Estimating Intent for Human-Robot Interaction

D. Kulić E. A. Croft

Department of Mechanical Engineering
University of British Columbia
2324 Main Mall
Vancouver, BC, V6T 1Z4, Canada

Abstract

This work proposes a methodology for estimating human intent during human-robot interaction (HRI). The intent signal is used within a planning and control strategy to improve safety and intuitiveness of the interaction. A method for estimating intent using physiological signals is described and preliminary test results are presented. A companion paper describes the planning and control strategy.

1. Introduction

As robots move from isolated work cells to unstructured and interactive environments, they will need to become better at acquiring and interpreting information about their environment [1]. In particular, in cases where robot-human interaction is planned, human monitoring provides valuable information, which can enhance the safety of the interaction by providing a feedback signal to robot planning and control system actions [2,3,4,5].

Monitoring is important during interperson interaction, where non-verbal cues such as eye-gaze direction, facial expression and gestures are all used as modes of communication [6]. Recently, the use of these modes is being considered in robotics research in order to improve the safety and intuitiveness of the interaction [6,7].

There are two aspects of "intent" that are of interest in terms of human – robot interaction:

(i) Attention/Focus –is the attention of the person focused (attentively occupied) on the robot's actions?

(ii) Expectation/Approval –does the person intend (want, expect) the robot's actions? Is the robot carrying out the intent (purpose) of the person? Does the person approve of the robot's actions?

During human-human interaction, non-verbal communication signals are frequently exchanged in order to assess each participant's emotional state, focus of attention and intent. Many of these signals are indirect; that is, they occur outside of conscious control. By monitoring and interpreting indirect signals during an interaction, significant cues about the intent of each participant can be recognized [6]. Furthermore, by using intent information, the robot can gauge user approval of its performance without requiring the user to continuously issue explicit feedback [2,3,4,5]. In addition, changes in some non-verbal signals precede a verbal signal from the user. Observation of intent information can allow the robot control system to anticipate command changes, creating a more responsive and intuitive human-robot interface.

2. Related Work

Existing robot-human monitoring systems can be classified by the type of monitoring utilized. One category of systems measures the mechanical forces and displacements during a physical interaction with the robot [8,9,10]. Another category of systems is concerned with monitoring communication signals from the human [7]. These systems can be further subdivided into visual monitoring or physiological monitoring systems. Visual monitoring systems capture video data of the human involved in robot-machine interaction and use this data to guide the machine response to the interaction. This can include visual tracking of the user's eye-gaze direction [2,3] and head position, or classifying of facial expressions [11] and hand and body gestures [12,13].

Physiological monitoring systems can also be used to extract information about the user's reaction. Signals proposed for use in human-computer interfaces include skin conductance, heart rate, pupil dilation and brain and muscle neural activity. Some of these interfaces have also been proposed for use in human-robot interaction. Bien et al. [7] advocate that soft computing methods are the most suitable methods for interpreting and classifying these

types of signals, because these methods can deal with imprecise and incomplete data.

Although physiological signals have the potential to provide objective measures of the human's emotional response, they are difficult to interpret. One problem is the large variability in physiological response from person to person. Another problem is that the same physiological signal is often triggered for a range of psychological states. Thus, it can be difficult for a controller to determine which emotional state the subject is in, or whether the response was caused by an action of the system, or by an external stimulus. For these reasons, researchers in psychophysiology recommend using more than one signal source or indicator, for example, both heart rate and galvanic skin response (GSR). However, human–robot interfaces developed thus far have most frequently used only one physiological mechanism, due in part to the difficulty of measuring physiological signals unobtrusively, e.g., [4, 5].

Sarkar proposes using multiple physiological signals to estimate emotional state, and using this estimate to modify robotic actions to make the user more comfortable [14]. In that work, biofeedback data are used exclusively for emotional state estimation, and it is assumed that all changes in physiological data are caused by robot actions, and not by external stimuli.

Outside of the domain of robot-human interaction, user monitoring has been researched extensively in human-computer interaction. Of particular interest is the work on perceptual intelligence. Some of the developed systems include technology to detect and track people in a room, facial expression and gesture recognition, as well as speech recognition that learns its vocabulary from the user [1]. In addition to this holistic approach, many researchers have focused on monitoring a specific expression modality, such as facial expression recognition [15], eye gaze tracking [16,17], physiological response [18], gesture recognition [19] and speech recognition [20,21].

