Towards Bi-Directional Affective Human-Machine Interaction

Andrea Kleinsmith, Tsuyoshi Fushimi, Hayato Takenaka, Nadia Bianchi-Berthouze
Database Systems Laboratory
Aizu University
Aizu Wakamatsu 965-8580
Japan
voice: [+81](242)37-2790; fax: [+81](242)37-2753
e·mail: {m5061202, m5071209, m5071207, nadia}@u-aizu.ac.jp

Abstract

As computers take on an ever-increasing importance in the function of society, effective incorporation as empathetic systems working with humans is important. This paper describes our experiments to understand the importance of body language in natural communication to achieve our goal of identifying the relevant features of body language necessary for associating emotional states with given body postures. We developed a computational model that uses incremental learning to recognize emotional states from body postures. Currently, we are expanding upon these models as we move from static body posture to dynamic motion. We now attempt to close the loop on bi-directional affective interactions between humans and machines by developing a system that can effectively recognize the user's emotional state and respond with an appropriate action.

Keywords: Human-Machine Interaction, Affective Body Language, Incremental Learning, Categorization, Non-verbal Communication

1 Introduction

As computers take on an ever-increasing importance in the function of society, knowing how to incorporate them more effectively and as empathetic systems able to work well with humans is important. This field of research is known as affective computing [Pic97], and refers to giving computers the human-like ability to recognize emotional cues and use them in interactions.

Human interactions consist of multiple channels of communication. Looking at a combination of these modes helps us to interpret messages and react accordingly to another person. For example, a person may state that nothing is wrong and they are perfectly happy while standing with their arms folded across their chest, a furrowed brow, and a frown on their face. In this case, their actions speak louder than their words, and their body language is probably more correct than their speech. [VSS02] claims that 93% of communication is nonverbal with 55% of it expressed through body language, and only 38% attributed to tone of voice, therefore providing evidence that the study of body language is important.

Gesture is purported to be an integral link between speech abilities and concept forming and understanding capabilities of humans according to psycholinguistic studies [McN92]. Use of gesture allows humans to express what is not always expressible through the modality of speech. According to [Ken83], the use of non-verbal communication can be separate from, and in theory, just as or more effective than speech. Socio-emotional content, described by [LDL01] as face-to-face interactions, facial expression, posture, and gesture is reported by [MCMO79] as more meaningful than verbal content, especially in demonstrating changes in emotional state. Recent neurological studies indicate that, in general, emotions are necessary in decision-making, problem solving, cognition, and intelligence [Dam94].

Most of the work in this field thus far has focused on the empathetic nature of humans in order to make them want to interact with the robot. This can be seen in Sony's AIBO [AIB03]; the Robota doll by [BDH98]; and a seal robot, PARO

[STT98]. All of these projects are similar in that all behaviors exhibited by the robot systems are codified to represent some emotional state using stereotypes, instead of learning through interaction, meaning that they are incapable of gaining "true" intelligence. However, these systems are similar to our research in that they also investigate changes within the system, and their effects on the human user's behavior. The difference is that we propose an adaptive model of the mechanisms involved instead of simply relying on descriptive evidence. Our attempt is for the system to establish an affective interaction with the human by broadening the communication channels.

To our knowledge, current models of affective body motions are not learned through natural interaction. In studies by Picard [Pic97], even if the computer learns to sense (through heart beat, galvanic skin response, etc.) the user, and then adapts its reaction models through interaction, it is not using the visual cues upon which humans rely. To be perceived by the computer, people would always have to wear sensors, which does not promote natural communication. Although these methods do prove to be useful at times (e.g., hospitals), they are difficult to reproduce on a daily basis, or in natural settings. Therefore, we focus on using natural channels of communication.

This paper describes experiments we carried out as a first step to understand the role that body language plays in natural communication. Our goal is to identify the relevant features of body language that cause human observers to correlate a specific emotional state with a given body posture. From these results we developed a computational model for recognizing emotional states from conveyed affective postures through incremental learning. The learning is performed by a categorization process in which each category is learned incrementally instead of being predefined.

