eight participants. The primary objective was to investigate the robot's ability to recognize user affect and adapt its own emotions based on the activity interaction.

3.2.1 Training of Body Language Affect Recognition Classifier

The body language affect classifier was trained on a database comprised of 362 samples of dynamic body language. The dynamic body language training samples were based on Wallbott's [220], and de Meijer's [221] affective body language, TABLE 3-1. Both valence and arousal were recorded based on a scale of -2 (high negative) to 2 (high positive).

In order to classify succeeding feature vectors into affect levels during real-time HRI, Weka Data Mining Software [228] was used to determine the appropriate classifiers to use. A number of classifiers were investigated: 1) Naïve Bayes, 2) AdaBoost with Naïve Bayes, 3) logistic regression, 4) multi-layer perceptron neural network, 5) *k*-nearest neighbors, 6) random forest decision tree, 7) support vector machine (SVM), and 8) radial basis function network. The classifiers selected encompasses a variety of learning techniques, including probabilistic learning, decision trees, lazy learning algorithms, meta-classifiers, neural networks, and non-linear models. In particular, Naïve Bayes, a probabilistic technique, was investigated due to its robustness to irrelevant features [229]. AdaBoost, a meta-classifier, was investigated due to its ability to re-weight a base classifier's misclassified samples and generate an updated classifier [230]. Logistic regression, a linear technique, was investigated due to its ability to handle data types that are not necessarily normally distributed, linearly related, or of equal variance [231]. Neural networks and SVM, non-linear techniques, were investigated due to their ability to respond well to feature vectors that have not been utilized in training [232]. *k*-nearest neighbors, a lazy learning technique, was investigated due to its ability to estimate complex target function [232]. Lastly, random forest, a decision tree technique, was investigated due to its robustness to outliers and over-fitting [233].

Ten-fold cross validation was used to test each candidate classifier. Random Forest decision tree achieved the highest classification rate of 93.6% and 95.8% for valence and arousal, respectively, TABLE 3-2.

TABLE 3-1. BODY LANGUAGE ASSOCIATED TO AFFECT LEVELS

Affect Levels (Valence, Arousal)	Body Movements and Postures
$(-2,-1)$	Bowing trunk, head forward, hanging arms, low movement dynamics
$(+2,+2)$	Stretching trunk, open arms, overall upwards motion, high movement activity and dynamics with expansive movements
$(-1,+2)$	Bowing trunk, high movement activity, high movement dynamics
$(-1,-2)$	Bowing trunk, head tilted back, hanging arms, low movement activity and inexpansive movements
$(+2,0)$	Stretching trunk, head forward, arms hanging, low movement dynamics
$(-2,+1)$	Collapsed upper body, downward head, arms crossed in front of chest, inexpansive movements
$(+1,0)$	Stretching trunk, opening arms, overall upward and forward motions, and low movement dynamics
$(0,+2)$	Stretching trunk, overall backwards motion, and high movement dynamics

TABLE 3-2. BODY LANGUAGE CLASSIFICATION RATES

Classifiers	Valence (%)	Arousal (%)
Naïve Bayes	88.4	93.6
AdaBoost with Naïve Bayes	88.4	93.1
Logistic Regression	90.6	91.2
Multi-layer Perceptron Neural Networks	90.1	92.5
k-Nearest Neighbours	91.7	95.3
Random Forest	**93.6**	**95.8**
SVM	80.7	83.7
Radial Basis Function Network	91.2	95.3

3.2.2 Training of Vocal Intonation Affect Recognition Classifier

This component of MARS was completed with Yuma Tsuboi. The vocal intonation affect classifier was trained on a database comprised of 362 samples of vocal intonation. The vocal intonation training samples were based on Scherer's affective voice classification [25], TABLE 3-3. The same classifiers above were investigated through the Weka Data Mining Software, with

the addition of Model Trees. Model Trees, a decision tree algorithm with linear regression at the leaves, was investigated due to its ability to handle numeric features [234]. Ten-fold cross validation using Model Tree achieved the highest classification rate of 55.6% and 61.9% for valence and arousal, respectively. Other classifiers have also been trained for comparison, TABLE 3-4.

TABLE 3-3. VOCAL INTONATION ASSOCIATED TO AFFECT LEVELS

Affect Levels (Valence, Arousal)	Voice Features
$(-2,-1)$	Down pitch contour, low pitch level, slow tempo, round envelope
$(+2,+2)$	Large pitch variation, small amplitude variation, fast tempo, sharp envelope, high pitch level
$(-1,+2)$	Small pitch variation, up pitch contour, high pitch level, fast tempo
$(-1,-2)$	Small pitch variation, low pitch level, slow tempo, round envelope
$(+2,0)$	Small amplitude variation, large pitch variation, fast tempo, sharp envelope
$(-2,+1)$	Small pitch variation, round envelope, slow tempo, up pitch contour
$(+1,0)$	Fast rate of speech and wider range of frequency
$(0,+2)$	Fast tempo, high pitch level pitch contour up, sharp envelope, large pitch variation

TABLE 3-4. VOCAL INTONATION CLASSIFICATION RATES

Classifiers	Valence (%)	Arousal (%)
Naïve Bayes	48.0	53.5
AdaBoost with Naïve Bayes	48.0	54.7
Logistic Regression	48.6	59.2
Multi-layer Perceptron Neural Networks	48.0	55.9
k-Nearest Neighbours	46.8	55.9
Random Forest	49.8	60.4
SVM	43.5	58.6
Radial Basis Function Network	46.5	53.7
Classification via Regression (Model Tree)	**55.6**	**61.9**

3.2.3 Training of Multimodal Affect Recognition Classifier

The classifiers within the multimodal affect recognition module were first trained using a body language and vocal intonation database that was created with 362 corresponding body and voice samples obtained from actors. Ten-fold cross validation using the multimodal affect Bayesian network achieved a classification rate of 93.4% and 97.2% for valence and arousal, respectively. Other classifiers have also been trained for comparison, TABLE 3-5.

TABLE 3-5. 10-FOLD CROSS VALIDATION MULTIMODAL AFFECT CLASSIFICATION RATES

Multimodal Classifiers	Valence Classification Rate (%)	Arousal Classification Rate (%)
Bayesian Network	**93.4**	**97.2**
Naïve Bayes	83.7	94.2
Logistic Regression	92.5	95.9
Random Forest	92.2	96.1
k-Nearest Neighbours	92.0	93.6

3.2.4 Multimodal HRI Experiments

HRI experiments were conducted with Nao for a diet and fitness planning HRI application. Eight, male and female, university students between the ages of 20 to 37 participated in the experiments. Participants were familiar with robots, but majority had not interacted directly with a social robot before. Each participant interacted with the Nao robot twice on the same day, i.e., the morning and evening interactions.

Nao is placed on top of a table at a distance of 0.75 m from the user for one-to-one standing interactions, Fig. 3-3. The microphone was placed in front of the robot on the table in order to capture a user's voice and the Kinect™ sensor was placed behind Nao in order to capture user body movements and poses during interaction.

During interactions between Nao and the users, the robot autonomously detected a user's affect in real-time, determined its own emotional states and expressions, and implemented its appropriate emotional behavior based on the activity. An operator was utilized only for user speech recognition during the experiments in order to minimize reliability issues of current speech recognition software. The microphone was used to provide audio output to the operator who was located outside of the interaction environment and not visible to the participants.

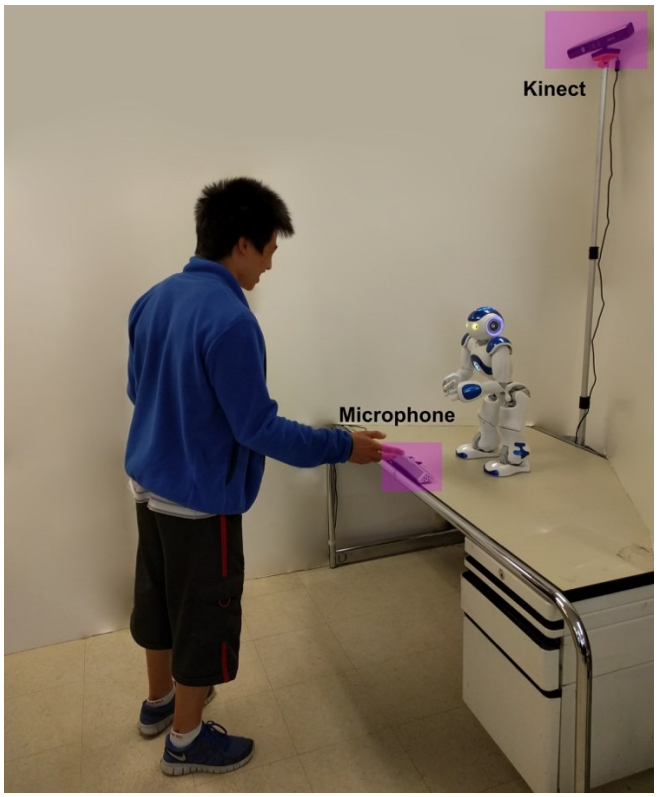


Fig. 3-3. Interaction Setup.

3.2.5 Results and Discussion

Fig. 3-4 and Fig. 3-5 show the affect levels of the participants during the morning and evening interactions. Both valence and arousal were recorded based on a scale of -2 (high negative) to 2 (high positive). Fig. 3-6 and Fig. 3-7 further expands on the interaction by showing the average affect across participants for each interaction stage. Fig. 3-8 and Fig. 3-9 show the corresponding robot emotion for each interaction stage.

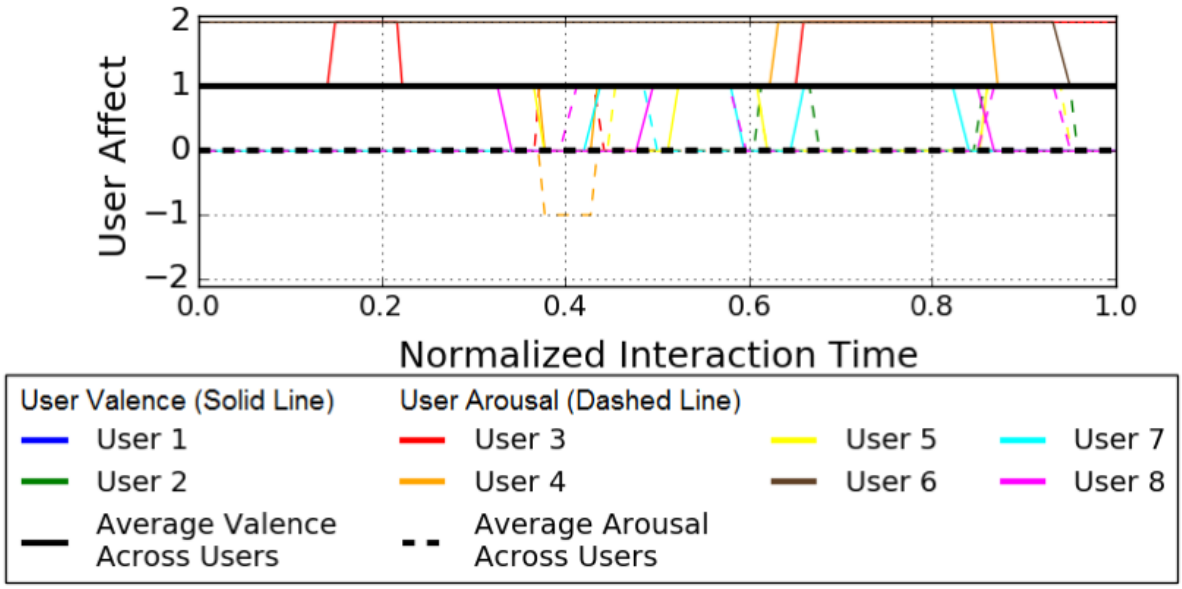


Fig. 3-4. User affect during the morning interactions for all participants.

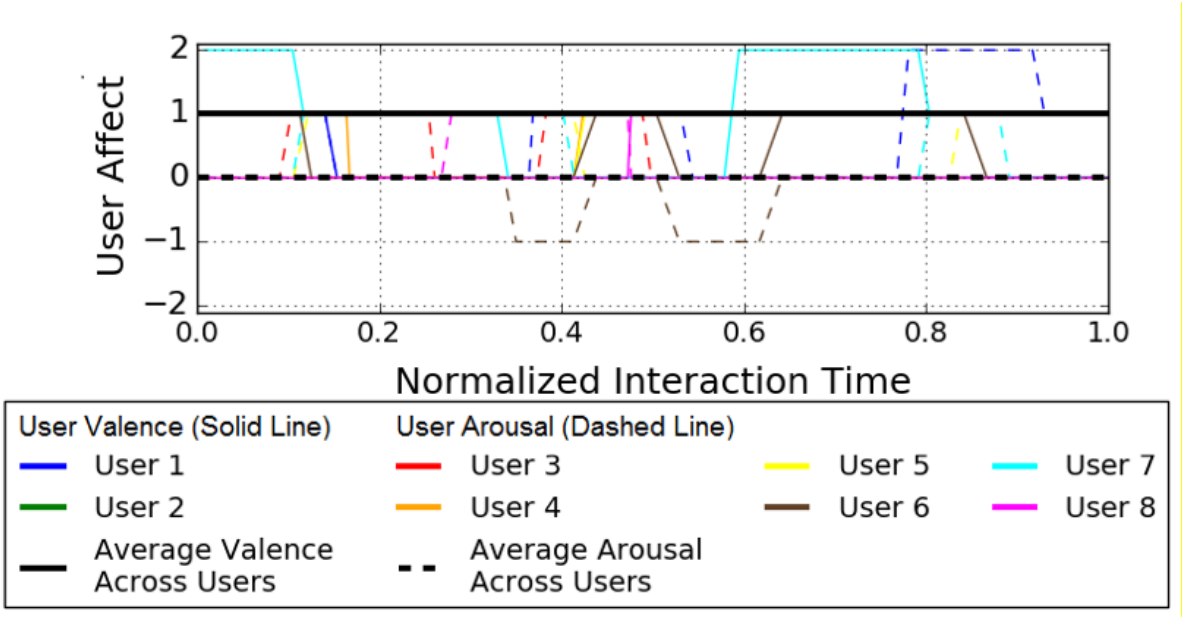


Fig. 3-5. User affect during the evening interactions for all participants.

On average, users had positive valence and neutral arousal during both interactions. A one-way ANOVA test was conducted for the user's affect during the different interaction stages for both morning (i.e., greet user, suggest a meal, suggest an exercise) and evening interactions (i.e., determine dietary requirements, check-in regarding meals, check-in regarding exercise). Examples of participant body language during HRI with different valence and arousal levels are

shown in Appendix D.

It was found that there were no significant differences between interaction stages ($F(12, 91) = 0.50, p = 0.908$ for valence and $F(12,91) = 0.58, p = 0.855$ for arousal).

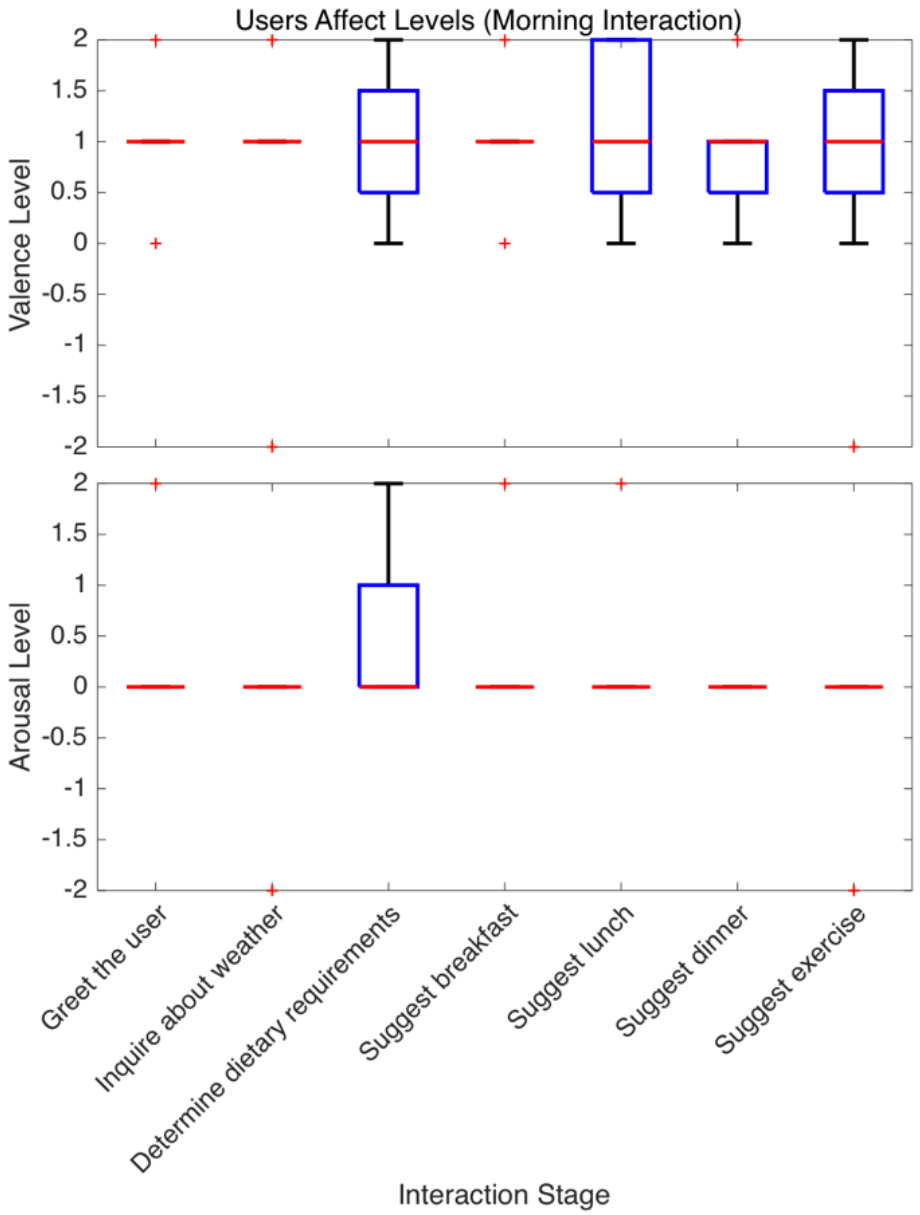


Fig. 3-6. Users affect level during morning interactions.

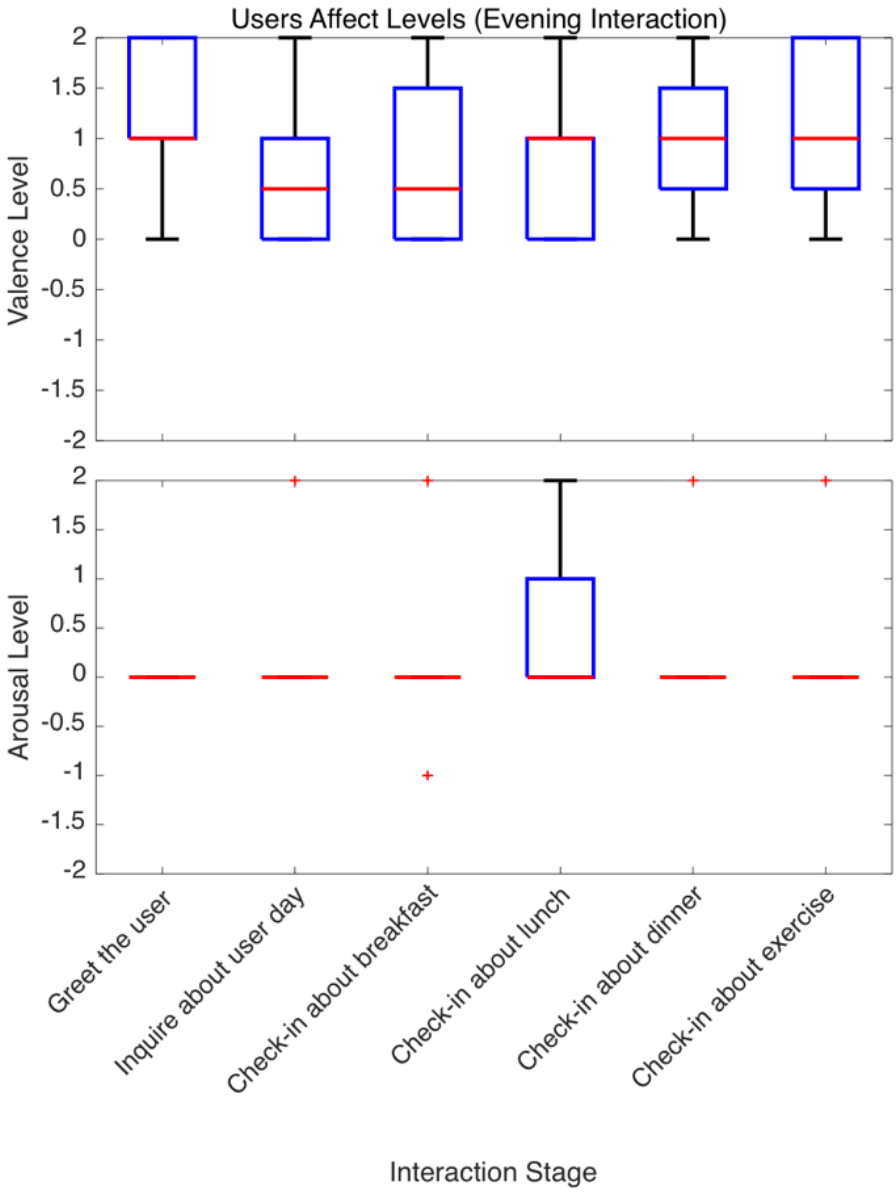


Fig. 3-7. Users affect level during the evening interactions.

