An Incremental and Interactive Affective Posture Recognition System

Andrea Kleinsmith
University of Aizu
Tsuruga Ikki Machi
Aizu Wakamatsu, Japan
andi@andisplanet.com

Tsuyoshi Fushimi
University of Aizu
Tsuruga Ikki Machi
Aizu Wakamatsu, Japan
m5071209@u-aizu.ac.jp

Nadia Bianchi-Berthouze
University of Aizu
Tsuruga Ikki Machi
Aizu Wakamatsu, Japan
nadia@u-aizu.ac.jp

ABSTRACT

The role of body posture in affect recognition, and the importance of emotion in the development and support of intelligent and social behavior have been accepted and researched within several fields. While posture is considered important, much research has focused on extracting emotion information from dance sequences. Instead, our focus is on creating an affective posture recognition system that incrementally learns to recognize and react to people's affective behaviors. In this paper, we examine a set of requirements for creating this system, and our proposed solutions. The first requirement is that the system is general and non-situation specific. Secondly, it should be able to handle explicit and implicit feedback. Finally, it must be able to incrementally learn the emotion categories without predefining them. We tested and compared the performance of our system using 182 standing postures described as a combination of form features and motion flow features, across several emotion categories, with a typical algorithm used for recognition, back-propagation, and with human observers in an aim to show the generalizability of the system. This initial testing showed positive results.

Keywords

Emotion recognition, affective posture, incremental lexicon, incremental learning, explicit feedback, implicit feedback

1. INTRODUCTION

According to Mehrabian and Friar [15], changes in a person's affective state (used as a general term for discussing mood, emotion, and feeling) are reflected by changes in her/his posture. The role of body posture in affect recognition, and the importance of emotion in the development and support of intelligent and social behavior have been accepted and researched within several fields including psychology, neurology, and biology.

In psychology, while there has been much research on understanding the importance of affective posture, there has been little research in the area of computer science to quantitatively model affective posture [4] [3]. In fact, as of yet there are no formal models for classifying affective whole body postures from low-level general features, as there are for classifying affective facial expressions (i.e., Facial Action Coding System (FACS) [6]).

Instead, computer scientists have mainly focused on endowing systems with the ability either to express affective behavior (e.g., Sony's Aibo [19]), or to use physiological methods (i.e., galvanic skin response, blood pressure, heart rate, etc.), e.g., Toyota's Pod car [17], to recognize affect. Moreover, within the field of affective computing, giving systems the ability to convey emotion through posture has progressed rapidly, while endowing systems with the ability to recognize the affective gestures of its user in varying forms such as body postures and motions, is quite original. While posture is considered important, much research has focused on extracting emotion information from dance sequences [21] [10]. Other affective research has concentrated on using information from facial and vocal expressions [18] [14].

Our focus in this paper is to present a discussion on the necessary requirements for creating an affectively aware, interactive system, along with our proposed method for satisfying each requirement. The complete architecture of our incremental affective posture recognition system is composed of several modules. In this paper we present the recognition part of the system, seen in Figure 1, which is composed of 3 modules: i) the posture description module; ii) the recognition module; and iii) the feedback module.

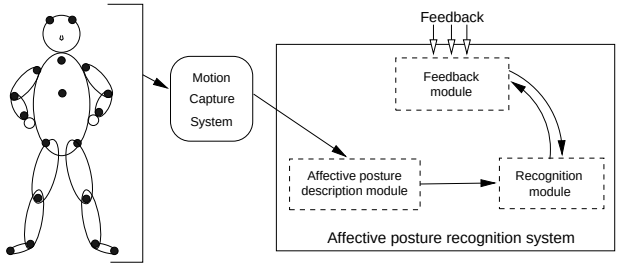

Figure 1: This figure shows our affective posture recognition system. Motion capture information of a human is input to the posture description module, which computes the 25 features. The computed vector is then sent to the recognition module which categorizes the posture. The output is an affective label which is sent to the feedback module. Feedback is used to trigger the adaptation process of the recognition module.