The remainder of this paper is organized as follows: first we report on our experiment to identify important features of body language. Section 3 explains the categorization process and an experiment to study the relation between posture and emotion. We end with a discussion of our continuing work in section 4.

2 Body Posture Features and Their Affective Messages

Studies on dancing motions [vL88] [CHST99] have shown that a movement conveys a different affective message when its features, such as spatial dimensions, are modified. Other studies, such as [Nak02], have shown that the same motion was accepted with different degrees of naturalness if its features, such as speed, were changed. These studies have been performed on non-anthropomorphic bodies, demonstrating the extent of humans' ability to empathize to non-human systems.

In this study, we carried out preliminary experiments to understand how humans recognize affective postures. Specifically, we aimed at identifying the features of body language that lead an observer to associate an emotional state to a given posture or motion. To this end, using the H-Anim specification [Spe01], we created two avatars with the same human-like body but with movements having different degrees of freedom. While the first avatar could perform human-like motions, the second had restricted abilities. We modeled the second avatar by implementing the same number of degrees of freedom as in the robot used in the project. This avatar could not bend or cross its arms, nor could it incline its head. Facial expressions were not used so as to focus only on body motion. The two avatars could be animated by using key-framing and inverse kinematics.

Three male and four female Japanese students were selected to evaluate the posture of both avatars. We selected five emotional states: happy, sad, angry, tantrum, and scared. For each emotion, we prepared three variations of the same animation by modifying the following features: speed, symmetry and amplitude. The same animations were reproduced with the robot-like avatar by reducing the number of degrees of freedom according to its motion abilities. In a questionnaire, the subjects were asked to:

- describe the affective state conveyed by each animation.
- identify a context or a situation that could have caused the affective state.
- evaluate the intelligibility of the affective state on a 5 degree scale (from very easy (1) to very difficult (5) to understand).

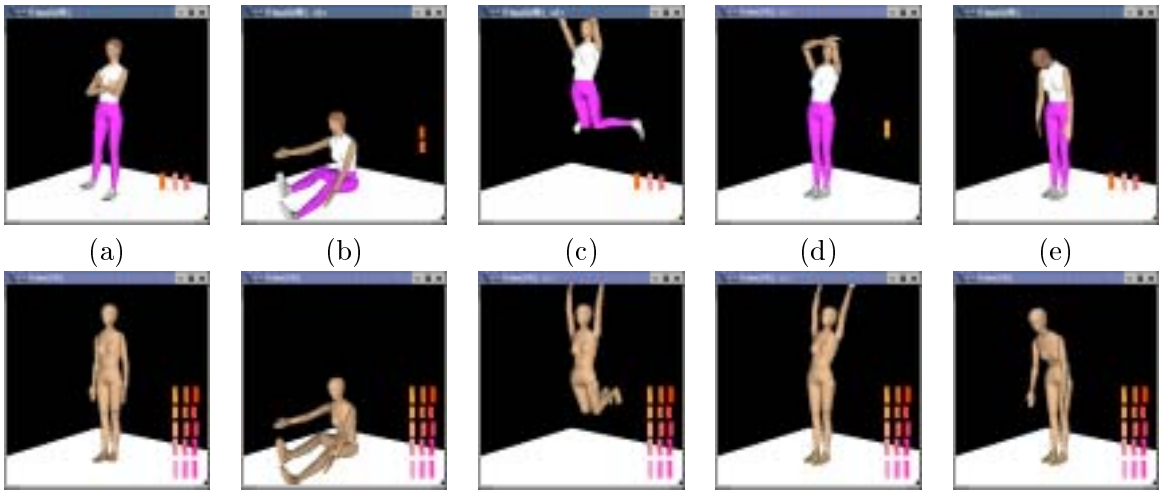

Figure 1: Emotional animations: the first row corresponds to human-like motions, the second row corresponds to robot-like motions. (a) *irritated-angry*, (b) *tantrum*, (c) *happy*, (d) *scared*, (e) *sad*.

- identify the most relevant features of each animation with respect to the conveyed affective state.