3. Approach

The proposed intent estimation functionality is developed within an overall controller architecture for human–robot interaction. The controller architecture is described in the companion paper [22]. During the entire interaction, the user is monitored to assess the level of safety of the interaction, as well as the level of approval of robot actions. This information is used for planning and safety control decisions. The focus of this paper is on the estimation of intent information and its potential for improving the safety of the interaction.

3.1. Modes of Intent

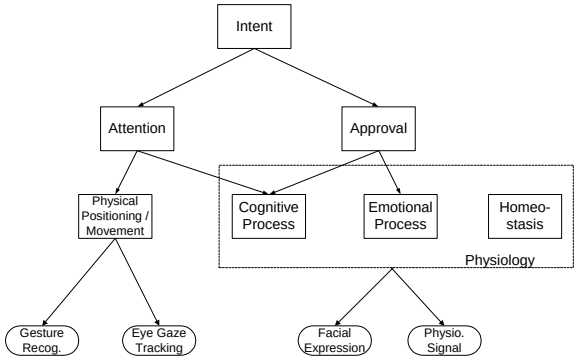

Figure 1 - Intent Signal Decomposition

Figure 1 shows an overview of how indirect intent can be communicated and captured. Intent is separated into two components: attention and approval. For example, the level of user attention can be assessed based on the physical posture of the user in relation to the robot; i.e. whether the user is facing the robot (stance) and looking at the robot (eye-gaze). However, static physical indicators of attention are not always sufficient to determine attention. It is quite possible for a user to face a robot and look directly at it, but be thinking about something else. To determine whether the user's attention is focused on the robot, the system must establish that the user is both physically and cognitively focused on the robot. This can be accomplished through the tracking of physiological signals, or by dynamically tracking the physical configuration of the user in relation to the robot. Approval can be measured using emotional interface technology such as facial expression recognition and physiological signal tracking [15,18]. However, most physiological signals are affected by cognitive and emotional processing, as well as maintaining the equilibrium of bodily functions (homeostasis). The measure of approval from robot actions must therefore be extracted by separating the signal component stemming both from homeostasis activities, as well as from emotional and cognitive processing which is not related to robot activity [23]. Approval estimation based on physiological signals must be validated by

physically based user attention information to ensure that the measured changes in physiology are the results of robot stimuli as observed by the user, and not a result of other external stimuli.

3.2. Intent as a Control Signal

In the proposed controller architecture, planning is divided in two stages. When a robot task is initiated, the planning module generates a safe geometric path region, represented as contiguous regions in space [22]. Dynamic planning is performed by the on-line trajectory planner. The trajectory planner plans a small number of waypoints at a time, so that changing conditions in the environment can be handled by local re-planning within the safe regions. The intent signal is one of the factors that constrains the trajectory planner. The trajectory planner used is described in [24]. If the planned trajectory is interactive (i.e. human – robot interaction is planned), the velocity specification is calculated based on the current estimate of the intent of the user, that is, the estimated level of approval the user expresses towards robot actions. This modulation acts on planned motions to be executed in the near future (1 –2 seconds), and is intended to respond to slow changes of the intent signal.

For sudden changes in the intent signal (for example, the user suddenly turns away), a faster response mechanism is needed. For this reason, the intent signal is also used as an input into the safety module. The safety module is responsible for controlling the danger level of the interaction in real time. This module is responsible for reacting to sudden changes in the environment, not anticipated during the planning stage. The safety module executes at every control step. The inputs into the safety module consist of the proposed next configuration of the robot from the trajectory planner, including the velocity and acceleration information, the current user configuration, and an estimate of the user's intent level [22].

3.3. Estimating Intent

In order to test the feasibility of our approach, a prototype intent estimator module was developed, using physiological signals only. As opposed to other signals such as facial expression, physiological signals are not under direct conscious control of the user and are relatively easy to measure. However, physiological data is highly variable, both between individuals and interaction contexts. In addition, quantitative descriptions for the relationship between physiological measures and emotional categories are not available. However, psychophysiological research [26,27,28,29,30,31] and recent work on emotion recognition [6,18,23] can provide qualitative relationships. In this work, a fuzzy inference engine was deemed suitable as an estimation method for these relationships. Another way of inferring emotional state is by using a model of the user [32]. However, such information may not be readily available for general applications of human-robot interaction.