The remainder of the paper is organized as follows: We identify context generalization as the first requirement. At a basic level, to create a general, non situation-specific posture recognition system, the postures need to be described according to low-level posture features. Section 3 exam-

ines explicit and implicit feedback, and incremental learning. Section 4 evaluates the testing of our system's performance. The final requirement, discussed in Section 5, is the need for an incremental lexicon. The system must be able to self-determine the number of categories to learn, thus eliminating a need for predefinition.

2. REQUIREMENT 1: CONTEXT GENERALIZATION

Research by Kapoor et al.[11] attempts to recognize a child's level of interest according to 3 categories (high interest, low interest, and "refreshing" (a short break)) from postures detected through the implementation of a chair embedded with pressure sensors while the child uses a computer to solve a puzzle. Their postures are defined by a set of 8 high-level (coarse-grained) posture features (i.e., leaning forward, slumping back, sitting on the edge), dependent on a computer task situation and the set of interest-level categories. While the work is interesting and recognition rates are positive, the generalization of their method to other emotions and contexts may be limited, since new high level postural descriptors may need to be added. In order to generalize the recognition procedure, we need a posture description framework that allows for the emergence of high-level postural features instead of defining it from the beginning.

A recent psychological study by Coulson [4] attempts to ground emotions into low-level static posture features. He uses computer generated avatars expressing 6 emotions (*angry, fear, happy, sad, surprise,* and *disgust* to examine the parameters necessary for attributing a specific emotional state to body posture. His proposed body description comprises 6 joint rotations (head bend, chest bend, abdomen twist, shoulder forward/backward, shoulder swing, and elbow bend). While the overall results were positive, it is interesting to note that the low recognition of some emotions such as fear indicate that features for describing motion, i.e., direction, velocity, and amplitude, also may be necessary.

2.1 Solution: A posture description module

As a solution for this requirement, we have extended our previous work [2] by combining a set of form features (a static instance of a posture) with a set of motion flow features (indicating direction of motion) to create a posture description module for our system.

In [2], we proposed a set of kinematic features for describing human posture suggested by Laban's [20] "sphere of movement" used to convey emotion. We focused mainly on global and upper body features as determined by our preliminary results indicating that the upper body is used most for displaying emotion. These features were computed in the frontal view by projecting 3D motion captured data on the 3 orthogonal planes to measure direction and volume of the body according to the lateral, frontal, and vertical extensions of the body, and body orientation. Refer to the top portion of Table 1 for a listing of these features.

The motion flow posture description features, listed in the lower portion of Table 1, were computed by measuring motion differences between 2 frames of motion capture data within a predetermined interval to show direction of motion. For example, the vertical motion of the right hand was computed by the ratio of the distance of the maximum vertical extension of the right hand along the z-axis. The forward and backward motions of the head and shoulders separately were computed by the ratio of the distance of the maximum frontal extension of these body parts along the y-axis.

Table 1: The table lists the set of form posture features and the set of motion flow features.

Form features
$Orientation_{XY}$: B.Head - F.Head axis
$Orientation_{YZ}$: B.Head - F.Head axis
$Distance_z$: R.Hand - R.Shoulder
$Distance_z$: L.Hand - L.Shoulder
$Distance_y$: R.Hand - R.Shoulder
$Distance_y$:L.Hand - L.Shoulder
$Distance_x$:R.Hand - L.Shoulder
$Distance_x$:L.Hand - R.Shoulder
$Distance_x$:R.Hand - R.Elbow
$Distance_x$:L.Hand - L.Elbow
$Distance_x$: R.Elbow - L.Shoulder
$Distance_x$: L.Elbow - R.Shoulder
$Distance_z$: R.Hand - R.Elbow
$Distance_z$:L.Hand - L.Elbow
$Distance_y$: R.Hand - R.Elbow
$Distance_y$:L.Hand - L.Elbow
Motion features
$MotionAmplitude_y$: Right hand
$MotionAmplitude_z$: Right hand
$MotionAmplitude_y$: Left hand
$MotionAmplitude_z$: Left hand
$MotionAmplitude_y$: Right shoulder
$MotionAmplitude_z$: Right shoulder
$MotionAmplitude_y$: Left shoulder
$MotionAmplitude_z$: Left shoulder
$MotionAmplitude_y$: Head

The output of this module is a pair of vectors that are sent to the recognition module (described in the following section) for determining the affective state of the person being monitored.