2.1 Results and Discussion

The *irritated-angry* animation (Figure 1-(a)) consisted of an avatar with arms crossed and head tilted slightly up, repeatedly tapping its foot on the ground. The set of *irritated-angry* animations differed only in the speed of tapping. Even if the animations could be recognized by the position assumed by the upper body, some subjects argued that the avatar was in a more pleasant state. Their explanations were mainly related to the rhythmic repetition of the foot tapping. The impossibility of manipulating the upper body of the robot-like avatar made it difficult for the animations to be recognized as such by the subjects.

The *tantrum* animation (Figure 1-(b)) was represented by the avatar laying down on the ground with the body slightly bent upwards, and showing repetitive arm and leg movements. The *happy* animation (Figure 1-(c)) was represented by jumps with arms moved up and down. Variations of both animations mainly involved their symmetry feature. In these animations, subjects showed a very strong rejection of symmetry and regularity of movements. Such characteristics were defined as unnatural. Even when the symmetry was removed, the regular repetition of the same movement was seen as awkward and considered more similar to a gymnastic exercise.

The *scared* animation (Figure 1-(d)) was represented by the avatar covering its face with its arms and bending the upper body slightly backward. In the robot-like avatar, the crossing of the arms was not possible. However, in both cases, the motion was not easily recognized because of the low speed and because the body was only slightly bent backwards (limited amplitude of the motion). The rigidity of the body in the second avatar was however not perceived as unnatural.

In the *sad* animation (Figure 1-(e)), the upper body of the avatar bows slowly. In the human-like avatar, the head was bent down as well, while that was not possible in the robot-like avatar. The motion was carried out slowly in all three variations but with varied amplitudes. While the affective state was easily recognized, the inclination (or amplitude) of the bow was the key point for associating the motion to a given affective state. Animations with high degrees of inclination were discarded as resembling a gymnastic exercise rather than expressing a natural affective gesture. Another important feature was the rigidity of the robot-like avatar's arms, hands and head: this appeared to explain why the robot-like avatar was not seen as natural.

Table 1 summarizes the observations made by the subjects on each animation. The first column lists the labels used by the subjects when classifying the affective state conveyed by each animation. From the user's observations, we identified as rel-

Labels	C	I	S	Sp	A	R
Angry	↗	↗	↘	↗	–	–
Tantrum	↗	↗	↘	–	–	↘
Happy	↗	↗	↘	–	↗	↘
Scared	–	–	–	↗	↗	–
Sad	↘	↘	–	↘	↘	↘

Table 1: This table summarizes the subjects' observations of each posture feature as either (↗): necessary, (–): irrelevant, (↘): not desired.

evant the following motion features: Complexity (C), Irregularity (I), Symmetry (S), Speed (Sp), Amplitude (A), Rigidity (R). The arrows indicate if the features were, on average, considered by the subjects as necessary (↗), irrelevant (–) or not desired (↘). The complexity feature refers to the number of body parts that were involved in the motion. It was observed that, in some cases, the simplification of the body movement was making the recognition of the affective state more difficult. The speed feature was pointed out as important in the case of the *angry* and *sad* animations. In the first case, a slow tapping of the foot was perceived as a rhythmical accompanying movement. The irregularity feature refers to arrhythmic motions (temporal randomness). Symmetry and irregularity were shown to be very important features both for the naturalness of the gesture and for effectively conveying an affective state. Rigidity not only resulted in the motion being effectively seen as unnatural but also decreased the intelligibleness of the affective state.

We computed the average intelligibility values associated by each subject to each animation of the first avatar. These values were: *angry* 1.43, *tantrum* 2.69, *happy* 1.86, *scared* 2.57 and *sad* 2.00. The *angry*, *happy*, and *sad* animations showed to be, on average, easy to recognize, while *scared* and *tantrum* animations showed to be more ambiguous. Instead, motions of the robot-like avatar were generally considered awkward. One explanation of this phenomenon is that of expectation. Because the robot-like avatar had a human-like body, subjects expected human-like motions. We believe that motions would have been better accepted if the avatar body looked more robot-like.