Output Representation

Two representations are frequently used in emotion and emotion detection research, one using discrete emotion categories (anger, happiness, etc.), and the other using a two-dimensional representation of valence and arousal. Valence measures the degree to which the emotion is positive (i.e. positive or negative), and arousal measures the strength of the emotion. The valence/arousal representation provides less data, but the amount of information retained appears adequate for the purposes of robotic control, and is easier to convert to a measure of user approval. In this paper, the valence/arousal representation is used. This representation system has been favored for use with physiological signals and in psychophysiological research [6,23,26,27].

Measured Variables

The measurement system utilized is the ProComp+ system from Thought Technologies [25]. The signals measured consist of: blood volume pressure, skin conductance, chest cavity expansion/contraction and corrugator muscle activity. The corrugator muscle is located on the forehead, it is used to control brow location; the activity of this muscle has been correlated to emotions such as frustration and anger [26,27,28]. The fuzzy inference engine does not use the measured signals directly; the signals are preprocessed to extract relevant features, which are then used for inference. Because the magnitudes of physiological signals vary widely between individuals, another important function of the preprocessing is to normalize the signal features so that a single inference engine can be used across individuals.

For the skin conductance sensor, the magnitude and the

rate of change of the signal are used as inference engine inputs. The skin conductance data is smoothed with a 1s averaging window prior to feature extraction. Both features are normalized for each subject, using the maximum and minimum data points recorded throughout the session. For the corrugator muscle electromyography (EMG) sensor, the average signal power over a 1s interval is extracted. The signal is normalized using baseline (pre-test, unstimulated) data extracted from each user during system calibration. The blood volume pressure data is used to calculate the heart rate and vasomotor activity. Prior to feature extraction, the blood volume pressure signal is low-pass filtered and smoothed over a 3s window. Both the heart rate and vasomotor activity are normalized using baseline data. The chest cavity sensor data is used to extract respiration rate and respiration depth. Prior to feature extraction, the respiration sensor data is low-pass filtered and smoothed over a 10s window.

The Inference Engine

The rulebase for the fuzzy inference engine was derived using data from psychophysiological research [26,27,28, 29,30,31]. Physiological responses can be highly variable between individuals, as well as vary for the same individual depending on the context of the response. Therefore, the rulebase was structured such that reliable outputs would be obtained even if a subject did not exibit all of the responses characterized by existing research. For this reason, each input was handled with separate rules, rather than combining indices. Five sets of rules were developed. The first set of rules encapsulates the relationship between the skin conductance response (SCR) and arousal, which are found to be linearly correlated for a majority of subjects [27,28]. The second set of rules describes the relationship between corrugator muscle EMG (i.e. frowning) and valence. The third set of rules relates heart activity to valence and arousal. Unlike the SCR and EMG muscle activity, the activity of the heart is governed by many variables, including physical fitness, posture, activity level in addition to emotional state [29]. It is more difficult to obtain significant correlation between heart activity and emotional state. In addition, heart rate activity is also dependent on context. In tests using external stimuli to generate the emotional response (such as picture viewing), heart rate response is initially decelerative, while tests using internal stimulus (recalling emotional imagery) showed an accelerative response [28]. Since our initial experiments utilized external stimuli (Section 3.4), and the intended application of this module will also be for external stimuli (robot actions), the external stimuli results were used. Using these results, heart rate deceleration is associated with the orienting response (i.e. increased arousal). Heart rate at the baseline, with no heart rate acceleration or deceleration is associated with low arousal, while high heart rate is associated with high arousal. The fourth set of rules relates vasomotor activity to arousal. Vasoconstriction occurs when the body is in the fight-or-flight response, which corresponds to highly negative valence. Vasodilation, on the other hand, occurs during relaxation or positive emotional responses [29]. However, the experimental setup used in this project (Section 3.4) was not able to elicit this response from any of the subjects, probably because the emotional stimuli presented were not powerful enough to elicit the fight-or-flight response. For this reason, rules related to vasomotor activity were not included in the preliminary rule base. The fifth set of rules relates respiratory activity to emotional state. Similar to heart activity, the regulation of respiratory activity is a highly complex process, which varies as a function of heart rate, physical fitness, posture, activity level in addition to emotional state [31]. One significant problem with respiratory activity features is its typically slow rate, delaying data acquisition. A 10 second delay was required before a reliable signal becomes available from this modality. In view of the real time requirements of the application, a 10 second response time is too long to be useful when estimating emotional state in real time. For this reason, respiratory activity rules were not used, and do not contribute to the results discussed in Section 4.