3. REQUIREMENT 2: FEEDBACK HANDLING

A system that can learn over time, or incrementally can be considered more human-like in its interaction, as humans also adapt to each other over time, through continued social interaction. Furthermore, a system that is incremental eliminates the need for, and difficulty of, creating a training set that covers the complete range of possible motions and the complete set of possible emotions that could occur. This requirement can be satisfied by using feedback to adapt the recognition model to each new user.

Explicit feedback may come directly from a student or a teacher, explicitly stating the student's emotion.

As we cannot expect the user to continuously give feedback to the system, the system should also be able to handle implicit feedback. By implicit feedback we mean an affective label or a set of affective labels indicating the most probable affective state of the user. This feedback could be inferred on the basis of contextual information such as the state of a game or of an e-learning session. Explicit feedback is generally more reliable, while implicit feedback may carry more uncertainty. Thus, this uncertainty should be taken into account in the adaptation process.

The incremental process should be considered also at the level of the categories to be learned. Refer to Section 5 for a discussion on implementing an incremental lexicon.

3.1 Solution: An adaptive posture recognition module

We see the mapping of posture description features into emotional labels as a categorization problem. We use a CALM [16] network, that can self-organize input into categories. A CALM network consists of several CALM modules, thus incorporating brain-like structural and functional constraints such as modularity and organization with excitatory and inhibitory connections. Figure 2(a) represents a single CALM module. Each module is a complex structure made up of different nodes, and is based on a competition mechanism. Competition is triggered by 2 external nodes that measure the novelty of the input pattern, and accordingly, generate more or less noise to maintain competition until one of the nodes wins. While the topology of a CALM architecture is fixed, connections between the modules (shown in Figure 2(b)), are learned. Novel input samples presented trigger the adaptation of the network by exploiting the unsupervised learning mechanism. The reader is directed to [16] for a complete discussion of a typical CALM network.

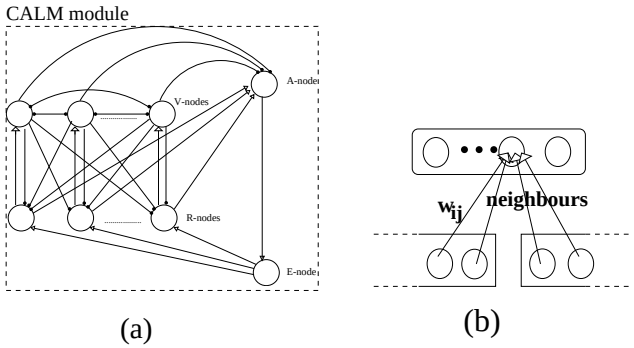

Figure 2: (a) shows the architecture of a basic CALM module. (b) shows the interconnectivity between R-nodes of connected modules.

We extended our CALM network topology proposed in [2] to handle two types of input. Shown in Figure 3, our topology consists of three layers. We use one input layer divided into 2 modules. The first module consists of the original 16 form features, and the second module is comprised of the 9 motion flow features. The division of information was determined based on a neurological study by Giese and Poggio [9] which provides evidence to show that two separate neural pathways in the brain are used for the recognition of biological motion, one for form information and one for motion information. Furthermore, integration of the feature information occurs between the modules at the intermediate layer through a horizontal connection. This is also a reflection of neurological studies stating that in the brain, information is integrated not at the input level, but at a higher level.