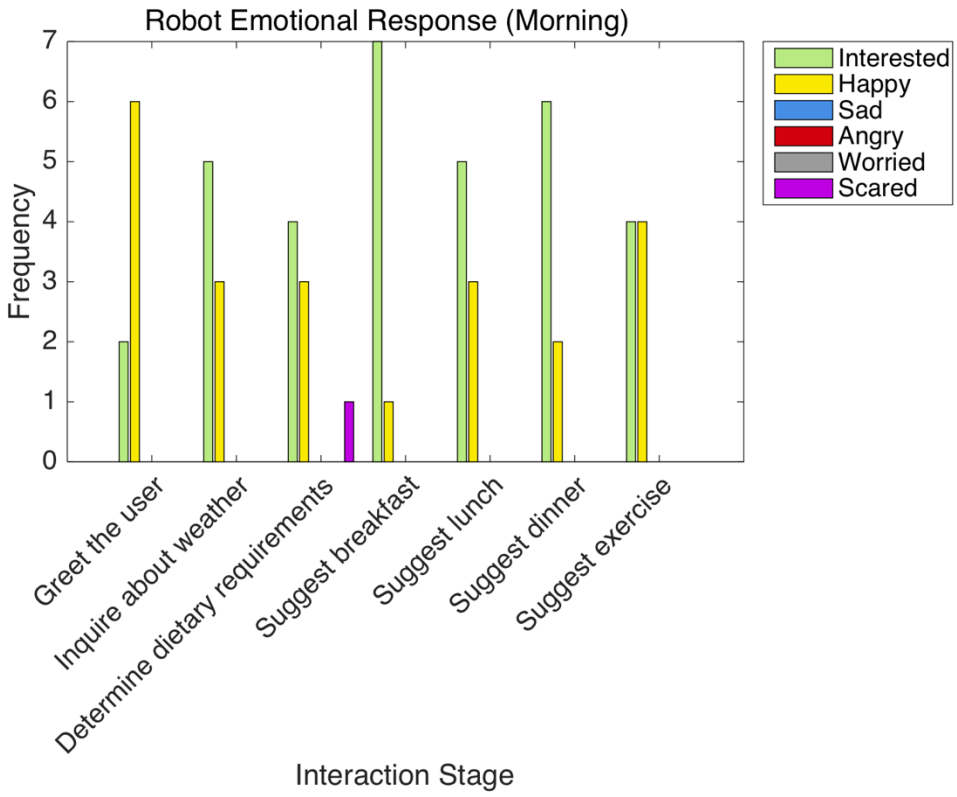


Fig. 3-8. Robot emotional response during the morning interactions for all participants.

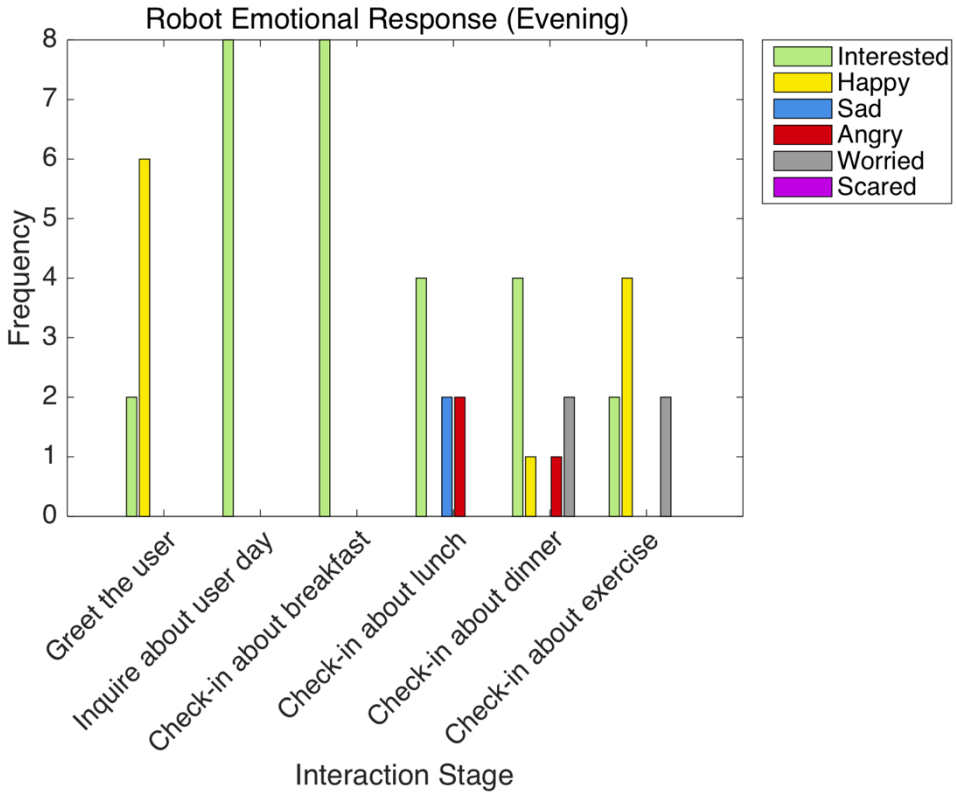


Fig. 3-9. Robot emotional response during the evening interactions for all participants.

The majority of the changes in the robot's emotional state during the morning interaction, Fig. 3-8, occurred while suggesting types of meals and exercises to each user, which the user agreed to. This was in accordance with the robot's desire to have the user comply with its suggestions. For one participant, the robot displayed a reactive emotion of scared as it was close to the ledge of the table (as detected by the robot's onboard camera), Fig. 3-8. The robot requested that the user assist it by moving it back away from the ledge. Once, the robot identified it was away from the ledge, it was no longer scared and transitioned back to the interested emotional state.

The majority of changes in the robot's emotional state during the evening interaction, Fig. 3-9, occurred while the robot was checking in to see if the user ate the suggested meals or completed the suggested exercise. This is consistent with whether the participants complied with the robot's suggestions. The robot either entered a happy emotional state when the users complied or a sad, worried or angry emotional state, when the users did not comply. Based on this input, the robot's desire to have the user comply with its suggestions was either succeeded or failed, resulting in either a positive or negative influence on its emotions.

As a detailed example, Fig. 3-10 and Fig. 3-11 show the interactions for User 7 during both the morning and evening, respectively. Throughout the morning interaction, Fig. 3-10, the robot transitioned between happy and interested states. During the morning interaction, the robot greeted the user in a low-intensity happy state. The robot transitioned to a high-intensity happy state when the user responded that the weather was nice outside. The robot was in an interested state when suggesting the meals and exercise plan to the user, and was in happy state when the user agreed to the robot's suggestions. During the first half of the morning interaction, the user displayed more high energy body movements and voice, while for the latter half, the user appeared to be more focused on the information being provided by the robot and leaned forward closer to the robot. Instances of positive valence were detected when the user liked and agreed to the lunch and dinner option meals.

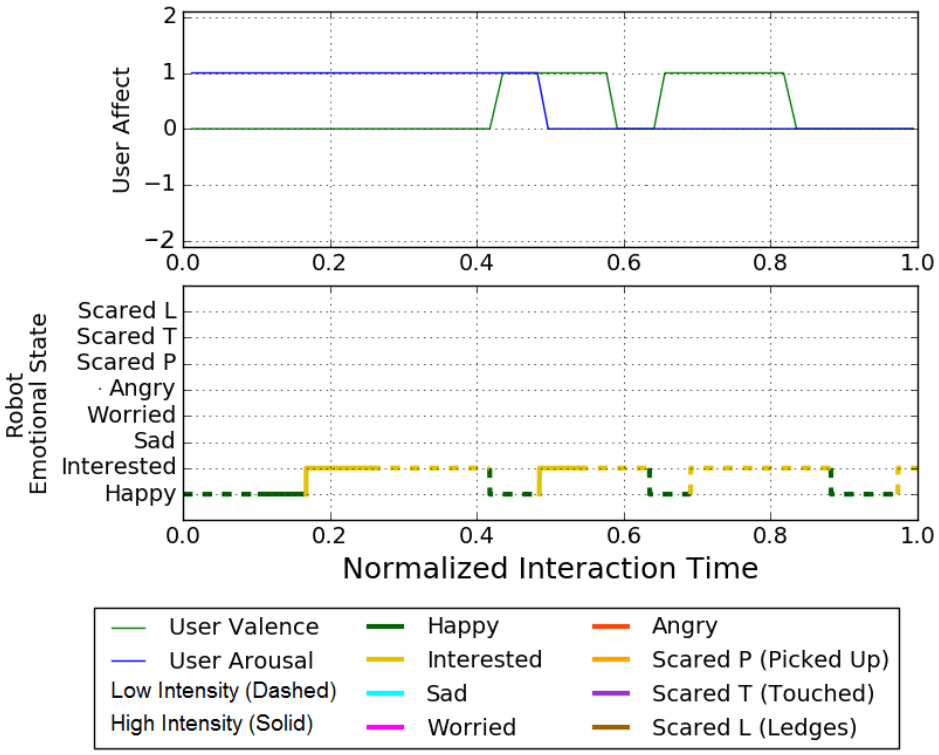


Fig. 3-10. Affect and robot expressions during morning interaction for User 7

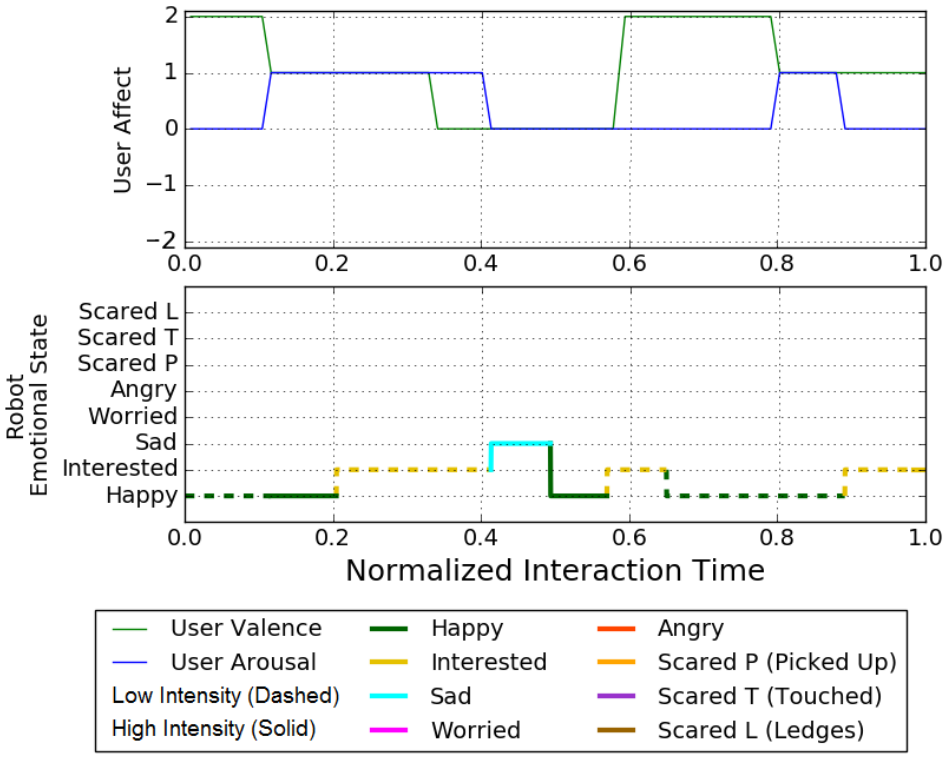


Fig. 3-11. Affect and robot expressions during evening interaction for User 7

During the evening interaction, Fig. 3-11, the robot greeted the user in a low-intensity happy state, and transitioned to the high-intensity happy state by displaying its dance when the user responded that his/her day was going great. Then, the robot transitioned to an interested state when checking-in about breakfast and finding out that the user ate the suggested meal, which the user did follow the meal suggestion. Next, the robot transitioned to a sad state when the user responded that they had not ate their suggested lunch. However, the robot transitioned back to a happy state when the user responded that they ate another healthy alternative instead. The robot was in the happy state when checking-in about dinner and exercise and transitioned to an interested state at the end of the interaction when user was discussing the level of difficulty of the exercise. The user displayed positive arousal (high energy body movements and voice) during the greeting in response to the robot's dance as well as when discussing his/her day and breakfast meal. Positive arousal was also detected during the checking-in regarding exercise stage when the user laughed at the robot when it told an exercise joke. Positive valence was detected (open and stretched body language, and high level of content in the voice) throughout the evening interaction, including during the greeting as the user was happy to see the robot and during the checking-in regarding breakfast, dinner and exercise.

3.3 Summary

In this Chapter, a novel multimodal affect recognition is presented to allow a robot to determine the user's affect during HRI using the unique combination of body language and vocal intonation. Integrating this system into a multimodal emotional HRI architecture allows for bi-directional communication between a user and a robot.

The architecture is verified with a small humanoid robot to investigate the robot's ability to detect affect, and adapt its emotion to changes in user affect and the progression of the interaction at hand during a diet and fitness counselling HRI scenario. Experimental results clearly showed that the robot can effectively recognize user affect, when compared to user self-reported affect, as well as adapt its emotions accordingly.

Chapter 4
Review on Human-Robot Teams in Urban Search & Rescue

Teams of semi-autonomous robots can provide valuable assistance in Urban Search and Rescue (USAR) by exploring cluttered environments while searching for potential victims. Their advantage over solely teleoperated robots is that they can address the task handling and situation awareness limitations of human operators by providing some level of autonomy to the multi-robot team. This Chapter presents a literature review of human-robot teams in USAR environments. In Section 4.1, robot exploration techniques are presented. In Section 4.2, the effect of operator-to-robot ratio on team performance is reviewed. Section 4.3 provides a literature review of task automation techniques in multi-robot teams. Section 4.4 reviews artificial intelligence (AI)-based approaches for controlling multi-robot teams. In Section 4.5, user interfaces for multi-robot control is reviewed. Section 4.6 provides a review of simulation environments used for USAR experiments. Finally, Section 4.7 summarizes the Chapter.

4.1 Robot Exploration

Exploration of USAR scenes is essential for finding victims. Two techniques that are used in USAR robot exploration are simultaneous localization and mapping, and Frontier-based exploration [26].

4.1.1 Simultaneous Localization and Mapping

Robot exploration and victim identification in USAR is a challenging task due to the sensing challenges in USAR environment. One component of robot exploration is mapping the environment and robots' location in the environment. Simultaneous Localization and Mapping (SLAM) techniques address the problem of robots navigating in an unknown environment. In SLAM, the robot seeks to acquire a map while navigating the environment, and at the same time, it wishes to localize itself using its map [235]. SLAM provides detailed environment models and accurate sense of a mobile robot's location. However, SLAM techniques mostly deal with static environments, yet nearly every actual robot environment is dynamic. More work is needed to

understand the interaction of moving and non-moving objects in SLAM [235]. Various SLAM approaches have been developed for USAR applications [26]. SLAM provides a way for robots to autonomously generate 3D maps of their environments and localize themselves as well as victims and objects of interest within these environments [26].

4.1.2 Frontier-Based Exploration

Navigation approaches in USAR, such as frontier-based exploration, have also been explored. Frontier-based exploration is a direction-based exploration technique based on the concepts of frontiers, regions on the boundary between open and unexplored space [236]. By moving to new frontiers, a mobile robot can extend its map into new territories until the entire environment has been explored. This exploration technique requires no previous knowledge of the map's topology.

The central idea behind frontier-based exploration is to gain the newest information about the world by moving to boundaries between open space and uncharted territory. The environment is mapped by employing occupancy grids to associate areas in the real world with grid cells in the map. Each grid cell is marked as either open, unknown, or occupied based on occupancy probability, the probability of the robot accessing that cell. An advantage to this approach is its ability to explore both large open spaces and narrow cluttered spaces, with walls and obstacles in arbitrary orientations [236]. This type of exploration has also been implemented with multiple robots [237]. Robots can share individual occupancy grids and perceptual information with each other. By doing this, a global grid can be created to be shared amongst group of robots. This approach enables robots to make use of information from other robots to explore more effectively, but it also allows exploration to be more robust to loss of individual robots [237].

This technique has been implemented on real robots, and demonstrated that they can explore and map office environments as a team [237]. This technique considers both the terrain information of an environment by classifying obstacle cells as climbable or non-climbable cells, and the direction of a robot to determine its ability to traverse a cell of interest. The performance of the semi-autonomous exploration approach has proved to have a significant increase in exploration coverage compared to autonomous exploration approach for this algorithm [44]. However, this exploration approach is limited as the implementation does not consider the

optimal path for exploration. This is important in USAR applications because time is of the essence.

4.2 Operator-to-Robot Ratio

Operator-to-robot ratios in USAR environment for teleoperated robots were extensively investigated. It was found that task performance increased as operator control increased from four to eight robots, but subsequently decreased as the number of robots was further increased from eight to twelve [40].

With the condition of live-streaming versus asynchronous video displays during teleoperation of multi-robot teams in USAR environment, greater performance was achieved in marking victim locations when using live-streaming [50]. However, with respect to overall performance, the two approaches were similar for all groups of robots, with the peak number of victims being found with eight robots when comparing between four, eight, and twelve-robot teams [51].

Multi-robot team structures of two operators and 24 robots were also investigated using both teleoperation and autonomous path planning control modes [43]. It was found that the team structure had no significant effect on the number of victims found, however, teleoperation exploration time was higher. Moreover, increasing the operator-to-robot ratio in teleoperation has been shown to increase effectiveness of task sharing [52].

4.3 Task Automation in Multi-Robot Teams

In order to increase team performance in USAR missions, numerous task automation methods have been investigated. In [54], two operators controlled 24 robots, using both semi-autonomous and teleoperation, to search for victims in a simulated USAR environment. Semi-autonomous control included low-level robot autonomy capabilities, such as path planning and navigation. Teleoperation consisted of robots navigating using operator assigned waypoints. The results showed that the operators could find more victims using the semi-autonomous controller. Furthermore, when comparing the regions that were explored, there was a substantial advantage for using semi-autonomous exploration over manual exploration when operators shared control of all the 24 robots.

Semi-autonomous controllers and full autonomous controllers were also compared. In [44],

USAR scene coverage using a semi-autonomous direction-based exploration approach for multiple cooperating robots by an operator was compared to full autonomous exploration. The robots were able to build a map of their local environments and, when needed, actively shared this information with other robots within the team in order to minimize exploration overlap. In contrast to the autonomous controller, the semi-autonomous controller would request human assistance when a robot was 'stuck' or needed help navigating through rubble piles. A comparative study showed that exploration coverage increased when the robot team size increased from one to four robots in both modes.

There have been automation algorithms that assist human operator with coordinating multi-robot teams for disaster response. In [55], an algorithm was presented to assist a human operator with coordinating a multi-robot team for disaster response. A decision tree algorithm was used in determining and adapting strategies to solve the complex problem via more manageable sub-problems. The algorithm was used to specify strategies, allocate agent to these strategies, and release agents in a timely manner to adapt these strategies. The algorithm was implemented in simulations for a team of up to twelve robots searching regions. The results showed that the algorithm was able to define an infinite number of alternative scenarios for human-defined strategies.

In [56], annunciator driven supervisory control was proposed to provide alarms to direct an operator's attention to sub-systems that need assistance within a larger more complex multi-robot system. A simulated USAR environment was used to test three conditions with six robots; no alarms, alarm condition for all robots, and a decision aid which only showed the highest priority alarm for a robot. The alarm condition reduced fault detection and victim detection times by having robots alert operators when they were in abnormal states, however, the select-to-mark times for victims was increased when using the alarms.

In [60], single-operator performance for a team of automated advising robots was compared to solely teleoperated robots. The advising robots assisted the operator by providing advice on which actions he/she should take for search and rescue tasks. Experiments conducted, in both simulations and physical environments, showed that the operator with feedback from advising robots covered more terrain, detected and correctly classified more desired objects, and reduced robots' idle time when compared to teleoperation.

Automation has also been shown to lower coordination demand for heterogeneous robot team.

In [57], two types of robots were considered; an *explorer robot* to build an environmental map and an *inspector robot* to search for victims. A simulated USAR environment was used with three of each type of robot to investigate three coordination demands (*i*) with varying sensor ranges for the explorer robot (Conditions 1 and 2), and (*ii*) when the explorer and inspector robots were paired in a sub-team (Condition 3). The results showed that operators can explore wider regions and find more victims when the explorer robots have a larger sensor range. Furthermore, the coordination demand for the explorers was found to be approximately twice that of the inspectors. When using the sub-team, the automatic coordination between the two robots resulted in lower coordination demand for the inspector robots than when the robots were independent, showing a potential benefit to automation. A simulated USAR environment was used with three of each type of robot to investigate coordination demands. It was found that the automation coordination between the two robots resulted in lower coordination demand compared to when the robots were independent.

In [45], a team management framework was presented to account for lost, failed, and new robots in a heterogeneous multi-robot team to be deployed in disaster zones. The framework also allowed for robot teams to be formed dynamically starting with a single robot. Task allocation was determined by the framework using the minimum requirements needed by a robot to complete a specific task and the suitability of a robot to complete the task. Experiments were conducted in simulated environments comparing the proposed framework against two baseline conditions: when a team structure is fixed and when robot tasks are fixed. The results showed the framework increased environment coverage and number of victims identified.

Foraging tasks of a multi-robot team were also investigated to determine which tasks can be automated to reduce operator workload [58]. Experiments were conducted in a simulated USAR environment with four, eight and twelve robots. Two conditions were tested: where operators had full control over each team of robots and performed the overall USAR foraging tasks, and where operators had independent control only over the exploration and perceptual search subtasks. Results showed that the perceptual search tasks when individually performed by the operators had better performance than the full control, especially, with increased number of robots. The overall results support the automation of the robot exploration tasks.

In [59], the performance of a single operator and a team of two cooperating operators in sharing control of a team of multiple robots were compared. All robots were equipped with a

semi-autonomous controller where they had the ability to follow paths set by the operator manually, or autonomously generate their own paths to user-specified go-to points. Experiments showed that mental activity and mental work fell, and time pressure felt was reduced when using the semi-autonomous robots.