3.4. Data Collection Process

In order to generate preliminary data for testing the developed rulebase, a test procedure was developed using the picture-based phychophysiological procedures by Lang et al. [27,28]. The test uses pictures of the human environment as stimuli to arouse an emotional response. The measurement system is connected to a subject. The subject then sits still for 10 seconds, while baseline physiological data is collected. After the baseline period, an emotionally arousing image is displayed on the screen. After 10 seconds of viewing the image, the subject is asked

to rate the emotional content of the image, using the valence and arousal scales. Once the user has finished evaluating the image, the screen is cleared for 10 seconds, to allow the subject to return to the baseline state. This process is repeated for 10 different images. The images used contained both “positive” and “negative” images. To evaluate the potential of this approach, four subjects were tested using the procedure.

4. Results

Table 1 summarizes the results for the four subjects tested. Arousal was considered successfully detected if an increase in arousal was observed within 3 seconds of stimulus onset. To compare the detected arousal with the subject reported arousal, the maximum value observed within the first 5 seconds of stimulus onset was used.

Table 1 - Arousal Estimation Results

	Arousal Detected [%]	Arousal Level Error [%]
Subject 1	100	25
Subject 2	100	25
Subject 3	75	33
Subject 4	100	22
AVERAGE	94	26

Because the valence data is based on one sensor input only (the EMG signal), valence estimation was less successful than arousal estimation. Table 2 summarizes the results for each subject. Results for Subject 3 were discarded as the EMG sensor became partially detached during the test. A valence reading was considered detected if a change in valence was detected within 5 seconds of stimulus onset. A valence direction was considered to be correct if both the reported and the detected valence were of the same sign (i.e. both are positive or both are negative). To compare the detected valence with the subject reported valence, the maximum value observed within the first 5 seconds of stimulus onset was used. A change in valence was detected 80% of the time. When valence was detected, the correct direction was detected 75% of the time.

Even with only the qualitative estimates available, the information generated by the estimator will be useful for improving the safety of human-robot interaction. One particular advantage of the estimator is the viability of real-time implementation, which makes it suitable for human-robot safety applications.

Table 2 - Valence Estimation Results

	Valence Detected [%]	Correct Valence Direction Detected [%]	Valence Level Error [%]
Subject 1	70	71	38
Subject 2	80	88	52
Subject 4	90	67	18
AVERAGE	80	75	36

Future work includes expanding the testing procedures to include more powerful stimuli, necessary to elicit the fight-or-flight response, as well as stimuli more related to robot interaction tasks. A more structured procedure for obtaining the baseline response is also required, and of course, a larger group of test subjects recruited. Finally, the estimator needs to be integrated with the other components of the user intent estimation module, including emotion estimation from other modalities, such as facial expression, as well as estimates of the focus of user attention. Further results of this work will be made available at the time of the conference presentation.

Acknowledgements

This work is supported by the National Scence and Engineering Research Council of Canada. The authors wish to acknowledge the technical assistance of Prof. Christina Conati and her lab group.

References

[1] A. Pentland. Perceptual Intelligence. Communications of the ACM, Vol. 43:3, pp. 35 – 44, March 2000.

[2] V. J. Traver et al. Making Service Robots Human-Safe. IROS 2000, pp. 696 – 701.

[3] Y. Matsumoto et al. The Essential Components of Human – Friendly Robot Systems. Int. Conf. on Field and Service Robotics, pp. 43 – 51, 1999.

[4] Y. Takahashi et al. Human Interface Using PC Display With Head Pointing Device for Eating Assist Robot. ICRA 2001, pp. 3674 – 3679.

[5] Y. Yamada et al. Proposal of a Psychophysiological Experiment System Applying the Reaction of Human Pupillary Dilation to Frightening Robot Motions. IEEE Int. Conf. on Systems, Man and Cybernetics, Vol. 2, pp. 1052 – 1057, 1999.

[6] R. Picard. Affective Computing. MIT Press.