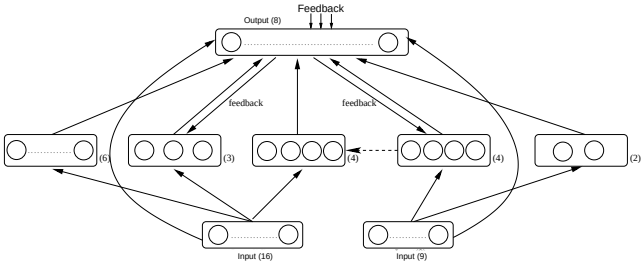

Figure 3: The topology of our system consists of 3 layers. One input layer with 2 modules, one intermediate layer with 5 modules, and one output layer with 1 module. The number of nodes is shown in brackets.

While originally CALM uses unsupervised learning, we use a novel version of CALM that integrates both unsupervised and supervised learning mechanisms [1]. The downward arrows in Figure 3 represent incoming feedback that was sent to the recognition module from the user/teacher and that flows from the output module, down to the intermediate layer. This feedback is used for triggering a supervised adaptation.

Our system utilizes 2 forms of feedback, explicit and implicit. The learning within the module is based on a competitive mechanism and only one R-node finally wins. To implement explicit feedback, when a wrong affective label is output, pulse information is sent back to the recognition module to reactivate competition. Explicit feedback (correct answer is known) triggers the competition module to favor the correct R-node. As we use a low amount of feedback, the output selection is not forced, but instead biases the self-organizing process.

To handle implicit feedback (correct answer is not known, but instead is a ranking of possible answers with an associated probability), we modified the feedback mechanism in order to weight the (biasing) pulse information that is sent to the R-node, thus reactivating competition. The weights reflect the probability of correctness of each emotion label. The probability of each emotion (R-node) could be derived from the context, e.g., state of the game, etc. While implicit and explicit feedback have been implemented, the implicit feedback mechanism has not yet been tested. Therefore, in this paper, we report on testing only with explicit feedback.

4. PERFORMANCE TESTING

We tested the performance of our affective posture recognition system on standing postures and four emotion categories: *angry, fear, happy,* and *sad.* 182 affective postures were collected using a motion capture system. Each subject, wearing the same motion capture suit, was asked to perform

postures expressing each of the 4 emotions. No constraints were placed on the subjects, thus allowing them to express the emotion postures in their own, individual way. For a more detailed description of the data collection techniques, please refer to [12][5].

50 learning trials were conducted. The learning process was stopped when the percentage of error ceased to decrease. Typically, this occurred at approximately 200 epochs. The high classification rates were positive at 79% for 4 emotion categories, versus 71% when employing a single input module (our previous implementation) containing the form features and motion flow features combined, thus providing further evidence to support dividing the 2 forms of input information. Categorization rates further declined to 65% when only the 16 form features were used for input.

Next, we tested the ability of our system to generalize by adding noise to the training set to create 15 testing sets. In looking at the results shown in Figure 4, we can clearly see that our new system, comprised of two input modules to separate form features from motion flow features, outperforms our previous implementation consisting of a single input module combining both feature types. Recognition rates at 10% variance (significant noise) were 60%. When doubling the amount of noise (20% variance), the success rate remained nearly the same.

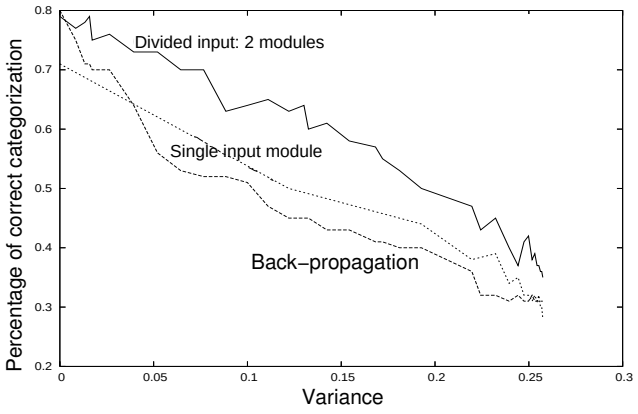

Figure 4: Correct categorization percentages for comparing back-propagation, our affective posture recognition system, and our former system with only one input module, using a training set of 182 affective postures. The horizontal axis denotes the variance for the testing set. The vertical axis denotes the correct percentage of categorization.