3 Categorization of Affective Posture

To facilitate the study of the relation between posture and emotion, we developed a computational model that maps an affective posture into an emotional label. A description of an affective posture of a human from either a video camera, or motion capture, is the input to the network. This can be seen as a categorization problem. We propose to use an association network, called Categorizing and Learning module (CALM) [MPW92], that can self-organize inputs into categories. A CALM network consists of several CALM modules. While the topology of a CALM module is fixed, connections between modules are learned over time. CALM incorporates structural and functional constraints related to the modular structure of both the human brain and its information processing, such as modularity and organization with excitatory and inhibitory connections.

The idea of modular architecture leads to an increased stability of representations and a reduced interference by subsequent learning because of the reduced plasticity. As shown in Figure 2-(a), each CALM module consists of four types of nodes: representation nodes (R-nodes), veto-nodes (V-nodes), an arousal node (A-node), and an external node (E-node). The number of R-nodes and V-nodes equals the number of inputs to the module. Information is sent to the R-nodes from other modules and the R-nodes activate the V-nodes and the A-node. The V-nodes receive activations from all R-nodes and inhibit other V-nodes, all R-nodes, and the A-node to trigger competition inside a module. These processes enable a system to autonomously categorize new patterns and hence, to implement unsupervised, competitive learning. CALM has an attentional control mechanism that is sensitive to the novelty of the input pattern. For more details, please refer to [MPW92] [BT02].

Our computational model consists of a CALM network whose topology is described in Figure 2-(b). It has 1 input module, 3 internal modules, 1 output module, and 1 feedback module. The input module has 16 nodes to which the posture description vector (a set of angles describing the relationships between body segments) is fed. The internal modules are of different sizes. The output module corresponds to the set of emotion cat-

egories that can be recognized, and initially has only 2 nodes but is incrementally grown as new categories are introduced. The number of modules has been decided experimentally. We use a modified version of CALM proposed by [BT02]. It includes a feedback module to supervise the learning process. With supervision, the winning nodes of the output module do not receive random pulses, but instead, the desired winner receives a positive pulse from the competitive mechanism, while the other nodes receive negative pulses.

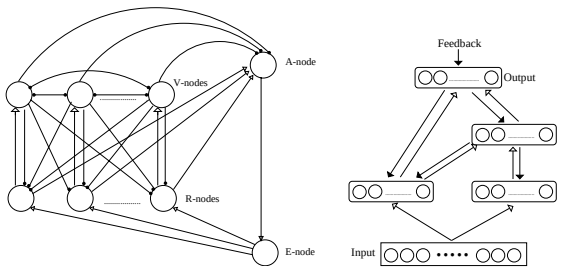

Figure 2: (a) A CALM module consists of four types of nodes: representation nodes (R-nodes), veto-nodes (V-nodes), an arousal node (A-node), and an external-node (E-node). (b) The architecture of the CALM network (consisting of many CALM modules) with a feedback module.

The learning process uses a short term memory of 10 patterns corresponding to the last 10 posture descriptions learned. Every time the model learns a new posture-emotion relation, it also rehearses the patterns in the short term memory. The new pattern is then added to the short term memory and the oldest pattern is removed. In other words, the mapping system incrementally learns new pattern-emotion relations and at the same time reinforces its closest past experiences without completely forgetting the oldest ones.

We tested the computational model with 108 images involving affective postures. The set included 36 different postures for each type of affective state, i.e. *happy, angry* and *sad* postures. The 108 images were split into two sets: 54 were used for the training phase, and the remaining 54 were used for the testing phase. 54 images were randomly fed to the computational model.

We tested the trained CALM network on the 54 posture descriptions of the testing set. Only 1 error occurred: the posture was related to a *sad* emotion but was classified as *angry* by the model.

4 Discussion: Closing the loop with the user

The goal of this research is to propose a shift in typical human-machine interaction towards more human-human-like (natural) interaction. While other research focuses on humans' ability to recognize gestures made by computer systems, we aim towards developing computer systems that are capable of recognizing, understanding and reacting to the gestures of their human partners based on models of empathy and affection. Therefore, the preliminary experiments discussed in this paper describe a first step in the recognition of affective body language. We began with static posture and now move toward dynamic motion in an effort to achieve our ultimate goal of creating a system capable of bi-directional affective communication.