As noted above, increased number of robots in a team has been shown to improve USAR mission performance in both exploration efficiency and victim identification. However, one of the many challenges a human operator could face in such environments is the simultaneous control of multiple robots. Thus far, in the literature, it has been shown that human operators can effectively control up to eight robots in a teleoperation mode before losses in performance [40], [50], [51]. Semi-autonomous controllers, on the other hand, have been shown to help reduce workload and, thus, increase the number of robots an operator can handle. These controllers, though, have mostly been shown to manage less than 1:12 operator-to-robot team ratios, and be limited in their capabilities to avoid obstacles and identify victims automatically.

4.4 AI-based Approaches

Markov Decision Process (MDP) based techniques have been proposed for multi-agent UAV navigation in semi-autonomous control [61], [62]. In [61], a Mixed Markov Decision Process (MIMDP) approach was utilized which was created from two MDP models, one for the autonomous system and one for the supervision unit. The approach allowed the autonomous system to decide what actions to take and when the supervision unit should be requested and control transferred to it. The MIMDP approach, however, was only implemented for a single UAV problem with experiments focusing on confirming the making of requests by the agent to an operator. In [62], a human-help provider MDP was presented for the control of UAVs by providing three different help requests to operators ranging from critical to non-critical. The system was tested with 1-15 UAV agents and 1-3 operators to determine how many requests from the agents were treated by the operators.

4.5 User Interfaces for Multi-Robot Control

Adaptive, adaptable, and adaptive delegation type human-computer interfaces have been developed to optimize workload, enhance user productivity, and increase user satisfaction.

Adaptive interfaces can automatically make decisions regarding the need for automation as well as change the degree of automation in order to decrease operator workload and increase situational awareness [238], [239]. For example, in [240], an interface was presented to dynamically modify objective function weightings of an automated planner for a group of unmanned vehicles, which resulted in enhanced SA, decreased workload, and fewer required operator interventions.

In adaptable interfaces, the operator determines the level of automation [239]. A type of adaptable interface is the delegation-type interface, where a trade-off between unpredictability of the system versus operator workload can be balanced [239]. Delegation-type interfaces allow operators to delegate tasks to automation at times of their own choosing, and receive feedback of the automation performance [239]. One main advantage of delegation-type interfaces is that they allow for flexible use of automation in response to unpredictable changes in task demands, while keeping the operator's mental workload within a manageable range [241]. For example, a delegation-type interface was presented in [239] using the "playbook" approach, where a hierarchical task model provided the ability for the human operator to communicate goals and plans to a robot task planning system in order to critique or complete plans provided by the operator to a group of robots searching for targets. The approach increased mission success rate and reduced mission completion time.

Adaptive delegation interfaces use a combination of both adaptive and delegation design. They allow both the operator and the system to define goals and plans to implement [242]. For example, in [242], an adaptive delegation interface was presented for controlling and monitoring multiple UAVs. Results showed that mental workload of operators was moderate and the interface did not overburden the operators, which is a potential concern when using adaptable interfaces.

4.6 Simulation Environment

In order to simulate USAR missions and test the performance of exploration algorithms, various software platforms have been used. Some of these platforms are based on 2D simulations, while others are more realistic and use a 3D game engine. MobileSim is a 2D simulation platform and has been used for testing autonomous navigation techniques, obstacle avoidance, and artificial

intelligence with robot teams [243]. However, two-dimensional simulations do not give a true representation of a real environment. In contrast, USARSim is a high-fidelity simulator that runs in a 3D game engine. Game engines are modular simulation code that can be used in creating a family of similar games [244]. For USAR missions to be as realistic as possible, USARSim uses an advanced 3D game engine known as Unreal Engine. This engine is the same engine used to build popular realistic games such as Unreal Tournament and Gears of War [244].

In the context of USAR applications, USARSim provides robot packages based on real-life models, and its virtual maps incorporate advance geometry and textures that faithfully simulate USAR environments [244]. In the past, USARSim has been used as the main platform for conducting performance evaluation in USAR robot competitions [26],[245]. In addition, USARSim has been used as a platform to test algorithms to find victims in USAR environments. Some of these algorithms include frontier-based exploration, and SLAM [245].

4.7 Summary

In Urban Search and Rescue (USAR), mobile robots can effectively explore disaster environments with minimum *a priori* knowledge about the location of victims and scene layout [27], [28]. The majority of past robotic USAR missions, however, have so far been based on the utilization of teleoperated single robots [28]-[30]. Operators of such robots have, typically, experienced perceptual difficulties in trying to understand the 3D cluttered environments via remote visual feedback [28]. Furthermore, the (single) rescue robots have experienced task-handling limitations. It has been found that operator-to-robot ratios in team can increase task performance, but can also decrease task performance after increasing the number of robots pass the operator's workload capacity. Task automation and AI-based approaches provide autonomy to relieve some workload off of operators.

The AI-based approaches, however, differs from the USAR problem addressed herein in that it involves more than simply navigation. Namely, the problem at hand comprises three sub-tasks: exploration, victim identification, and navigation. This significantly increases the state space of the USAR problem, limiting the use of traditional modeling techniques, such as MDPs, POMDPs (Partially Observable MDPs), or DCOPs (Dynamic Distributed Constraint Optimization Problems). The main drawback of these traditional techniques is that they often fail to scale up to

large numbers of sub-tasks and agents. Furthermore, both MDPs and POMDPs suffer from the curse of dimensionality, where state parameter dimensionality can increase exponentially with team size [246]. The technique to be presented, MAXQ, on the other hand, utilizes a hierarchical organizational structure by decomposing an overall task into a finite set of sub-tasks recursively, where each sub-task is modelled as a MDP. MAXQ uniquely supports temporal, subtask and state abstraction which can significantly reduce the number of state variables needed and speed up the overall learning process for real-world problems [247]. Furthermore, it has fewer constraints on its policies (i.e., mapping of states to possible actions) and, thus, is generally known to require less prior knowledge about the environment.

In contrast to the existing controllers, the hierarchical learning semi-autonomous controller, presented in the next Chapter, manages task allocation between robots and human operators effectively, while learning from the cluttered USAR environment to increase performance in exploration and victim identification, thus, allowing higher robot-to-operator ratios without significant performance loss.

The proposed approach is one of adaptive interface type, where the system assigns robot tasks, and the operator assists the robots with completing tasks when the system requests for assistance. What is unique about the approach is that the system can learn how to allocate and execute tasks as well as learn from the experience of an operator in order to further minimize operator workload.

Chapter 5
Human-Robot Teams for Learning-based Semi-Autonomous Control in Urban Search & Rescue Environments

This Chapter investigates, specifically, the influence of the operator-to-robot ratio on the performance of a proposed MAXQ hierarchical reinforcement learning based semi-autonomous controller for USAR missions. In particular, a unique system architecture that allows operator control of the rescue robots in a team as well as effective information sharing between the robots is proposed. A rigorous comparative study of the proposed semi-autonomous control-based system versus a fully teleoperation-based system was also implemented in the high-fidelity 3D simulation environment USARSim. The results showed that, for both semi-autonomous and teleoperation modes, the total scene exploration time increases as the number of robots utilized increases for larger USAR scenes. However, the rate of time increase is significantly less for semi-autonomous mode, thus, justifying the use of teams of semi-autonomous rescue robots. Section 5.1 presents the proposed multi-robot single-operator system architecture. Section 5.2 presents USAR mission experiments using the proposed architecture. Finally, Section 5.3 summarizes this Chapter.

5.1 Proposed Multi-Robot Single-Operator System Architecture

The proposed system architecture for semi-autonomous control of a multi-robot team is shown in Fig 5-1. The system encompasses both a user interface and a MAXQ HRL-based deliberation layer. In the teleoperation mode, the MAXQ HRL-based deliberation layer is not present, and the user interface is used to directly control the robots individually, Fig 5-2.

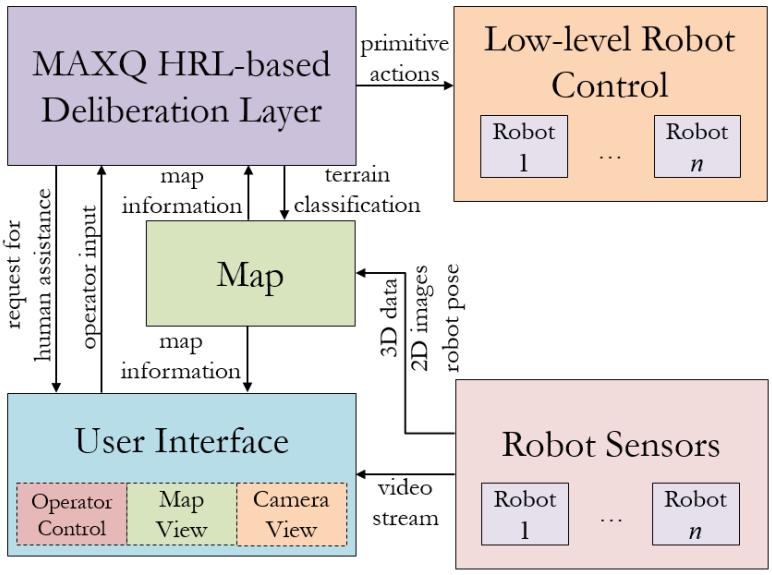


Fig 5-1. System architecture for multi-robot rescue team in semi-autonomous mode.

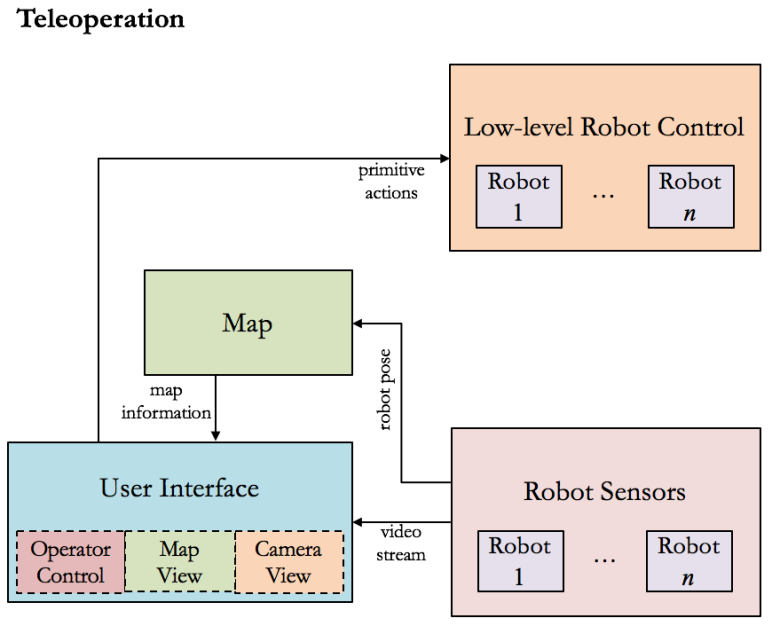


Fig 5-2. System architecture for multi-robot rescue teams in teleoperation mode.

5.1.1 Robot Sensors

Each robot is equipped with four 3D sensors used to provide depth information about its surroundings. This information is used to classify the terrain as open space, climbable, or non-climbable obstacles, as well as to build a map of the cluttered environment. Each robot has an

inertial navigation system (INS) for tracking the robot's position and orientation within the environment, and a 2D camera which provides video streaming to the operator and is also used for victim identification by the robot. In semi-autonomous mode, victim identification is implemented by analyzing 2D images provided by the camera using a skin-detection method [249]. The victim's location is, then, tagged on the map. In teleoperation mode, victim identification is achieved by the operator using the 2D video stream.

5.1.2 Mapping

The mapping module receives 3D information of the USAR environment from the 3D sensors mounted on the robot, and uses this information to classify the terrain. Accessible regions are classified as open or climbable obstacles. Inaccessible regions are classified as non-climbable obstacles. Terrain classification is accomplished by fitting a plane to the 3D data using a least-squares method. The slope of the plane is used to determine whether regions are traversable (i.e., climbable or non-climbable). This module also uses the information from the INS to localize the team of robots within the map.

A 2D occupancy grid map is used to represent terrain information as well as the locations of the victims in the environment. Grid cells are labelled as open, climbable, non-climbable, and victim cells. The accessibility of a cell is determined by the terrain properties of the USAR environment (more details are provided in [48]). This approach allows detailed mapping of 3D cluttered environments by providing information about the traversability of the cells the robots are exploring. The 2D occupancy map information is sent to both the semi-autonomous controller and the user interface. The global map of the USAR scene can be viewed by combining together all the individual sub-scene maps generated by each robot in the USAR environment.

In teleoperation mode, terrain classification is not available. A 2D occupancy grid is provided, only consisting of visited regions and victim locations.

5.1.3 MAXQ HRL-based Deliberation Layer

The objective of introducing the MAXQ HRL technique into the Deliberation Layer is to have the robot team learn from its own experiences and those of an operator in order to effectively perform tasks in USAR environments [26]. By introducing the MAXQ HRL technique, a rescue

robot team can cooperatively learn and determine which tasks should be executed at a given time, and to decide whether a rescue robot or an operator should carry out those tasks to achieve optimal performance. This was proposed in [47]. The theory used for this component of the architecture was completed by Yugang Liu. This thesis focuses on the incorporation and implementation of the theory into the system architecture.

The fundamental principle of MAXQ HRL is to decompose the decision-making problem modeled as an MDP, M_0, into a finite set of smaller and easier to resolve subtasks, $M_1, M_2, \cdots, M_n$, and to derive the optimal policies for these subtasks in order to achieve a hierarchical optimal policy for the overall task, M_0. The purpose of MAXQ learning is to determine this hierarchical optimal policy in order to maximize the expected cumulative reward for M_0, defined as the action-value function, namely the Q function. For every subtask, M_p, a policy, π_p, which maps all possible states of the subtask to a child task is defined. The child task can be either a primitive action or another subtask to execute. Subsequently, the hierarchical optimal policy, π, is the set containing all the policies for all subtasks. More details on MAXQ learning can be found in [247].

The MAXQ task hierarchy for the multi-robot USAR problem at hand is presented in Fig 5-3. Herein, sub-scenes are defined as isolated regions of the USAR environment. The $Root$ task represents the overall USAR problem of scene exploration and victim identification. The $Root$ task is divided into five different subtasks: Search Sub-scene (SSS_i, where i represents the index of the sub-scene), Navigate to Unvisited Regions (NUR), Victim Identification (VI), Navigate (NG), and Human Control (HC). Cooperation is achieved by providing the robots with their own copy of the task hierarchy while sharing the same common $Root$ task.

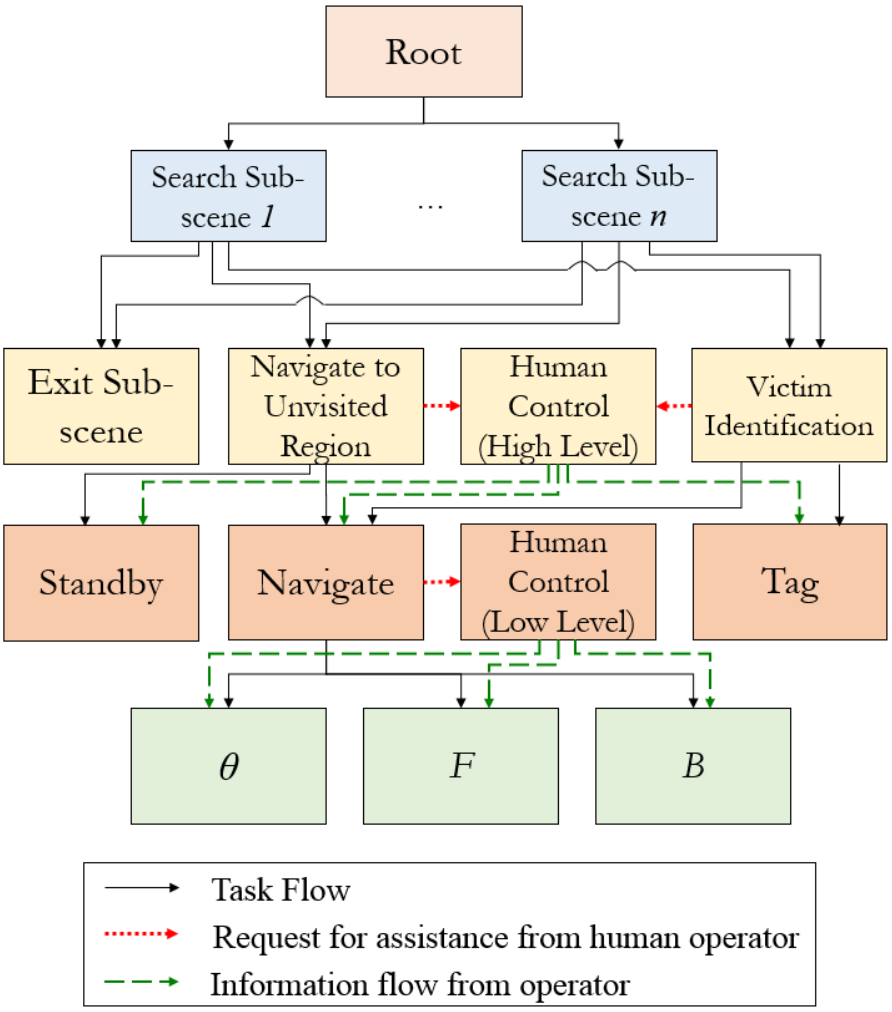


Fig 5-3. MAXQ task hierarchy [48].

The MAXQ state function of the $Root$ task is defined as $S_{Root}(V, S_S, M_G)$. Herein, V represents the presence of potential victims; S_S denotes the sub-scene status (i.e., unexplored, being explored, or explored); and M_G is a collection of 2D occupancy maps of USAR sub-scenes. The purpose of the SSS_i subtask is to allocate rescue robots to sub-scenes. The state function of this subtask is defined by $S_{SSSi}(V_i, L_R, M_{G,i}, A_{O,SSSi})$, where $L_R = \{L_R^1, L_R^2, \cdots, L_R^k\}$ denote the robots' locations within the same sub-scene SS_i with respect to the starting location of the first robot, which is defined as the origin of the local coordinate frame, and number of robots deployed in the same sub-scene is depicted by k. $M_{G,i}$ is the 2D occupancy map of the sub-scene obtained by merging 2D maps generated by each individual robot, j, deployed into the same sub-scene. $A_{O,SSSi} = \{A_{O,SSSi}^1, \cdots, A_{O,SSSi}^{j-1}, A_{O,SSSi}^{j+1}, \cdots, A_{O,SSSi}^k\}$ represents the other robots'

actions/subtasks.

The primitive action $Exit\ Sub\text{-}scene$ (ESS) is proposed to terminate the SSS_i subtask and guide the robot out of an explored sub-scene. The purpose of the NUR subtask is to control the robot to explore unvisited regions within the sub-scene through cooperation with other potential robots. The state function of NUR is defined as $S_{NUR}(L_R, M_{G,i}, A_{O,NUR,i})$, where $A_{O,NUR,i} = \{A^1_{O,NUR,i}, \cdots A^{j-1}_{O,NUR,i}, A^{j+1}_{O,NUR,i}, \cdots, A^k_{O,NUR,i}\}$ represents the other robots' actions/tasks while cooperatively executing the NUR subtask with $Robot_j$. A direction-based exploration strategy based on frontiers is implemented to effectively explore a sub-scene utilizing the 3D cluttered terrain information of the environment in this subtask [48]. The primitive action $Standby$ is used to end exploration of a sub-scene for all robots in the sub-scene.

The state function of the VI subtask used to identify potential victims in each sub-scene is $S_{VI}(L^j_{V/R}, M^j_{G,i})$, where the potential victim's location is marked as $L^j_{V/R}$ in the scene. When a victim is identified, the primitive action Tag is executed to tag the victim's location within $M^j_{G,i}$. The NG subtask is proposed for local navigation and obstacle avoidance, which utilizes 2D grid map information of the robot's surrounding cells [48]. The state function of the NG subtask is $S_{NG}(C^j_l, D^j_E, D^j_{xy}, L_{V/R})$, where $Robot^j_s$ surrounding cells C^j_l, $l = 1\ to\ 8$, can be categorized according to the depth profile information D^j_{xy} of the rubble pile in the robot's surrounding environment; the desired exploration direction (determined by NUR) is depicted by D^j_E. The primitive actions, rotate $Robot_j$ by an angle (θ), and move $Robot_j$ forward (F) or backwards (B), are determined by the status of the robot's surrounding cells and sent to the robot's low-level controller to execute into motion commands.

The HC subtasks are used to request for human assistance and allow the operator to intervene when a robot cannot execute any of the aforementioned tasks autonomously. In order to minimize the workload of the user, MAXQ only requests for human assistance of a subtask when required (i.e., when the robot is stuck or there is uncertainty in victim identification).

MAXQ decomposes the $Root$ task into a finite set of subtasks or primitive actions recursively. In a MAXQ task hierarchy, the possible states of each task are mapped to a child (either a primitive action or another subtask) through a policy π. In the proposed MAXQ task hierarchy, the Q value (action-value function) for the $Root$ task is defined as follows:

$$Q(Root, s, SSS_i) = V(SSS_i, s) + C(Root, s, SSS_i), \quad (5\text{-}1)$$

where $V(SSS_i, s)$, the projected value function of executing the SSS_i subtask in state s, and $C(Root, s, SSS_i)$, the completion function representing the discounted cumulative reward of executing the SSS_i subtask, can be defined as:

$$V(SSS_i, s) = Q\left(SSS_i, s, \pi_{SSS_i}(s)\right), \text{ and} \quad (5\text{-}2)$$

$$C(Root, s, SSS_i) = \sum_{s' \in S_{Root}, N}\{P_{Root}(s', N|s, SSS_i)\gamma^N Q(Root, s', \pi_{Root}(s'))\}, \quad (5\text{-}3)$$

where $\pi_{SSS_i} \in \{ESS, NUR, VI\}$ and $\pi_{Root} \in \{SSS_1, \cdots, SSS_n\}$ represent the policies for the SSS_i subtask and $Root$ task, respectively. S_{Root} is the state function of the $Root$ task, γ is the discount factor and N denotes the number of transition steps from state s to the next state s'. P_{Root} is the probability transition function for the $Root$ task.