For further testing, our categorization system was compared with a back-propagation algorithm [7], one of the most typical neural networks used for recognition tasks. To do so, we repeated the training on the 182 postures 10 times for each method. On average, there was an 80% classification rate for back-propagation; almost the same as with our system (79%). Next, we created 35 testing sets by adding noise to the training set to examine the generalization capabilities of back-propagation. However, while the training results of the 2 methods were quite similar, again referring to Figure 4, we can see that during testing the recognition rate using the back-propagation algorithm falls to approximately 56% by adding only 5% variance, whereas the recognition rate for our system is almost 73%. While these results are clearly positive, further analysis using a variety of back-propagation topologies is necessary to definitively conclude these results. However, it is our belief that in our recognition system, since the supervised mechanism is used only to bias the learning, the generalization capability appears to be more significant. This is quite important due to the intrinsic variability present in emotion expression.

4.1 Comparison with human performance

In order to evaluate our system, we compared its performance with human discriminative performance by conducting a series of psychological experiments. The same 182 postures used in the previous experiments were used to build a set of avatars from the original motion capture data. For each posture, a single frame was chosen that the actor evaluated to be the most affectively expressive instant. Viewing a series of single posture webpages, subjects were asked to evaluate each posture by choosing from a list of 4 emotion categories, *angry, fear, happy*, and *sad*. 143 Japanese university students participated.

After determining the most frequent label associated to each posture by the observers, we see that the recognition rate for observers is significantly lower (69%) than the recognition rate of our system (79%) for 4 affective categories. Reasons for these misclassifications by the human observers could be due to several factors.

For example, the results of a study by Feldman Barrett et al [8] state that people tend to differ in their ability to differentiate between the specific emotions they experience. Instead, they may be able to indicate only whether or not the emotion is "good" or "bad", or they may group together emotions according to other distinguishing factors such as arousal or action tendency. In fact, an examinination of our data shows that all of the misclassifications can be accounted for when considering 3 typical dimensions used to evaluate emotion: arousal, valence, and action tendency.

Another factor may be that some features appear to still be missing. Specifically, we are missing a more complete description of the hands and fingers due to the inability to capture positions of such detailed information with our current motion capture system. Another factor appears to be due to posture ambiguity, indicating that more clues, e.g., facial expression and voice, may be necessary. Furthermore, recognition of some affective states may require knowledge about the relation between the hands and eyes as well as the inclination of the body.

In evaluating the various trainings we performed, we identified the postures that were consistently misclassified by our affective posture recognition system. A total of 25 postures were identified. For these misclassifications, we compared the evaluations of the system with the evaluations of the observers and observed 3 distinct cases. One, the system *agreed* with the most frequent label assigned by the observers (25% of misclassifications). Two, our system *agreed* with the observers' second most frequently chosen label (37% of mis-

classifications). Three, our system *disagreed* with either the actor or the observer, meaning that the system completely failed (37% of misclassifications). The general conclusion here is that when the system makes a mistake in the recognition of an affective posture, it may act as an observer.

5. REQUIREMENT 3: AN INCREMENTAL LEXICON

Ultimately, an affectively aware system should have the ability to incrementally learn to recognize and react to (interact with) the affective behaviors of people, detected through posture, as one of several modalities. Therefore, the system should be able to be used in an interactive and continuous learning situation where new emotion categories (or nuances) could appear. Each individual has her/his own way to interpret and express each type of emotion. According to this reasoning, the emotion lexicon of the system should not be defined a priori but instead should emerge through interaction with the environment.

5.1 Solution: Emotion category emergence

Our recognition module allows for the emergence of the lexicon. An incremental process of the topology of the output module of the CALM network occurs when new emotion concepts are encountered. The R-nodes of the output module are named through the use of explicit feedback. Specifically, each time a new label is given through explicit feedback by the user, it is used to name one of the R-nodes (shown in (a) and (b) of Figure 2) that has not yet been named. Later, the name of that R-node is used to compare the output of the network with the feedback of the users. If all the existing R-nodes are already named, a new pair of R- and V-nodes is created, and eventually named if the new label is available. Indeed, the number of output nodes is not decided a priori, but instead new nodes are added when new emotions are encountered.