Interactive software currently lacks the ability to handle affective feedback from the user even though it carries important information to be used by the software to direct the selection of actions, as well as for the software's adaptability itself. An example is e-learning software, in which the application, if personalizable, adapts its interaction with the students according to the number of errors the students make and the answering time. The number of errors is not always a good measure of learning: for example, if the student is bored, her attention level decreases and hence her performance.

Integrating the affective gesture recognition (AGM) within a software application could broaden the type of feedback it can handle from its users and hence help select more appropriate actions to perform. Consider an e-learning application endowed with our AGM. A student sits in front of the computer with a camera attached to it that captures her body postures and motions, and sends the information to the AGM. The AGM then maps the body language into an emotional label and sends this information to the software application to determine which level of questions or which topic would be most useful to the user's learning process based on her emotional state at that time. The AGM acts also as a confirmation feedback source. The software application is continuously receiving information about the user's affective state, and based upon transitions of this state, can decide whether or not the correct action

was taken.

While the AGM acts as a source of feedback for the software application, the software application itself acts as a feedback source for the AGM so that it incrementally learns and adapts the affective-gesture state mapping to the user's body language. In fact, while the AGM continuously captures variations in the user's gesture, the evaluation of the correct gesture interpretation is determined only by the software application, as only it can assess if the obtained mood transition was the predicted one, and if the wrong mood transition was due to a wrong initial affective state.

The AGM module now closes the loop with the user as it is not only able to recognize the user's affective state, but also to trigger actions based on this state, as well as to adapt according to a contextualized transition in the user's mood.

References

[AIB03] AIBO, *Entertainment robot*, http://www.aibo.com/, 2003.

[BDH98] A. Billard, K. Dautenhahn, and G. Hayes, *Experiments on human-robot communication with robota, an imitative learning and communication doll robot*, Tech report CPM-98-38, Centre for Policy Modelling, Manchester Metropolitan University, Zurich, 1998.

[BT02] L. Berthouze and A. Tijsseling, *Acquiring ontological categories through interaction*, Proc. of the Fifth Int'l Conf. Human and Computer, 2002, pp. 160–166.

[CHST99] A. Camurri, S. Hashimoto, K. Suzuki, and R. Trocca, *Kansei analysis of dance performance*, IEEE Int'l Conf. on Systems, Man and Cybernetics, 1999, pp. 327–332.

[Dam94] A.R. Damasio, *Descarte's error: Emotion, reason and the human brain*, Avon Books, New York, 1994.

[Ken83] A. Kendon, *Gesture and speech: How they interact, in nonverbal interaction*, Sage Publications, 1983.

[LDL01] C.L. Lisetti, M. Douglas, and C. LeRouge, *Intelligent affective interfaces: A user-modeling approach for telemedecine*, Proc. of the Int'l Conf. on Universal Access in HCI (UAHCI), Elsevier Science Publishers B.V., 2001.

[MCMO79] D. Morris, P. Collett, P. Marsh, and M. O'Shaughnessy, *Gestures, their origin and distribution*, Jonathan Cape Ltd, London, 1979.

[McN92] D. McNeill, *Hand and mind*, The University of Chicago Press, 1992.

[MPW92] J.M.J. Murre, R.H. Phaf, and G. Wolters, *Calm: Categorizing and learning module*, Neural Networks **5** (1992), 55–82.

[Nak02] T. Nakata, *Generation of whole-body expressive movement based on somatical theories*, Proc. of the 2nd Int'l Workshop on Epigenetic Robotics, 2002, pp. 105–114.

[Pic97] R. Picard, *Affective computing*, MIT Press, Cambridge, 1997.

[Spe01] H-Anim Specification, http://www.h-anim.org/, 2001.

[STT98] T. Shibata, T. Tashima, and K. Tanie, *Emergence of emotional behavior through physical interaction between human and pet robot*, Proc. of the Int'l Workshop on Humanoid and Human Friendly Robotics, 1998, pp. 1–6.

[vL88] R. von Laban, *The mastery of movement*, Princeton, 1988.

[VSS02] V. Vinayagamoorthy, M. Slater, and A. Steed, *Emotional personification of humanoids in immersive virtual environments*, Tech report Equator-02-029, Dept. Computer Science, Univ. College London, Sept 2002.