The action-value function of the $Root$ task is recursively decomposed into the summation of action-value functions of its subtasks. For example, the action-value function for SSS_i can be further decomposed as follows:

$$\begin{gathered} Q(SSS_i, s, ESS) = V(ESS, s) + C(SSS_i, s, ESS), \\ Q(SSS_i, s, NUR) = V(NUR, s) + C(SSS_i, s, NUR), \\ Q(SSS_i, s, VI) = V(VI, s) + C(SSS_i, s, VI), \end{gathered} \quad (5\text{-}4)$$

where $V(ESS, s)$, $V(NUR, s)$, and $V(VI, s)$ are the projected value functions and $C(SSS_i, s, ESS)$, $C(SSS_i, s, NUR)$ and $C(SSS_i, s, VI)$ are the completion functions. It should be noted that *ESS* is a primitive action and its projected value function is defined by:

$$V(ESS, s) = \sum_{s'} P(s'|s, ESS)\ R\ (s'|s, ESS), \quad (5\text{-}5)$$

where P and R represent the probability transition function and the expected reward function, respectively. The action-value functions for the remaining subtasks can be defined in a similar manner.

When multiple robots are deployed to search the exact same sub-scene, each robot first has its

own action-value functions and receives rewards for its own contribution to the relevant sub-tasks. This information is, then, utilized with similar information from the other robots in the sub-scene in order to determine the overall action-value function for the corresponding subtask. Cooperative learning occurs by each robot considering the actions of the other robots while updating its own projected value and completion functions. For example, when a sub-team of rescue robots $\{R_i^1, \cdots, R_i^k\}$ cooperatively search sub-scene i, the projected value function and completion function for robot R_i^j can be defined as:

$$V^j(SSS_i, s_{SSS_i}, A_{o,SSS_i}) = Q^j(SSS_i, s_{SSS_i}, A_{o,SSS_i}, \pi_{SSS_i}), \quad (5\text{-}6)$$

$$C^j(Root, s, SSS_i) = \sum_{s' \in S_{Root}, N}\{P_{Root}(s', N|s, SSS_i)\gamma^N Q^j(Root, s', \pi_{Root}(s'))\}, \quad (5\text{-}7)$$

where $A_{o,SSS_i}^l \in \{NUR, VI, ESS\}\ and\ (l = 1, \cdots, k; l \neq j)$.

The implementation of the semi-autonomous controller allows the operator to provide more than one primitive action when the controller requests for help. Actions by the operator are recorded into the existing 2D occupancy map. When the operator hands back control over to the semi-autonomous controller, the latter knows and learns from the information gathered by the operator. This provides a more robust approach in searching USAR scenes with human assistance.

A reward system is implemented to encourage the robots to learn positive actions that lead to transitions from their current states to desired states [47]. Negative rewards are given to actions that result in transitions to undesirable states. Appendix E provides further details on the reward system. For this work, the MAXQ semi-autonomous controller was trained with over 25,000 training episodes. During USAR experiments, the trained model is updated online to adapt to new unknown USAR environments.

5.1.4 User Interface

The user interface of the system was completed with Onome Igharoro. It was developed for handling communication between the operator and the multi-robot team in both semi-autonomous and teleoperation modes, Fig 5-4. The interface comprises three main modules: (1) multi-robot operator control (bottom of interface), (2) map view (top right corner of interface), and (3) 2D camera view (top left corner of interface).

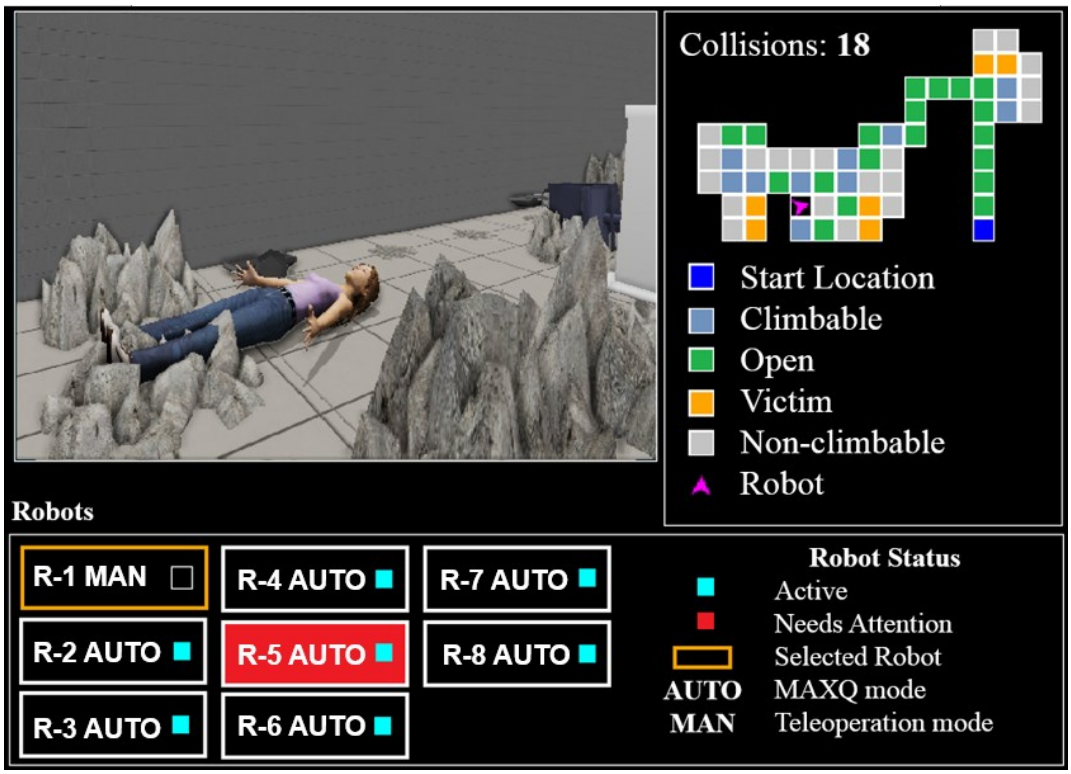


Fig 5-4. User interface for human operator.

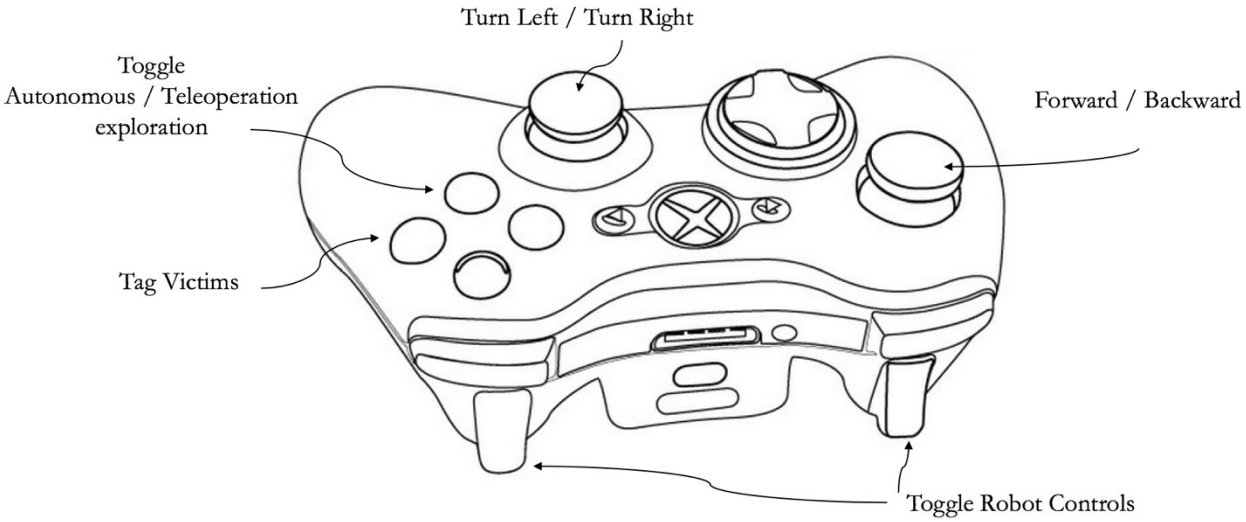


Fig 5-5. Control inputs for multi-robot team in USAR (image modified from [248]).

The operator control input module handles various user inputs from an XBOX gamepad that is used to teleoperate the robots, Fig 5-5. These control inputs include moving a robot forward/backward, turning the robot, tagging a victim, and switching between different robots in

a team. In addition, the operator can also return control back to a robot in the semi-autonomous mode.

The map view module receives map information from the map (generation) module and displays the map of already visited regions and victim locations in the sub-scenes. The camera view output module displays a live video stream for each individual robot. For the semi-autonomous mode, the user interface also alerts the operator when a robot needs human assistance.

The robot team view also provides the status of the robots in the team. The operator is limited to controlling one robot at a time. The team view also indicates whether a particular robot is being controlled manually (MAN) via teleoperation or by the semi-autonomous (AUTO) controller. The user interface improves the operator's spatial awareness of the USAR scene by providing the ability to generate an occupancy grid map, in addition to the camera view of the robot.

In semi-autonomous mode, the user interface is interconnected with the MAXQ HRL-based deliberation layer. The team of robots moves autonomously until a robot requires human assistance. As previously mentioned this follows an adaptive interface design. During operator input, the robot control status switches to MAN mode to indicate that the robot is under the operator's control. In teleoperation mode, the MAXQ HRL-based deliberation layer is not present. The user interface directly sends operator inputs to control the team of robots. The operator has full control over each individual robot in the team.

5.1.5 Low-Level Robot Control

In the semi-autonomous mode, the primitive actions (i.e., move forward, move backward, turn) from the MAXQ HRL-based deliberation layer are converted into motion commands for the team of robots. In teleoperation mode, the operator has direct control of the motion of the robots and the low-level controller is used to process operator commands into low-level motion commands.

5.1.6 USARSim

USARSim was used as the 3D simulation environment [250]. The USARSim platform was used together with the Unreal Developer's Kit (UDK) game engine [251]. This provided the capability

to create realistic unstructured USAR environments consisting of both rubble and victims. Herein, rubble is defined by concrete piles and overturned furniture. The aforementioned system architecture contains a library of extensive functions for communicating with USARSim.

5.1.7 Software Implementation

The software components included the UDK game engine running USARSim, Multi-Robot Operator Team (MROT), and the MAXQ HRL-based deliberation program, Fig 5-6. MROT, the custom multi-robot remote control application for USARSim, was integrated with the MAXQ program, the semi-autonomous controller, to build an effective command console for operators monitoring multiple semi-autonomous robots in real-time within USARSim. With this implementation, operators can, at a glance, monitor the status of all robots in the team. Operator intervention is made seamless by detecting when an operator intends to take control of a robot, or when the MAXQ program requests assistance. Following an operator intervention, control is transferred back to the MAXQ program with a single input.

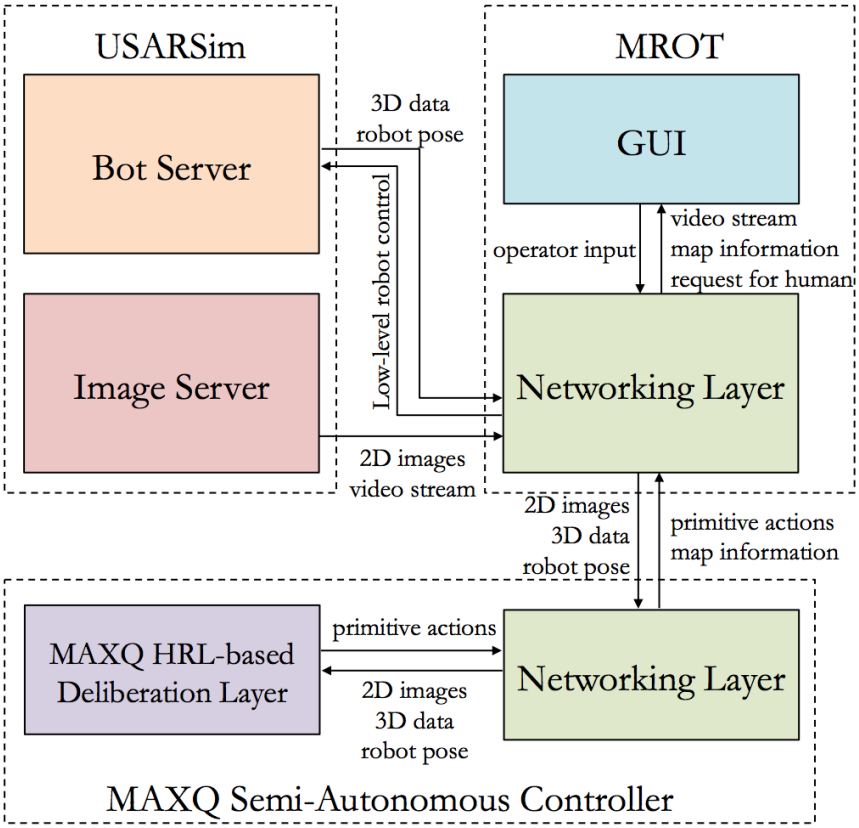


Fig 5-6. Software schematic diagram.

5.2 Experiments

Experiments were conducted with operators controlling multi-robot rescue teams in USARSim to investigate the influence of the operator-to-robot ratio on the performance of the HRL-based semi-autonomous controller as well as on the full teleoperation control of the robots. The performance metrics used were: (1) percentage of scene coverage, (2) percentage of victims found, (3) number of robot team collisions in the environment, and (4) total exploration time.

For experiments, human performance metrics of Interaction Effort (IE) using operator control time [252], [253], and Task Performance (TP) [254] was measured. IE is an important measure since it shows the demands on an operator based on the resources at hand. Namely, this measurement enables identification of bottlenecks in the system (i.e., robot team size), in which performance can be negatively impacted [255]. TP provides a metric for the overall performance of a robot team, and reflects the operator's awareness of the system and environment [253]. It is important to determine how automation can affect the operator's SA in a USAR environment [256].

5.2.1 Procedure

Twenty-one people participated in the trials, ranging from 23 to 36 years in age ($\mu = 25.9, \sigma = 3.6$). All participants were engineering students. None had prior experience with controlling a USAR robot, however, they had varying expertise in playing 3D video games, ranging from none to with little experience (43%), as well as moderate to more experienced (57%).

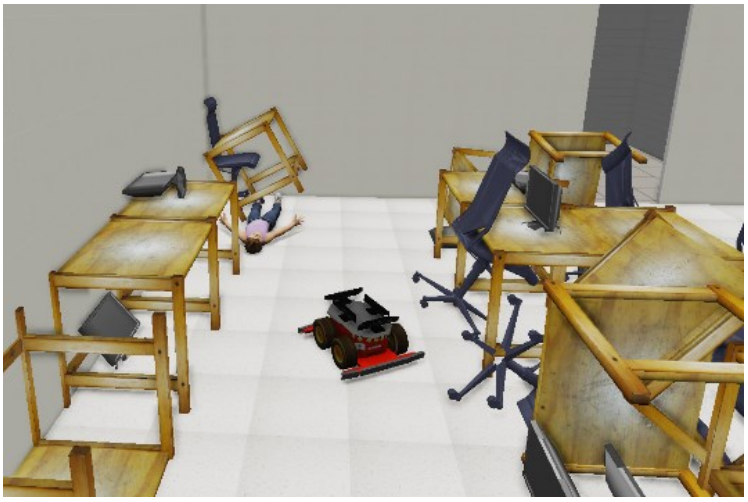
Each trial consisted of having the participant control a team of 5, 10, 15, and 20 robots in both semi-autonomous mode and teleoperation mode, respectively. The USAR scenes used for the experiments occupied 288 m^2, 544 m^2, 944 m^2, and 1184 m^2 for the 5, 10, 15, and 20 robot teams, respectively. Each USAR scene was divided into smaller sub-scenes. The overall size of each sub-scene varied, ranging from 32 to 80 m^2, the amount of clutter ranged from 60% to 75% of the overall sub-scene, and the number of victims ranged from 1 to 4.

Fig 5-7 provides examples of rubble pile and victim configurations within the sub-scenes. The Pioneer P3AT mobile robotic platforms were used, which contained four 3D sensors located to scan the front, left, right, and back of each robot for terrain classification, an INS, and a 2D camera with 320×240 pixel resolution.

A counter-balance approach was used; half of the participants started each trial in teleoperation mode, and the remaining in semi-autonomous mode. After finishing the trials in each mode, participants switched to the next mode. The robot team size configurations in each mode were randomized for each participant (e.g., 5-20-15-10, 15-20-5-10). Each participant had ten minutes of training with respect to the gamepad control inputs and the user interface prior to the experiments. The objective for the operator was to explore the USAR environment to cover as much area as possible and to identify as many victims within the overall environment with no time limits. After the experiments were completed, the participants were asked to complete a 5-point Likert questionnaire (5 – Strongly Agree, 1 – Strongly Disagree) based on their experiences, Appendix F.

5.2.2 Results and Discussion

A statistical power analysis was first conducted to confirm that the sample size was sufficient with respect to the performance metrics, achieving $p < 0.05$, with powers greater than 0.99 for all performance metrics (one-tailed). During semi-autonomous operations, all robots in the team worked in parallel and asked for assistance from the operator only when required. Thus, for example, it is expected for some increase in exploration time with increased environment size and robot team size. In teleoperation mode, however, the hypothesis is that a more significant increase in exploration time would occur with increased robot team size since each individual robot requires the operator's attention at all times. TABLE 5-1 shows the average values and ranges for the collected performance metrics for the experiments.



(*a*)

(*b*)

(*c*)

Fig 5-7. Example sub-scenes: (*a*) with overturned furniture, (*b*) with climbable and non-climbable concrete piles, and (*c*) a combination of both furniture and concrete piles.

TABLE 5-1. VARYING ROBOT-TEAM SIZE PERFORMANCE METRICS

		Average Metric Value (Range for all Participants)				
Exploration Mode	# of Robots on a Team	# of Victims	% of Scenes Explored	# of Collisions	% of Victims found	Total Exploration Time (sec)
Semi-Autonomous	5	12	100 (100-100)	0.3 (0-7)	100 (100-100)	101 (76-162)
	10	20	100 (100-100)	0.5 (0-4)	100 (100-100)	129 (90-192)
	15	30	100 (100-100)	6.2 (1-15)	100 (100-100)	184 (109-344)
	20	37	100 (100-100)	9.5 (0-27)	100 (100-100)	195 (126-284)
Tele-operation	5	12	93 (81-100)	10.8 (0-48)	98 (83-100)	319 (168-811)
	10	20	88 (76-100)	17.8 (0-57)	98 (90-100)	516 (277-1082)
	15	30	89 (77-100)	28.4 (4-77)	98 (93-100)	916 (552-1688)
	20	37	88 (81-97)	55.1 (16-174)	97 (92-100)	1405 (737-3386)

As expected, in both the semi-autonomous and the teleoperation modes, the total exploration time increases as the number of robots increases for larger USAR scenes. However, the rate of increase is significantly higher for teleoperation as it was hypothesized. This mode also results in a greater number of robot collisions with the environment. When a second-order-polynomial least-squares fit was utilized for both cases, the plots shown in Fig 5-8 were obtained.

For easier comparison, when linear least-squares was utilized, the slopes were determined to be about 6.7 s/robot versus 73.2 s/robot, with a confidence level of more than 94.64%. Namely, the results confirm the difficulty an operator would face when trying to control a large team of rescue robots, exploring the scenes sequentially. Furthermore, when using the semi-autonomous controller, robot teams were able to explore 100% of the scenes and identify all the victims while minimizing the number of collisions they had in the environment.

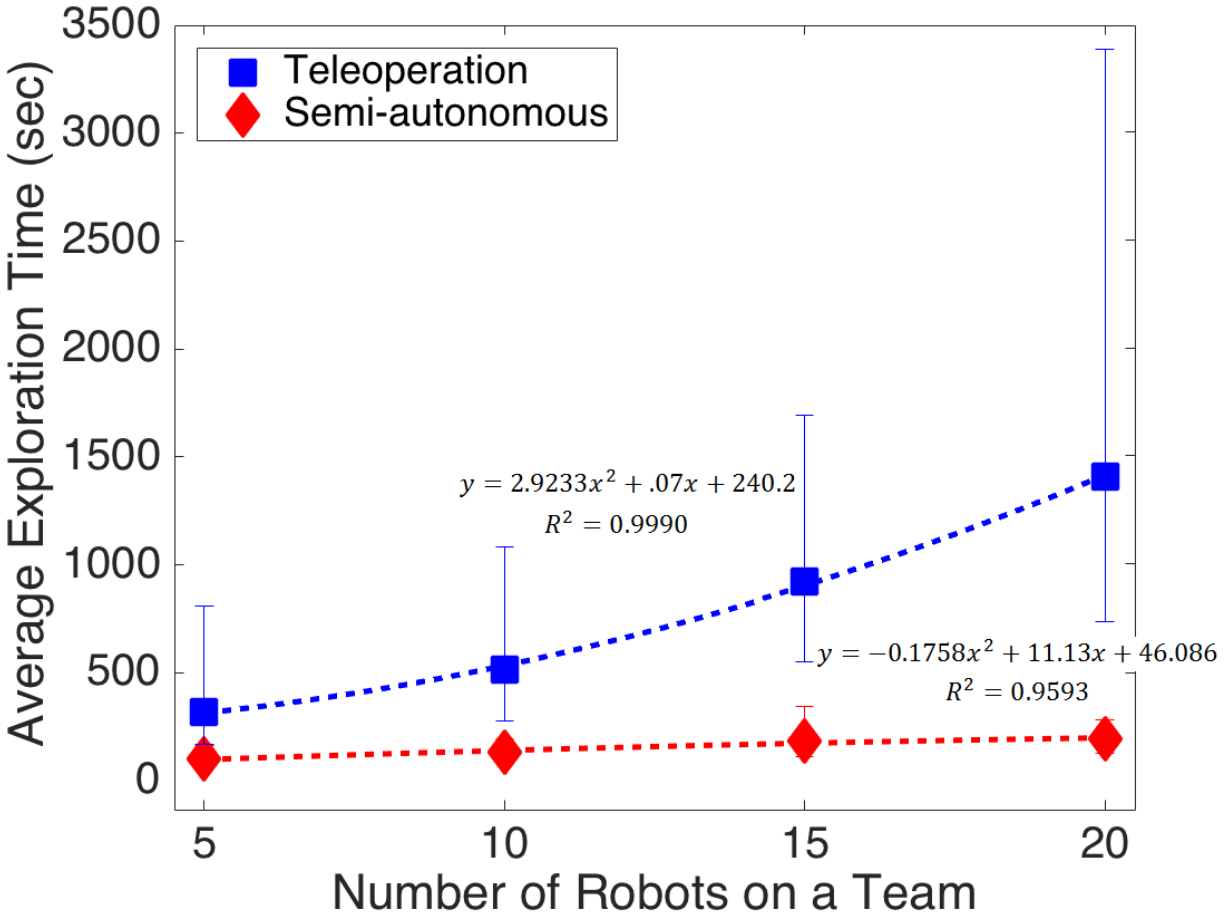


Fig 5-8. Total exploration time for all participants controlling 5, 10, 15, and 20 robots in both control modes.