Cambridge, Masssachusetts, 1997.
[7] Z. Z. Bien et al. Soft Computing Based Emotion / Intention Reading for Service Robot. Lecture Notes in Computer Science, Vol. 2275, pp. 121 – 128, 2002.
[8] Y. Maeda et al. Human-Robot Cooperation with Mechanical Interaction Based on Rhythm Entrainment. ICRA 2001, pp. 3477 – 3482, 2001.
[9] V. Fernandez et al. Active Human-Mobile Manipulator Cooperation Through Intention Recognition. ICAR 2001, pp 2668 – 2673, 2001.
[10] Y. Yamada et al. Construction of a Human/Robot Coexistence System Based on A Model of Human Will – Intention and Desire. ICRA 1999, pp. 2861 – 2867.
[11] W. K. Song et al. Visual Servoing for a User's Mouth with Effective Intention Reading in a Wheelchair-based Robotic Arm. ICRA 2001, pp. 3662 – 3667.
[12] T. Yamaguchi and N. Ando. Intelligent Robot System Using "Model of Knowledge, Emotion and Intention" and "Information Sharing Architecture". Conf. of the IEEE Industrial Electronics Society, pp. 2121 – 2125, 2001.
[13] M. A. T. Ho et al. An HMM-based Temporal Difference Learning with Model-Updating Capability for Visual Tracking of Human Communicational Behaviours. IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 163 – 168, 2002.
[14] N. Sarkar. Psychophysiological Control Architecture for Human-Robot Coordination – Concepts and Initial Experiments. ICRA 2002.
[15] M. Pantic & L. J. M. Rothkrantz. Automatic Analysis of Facial Expressions: The State of the Art. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22(12): 1424 – 1445, 2000.
[16] J. G. Wang & E. Sung. Study on Eye Gaze Estimation. IEEE Trans. on Systems, Man, and Cybernetics – Part B: Cybernetics, Vol. 32(3): 332 – 350, 2002.
[17] T. Ohno et al. FreeGaze: A Gaze Tracking System for Everyday Gaze Interaction. Eye Tracking Research and Applications Symp. 2002, pp. 125 – 132.
[18] R. W. Picard et al. Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 23(10): 1175 – 1191, 2001.
[19] V. I. Pavlovic et al. Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.19(7): 677 – 695, 1997.
[20] M. F. McTear. Spoken Dialogue Technology: Enabling the Conversational User Interface. ACM Computing Surveys, Vol. 34 (1): 90 – 169, 2002.
[21] B. H. Juang & S. Furui. Automatic Recognition and Understanding of Spoken Language – A First Step Toward Natural Human – Machine Communication. Proc. of the IEEE, Vol. 88 (8): 1142 – 1165, 2000.
[22] D. Kulic and E. A. Croft. Strategies for Safety in Human-Robot Interaction. ICAR 2003.
[23] E. Hudlicka and M. D. McNeese. Assessment of User Affective and Belief States for Interface Adaptation: Application to an Air Force Pilot Task. User Modeling and User-Adapted Interaction, Vol. 12: 1 – 47, 2002.
[24] S. Macfarlane and E. A. Croft. Jerk-Bounded Robot Trajectory Planning - Design for Real-Time Applications. IEEE Trans. on Robotics and Automation, in press for Feb 2003 issue.
[25] http://www.thoughttechnology.com
[26] J. T. Cacioppo and L. G. Tassinary. Inferring Psychological Significance From Physiological Signals. American Psychologist. Vol. 45 (1): 16 – 28, 1990.
[27] P. J. Lang. The Emotion Probe: Studies of Motivation and Attention. American Psychologist. Vol. 50 (5): 372 – 385, 1995.
[28] M. M. Bradley and P. J. Lang. Measuring Emotion: Behavior, Feeling and Physiology. In R. D. Lane and L. Nadel (Eds.). Cognitive Neuroscience of Emotion, Oxford University Press, New York, 2000.
[29] K. A. Brownley et al. Cardiovascular Psychophysiology. In J. T. Cacioppo, L. G. Tassinary and G. G. Berntson (Eds.). Handbook of Psychophysiology. Cambridge University Press, Cambridge, 2000.
[30] M. E. Dawson et al. The Electrodermal System. In J. T. Cacioppo, L. G. Tassinary and G. G. Berntson (Eds.). Handbook of Psychophysiology. Cambridge University Press, Cambridge, 2000.
[31] A. Harver and T. S. Lorig. Respiration. In J. T. Cacioppo, L. G. Tassinary and G. G. Berntson (Eds.). Handbook of Psychophysiology. Cambridge University Press, Cambridge, 2000.
[32] A. Ortony et al. The Cognitive Structure of Emotions. Cambridge University Press, 1998.