6. DISCUSSION

To explore the emergence of the lexicon and how the emotion categories interact, we simulated a real-time situation in which the system learns while being presented with new postures to recognize. In this situation, the system begins without knowing any words. Postures are presented in a random order from a database of postures, and periodically, explicit feedback is sent to the system to give the correct name to an emotion. The system uses this feedback to name a node whenever the emotion label given is new. This feedback also triggers adaptation. Moreover, postures representing one of the emotion labels may not appear for a prolonged period of time. This is different from a normal training session in which the system is presented with a predefined and well organized (and balanced) set of postures, etc.

In total, 212 postures were used (the 182 standing postures discussed in the previous section, plus 30 new sitting postures, as this is the direction of our research), across 9 affective categories (*angry, confused, fear, happy, interest, relaxed, sad, startled*, and *surprised*) chosen to represent different types of emotion situations. For this case, we attained recognition rates of at least 70% correct.

The competitive nature of the self-organizing process of our

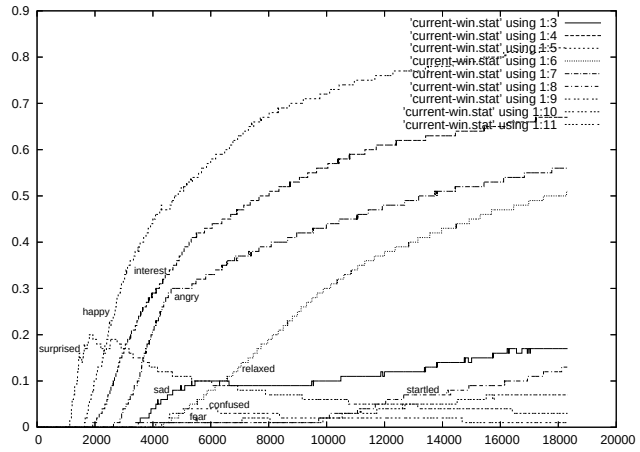

Figure 5: The emergence of affective concepts. The vertical axis denotes the percentage of classification for each affective state. The horizontal axis denotes the number of posture presentations to the system.

system can be seen by looking at the emergence of the affective concepts of the above simulation, represented in Figure 5. This figure shows the cumulated ratio, i.e., the cumulated number of correct classifications over the total number of presentations per category, during learning. What is important here is not the percentage value reached by each curve, but the trend of each curve. In fact, the percentage of correct classification for each emotion category depends on the number of presentations of postures for that emotion. If few postures for an emotion have been presented, the percentage for that curve will remain low.

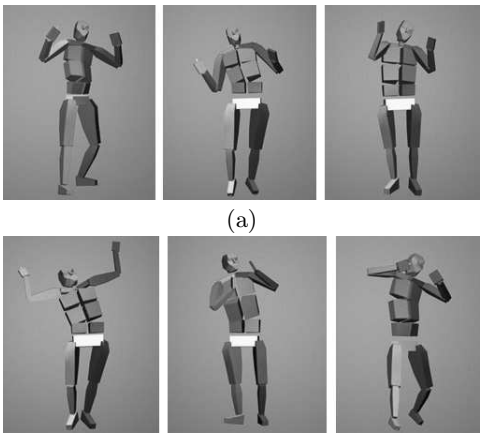

Figure 6: Examples of affective avatars expressing (a) happy and (b) surprise.

More important than the percentage is the trend of the curve. If the curve continues to climb, it means that overall, the postures for that emotion are recognized and hence, that that category is well learned. If the curve is flat, it means that no postures for that emotion have been presented. However, if the curve descends, it means that over-

all, the postures for that emotion are not recognized, i.e., the concept is not well learned. We can see from Figure 5 that the first emotion category to emerge is *surprised*. While new words are used as feedback and learning for those categories begins, new curves start to emerge. When a curve emerges, generally the previously learned category shows a decrease as confusion occurs between the learned categories and the new category.