Operator Interaction Effort, IE_h, for each robot team size for participant h was defined as:

$$IE_h = \frac{O_{t_h}}{E_{t_h}}, \tag{5-8}$$

where O_{t_h} represents the total time participant h was controlling the robots, and E_{t_h} represents the total exploration time for that participant.

Task Performance, TP_h, for participant h was defined as:

$$TP_h = w_1 S_h + w_2 C_h + w_3 V_h, \tag{5-9}$$

where S_h is the percentage of scene explored, C_h is the number of collisions, V_h is the percentage

of victims found and w_1, w_2, and w_3 are performance weights ($\sum w_i = 1$):

$$S_h = \frac{S_{t_h}}{\max S_{t_h}}, \tag{5-10}$$

$$C_h = \frac{\max C_{t_h} / C_{t_h}}{\max\left(\max C_{t_h} / C_{t_h}\right)}, \tag{5-11}$$

$$V_h = \frac{V_{t_h}}{\max V_{t_h}}, \tag{5-12}$$

where S_{t_h}, C_{t_h},and V_{t_h} represent the percentage of scenes explored, number of collisions, and percentage of victims found over E_{t_h}, respectively. The average operator IE and TE per robot team size is presented in Fig 5-9 and Fig 5-10, respectively.

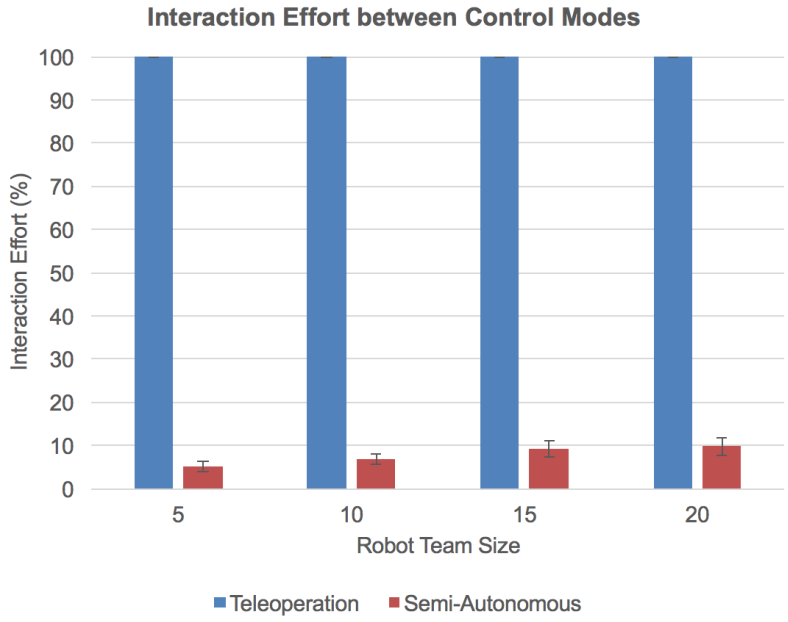


Fig 5-9. Interaction effort between control modes.

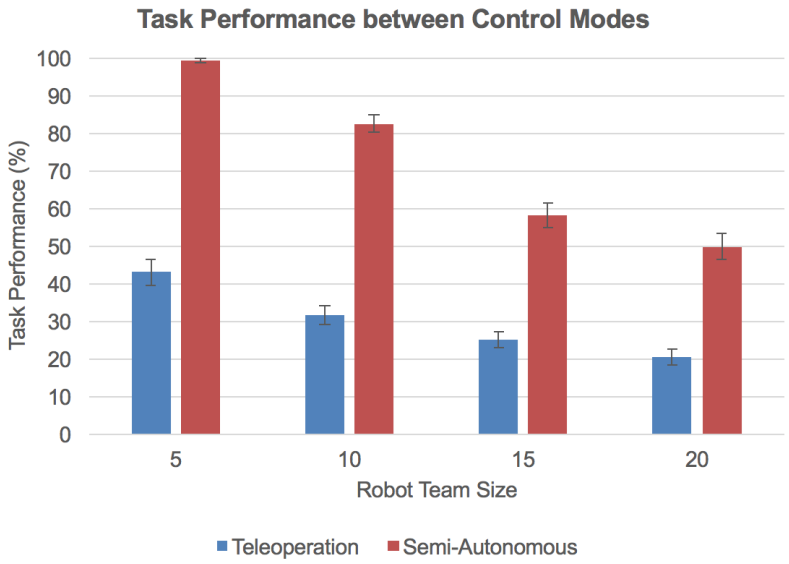


Fig 5-10. Task performance between control modes.

An analysis of variance (ANOVA) test was performed to determine statistical significance for all performance metrics. The results showed that semi-autonomous mode was significantly better compared to teleoperation mode in all performance metrics regardless of the robot team size: (1) percentage of scene exploration, $F(1,160) = 279.0,\ p < 0.001$; (2) percentage of victims found, $F(1,160) = 40.42,\ p < 0.001$; (3) total exploration time $F(1,160) = 201.7,\ p < 0.001$; and (4) total number of collisions, $F(1,160) = 77.79,\ p < 0.001$.

With respect to human performance metrics, statistical significance was also determined between the control modes regardless of the robot team size for: (1) IE, $F(1,160) = 12740, p < 0.001$; and, (2) TE, $F(1,160) = 506.2,\ p < 0.001$. Namely, both interaction effort and task effectiveness were significantly better for the semi-autonomous mode compared to the teleoperation mode. In addition, there was no significant correlation between 3D video games experience of the participants and TE, ($\rho = -0.1, p = 0.527$).

5.2.3 Exploration With and Without Learning

A comparison was also done between the learning-based MAXQ technique with a non-learning technique for direction-based exploration [44] of a 96 m^2 highly cluttered USAR scene by a robot through USARSim simulation. The percentage of scene coverage for this scene was 100% using MAXQ and 24% using the non-learning technique. In the case of the non-learning technique, the robot got into situations where it became trapped in corners or rubble piles,

whereas when using MAXQ, the robot was able to avoid such situations through learning.

5.3 Summary

This Chapter investigated the influence of the operator-to-robot ratio on the performance of the unique system architecture for using a semi-autonomous controller to aid an operator in a multi-robot team USAR. Experiments showed that operator performance improved significantly when aided by the semi-autonomous controller. With the semi-autonomous controller, operators can cover more area in a shorter time, and exhibit greater patience in exploration. Further investigation on human-performance metrics such as workload and situation awareness is recommended to examine the influences of the architecture on human factors. The semi-autonomous controller architecture proved to be more effective when the operator is controlling a large robot team compared to (pure) teleoperation. In addition, operator workload decreased significantly when compared to teleoperation.

Chapter 6
Conclusions & Recommendation for Future Work

This Chapter presents a summary of the research challenges addressed in this thesis, as well as the architectures developed to address them. The work encompasses a multimodal system for detecting affect from human body language and vocal intonation during HRI, and a learning-based semi-autonomous control architecture for multi-robot USAR missions. In addition, possible directions for future research are also discussed in this Chapter. In Section 6.1, the summary of contributions of this thesis is presented. Sections 6.1.1 and 6.1.2 summarize the contribution of the proposed multimodal affect recognition system, and multi-robot single-operator USAR architecture, respectively. Section 6.2 provides possible directions for future work.

6.1. Summary of Contributions

The two main contributions of this thesis are:

1. The development of a real-time multimodal affect recognition system that combines the unique inputs of human body language and vocal intonation to infer a person's affect during socially assistive HRI, and
2. The development of a multi-robot single-operator learning-based semi-autonomous architecture for higher performance in scene exploration and victim identification task during USAR.

The multimodal affect recognition system utilizes both dynamic body language and vocal intonation to determine user affect, and, in turn, can be used to determine a robot's appropriate emotion and response. This leads to better bi-directional HRI between humans and robots, as robots are able to recognize, interpret, and respond effectively to social cues. Such robots would promote more effective and engaging interactions with the user. These social robots developed can be used as assistive robots for elderly to provide social interaction and cognitive assistance with activities of daily living.

The multi-robot single-operator USAR architecture uses hierarchical reinforcement learning to learn the USAR environment in order to increase performance in scene exploration and victim

identification. The architecture provides the ability to effectively allocate sub-tasks to robots in order to complete the overall USAR mission. This architecture allows the operator to handle a greater number of robots compared to current methods without significant performance loss due to the controller's ability to only request human assistance when the robot is stuck or when there is uncertainty in human identification, and reduces the amount of time the operator spends with each robot on the team.

6.1.1 Multimodal Affect Recognition System

To date, only a handful of multimodal affect recognition systems have been used for HRI, and these systems have primarily focussed on using facial and vocal expressions as inputs. Body language, however, plays an important role in conveying changes in human emotions during social interactions and has yet to be implemented in a multimodal system. Vocal intonation also plays an important role in conveying changes in emotion through vocal properties of pitch, tempo, and loudness during HRI. The multimodal affect recognition system proposed classifies a person's affect based on body language and vocal intonation in a 2D valence-arousal scale (circumplex dimensional affect model) in real-time.

Body language features are acquired using a Kinect™ 3D sensor to extract position coordinate points of body parts. These points are then used to calculate dynamic body language features. Vocal intonation features are acquired through an environmental microphone. These features are based on the peaks and plateaus of the vocal signals. The features from their respective modalities are then classified into affect. Finally, a Bayesian network is used to combine the affect from body language and vocal intonation via decision-level fusion.

The multimodal affect can then be used as input to a robot's emotional model to determine its own emotion, and appropriate response during HRI. The system has been implemented on the Nao humanoid robot for the application of fitness and nutrition counselling. Experimental results showed that the robot can effectively recognize user affect during real-time HRI, as well as adapt its emotions accordingly.

Affective voice recognition of older adults has also been investigated. This is a challenging problem as aging in humans directly affects the quality of voice. Current research of affect recognition has not yet targeted the elderly population. The affect detection developed, herein, is

based on a categorical model where affective states of happy, sadness, anger, and neutral are detected. The affect voice recognition system for elderly achieve an affect classification rate of approximately 68%.

6.1.2 Multi-robot Single Operator USAR Architecture

Multi-robot teams can provide valuable assistance in USAR missions as they can increase efficiency and system robustness. Past controllers have been shown to control operator-to-robot teams with 1:12 ratio and have been limited in their capabilities to avoid obstacles and identify victims automatically. The proposed multi-robot single operator architecture uses a learning-based semi-autonomous controller and provides two primary advantages, when compared to methods currently in use: (*i*) the operator can handle greater number of robots in the multi-robot team without significant performance loss, and (*ii*) the interaction effort of operators is reduced significantly. The proposed approach uses an adaptive user interface, where the system assigns robot tasks, and the operator assists the robot with completing tasks when the system request for human assistance. The unique aspect of the approach is the ability of the system to learn how to allocate and execute tasks, as well as learn from experience of an operator in order to further minimize operator workload. Compared to traditional AI-based approaches, the hierarchical reinforcement learning architecture can scale up to large numbers of subtasks. Since USAR missions are comprised of many subtasks, the state space is large, and traditional modelling techniques fail to scale.

In order to effectively implement such semi-autonomous controller for cooperative multi-robot teams, the impact of increased number of robots on system performance has been investigated. Experiments were conducted with operators controlling multi-robot rescue teams in a simulator using the HRL-based semi-autonomous controller and full teleoperation control of robots. Performance metrics of percentage of scene coverage, percentage of victims found, number of robot team collisions in the environment, and total exploration time were investigated. In addition, human performance metrics of interaction effort and task performance were analyzed. Results showed that operator performance improved significantly when aided by the semi-autonomous controller. With the semi-autonomous controller, operators are able to cover more area in a shorter amount of time, and require less interaction effort for controlling a multi-

robot team.

6.2. Recommendations for Future Work

Future research direction should include investigation and implementation of additional affective communication modes to the multimodal affect recognition system. These communication modes can be from facial expression and physiological signals. Facial expression is a primary mode of communication of affective information due to the inherently natural face-to-face communication in human-to-human interactions. Facial expressions can convey emotions and be recognized across all cultures. Facial features can be extracted from the eyes, eyebrows, lips, and nose region according the FACS using a 2D camera. The challenge with facial affect recognition, however, is that it only performs well when the human frontal face is positioned directly in front of the camera, which is not always the case in HRI. The 3D sensor currently in the system can also be utilized to gather facial feature information. Moreover, physiological signals can be a strong indication of a person's affect as human affect influences the body in many ways (e.g., changing a person's heart rate, skin conductance, breathing rate). Physiological signal features can be extracted from heart rate, facial muscle activity, and skin conductance using EDR and ECG sensors.

Future work should also be invested in improving the affective vocal intonation recognition system. Compared to the affective body language recognition rate of 93.6% and 95.8% for valence and arousal respectively, the current voice system has a recognition rate of 55.6% and 61.9% for valence and arousal respectively. Improving the voice system can lead to a higher multimodal detection rate. A suggestion in improving the voice system is adding additional features detected from Scherer's affective voice classification [25] to the Nemesysco features defined.

Future research should also focus on improving the long-term acceptance of the multimodal robot in homes. Experiments should be further conducted with users to see if they maintain their engagement, compliance, and enjoyment over a large number of interactions. The multimodal system should also be evaluated in elderly care facilities as well. The study should include if the system improves social activity, and cognitive function in older adults. Future work could also include the ability for the robot to interact with more than one user during HRI. This enables

more natural communication as humans can interact with more than one user at a time.

For the multi-robot single-operator USAR architecture, future work should include experiments using physical robots in a USAR-like environment. This will further justify the system for usage in USAR missions, as experiments are physical and not just simulated. Future work should also include the usage and implementation of heterogeneous robot teams, as USAR missions typically have more than one type of robot deployed. Example robot teams can include UAVs in addition to ground vehicles. This will lead to a more flexible system as the system can determine the optimal actions for the team using the different types of robot to complete the USAR mission. Lastly, further investigation of the influence of operator-to-robot ratio on situational awareness and operator workload should also be conducted. Situational awareness and operator workload are important to measure when using automation. This can evaluate human trust and reliance on the system.

Appendices

A. Dynamic Body Language Features

The equations for calculating dynamic body language are presented herein and have been obtained from [222].

TABLE A-1. DYNAMIC BODY LANGUAGE FEATURES [222]

Features	Description	Equation
Bowing / Stretching of the Trunk	Average trunk lean angle towards or away form the robot during the body language display	$\frac{1}{N}\sum_{f=1}^{N}\arctan\left(\frac{p_{y_{shoulder,f}}-p_{y_{hip,f}}}{p_{z_{shoulder,f}}-p_{z_{hip,f}}}\right)$, where N is the total number of 3D data frames in a body language display, and $p_{shoulder,f}=\frac{1}{2}\left(p_{leftshoulder,f}+p_{rightshoulder,f}\right)$, and $p_{hip,f}=\frac{1}{2}\left(p_{lefthip,f}+p_{righthip,f}\right)$
Opening / Closing of the Arms	Average distance between the hands and the center of the trunk during the body language display	$\frac{1}{N}\sum_{f=1}^{N}\left(\frac{1}{2}\left\|p_{lefthand,f}-p_{trunkcenter,f}\right\|+\frac{1}{2}\left\|p_{righthand,f}-p_{trunkcenter,f}\right\|\right)$, where $p_{trunkcenter,f}$ is the centroid with respect to $p_{leftshoulder,f}$, $p_{rightshoulder,f}$, $p_{lefthip,f}$, and $p_{righthip,f}$ points at frame f.
Vertical Head Position	Average relative height of the head with respect to the neck during the body language display	$\frac{1}{N}\sum_{f=1}^{N}\left(p_{y_{head,f}}-p_{y_{neck,f}}\right)$
Forward / Backwards Head Position	Average distance between the head and the neck towards or away from the robot during the body language display	$\frac{1}{N}\sum_{f=1}^{N}\left(p_{z_{head,f}}-p_{z_{neck,f}}\right)$
Vertical Motion of the Body	Average upwards/downwards movement of the body during the body language display	$\frac{1}{N-1}\sum_{f=1}^{N-1}\left(\frac{1}{S}\sum_{i=1}^{S}p_{y_{i,f+1}}-p_{y_{i,f}}\right)$, where S is the total number of points on the skeleton model.
Forward / Backwards Motion of the Body	Average towards or away movement of the body with respect to the robot during the body language display	$\frac{1}{N-1}\sum_{f=1}^{N-1}\left(\frac{1}{S}\sum_{i=1}^{S}p_{z_{i,f+1}}-p_{z_{i,f}}\right)$

Features	Description	Equation
Expansiveness of the Body	Average spatial extension of the body during the body language display	$\frac{1}{N-1}\sum_{f=1}^{N-1}\left(\left(\max_i p_{x_{i,f}} - \min_i p_{x_{i,f}}\right)\left(\max_i p_{y_{i,f}} - \min_i p_{y_{i,f}}\right)\left(\max_i p_{z_{i,f}} - \min_i p_{z_{i,f}}\right)\right)$
Speed of the Body	Average velocity of the movement of the body during the body language display	$\frac{1}{N-1}\sum_{f=1}^{N-1}\left(\frac{1}{S}\sum_{i=1}^{S}\left(\frac{\|p_{i,f+1}-p_{i,f}\|}{T_{f+1}-T_f}\right)\right)$, where T_f is the time at frame f.

B. Vocal Intonation Features

The description for each vocal intonation feature are presented herein.

TABLE B-1. VOCAL INTONATION FEATURES [225]

Features	Description
Anger	Indicates level of anger.
Excitement	Indicates positive or negative excitement.
Upset	Indicates level of unpleasantness or sadness.
Energy	Indicates conversation energy. Low values (< 5) indicates sad or tired, mid-values (5-9) indicates comfortable, and high values (> 9) indicates high energy. Very low values (0-1) may also indicate boredom.
Hesitation	Indicates level of comfort. Below 14 indicates comfort, above 17 indicates regretting.
Embarrassment	Indicates how uncomfortable the user is.
Stress	Indicates nervousness.
Extreme State	Indicates how extreme the overall emotional activity is.
Arousal Factor	Indicates deep and profound interest in conversation topic.
Imagination Activity	Indicates the user is either recalling information from memory or visualizing something.
Intensive Thinking	Indicates user is thinking intensively while speaking.
Concentration Level	Indicates level of concentration.
Uncertainty	Indicates level of certainty. Below 15, user is more certain; above 15, user is more uncertain.
Brain Power	Indicates emotional and cognitive processes in the brain.
Max Amplitude	Indicates max amplitude in sound signal.
Volume	Indicates volume level in sound signal.
Voice Energy	Measures frequency in sound signal.
Emotion-cognition ratio	Indicates rationality of user. Above 100, the user is more emotional; below 100, the user is more logical.

C. Affective Voice Recognition of Older Adults

It has been shown that classifying affective states through voice is challenging, particularly for person-independent recognition and, furthermore, that recognition rates for older adults are lower compared to younger age groups [257]. The aging process directly affects the quality of the voice, as well as its production as a result of various physiological and anatomical changes on the vocal system [258]. For example, a valence detector was investigated in [259] using elderly voices. However, overall, with respect to automated recognition and classification of affect encompassing states of both arousal and valence during HRI scenarios, current research has not targeted the elderly population [5].

Herein, the recognition and classification of the following combination of positive, neutral and negative affective states: happy, sadness, anger, and neutral is investigated. Happiness is important to detect as for older adults it can indicate well-being, health, and longevity [6]. Sadness and anger are important to detect as they can be signs of depression as a result of aging, for example, they are often observed in people suffering from dementia [7]. Neutral, which represents an experience of little or no noticeable feelings, is also useful to detect as a baseline for comparing other affective states.

The proposed automated vocal affect detection and classification architecture consists of three main modules: voice recognition, affect feature extraction, and affect classification, Fig. C-1.

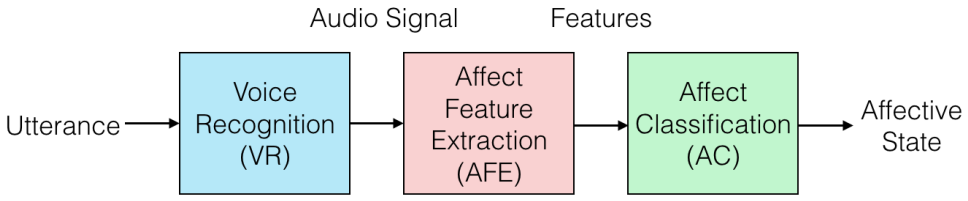


Fig. C-1. Vocal affect recognition and classification architecture.

Voice Recognition (VR) Module

The VR module is responsible for capturing the audio signal of the elderly speaker and processing it into a file to be used by the affect feature extraction module in order to extract voice features from the signal (in this case, a 16-bit 11025 Hz .wav file). This process was automated for real-time analysis by the robot. Each audio clip is 2 to 3 s in duration.

Affect Feature Extraction (AFE) Module

The AFE module determines the vocal features used to classify the affective states of the elderly. In this work, the QA5 SDK Version 5.5 software by Nemesysco is utilized to identify these features. The .wav files are analyzed based on signal features such as thorns (which are local extrema in amplitude found in the second voice sample in three consecutive voice samples in a clip) and plateaus (local flatness in the voice in the clip) [260]. The output use from the software is 18 emotion features, which are identified in the audio clip. These include content, angry, excitement, upset, energy, hesitation, embarrassment, stress, extreme state, emotion-cognition ratio, arousal factor, imagination activity, intensive thinking, concentration level, uncertainty, brain power, max amplitude volume, and voice energy.

Affect Classification (AC) Module

The 18 features determined are used to classify affective states. Namely, within the AC module, the relationship between the affective states and the features can be identified using a machine learning technique. The following learning-based classifiers were investigated in this work: Naïve Bayes probabilistic classifier, logistic regression linear classifier, random forest decision tree, *k*-nearest neighbors lazy learning-based classifier, multi-perceptron neural network, and non-linear support vector machines (SVM). These techniques were considered based on their robustness to handle a wide variety of features needed to determine the affective states.

Experiments

In order to validate the proposed architecture, 123 audio clips from 57 older adult speakers were obtained. The participants were both males and females (≥ 58 years old) engaged in conversation with different intonations. The audio clips were obtained from numerous sources, including YouTube videos, talk-show interviews, news broadcasts, and the SEMAINE database [261]. Two coders were used to code the baseline affect classifications for each clip. The clips for which consensus was obtained between the two coders were used as the input dataset into the proposed automated vocal affect detection and classification system.

A 10-fold cross-validation approach was used to both train and test each classifier using the aforementioned 123 audio clips. The results are presented in TABLE C-1. The random forest

decision tree and logistic regression linear classifier provided the highest classification rate of 68.3%.

TABLE C-1. 10-FOLD CROSS-VALIDATION RESULTS.

Classifier	Classification Rate
Naïve Bayes	63.4%
Logistic Regression	68.3%
Random Forest	68.3%
K-nearest neighbors	63.4%
Multi-layer perceptron	62.6%
SVM	63.4%

The confusion matrix for the affective states for both classifiers are presented in TABLE C-2. The highest classification rate for both classifiers was for anger (78%). The lowest classification rate was for sadness (56% for random forest and 64% for logistic regression, respectively). Sadness was challenging to recognize for all the classifiers as Nemesysco does not provide a distinctive feature to illustrate a sadness affective state.

TABLE C-2. CONFUSION MATRIX FOR RANDOM FOREST (RF) AND LOGISTIC REGRESSION (LR) CLASSIFIERS.

Coded Affective States	Classified Affective States (RF-LR)			
	Neutral	Happy	Sadness	Anger
Neutral	27-25	9-8	0-2	0-1
Happy	4-3	30-29	7-10	1-0
Sadness	8-7	8-6	20-23	0-0
Anger	2-1	0-1	0-0	7-7

The appendix presents an automated vocal affect recognition and classification architecture for estimating the affective states of older adults. The results show that by using random forest and logistic regression classifiers, one can classify the affective states of happy, sadness, anger, and neutral at a rate of approximately 68%. In contrast, compared to [259], where elderly valence was classified at a rate of 55% or lower. Future work will consist of investigating and comparing the features to psychoacoustic features (i.e., loudness, tempo, contour, sharpness), which have been directly linked to affective states [262].

D. Body Language Examples

Examples of participant body language during the one-on-one HRI experiments are presented herein.

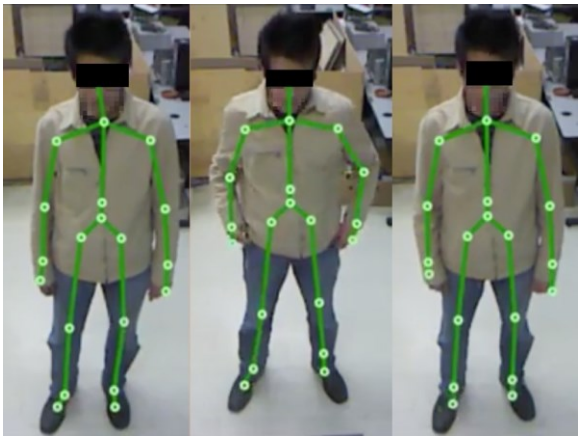


Fig. D-1. Body language with low movement dynamics, 0-valence and 0-arousal.

Fig. D-2. Body language with body leaning towards the robot, 1-valence and 1-arousal.

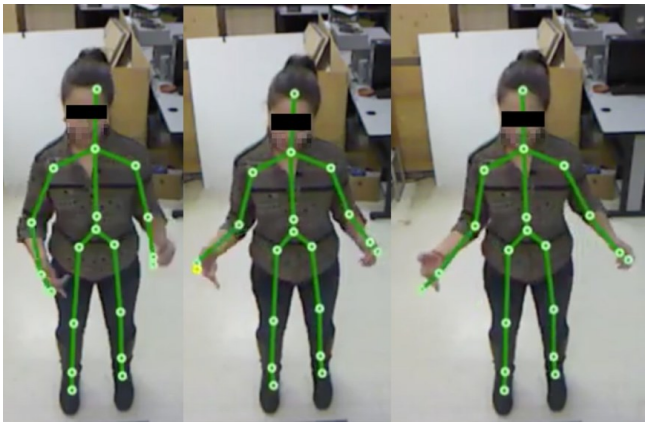


Fig. D-3. Body language with opening and stretching the trunk, 2-valence:, 1-arousal.

Fig. D-4. Body language with static posture, 0-valence: 0, -1-arousal.

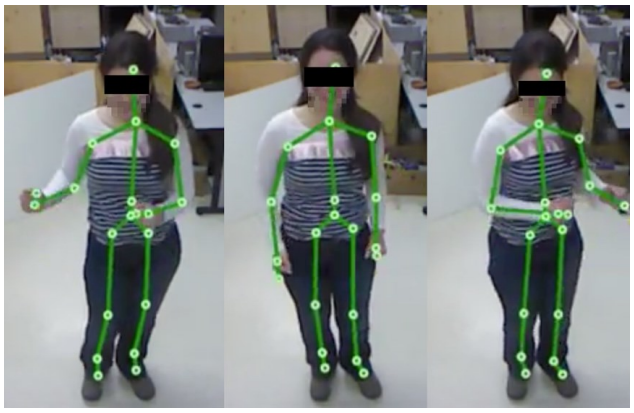


Fig. D-5. Body language with high movement dynamics, 0-valence and 2-arousal.

E. Rewards for MAXQ Hierarchical Reinforcement Learning

The reward system used herein for MAXQ is based on previous work by our group, TABLE E-1, [47]. Positive rewards are given to encourage transitions from the robot's current state to desirable states. Negative rewards are given when a transition is made from the robot's current state to an undesirable state. The reward values are chosen based on the two criteria: (1) the rewards should encourage transitions from the robot's current state to desirable states, and to avoid transitions to undesirable states, and (2) potential benefits and costs should be used to determine the magnitudes of the rewards in order to promote convergence to optimal policies. For example, successfully exiting an explored sub-scene is given a positive reward of $+50$. However, if the sub-scene is exited prior to all accessible unknown cells being explored, a negative reward of -10 is given.

TABLE E-1. MAXQ TRANSITION REWARDS FOR MULTI-ROBOT USAR [47].

Subtask	***Robot state transition***	***Reward***
Root	The mission is completed successfully	$+100$
Search Sub-scene	Exit a sub-scene after it has been successfully explored	$+50$
Search Sub-scene	Exit a sub-scene when there are still accessible unknown cells.	-10
Navigate to Unvisited Regions	Exit into Standby after exploring all unvisited regions in the sub-scene	$+10$
Navigate to Unvisited Regions	Exit into Standby when there are still accessible unvisited regions	-10
Victim Identification	Tag a victim correctly	$+10$
Victim Identification	False identification by tagging an object that is not a victim	-10
Navigate	Move into an unvisited region in the desired global exploration direction	$+15$
Navigate	Avoid an obstacle	$+10$
Navigate	Collide with an obstacle, a victim or another robot in the team	-20
Navigate	Repeatedly revisit an explored region	-1
Human Control	Human Control is requested when necessary	$+10$
Human Control	Human Control is unnecessarily requested	-10

F. USAR Post-Experimental Questionnaire

The results from the 5-point Likert questionnaire showed that participants had a better overall experience using the semi-autonomous controller (with a mean of 4.1 for this question versus a mean of 3.1 for teleoperation) and felt less stressed during the USAR mission (with a mean of 1.9 versus a mean of 3.1 for teleoperation), TABLE F-1. Participants preferred the semi-autonomous controller over solely teleoperated robots due to the task handling capabilities of the former – with no apparent a priori design bias toward either mode of operation (e.g., Questions 1 and 2). Fig. F-1 provides a graphical representation of the direct comparison of the statements 1, 2, 5, 6, and 8 for both the teleoperation and semi-autonomous modes.

TABLE F-1. POST-EXPERIMENT QUESTIONNAIRE.

Statement	Mean	Standard Error
1. With the information provided, I was able to visualize the layout of the environment in:		
a. Teleoperation mode.	4.5	0.1
b. Semi-autonomous mode.	3.9	0.2
2. I had a difficult time monitoring all of the sensory information in:		
a. Teleoperation mode.	2.4	0.3
b. Semi-autonomous mode.	2.6	0.3
3. The user interface was easy to use.	4.4	0.1
4. I had confidence in the robots performing their tasks in semi-autonomous control.	3.8	0.2
5. I had an easy time controlling all the robots in:		
a. Teleoperation mode.	3.5	0.3
b. Semi-autonomous mode.	4.3	0.2
6. I felt stressed during:		
a. Teleoperation mode.	3.1	0.2
b. Semi-autonomous mode.	1.9	0.2
7. In general, the semi-autonomous mode was more stressful for me.	2.1	0.2
8. I had a better overall experience in:		
a. Teleoperation mode.	3.1	0.3
b. Semi-autonomous mode.	4.1	0.2
9. Given a choice, I would choose manual teleoperation over semi-autonomous control.	1.7	0.2

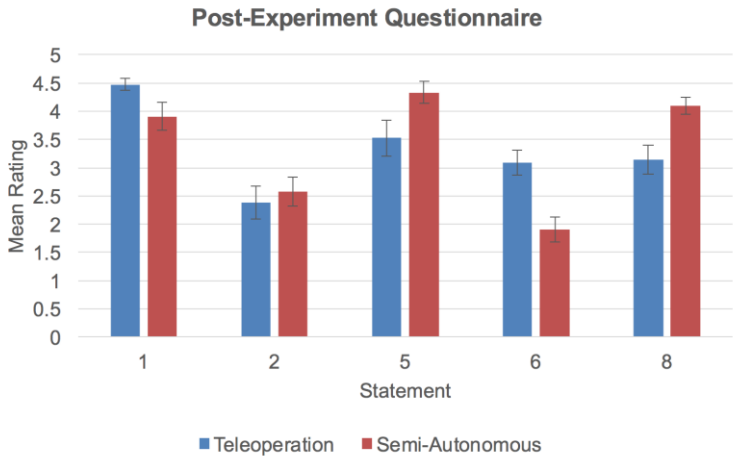


Fig. F-1. Mean ratings of post-experiment questionnaire directly comparing the teleoperation and semi-autonomous modes

G. A List of My Publications

Journal Publications

[1] A. Hong, Y. Tsuboi, G. Nejat, and B. Benhabib, "Affective Voice Recognition of Older Adults," *Journal of Medical Devices-Transactions of the ASME*, vol. 10, no. 2, 020931, 2016.

[2] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, B.Benhabib "*A Survey of Autonomous Human Affect Detection Methods for Social Robots Engaged in Natural HRI*," Journal of Intelligent Robotic Systems, vol. 82, pp. 101-133, 2016.

[3] E. Zhou, M. Zhu, A. Hong, G. Nejat, B. Benhabib, "Line-of-sight based 3D localization of parallel kinematic mechanisms," *International Journal of Smart Sensing and Intelligent Systems*, vol. 8, no. 2, pp. 842-868, 2015.

Conference Proceedings

[1] A. Hong, N. Lunscher, T. Hu, Y. Tsuboi, G. Nejat, B. Benhabib, "A Multimodal Human-Robot Interaction Architecture to Promote Natural Emotional Bi-directional Communication," in *Proc. IEEE International Symposium on Robot and Human Interactive Communication*, New York, NY, 2016, In Print.

References

[1] Goodrich, M., Schultz, A.: Human-robot interaction: a survey. J. Foundations and Trends in Human-Computer Interaction **1**(3), 203-275 (2007)

[2] Valero, A., Randelli, G., Botta, F.: Operator performance in exploration robotics. J. Intell. Robot. Syst. **64**(3-4), 365-385 (2011)

[3] Rosenthal, S., Veloso, M.: Is Someone in this Office Available to Help Me? J. Intell. Robot. Syst. **66**(2), 205–221 (2011).

[4] Swangnetr, M., Kaber, D.: Emotional state classification in patient–robot interaction using wavelet analysis and statistics-based feature selection. IEEE Trans. Human-Machine Syst. **43**(1), 63-75 (2013)

[5] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, B. Benhabib "A Survey of Autonomous Human Affect Detection Methods for Social Robots Engaged in Natural HRI," *Journal of Intelligent Robotic Systems*, vol. 82, pp. 101-133, 2016.

[6] Post, S.G., 2005, "Altruism, happiness, and health: it's good to be good," *Int. J. Behav. Med*., 12(2), pp. 66-77.

[7] Rojas, V., Ochoa, S.F., and Hervás, R., 2014, "Monitoring moods in elderly people through voice processing," Ambient Assisted Living and Daily Activities, Springer International Publishing, pp. 139-146.

[8] Breazeal, C.: Social interactions in HRI: the robot view. IEEE Trans. Syst. Man Cybern. C, Appl. Rev. **34**(2), 181-186 (2004)

[9] Tilvis, R.S., *et al*., 2012, "Social isolation, social activity and loneliness as survival indicators in old age: a nation-wide survey with a 7-year follow-up," Eur. Geriatr. Med., 3(1), pp. 18-22.

[10] Terao, J., Trejos, L., Zhang, Z., and Nejat, G., 2008, "An intelligent socially assistive robot for health care," Proceedings of the IMECE, Boston, MA, pp. 69-78.

[11] Lawton, M.P., Haitsma, K.V., and Klapper, J., 1996, "Observed affect in nursing home residents with Alzheimer's disease," J. Gerontol. B Psychol. Sci. Soc. Sci., **51B**(1), pp. P3-P14.

[12] Keltner, D., Haidt, J.: Social functions of emotions at four levels of analysis. Cognition and Emotion **13**(5), 505-521 (1999)

[13] Scherer, K.: Psychological models of emotion. The neuropsychology of emotion, pp. 137-162 (2000)

[14] Picard, R.: Affective computing. MIT Press (2000)

[15] Sorbello, R., Chella, A., Calí, C.: Telenoid android robot as an embodied perceptual social regulation medium engaging natural human–humanoid interaction. Robotics and Autonomous Syst. **62**(9), 1329-1341 (2014)

[16] R. Picard, E. Vyzas and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state", IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 10, pp. 1175-1191, 2001.

[17] R. Calvo and S. D'Mello, "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications", IEEE Trans. Affect. Comput., vol. 1, no. 1, pp. 18-37, 2010.

[18] M. Paleari, R. Chellali and B. Huet, "Bimodal Emotion Recognition," Social Robotics, vol. 6414, pp. 305-314, 2010.

[19] F. Cid, J. Moreno, P. Bustos and P. Núñez, “Muecas: A Multi-Sensor Robotic Head for Affective Human Robot Interaction and Imitation,” Sensors, vol. 14, no. 5, pp. 7711-7737, 2014.

[20] D. Limbu, W. Anthony, T. Adrian, T. Dung, T. Kee, T. Dat, W. Alvin, N. Terence, J. Ridong and L. Jun, “Affective social interaction with CuDDler robot,” in Proc. IEEE Int. Conf. Robotics, Automation and Mechatronics, Manila, Philippines, 2013, pp. 179-184.

[21] H.-W. Jung, Y.-H. Seo, M. Ryoo and H. Yang, “Affective communication system with multimodality for a humanoid robot, AMI,” in *Proc. IEEE/RAS Int. Conf. Humanoid Robots*, Santa Monica, CA, 2004, pp. 690-706.

[22] J. Prado, C. Simplício, N. Lori and J. Dias, “Visuo-auditory Multimodal Emotional Structure to Improve Human-Robot-Interaction.” *Int. J. Soc. Robot.*, vol. 4, no. 1, pp. 29-51, 2011.

[23] C. Shan, S. Gong and P. McOwan, “Beyond Facial Expressions: Learning Human Emotion from Body Gestures,” in *Proc. British Machine Vision Conference*, Warwick, United Kingdom, 2007, pp 1-10.

[24] J. Van den Stock, R. Righart and B. de Gelder, “Body expressions influence recognition of emotions in the face and voice.”, *Emotion*, vol. 7, no. 3, pp. 487-494, 2007.

[25] K. R. Scherer, “Vocal affect expression: A review and a model for future research.” *Psychol. Bulletin*, vol. 99, no. 2, p. 143, 1986.

[26] Y. Liu and G. Nejat, "Robotic urban search and rescue: a survey from the control perspective," *J. Intell. Robot. Syst.*, vol. 72, no. 2, pp. 147-165, 2013.

[27] J. Casper, M. Micire, and R. Murphy, "Issues in intelligent robots for search and rescue," in *Proc. SPIE Unmanned Ground Vehicle Technology II*, Orlando, FL, 2000, pp. 292-302.

[28] J. Casper and R. Murphy, "Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center," *IEEE Trans. Syst. Man Cybern. B*, vol. 33, no. 3, pp. 367-385, 2003.

[29] C. Wong, G. Seet, and S. Sim, "Multiple-robot systems for USAR: key design attributes and deployment issues," *Int. J. Adv. Robot. Syst.*, vol. 8, no. 1, pp. 85-101, 2011.

[30] Y. Gatsoulis, G. S. Virk, and A. A. Dehghani-Sanij, "On the measurement of situation awareness for effective human-robot interaction in teleoperated systems," *J. Cogn. Eng. Decis. Mak.*, vol. 4, no. 1, pp. 69-98, 2010.

[31] J. A. Adams, "Multiple robot/single human interaction: effects on perceived workload," *Behav. Inf. Technol.*, vol. 28, no. 2, pp. 183-198, 2009.

[32] E. de Visser, and R. Parasuraman, "Adaptive aiding of human-robot teaming effects of imperfect automation on performance, trust, and workload," *J. Cogn. Eng. Decis. Mak.*, vol. 5, no. 2, pp. 209-231, 2011.

[33] J. Y. C. Chen, P. J. Durlach, J. A. Sloan, and L. D. Bowens, "Human-robot interaction in the context of simulated route reconnaissance missions," *Mil. Psychol.*, vol. 20, no. 3, pp. 135-149, 2008.

[34] L. A. Breslow, D. Gartenberg, J. M. McCurry, and J. G. Trafton, "Dynamic operator overload: A model for predicting workload during supervisory control," *IEEE Trans. Hum. Mach. Syst.*, vol. 44, no. 1, pp. 30-40, 2014.

[35] R. McKendrick, T. Shaw, E. de Visser, H. Sager, B. Kidwell, and R. Parasuraman, "Team performance in networked supervisory control of unmanned air vehicles effects of

automation, working memory, and communication content," *Hum. Factors*, vol. 56, no. 3, pp. 463-475, 2014.

[36] T. D. Fincannon, A. W. Evans, E. Phillips, F. Jentsch, and J. Keebler, "The influence of team size and communication modality on team effectiveness with unmanned systems" in *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, San Antonio, TX, 2009, pp. 419-423.

[37] E. de Visser, B. Kidwell, J. Payne, L. Lu, J. Parker, N. Brooks, T. Chabuk, S. Spriggs, A. Freedy, P. Scerri, R. Parasuraman, "Best of both worlds design and evaluation of an adaptive delegation interface," in *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, San Diego, CA, 2013, pp. 255-259.

[38] M. L. Cummings, L. F. Bertucelli, J. Macbeth, and A. Surana, "Task versus vehicle-based control paradigms in multiple unmanned vehicle supervision by a single operator," *IEEE Trans. Hum. Mach. Syst.*, vol. 44, no. 3, pp. 353-361, 2014.

[39] M. S. Prewett, K. N. Saboe, R. C. Johnson, M. D. Coovert, and L. R. Elliot, "Workload in human-robot interaction: a review of manipulations and outcomes," in *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, San Antonio, TX, 2009, pp. 1393-1397.

[40] P. Velagapudi, P. Scerri, K. Sycara, H. Wang, M. Lewis, and J. Wang, "Scaling effects in multi-robot control," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nice, France, 2008, pp. 2121-2126.

[41] M. Hsieh, A. Cowley, J. F. Keller, L. Chaimowicz, B. Grocholsky, V. Kumar, C. J. Taylor, Y. Endo, R. C. Arkin, B. Jung, D. F. Wolf, G. S. Sukhatme, and D. C. MacKenzie, "Adaptive teams of autonomous aerial and ground robots for situation awareness," *J. Field Robot.*, vol. 24, no. 11, pp. 991-1014, 2007.

[42] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman. "A meta-analysis of factors affecting trust in human-robot interaction", *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517-527, 2011.

[43] F. Gao, M. L. Cummings, and E. T. Solovey, "Modeling teamwork in supervisory control of multiple robots." *IEEE Trans. Hum. Mach. Syst.*, vol 44, no. 4, pp. 441-453, 2014.