Little by little, all the trends should begin to climb unless the set of postures within one of the categories are not completely separable. For example, we can see that the curve for *surprised* starts to decrease as soon as the *happy* category starts to appear. In fact, *surprised* and *happy* share many postural features (shown in Figure 6(a) and (b)), i.e. arms up above the shoulders and head straight up [5]. Moreover, according to Coulson's study [4], the postures generated for *happy* and *surprise* were visually similar. As a future step, this trend could be used to automatically identify difficulties in the discrimination of emotions, and to indicate a need for refining the description process.

7. ACKNOWLEDGMENT

This study was supported by a Grants-in-Aid for Scientific Research from the Japanese Society for the Promotion of Science granted to N. Bianchi-Berthouze.

8. REFERENCES

[1] L. Berthouze and A. Tijsseling. Acquiring ontological categories through interaction. *The Journal of Three Dimensional Images*, 16(4):141–147, December, 2002.

[2] N. Bianchi-Berthouze and A. Kleinsmith. A categorical approach to affective gesture recognition. *Connection Science special issue on Epigenetic Robotics*, 15(4):259–269, 2003.

[3] E. P. Bull. *Posture and Gesture*. Pergamon, Oxford, 1987.

[4] M. Coulson. Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence. *Journ. of Nonver. Behav.*, 28:117–139, 2004.

[5] P.R. de Silva and N. Bianchi-Berthouze. Modeling human affective postures: An information theoretic characterization of posture features. *Journal of Computer Animation and Virtual Worlds*, 15:269—276, 2004.

[6] P. Ekman and W. Friesen. *Manual for the facial action coding system*. Consulting Psychology Press, Palo Alto, Ca, 1978.

[7] L. A. Fausett. *Fundamentals of neural networks*. Prentice HAll, London, 1994.

[8] L. Feldman Barrett, J. Gross, T. Christensen, and M. Benvenuto. Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition and Emotion*, 15(6):731–724, 2001.

[9] M. Giese and T. Poggio. Neural mechanisms for the recognition of biological movements. *Neuroscience*, 4:179–191, 2003.

[10] S. Kamisato, S. Odo, Y. Ishikawa, and K. Hoshino. Extraction of motion characteristics corresponding to sensitivity information using dance movement. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 8(2):167–178, 2004.

[11] A. Kapoor, R. Picard, and Y. Ivanov. Probabilistic combination of multiple modalities to detect interest. *Proc. of the 17th International Conference on Pattern Recognition*, 3:969–972, August 2004.

[12] A. Kleinsmith, P. de Silva, and N. Bianchi-Berthouze. Recognizing emotion from postures: Cross-cultural differences in user modeling. In *to be published in UM2005 User Modeling: Proceedings of the Tenth International Conference*. Springer-Verlag, 2005.

[13] J. Larsen, A. McGraw, and J. Cacioppo. an people feel happy and sad at the same time? *Journal of Personality and Social Psych.*, 81(4):684–696, 2001.

[14] D. Massaro and P. Egan. Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review*, 3(2):215–221, 1996.

[15] A. Mehrabian and J. Friar. Encoding of attitude by a seated communicator via posture and position cues. *Journal of Consulting and Clinical Psychology*, 33:330–336, 1969.

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

[17] T. O. Road. *Toyota's Telematic, Telepathic "POD" Smiles, Cries And Wags Its Tail At Chicago Auto Show*. http://www.toyotaoffroad.com/Articles/Toyota /POD/toyota_pod.htm, Chicago, 2002.

[18] J. Russell. Is there universal recognition of emotion from facial expressions? a review of the cross-cultural studies. *Psychological Bulletin*, 115:102–141, 1994.

[19] Sony. *Entertainment Robot*. http://www.aibo.com/, 2003.

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

[21] W. Woo, J. Park, and Y. Iwadate. Emotion analysis from dance performance using time-delay neural networks. In *Proceedings of JCIS-CVPRIP'00*, pages 374–377, 2000.