[44] J. Vilela, Y. Liu, and G. Nejat, "Semi-autonomous exploration with robot teams in urban search and rescue," in *Proc. IEEE Int. Symp. Safety, Security, and Rescue Robotics*, Linkoping, Sweden, 2013, pp. 1-6.

[45] T. Gunn, and J. Anderson, "Dynamic heterogeneous team formation for robotic urban search and rescue," *J. Comp. Syst. Sci.*, vol. 81, no. 3, pp. 553-567, 2015.

[46] J. M. Whetten and M. A. Goodrich, "Specialization, fan-out and multi-human/multi-robot supervisory control," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction*, Osaka, Japan, 2010, pp. 147-148.

[47] Y. Liu, and G. Nejat, "Multirobot Cooperative Learning for Semiautonomous Control in Urban Search and Rescue Applications," *J. Field Robot.*, doi: 10.1002/rob.21597, 2015.

[48] B. Doroodgar, Y. Liu, and G. Nejat, "A Learning-Based Semi-Autonomous Controller for Robotic Exploration of Unknown Disaster Scenes While Searching for Victims," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2719-2732, 2014.

[49] Y. Liu, G. Nejat, and B. Doroodgar, "Learning-based semi-autonomous control for robots in urban search and rescue," in *Proc. IEEE Int. Symp. Safety, Security, and Rescue*

Robotics, College St., TX, 2012, pp. 1-6.

[50] N. Brooks, P. Scerri, and K. Sycara, "Asynchronous control with ATR for large robot teams," in *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, Las Vegas, Nevada, 2011, pp. 444-448.

[51] H. Wang, M. Lewis, S. Chien, and P. Velagapudi, "Scaling effects for synchronous vs. asynchronous video in multi-robot search," in *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, San Antonio, TX, 2009, pp. 364-368.

[52] N. Sato, F. Matsuno, T. Yamasaki, T. Kamegawa, N. Shiroma, and H. Igarashi, "Cooperative task execution by a multiple robot team and its operators in search and rescue operations," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sendai, Japan, 2004, pp. 1083-1088.

[53] J. L. Burke, and R. R. Murphy "Using Mobile Robots as a Shared Visual Presence in USAR Environments." *AAAI Spring Symposium: AI Technologies for Homeland Security.* 2005, pp. 111-116.

[54] M. Lewis, H. Wang, and S. Y. Chien, "Process and performance in human-robot teams," *J. Cogn. Eng. Decis. Mak.*, vol. 5, no. 2, pp. 186-208, 2011.

[55] R. Nourjou, S. F. Smith, M. Hatayama, and P. Szekely. "Intelligent algorithm for assignment of agents to human strategy in centralized multi-agent coordination," *Journal of Software*, vol. 9, no. 10, pp. 2586-2597, 2014.

[56] S. Chien, H. Wang, and M. Lewis, "Effects of alarms on control of robot teams," in *Proc. Hum, Factors Ergon. Soc. Annu. Meet.*, Las Vegas, Nevada, 2011, pp. 434-438.

[57] M. Lewis, and J. Wang. "Measuring coordination demand in multirobot teams," in *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, San Antonio, TX, 2009, pp. 779-783.

[58] M. Lewis, H. Wang, and S. Y. Chien, "Choosing autonomy modes for multirobot search," *Hum. Factors*, vol 52, no. 2, pp. 225-233, 2010.

[59] J. M. Whetten and M. A. Goodrich, "Specialization, fan-out and multi-human/multi-robot supervisory control," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction*, Osaka, Japan, 2010, pp. 147-148.

[60] A. Rosenfeld, N. Agmon, O. Maksimov, A. Azaria, and S. Kraus, "Intelligent agent supporting human-multi-robot team collaboration," in *Proc. AAMAS workshop ARMS*, Istanbul, Turkey, 2015.

[61] A. Mouaddib, S. Zilberstein, A. Beynier, and L. Jeanpierre, “A decision-theoretic approach to cooperative control and adjustable autonomy,” in *Proc. European Conference on Artificial Intelligence*, Lisbon, Portugal, 2010, pp. 971-972.

[62] N. Côté, A. Canu, M. Bouzid, and A. Mouaddib, “Humans-robots sliding collaboration control in complex environments with adjustable autonomy,” in *Proc IEEE/WIC/ACM Int. Conf. Web Intelligence and Intelligent Agent Technology*, Macau, China, 2012, pp. 146-153.

[63] Martinez, A., Du, S.: A Model of the Perception of Facial Expressions of Emotion by Humans: Research Overview and Perspectives. J.Machine Learning Research **13**(1), 1589-1608 (2012)

[64] Morris, J.D.: Observations: SAM: The self-assessment manikin An efficient cross-cultural measurement of emotional response. J. Advertizing Research **35**(8), 63-68 (1995)

[65] Niedenthal, P.M., Halberstadt, J.B., Setterlund, M.B.: Being happy and seeing "happy" emotional state mediates visual word recognition. Cognition and Emotion **11**(4), 403–432 (1997)
[66] Russell, J.A., Fernández-Dols, J. M.: The psychology of facial expression (1997)
[67] Y. Matsuda, T. Fujimura, K. Katahira, M. Okada, K. Ueno, K. Cheng and K. Okanoya, "The implicit processing of categorical and dimensional strategies: an fMRI study of facial emotion perception", *Frontiers in Human Neuroscience*, vol. 7, 2013.
[68] Darwin, C.: The expression of the emotions in man and animals. The Amer. J. Med. Sci. **232**(4) 477 (1956)
[69] Hegel, F., Spexard, T.: Playing a different imitation game: Interaction with an Empathic Android Robot. In: Proceedings of the IEEE-RAS Int. Conf. Humanoid Robots, pp. 56-61 (2006)
[70] Tomkins, S: Affect, imagery, consciousness: Vol. I. The positive affects, Oxford, England (1962)
[71] Tomkins, S: Affect, imagery, consciousness: Vol. II. The negative affects, Oxford, England (1963)
[72] Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Personality and Social Psychology **17**(2), 124-129 (1971)
[73] Ekman, P., Friesen, W., Ellsworth, P.: Emotion in the human face: Guidelines for research and an integration of findings. Pergamon Press (1972)
[74] Bann, E.Y., Bryson, J.J.: The Conceptualisation of Emotion Qualia: Semantic Clustering of Emotional Tweets. Progress in Neural Processing **21**, 249-263 (2012)
[75] Harish, R., Khan, S., Ali, S., Jain, V.: Human computer interaction-A brief study. Int. J. of Managment, IT and Eng. **3**(7), 390-401 (2013)
[76] Barrett, L.F., Gendron, M., Huang, Y. M.: Do discrete emotions exist? Philos. Psychol. **22**(4), 427–437 (2009)
[77] Wundt, W.: Outlines of psychology. In: Wilhelm Wundt and the Making of a Scientific Psychology, pp. 179–195 (1980)
[78] Schlosberg, H.: Three dimensions of emotion. Psychological Review **61**(2), 81-88 (1954)
[79] Trnka, R., Balcar, K., Kuska, M.: Re-constructing Emotional Spaces: From Experience to Regulation. Prague Psychosocial Press (2011)
[80] Plutchik, R., Conte, H.: Circumplex models of personality and emotions, Washington, DC (1997)
[81] Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology **14**(4), 261-292 (1996)
[82] Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**(6), 1161-1178 (1980)
[83] Remmington, N. A., Fabrigar, L .R., Visser, P. S.: Reexamining the circumplex model of affect. J. Personality and Social Psychology **79**(2), 286–300 (2000)
[84] Rubin, D.C., Talarico, J.M.: A comparison of dimensional models of emotion: evidence from emotions, prototypical events, autobiographical memories, and words. Memory **17**(8), 802–808 (2009)
[85] Watson, D., Tellegen, A.: Toward a consensual structure of mood. Psychological Bulletin **98**(2), 219-235 (1985)
[86] Barrett, L.F.: Discrete emotions or dimensions? The role of valence focus and arousal focus. Cognition and Emotion **12**(4), 579–599 (1998)

[87] Kobayashi, H., Hara, F.: The recognition of basic facial expressions by neural network. In: Proceedings of the IEEE Int. Joint Conf. Neural Networks, pp. 460-466 (1991)
[88] Wimmer, M., MacDonald, B.A., Jayamuni, D., Yadav, A.: Facial Expression Recognition for Human-robot Interaction–A Prototype. Robot Vision **4931**, 139-152 (2008)
[89] Luo, R.C., Lin, P.H., Wu, Y.C., Huang, C.Y.: Dynamic Face Recognition System in Recognizing Facial Expressions for Service Robotics. In: Proceedings of the IEEE/ASME Int. Conf. on Advanced Intell. Mechatronics, pp. 879-884 (2012)
[90] Tscherepanow, M., Hillebrand, M., Hegel, F., Wrede, B., Kummert, F.: Direct imitation of human facial expressions by a user-interface robot. In: Proceedings of the IEEE-RAS Int. Conf. on Humanoid Robots, pp. 154-160 (2009)
[91] Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P.W., Paiva, A.: Automatic analysis of affective postures and body motion to detect engagement with a game companion. In: Proceedings of the ACM/IEEE Int. Conf. on Human-Robot Interaction, pp. 305-311 (2011)
[92] Barakova, E., Lourens, T.: Expressing and interpreting emotional movements in social games with robots. Personal Ubiquitous Comput. **14**(5), 457-467 (2010)
[93] Xiao, Y., Zhang, Z., Beck, A., Yuan, J., Thalmann, D.: Human-virtual human interaction by upper body gesture understanding. In: Proceeding of the ACM Symp. on Virtual Reality Software and Technology, pp. 133-142 (2013)
[94] Cooney, M., Nishio, S., Ishiguro, H.: Recognizing affection for a touch-based interaction with a humanoid robot. In: Proceedings of the IEEE/RSJ Int. Conf. on Intell. Robots and Syst., pp. 1420-1427 (2012)
[95] Scheutz, M., Dame, N., Schermerhorn, P., Kramer, J.: The Utility of Affect Expression in Natural Language Interactions in Joint Human-Robot Tasks. In: Proceedings of ACM SIGCHI/SIGART Conf. Human-Robot Interaction. pp. 226–233 (2006)
[96] Kim, H.R., Kwon, D.S.: Computational model of emotion generation for human–robot interaction based on the cognitive appraisal theory. J. Intell. Robot. Syst. **60**(2), 263-283 (2010)
[97] Lin, Y.Y., Le, Z., Becker, E., Makedon, F.: Acoustical implicit communication in human-robot interaction. In: Proceedings of the Conf. on Pervasive Technologies Related to Assistive Environments, pp. 5 (2010)
[98] Hyun, K.H., Kim, E.H., Kwak, Y.K.: Emotional feature extraction method based on the concentration of phoneme influence for human–robot interaction. Advanced Robotics **24**(1-2), 47-67 (2010)
[99] Yun, S., Yoo, C.D.: Speech emotion recognition via a max-margin framework incorporating a loss function based on the Watson and Tellegen's emotion model. In: Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 4169-4172 (2009)
[100] Kim, E. H., Hyun, K. H., Kim, S. H.: Improved emotion recognition with a novel speaker-independent feature. IEEE/ASME Trans. Mechatronics, **14**(3), 317-325 (2009)
[101] Kulic, D., Croft, E.: Anxiety detection during human-robot interaction. In: Proceedings of the IEEE/RSJ Int. Conf. on Intell. Robots and Syst., pp. 616-621 (2005)
[102] Rani, P., Liu, C., Sarkar, N., Vanman, E.: An empirical study of machine learning techniques for affect recognition in human–robot interaction. Pattern Anal. Applicat.. **9**(1), 58–69 (2006)
[103] Strait, M., Scheutz, M.: Using near infrared spectroscopy to index temporal changes in

affect in realistic human–robot interactions. In: Physiological Computing Syst., Special Session on Affect Recogntion from Physiological Data for Social Robots (2014)
[104] N. Lazzeri, D. Mazzei and D. De Rossi, "Development and Testing of a Multimodal Acquisition Platform for Human-Robot Interaction Affective Studies", *Journal of Human-Robot Interaction*, vol. 3, no. 2, pp. 1-24, 2014.
[105] Paleari, M., Chellali, R., Huet, B.: Bimodal emotion recognition. Social Robotics **6414**, 305-314 (2010)
[106] Cid, F., Prado, J.A., Bustos, P., Nunez, P.: A real time and robust facial expression recognition and imitation approach for affective human-robot interaction using Gabor filtering. In: Proceedings of the IEEE/RSJ Int. Conf. on Intell. Robots and Syst., pp. 2188–2193 (2013)
[107] Schacter, D., Wang, C., Nejat, G., Benhabib, B.: A two-dimensional facial-affect estimation system for human–robot interaction using facial expression parameters. Advanced Robot. **27**(4), 259–273 (2013)
[108] Tielman, M., Neerincx, M., Meyer, J. J., Looije, R.: Adaptive emotional expression in robot-child interaction. In: Proceedings of the ACM/IEEE Int. Conf. on Human-robot interaction, pp. 407–414 (2014)
[109] Leite, I., Castellano, G., Pereira, A., Martinho, C., Paiva, A.: Modelling Empathic Behaviour in a Robotic Game Companion for Children : an Ethnographic Study in Real-World Settings. In: Proceedings of the ACM/IEEE Int. Conf. on Human-Robot Interaction, pp. 367-374 (2012)
[110] McColl, D., Nejat, G.: Determining the affective body language of older adults during socially assistive HRI. In: Proceedings of the IEEE Int. Conf. on Intell. Robots Syst. pp. 2633-2638 (2014)
[111] McColl, D., Nejat, G.: Affect detection from body language during social HRI. In: Proceedings of the IEEE Int. Symp. on Robot and Human Interactive Communication, pp. 1013-1018 (2012)
[112] Xu, J., Broekens, J., Hindriks, K., Neerincx, M.: Robot mood is contagious: effects of robot body language in the imitation game. In: Proceedings of the Int. Conf. on Autonomous agents and multi-agent Syst., pp. 973-980 (2014)
[113] Iengo, S., Origlia, A., Staffa, M., Finzi, A.: Attentional and emotional regulation in human-robot interaction In: Proceedings of the IEEE Int. Symp. on Robot and Human Interactive Communication, pp. 1135-1140 (2012)
[114] Tahon, M.: Usual voice quality features and glottal features for emotional valence detection. In: Proceedings of Speech Prosody, pp. 1-8 (2012)
[115] Conn, K., Liu, C., Sarkar, N.: Towards affect-sensitive assistive intervention technologies for children with autism. Affective Computing: Focus on Emotion Expression, Synthesis and Recognition 365-390 (2008)
[116] Kulic, D., Croft, E. A.: Affective State Estimation for Human-Robot Interaction. IEEE Trans. Robotics **23**(5), 991–1000 (2007)
[117] Rani, P., Liu, C., Sarkar, N.: Affective feedback in closed loop human-robot interaction. In: Proceedings of the ACM SIGCHI/SIGART Conf. on Human-robot interaction, pp. 335-336 (2006)
[118] Saulnier, P., Sharlin, E., Greenberg, S.: Using bio-electrical signals to influence the social behaviours of domesticated robots In: Proceedings of the ACM/IEEE Int. Conf. on Human robot interaction, pp. 263-264 (2009)

[119] Broadbent, E., Lee, Y. I., Stafford, R. Q., Kuo, I. H.: Mental schemas of robots as more human-like are associated with higher blood pressure and negative emotions in a human-robot interaction. Int. J. Soc. Robot. **3**(3), 291-297 (2011)
[120] Schaaff, K., Schultz, T.: Towards an EEG-based emotion recognizer for humanoid robots. In: Proceedings of the IEEE Int. Symp. on Robot and Human Interactive Communication, pp. 792-796 (2009)
[121] Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., Mcowan, P.W.: Multimodal affect modeling and recognition for empathic robot companions. Int. J. Humanoid Robotics **10**(1), 1–23 (2013)
[122] Gonsior, B., Sosnowski, S., Buss, M., Wollherr, D., Kuhnlenz, K.: An Emotional Adaption Approach to increase Helpfulness towards a Robot. In: Proceedings of the IEEE/RSJ Int. Conf. on Intell. Robots and Syst., pp. 2429-2436 (2012)
[123] Jung, H., Seo, Y., Ryoo, M.S., Yang, H.S.: Affective communication system with multimodality for a humanoid robot, AMI. In: Proceedings of the IEEE/RAS Int. Conf. on Humanoid Robots, 690-706 (2004)
[124] Lim, A., Ogata, T., Okuno, H.G.: Towards expressive musical robots: a cross-modal framework for emotional gesture, voice and music. EURASIP J. Audio, Speech, and Music Processing **2012**(1), 1-12 (2012)
[125] Breuer, T., Giorgana Macedo, G. R., Hartanto, R., Hochgeschwender, N., Holz, D., Hegger, F., Jin, Z., Müller, C., Paulus, J., Reckhaus, M., Álvarez Ruiz, J.A., Plöger, P.G., Kraetzschmar, G.K.: Johnny: An Autonomous Service Robot for Domestic Environments. J. Intell. Robot. Syst. **66**(1-2), 245–272 (2011)
[126] Littlewort, G., Bartlett, M.S., Fasel, I., Chenu, J., Kanda, T., Ishiguro, H., Movellan, J.R.: Towards Social Robots: Automatic Evaluation of Human-robot Interaction by Face Detection and Expression Classification. In: Advances in Neural Information Processing Syst., vol. 16, MIT Press (2003)
[127] Boucenna, S., Gaussier, P., Andry, P., Hafemeister, L.: Imitation as a communication tool for online facial expression learning and recognition. In: Proceedings of the IEEE/RSJ Int. Conf. on Intell. Robots and Syst., pp. 5323-5328 (2010)
[128] Kobayashi, H., Hara, F.: Facial Interaction between Animated 3D Face Robot and Human Beings. In: Proceedings of the IEEE Int. Conf. on Syst., Man, and Cybernetics, vol. 4, pp. 3732-3737 (1997)
[129] Garcíia Bueno, J., González-Fierro, M., Moreno, L., Balaguer, C.: Facial Emotion Recognition and Adaptative Postural Reaction by a Humanoid based on Neural Evolution. Int. J. Advanced Computer Sci.. **3**(10), 481–493 (2013)
[130] Garcíia Bueno, J., González-Fierro, M., Moreno, L., Balaguer, C.: Facial gesture recognition using active appearance models based on neural evolution. In: Proceedings of the ACM/IEEE Int. Conf. on Human-Robot Interaction, pp. 133-134 (2012)
[131] Dornaika, F., Raducanu, B.: Efficient Facial Expression Recognition for Human Robot Interaction. In: Computational and Ambient Intelligence, pp. 700-708 (2007)
[132] Luo, R.C., Lin, P. H., Chang, L. W.: Confidence fusion based emotion recognition of multiple persons for human-robot interaction. In: Proceedings of the IEEE/RSJ Int. Conf. on Intell. Robots and Syst., pp. 4590-4595 (2012)
[133] Cattinelli, I., Goldwurm, M., Borghese, N.A.: Interacting with an artificial partner: modeling the role of emotional aspects. Biological Cybernetics **99**(6), 473–89 (2008)
[134] Hyun, K., Kim, E., Kwak, Y.: Emotional feature extraction based on phoneme information

for speech emotion recognition. In: Proceedings of the IEEE Int. Symp. Robot and Human Interactive Comm. pp. 802-806 (2007)
[135] Song, K., Han, M., Wang, S.: Speech signal-based emotion recognition and its application to entertainment robots. J. Chinese Inst. of Eng. **37**(1), 14-25 (2014)
[136] Cid, F., Moreno, J., Bustos, P., Núñez, P.: Muecas: a multi-sensor robotic head for affective human robot interaction and imitation. Sensors **14**(5), 7711–7737 (2014)
[137] F. Alonso-Martín, M. Malfaz, J. Sequeira, J. Gorostiza and M. Salichs, "A Multimodal Emotion Detection System during Human–Robot Interaction." *Sensors*, vol. 13, no. 11, pp. 15549-15581, 2013.
[138] Limbu, D.K., Anthony, W.C.Y., Adrian, T.H.J., Dung, T.A., Kee, T.Y., Dat, T.H., Alvin, W.H.Y., Terence, N.W.Z., Ridong, J., Jun, L.: Affective social interaction with CuDDler robot. In: Proceedings of the IEEE Int. Conf. on Robotics, Automation and Mechatronics, pp. 179-184 (2013)
[139] Prado, J.A., Simplício, C., Lori, N.F., Dias, J.: Visuo-auditory Multimodal Emotional Structure to Improve Human-Robot-Interaction. Int. J. Soc. Robot. **4**(1), 29–51 (2011)
[140] Kulic, D., Croft, E.: Affective state estimation for human–robot interaction. IEEE Trans. Robot. **23**(5), 991-1000 (2007)
[141] Lim, A., Member, S., Okuno, H. G.: The MEI Robot : Towards Using Motherese to Develop Multimodal Emotional Intelligence. IEEE Trans. Autonomous Mental Development **6**(2), 126–138 (2014)
[142] Rani, P., Sarkar, N., Smith, C., Kirby, L.: Anxiety detecting robotic system–towards implicit human-robot collaboration. Robotica **22**(1), 85-95 (2004)
[143] Strait, M., Scheutz, M.: Measuring users' responses to humans, robots, and human-like robots with functional near infrared spectroscopy. In Proceedings of the IEEE Int. Symp. Robot and Human Interactive Communication pp. 1128-1133 (2014)
[144] Keltner, D., Ekman, P., Gonzaga, G.C., Beer, J.: Facial expression of emotion. Handbook of affective sciences. Series in affective science, pp. 415-432 (2003)
[145] Schiano, D.J., Ehrlich, S. M., Rahardja, K., Sheridan, K: Face to interface: facial affect in (hu)man and machine. In: Proceedings of the SIGCHI Conf. on Human Factors in Computing Systems, pp. 193–200 (2000)
[146] Fridlund, A.J., Ekman, P., Oster, H.: Facial expressions of emotion. Nonverbal Behavior and Communication (2nd ed.), pp. 143-223 (1987)
[147] Fridlund, A.J.: Human facial expression: An evolutionary view. Academic Press (1994)
[148] Fridlund, A.J.: The new ethology of human facial expressions. In: The psychology of facial expression, pp. 103-127. Cambridge University Press (1997)
[149] Yang, Y., Ge, S.S., Lee, T.H., Wang, C.: Facial expression recognition and tracking for intelligent human-robot interaction. J. Intell. Serv. Robot. **1**(2), 143–157 (2008)
[150] Tapus, A., Maja, M., Scassellatti, B.: The grand challenges in socially assistive robotics. IEEE Robotics and Automation Mag. **14**(1) 1–7 (2007)
[151] Bartlett, M. S., Littlewort, G., Fasel, I., Movellan, J. R.: Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. In: Proceedings of the CVPRW Conf. Computer Vision and Pattern Recognition, vol. 5, pp. 53-53 (2003)
[152] Castellano, G., Caridakis, G., Camurri, A., Karpouzis, K., Volpe, G., Kollias, S.: Body gesture and facial expression analysis for automatic affect recognition. Blueprint for affective computing: A sourcebook, pp. 245-255 (2010)

[153] Fong, T., Nourbakhsh, I., Dautenhahn, K.: A Survey of Socially Interactive Robots Robotics and Autonomous Syst. **42**(3), 143-166 (2003)
[154] Strupp, S., Schmitz, N., Berns, K.: Visual-based emotion detection for natural man-machine interaction. In: Advanced Artificial Intell. pp. 356-363 (2008)
[155] Li, Y., Hashimoto, M.: Effect of Emotional Synchronization using Facial Expression. In: Proceedings of the IEEE Int. Conf. on Robotics and Biomimetics, pp. 2872-2877 (2011)
[156] Ekman, P., Friesen, W.: Facial action coding system. Consulting Psychologists Press Inc (1977)
[157] Shan, C., Gong, S., McOwan, P.: Beyond Facial Expressions: Learning Human Emotion from Body Gestures. In: Proceedings of the British Mach. Vision Conf. pp. 1-10 (2007)
[158] Mehrabian, A.: Significance of posture and position in the communication of attitude and status relationships. Psychology Bulletin **71**(5), 359-372 (1969)
[159] Montepare, J., Koff, E., Zaitchik, D., Albert, M.: The use of body movements and gestures as cues to emotions in younger and older adults. J. Nonverbal Behav. **23**(2), 133-152 (1999)
[160] Wallbott, H.: Bodily expression of emotion. Eur. J. Soc. Psychology **28**(6), 879-896 (1998)
[161] Kleinsmith, A., Bianchi-Berthouze, N.: Recognizing affective dimensions from body posture. Affective Computing and Intell. Interaction **4738**, 48-58 (2007)
[162] Gross, M., Crane, E., Fredrickson, B.: Effort-shape and kinematic assessment of bodily expression of emotion during gait. Human Movement Sci. **31**(1), 202-221 (2012)
[163] Xu, D., Wu, X., Chen, Y., Xu, Y.: Online Dynamic Gesture Recognition for Human Robot Interaction. J. Intell. Robot. Syst. doi: 10.1007/s10846-014-0039-4 (2014)
[164] Hasanuzzaman, M.: Gesture-based human-robot interaction using a knowledge-based software platform. Industrial Robot An Int. J. **33**(1), 37-49 (2006)
[165] Yan, R., Tee, K., Chua, Y.: Gesture Recognition Based on Localist Attractor Networks with Application to Robot Control. In: IEEE Computational Intell. Mag., pp. 64-74 (2012)
[166] Suryawanshi, D., Khandelwal, C.: An Integrated Color and Hand Gesture Recognition Control for Wireless Robot. Int. J. Advances in Eng. & Tech. **3**(1), 427-435 (2012)
[167] Obaid, M., Kistler, F., Häring, M.: A Framework for User-Defined Body Gestures to Control a Humanoid Robot. Int. J. Soc. Robot. **6**(3), 383-396 (2014)
[168] Malima, A., Ozgur, E., Çetin, M.: A fast algorithm for vision-based hand gesture recognition for robot control. In: Proceedings of the IEEE Signal Processing and Communication Applic. pp. 1-4 (2006)
[169] Waldherr, S., Romero, R., Thrun, S.: A gesture based interface for human-robot interaction. Autonomous Robots **9**(2), 151-173 (2000)
[170] Corradini, A., Gross, H.: Camera-based gesture recognition for robot control. In: Proceedings of the IEEE-INNS-ENNS Int. Joint Conf. Neural Networks, vol. 4, pp. 133-138 (2000)
[171] Boehme, H.: Neural networks for gesture-based remote control of a mobile robot. In: Proceedings of the IEEE World Congr. on Computer Intell. and IEEE Int. Joint Conf. Neural Networks, vol. 1, pp. 372-377 (1998)
[172] Burger, B., Ferrané, I., Lerasle, F.: Multimodal interaction abilities for a robot companion. Computer Vision Syst., vol. 5008, pp. 549-558 (2008)
[173] Rogalla, O., Ehrenmann, M.: Using gesture and speech control for commanding a robot assistant. In: Proceedings of the IEEE Int. Workshop Robot and Human Interactive Comm. pp. 454-459 (2002)

[174] Becker, M., Kefalea, E., Maël, E.: GripSee: A gesture-controlled robot for object perception and manipulation. Autonomous Robot. **6**(2), 203-221 (1999)
[175] Gerlich, L., Parsons, B., White, A.: Gesture recognition for control of rehabilitation robots. Cognition Technol. & Work, **9**(4), 189-207 (2007)
[176] Raheja, J., Shyam, R.: Real-time robotic hand control using hand gestures. In: Proceedings of the Second Int. Conf. Machine Learning and Computing pp. 12-16 (2010)
[177] Kanda, T., Ishiguro, H., Imai, M., Ono, T.: Development and evaluation of interactive humanoid robots. Proc. of the IEEE, **92**(11), 1839-1850 (2004)
[178] Scherer, K., Bänziger, T.: Emotional expression in prosody: a review and an agenda for future research. In: Proceedings of the Speech Prosody, pp. 359-366 (2004)
[179] Kreibig, S.: Autonomic nervous system activity in emotion: A review. Biological Psychology **84**(3), 394-421 (2010)
[180] Smith, C.: Dimensions of appraisal and physiological response in emotion. J. Personality and Soc. Psychology **56**(3), 339-353(1989)
[181] Fernández, C., Pascual, J., Soler, J.: Physiological responses induced by emotion-eliciting films. Applied Psychophysiology and Biofeedback **37**(2), 73-79 (2012)
[182] Liu, C., Conn, K., Sarkar, N., Stone, W.: Online affect detection and robot behavior adaptation for intervention of children with autism. IEEE Trans. Robotics **24**(4), 883-896 (2008)
[183] Feil-Seifer, D., Mataric, M.J.: Socially Assistive Robotics. IEEE Robotics & Automation Mag. **18**(1), 24-31. (2011)
[184] Valenza, G., Lanata, A., Scilingo, E. P.: The role of nonlinear dynamics in affective valence and arousal recognition. IEEE Trans. Affective Comput., **3**(2), 237-249 (2012)
[185] Littlewort, G., Whitehill, J., Wu, T. F., Butko, N., Ruvolo, P., Movellan, J., Bartlett, M.: The Motion in Emotion—A CERT based approach to the FERA emotion challenge. In: Proceedings of the IEEE Int. Conf. on Automatic Face & Gesture Recognition and Workshops, pp. 897-902 (2011)
[186] Ma, Y., Paterson, H., Pollick, F.: A motion capture library for the study of identity, gender, and emotion perception from biological motion. Behavior Research Methods **38**(1), 134-141 (2006)
[187] S. Hussain, R. A. Calvo, "A Framework for Multimodal Affect Recognition," Poster [Online]: http://sydney.edu.au/education_social_work/coco/events/research_fest/posters/hussain.pdf
[188] S. Haq, and P. J. B. Jackson, "Multimodal Emotion Recognition," Machine Audition: Principles, Algorithms, and Systems, pp. 398-423.
[189] Rabie, A., Handmann, U.: Fusion of audio-and visual cues for real-life emotional human robot interaction. In: Pattern Recognition, vol. 6835, pp. 346-355 (2011)
[190] Yoshitomi, Y., Kim, S.I., Kawano, T., Kilazoe, T.: Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In: Proceedings of the IEEE Int. Workshop on Robot and Human Interactive Communication, pp. 178-183 (2000)
[191] S. Hoch, F. Althoff, "Bimodal Fusion of Emotional data in an automotive environment," in *Proc. Acoustics, Speech, and Signal Processing*, Philadelphia, PA, 2005, pp. 1085-1088
[192] Kwon, D., Kwak, Y.K., Park, J. C., Chung, M. J., Jee, E., Park, K., Kim, H., Kim, Y., Park, J., Kim, E., Hyun, K.H., Min, H., Lee, H.S., Park, J.W., Jo, S.H., Park, S., Lee, K.: Emotion interaction system for a service robot. In: Proceedings of the IEEE Int. Symp. on

Robot and Human interactive Communication, pp. 351-356 (2007)
[193] Castrillón, M., Déniz, O., Guerra, C., Hernández, M.: ENCARA2: Real-time detection of multiple faces at different resolutions in video streams. J. Visual Communication and Image Representation **18**(2), 130-140 (2007)
[194] Vogt, T., André, E., Bee, N.: EmoVoice—A framework for online recognition of emotions from voice. In: Perception in Multimodal Dialogue Syst., pp. 188-199 (2008)
[195] Battocchi, A., Pianesi, F., Goren-Bar, D.: A first evaluation study of a database of kinetic facial expressions (dafex). In: Proceedings of the Int. Conf. on Multimodal Interfaces, pp. 214-221 (2005)
[196] Breazeal C, Brooks R (2005) Robot emotion: A functional perspective. In: Fellous JM, Arbib MA (ed) Who needs emotions? The brain meets the robot, Oxford: Oxford University Press, pp 271-310.
[197] A. Lim and H. Okuno, "The MEI Robot: Towards Using Motherese to Develop Multimodal Emotional Intelligence." IEEE Trans. Auton. Ment. Dev., vol. 6, no. 2, pp. 126-138, 2014.
[198] H. Yang, Z. Pan, M. Zhang, and C. Ju. "Modeling emotional action for social characters." The Knowledge Engineering Review, vol. 23, no. 4, pp. 321-337, 2008.
[199] R. Kirby, J. Forlizzi and R. Simmons, "Affective social robots," Rob. Auton. Syst., vol. 58, no. 3, 2010, pp. 322-332.
[200] X. Hu, L. Xie, X. Liu, and Z. Wang, "Emotion Expression of Robot with Personality." *Mathematical Problems in Engineering*, vol. 2013, 2013.
[201] L. Xin, X. Lun, W. Zhi-liang, and F. Dong-mei, "Robot Emotion and Performance Regulation Based on HMM." *Int. J. Adv. Robot. Syst.*, vol. 10, no. 160, pp. 1-10, 2013.
[202] Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis. John Wiley & Sons (2004)
[203] Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: A survey. IEEE Trans. Affective Comput. **4**(1), 15-33 (2013)
[204] Park, H., Howard, A.: Providing tablets as collaborative-task workspace for human-robot interaction. In: Proceedings of the ACM/IEEE Int. Conf. on Human-Robot Interaction, pp. 207-208 (2013)
[205] McColl, D., Nejat, G.: Meal-time with a socially assistive robot and older adults at a long-term care facility. J. Human-Robot Interaction, **2**(1), 152-171 (2013)
[206] Ishiguro, H., Ono, T., Imai, M., Maeda, T., Kanda, T., Nakatsu, R.: Robovie: an interactive humanoid robot. Industrial Robot An Int. J. **28**(6), 498–504 (2001)
[207] Hinds, P. J., Roberts, T. L., Jones, H.: Whose Job Is It Anyway? A Study of Human-Robot Interaction in a Collaborative Task. J. Human-Computer Interaction **19**(1), 151-181 (2004)
[208] Längle, T., Wörn, H.: Human–Robot Cooperation Using Multi-Agent-Systems. J. Intell. Robot. Syst. **32**(2), 143-160 (2001)
[209] Ettelt, E., Furtwängler, R.: Design issues of a semi-autonomous robotic assistant for the health care environment. J. Intell. Robot. Syst. **22**(3-4), 191-209 (1998)
[210] Nejat, G., Ficocelli, M.: Can I be of assistance? The intelligence behind an assistive robot. In Proceedings of the IEEE Int. Conf. Robotics and Automation, pp. 3564-3569 (2008)
[211] Breazeal, C., Scassellati, B.: Robots that imitate humans. Trends Cognitive Sci. **6**(11), 481-487 (2002)
[212] Bourgeois, P., Hess, U.: The impact of social context on mimicry. Biological Psychology **77**(3), 343-352 (2008)

[213] Kanade, T., Cohn, J. F., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of the IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 46-53 (2000)
[214] Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression database. IEEE Trans. Pattern Analysis and Machine Intelligence **25**(12), 1615-1618 (2003)
[215] Lyons, M.J., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding Facial Expressions with Gabor Wavelets In: Proceedings of the IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 200-205 (1998)
[216] Burkhardt, F., Paeschke, A., Rolfes, M.: A database of German emotional speech. Interspeech **5**, 1517-1520 (2005)
[217] J. Russell, A. Weiss and G. Mendelsohn, "Affect Grid: A single-item scale of pleasure and arousal." *J. Pers. and Soc. Psychol.*, vol. 57, no. 3, pp. 493-502, 1989.
[218] L. Barrett, "Discrete Emotions or Dimensions? The Role of Valence Focus and Arousal Focus." *Cogn. Emot.*, vol. 12, no. 4, pp. 579-599, 1998.
[219] J. Posner, J. Russell and B. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology." *Develop. Psychopathol.,* vol. 17, no. 3, 2005.
[220] H. Wallbott, "Bodily expression of emotion." *Eur. J. Soc. Psychol.*, vol. 28, no. 6, pp. 879-896, 1998.
[221] M. de Meijer, "The contribution of general features of body movement to the attribution of emotions." *J. Nonverbal Behav.*, vol. 13, no. 4, pp. 247-268, 1989.
[222] D. McColl and G. Nejat, "Determining the Affective Body Language of Older Adults during Socially Assistive HRI," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Chicago, IL*,* 2014, pp. 2633-2638.
[223] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook and R. Moore, "Real-time human pose recognition in parts from single depth images." *Commun. ACM*, vol. 56, no. 1, pp. 116-124, 2013.
[224] A. Liberman, "Apparatus and methods for detecting emotions", US 6638217 B1, October 28, 2003.
[225] *QA5 SDK v. 5.5 Product Description & User Guide*, Nemesysco, Ltd., Netanya, Isreal, 2009.
[226] C. D. Kidd and C. Breazeal, "Robots at home: Understanding long-term human-robot interaction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nice, France, 2008, pp. 3230-3235.
[227] M. Häring, N. Bee and E. André, "Creation and Evaluation of emotion expression with body movement, sound and eye color for humanoid robots," in *Proc. IEEE RO-MAN*, Atlanta, GA, pp. 204-209, 2011.
[228] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, pp. 10-18, 2009.
[229] M. Karg, K. Kuhnlenz and M. Buss, "Recognition of Affect Based on Gait Patterns," *IEEE Trans. Syst., Man, Cybern., Part B: Cybern.*, vol. 40, pp. 1050-1061, 2010.153
[230] Y. Gao, G. Ji, Z. Yang and J. Pan, "A Dynamic AdaBoost Algorithm With Adaptive Changes of Loss Function," *IEEE Trans. Syst., Man, Cybern., C: Appl. and Reviews*, vol. 42, pp. 1828-1841, 2012.

[231] A. B. Acharya, S. Prabhu and M. V. Muddapur, "Odontometric sex assessment from logistic regression analysis," *Int. J. Legal Med.*, vol. 125, pp. 199-204, 2011.
[232] T. M. Mitchell, *Machine Learning*. WCB/McGraw Hill, 1997.
[233] V. Y. Kulkarni and P. K. Sinha, "Random Forest Classifiers: A Survey and Future Research Directions," *Int. J. Advanced Computing*, vol. 36, pp. 1144-1153, 2013.
[234] E. Frank, Y. Wang, S. Inglis, G. Holmes and I. Witten, "Using Model trees for Classification", *Machine Learning*, vol. 32, no. 1, pp. 63-76, 1998.
[235] S. Thrun, "Simultaneous localization and mapping," in *Robotics and Cognitive Approaches to Spatial Mapping*, Springer-Verlag, Berlin, 2008, pp. 13-41.
[236] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proc. Computational Intelligence in Robotics and Automation*, Monterey, CA, 1997, pp. 146–151.
[237] B. Yamauchi, "Frontier-based exploration using multiple robots," in *Proc. Autonomous Agents,* New York, NY, 1998, pp. 47-53.
[238] R. Oppermann, Adaptive user support: ergonomic design of manually and automatically adaptable software. *CRC Press*, 1994.
[239] R. Parasuraman, S. Galster, P. Squire, H. Furukawa and C. Miller, "A Flexible Delegation-Type Interface Enhances System Performance in Human Supervision of Multiple Robots: Empirical Studies With RoboFlag", *IEEE Trans. Cybern.*, vol. 35, no. 4, pp. 481-493, 2005.
[240] A. Clare, M. Cummings, J. How, A. Whitten and O. Toupet, "Operator Object Function Guidance for a Real-Time Unmanned Vehicle Scheduling Algorithm", *Journal of Aerospace Computing, Information, and Communication*, vol. 9, no. 4, pp. 161-173, 2012.
[241] C. Miller and R. Parasuraman, "Designing for Flexible Interaction Between Humans and Automation: Delegation Interfaces for Supervisory Control", *Hum. Factors*, vol. 49, no. 1, pp. 57-75, 2007.
[242] E. de Visser, B. Kidwell, J. Payne, L. Lu, J. Parker, N. Brooks, T. Chabuk, S. Spriggs, A. Freedy, P. Scerri, and R. Parasuraman, "Best of both worlds: design and evaluation of an adaptive delegation interface," in *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, San Diego, CA, 2013, pp. 255-259.
[243] Adept. MobileSim. [Online]. http://robots.mobilerobots.com/wiki/MobileSim.
[244] J. Wang, M. Lewis, and J. Gennari, "A game engine based simulation of the NIST urban search and rescue arenas," in Proc. Simulation Conference, New Orleans, LA, 2003, pp. 1039-1045.
[245] G. Polverari, D. Calisi, A. Farinelli, and D. Nardi. "Development of an autonomous rescue robot within the USARSim 3D virtual environment," in *RoboCup 2006: Robot Soccer World Cup X*, Springer-Verlag, Berlin, 2006, pp. 491-498.
[246] Z. Zhang, M. Littman and X. Chen "Covering Number as a Complexity Measure for POMDP Planning and Learning," in *Proc. AAAI Conf. Artificial Intelligence*, 2012, pp. 1853-1859.
[247] D. Dietterich, "Hierarchical reinforcement learning with MAXQ value function decomposition," *J. Artif. Intell. Res.*, vol. 13, pp. 227-303, 2000.
[248] WKJonesnet, *Xbox Controller Outline* [Online Image], 2016. Available from: http://wkjonesnet.deviantart.com/art/Xbox-Controller-Outline-155064341
[249] B. Doroodgar and G. Nejat, "A hierarchical reinforcement learning based control

architecture for semi-autonomous rescue robots in cluttered environments," in *Proc. IEEE Int. Conf. Autom. Sci. Eng.*, Toronto, Canada, 2010, pp. 948-953.

[250] Sourceforge. USARSim. [Online]. http://sourceforge.net/projects/usarsim/

[251] Epic Games Inc. UDK. [Online]. http://www.unrealengine.com/udk.

[252] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich. "Common metrics for human-robot interaction," in *Proc. ACM SIGCHI/SIGART Conf. on Human-robot Interaction*, New York, NY, 2006, pp. 33-40.

[253] D. R. Olsen and M. A. Goodrich, "Metrics for evaluating human–robot interactions," in *Proc. PERMIS*, Gaithersburg, MA, 2003.

[254] Y. Gatsoulis, G. Virk and A. Dehghani-Sanij, "On the Measurement of Situation Awareness for Effective Human-Robot Interaction in Teleoperated Systems", *J. Cogn. Engng. Dec. Making*, vol. 4, no. 1, pp. 69-98, 2010.

[255] B. Donmez, P. E. Pina, and M. L. Cummings, "Evaluation criteria for human-automation performance metrics." in *Performance Evaluation and Benchmarking of Intelligent Systems*, 2009, ch. 2, pp. 21-40.

[256] R. Parasuraman, T. Sheridan and C. Wickens, "A model for types and levels of human interaction with automation", *IEEE Trans. Cybern.*, vol. 30, no. 3, pp. 286-297, 2000.

[257] Wang, K., An, K., Li, B.N., Zhang, Y., Li, L., 2015, "Speech emotion recognition using Fourier parameters," IEEE Trans. Affect. Comput., **6**(1), pp. 69-75.

[258] Mazzetto de Menezes, K.S., *et al*., 2014, "Differences in acoustic and perceptual parameters of the voice between elderly and young women at habitual and high intensity," Acta Otorrinolaringologica, **65**(2), pp. 76-84.

[259] Tahon, M., Degottex, G., and Devillers, L., 2012, "Usual voice quality features and glottal features for emotional valence detection," *Proceedings of Speech Prosody*, Shanghai, China.

[260] Amir Liberman, 2003, "Apparatus and methods for detecting emotions," U.S. Patent 6638217.

[261] McKeown, G., 2011, "The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent," IEEE Trans. Affect. Comput., **3**(1), pp. 5-17.

[262] Coutinho, E., and Dibben, N., 2013, "Psychoacoustic cues to emotion in speech prosody and music," Cogn. Emot., **27**(4), pp. 658-684.