SENSITIVE SYSTEMS
INCORPORATING AFFECT RECOGNITION INTO COMPUTER BASED LEARNING

Ken Newman

ABSTRACT
Advances in video and audio feature extraction methodology combined with adaptive processing techniques using pattern analysis models have profound implications for incorporating affect recognition into future tutoring agents. This report gives an overview of current affect recognition methodologies and the type of features that can be extracted from digital signals in the form of movement and sound analysis. There is some discussion of translating features into a set of observable human sentic data, which can be in turn 'fuzzy-mapped' to internal affective states. Issues concerning the digital representation and the resolution granularity of affective states are discussed. The final discussion concerns issues of implementing affect recognition methodology into tutoring agents.

KEYWORDS
Recognition methodologies, tutoring agents

INTRODUCTION

Promising work in affect recognition systems (for example, Huang, 1998; Grammer, 1998; Camuri, 1999) from video and audio analysis suggests there are a range of extraction features that can be used to construct an understanding of the affective states of the subject. As future, sensitive tutoring agents become aware of the emotional state of their human users the possibility exists for them to have the potential to adapt and modify their behavior, to try to optimize the individual users' learning experience.

The discussion begins with a general introduction to affective computing including the relationships between affective states and sentic expression, representing affective states and granularity. The discussion then moves on to describe a number of affect recognition systems, and demonstrates some examples of features extraction. Finally, the discussion turns to the application of affect recognition into tutoring agents including some ethical question arising from the use of emotionally aware machines.

Affective Computing

This section is a general introduction to affective computing, and outlines key issues of affect recognition, affect granularity, and the representation of affective states.

Underlying assumptions

The central argument for affective computing is that a system cannot be truly intelligent without emotion (Picard 1995). This argument would appear to run counter to what might be called a traditional scientific philosophy, which rigidly adheres to logic, rational thought, and provable theory. On the other hand, it is entirely consistent with both artistic creativity and is probably closer to everyday human experience than scientific philosophy. Behind human intelligence and decision-making is a complex tangle of emotional and rational processes. Neurological studies (Damasio, 1994) demonstrate that the efficiency of humans to make decisions reduces significantly without emotion. Affective computing advocates would claim that computational systems that aspire to be intelligent, must also be designed with "affective intelligence" especially if they are required to deal with humans or find creative solutions.

Affective States and Sentic Expression

It has become common in the semantics of affective computing, to distinguish between internal and external states of emotion. The common practice in the literature is to refer to the internal state as 'affect' or 'an affective state' and the external state as a 'sentic expression' (Picard, 1995). Fear, for example, is an affective state, while trembling is a sentic expression, which is sometimes also an indicator of a fear state. As humans we are only ever able to observe sentic expression in another person unless we are able to 'read their minds' in some extra-sensory kind of way - but for the purpose of this paper I will assume no special abilities in this area. Our understanding of another person's affective state, then, is always a conjecture based on observing sentic expression; our understanding of how sentic data maps to affective states; our knowledge of the individual we are observing; and our knowledge of their situation. From these inputs, we attempt to discern whether a person is angry, scared, pleased, surprised, amused, frustrated and so on. We can never be totally sure that we have discerned the other's affective state correctly of course, and it is a common human experience to learn after an event that we misjudged a person's feelings at the time. Obviously any affective system we build works under the same constraints.

One of the complicating factors in discerning affective state is the 'fuzziness' involved in mapping sentic indicators to a related affective state and, conversely, mapping the affective state back to the sentic expression. While trembling is a powerful indicator of a state of fear, it may actually be indicating other states such as excitement, rage, anticipation, delight or simply cold. Conversely an internal state of fear may map to a range of sentic expression other than trembling, for example, a drop in body temperature; reduced blood flow to the skin (turning pale); changes in vocal pitch and inflections, and in extreme cases, hysterical screaming.

The Three Fundamental Problems of Affective Computing

Affective computing research divides comfortably into 3 discrete problems (Picard,1995);

- Recognizing affect
- Synthesizing affect
- Experiencing affect.

As human's we accomplish all three tasks with apparent ease. Consider an ordinary 2-year-old who recognizes affective states in others and expresses a wide-ranging repertoire of affective states. Some states were present at birth (fear, being startled), while many have been learned since (pleasure, anger, shyness). This child synthesizes or expresses emotions in unambiguous ways to family and neighbours and uses emotions to focus attention, filter irrelevant thought processes, and efficiently direct decision-making. To create an artificial system with such versatility is the ultimate goal of affective computing. There is some significant research effort going into solving problem 2, synthesizing affect (for example Elliott 1997, Prendinger 2002, Baillie 2002), There is, however, rather less work being done on problem 1, recognizing emotion, which is of course, the focus of this paper - creating systems that can recognize sentic expressions and can make useful assumptions about the user's affective state.

Recognizing affect as pattern matching

Research shows that the ability to recognize emotions in others is a vital social skill for humans, and probably more important than IQ in determining success in life (Goleman, 1995). Picard (1995) argues that recognizing affect is a pattern matching exercise although this could be seen as something of an over-simplification – in the same way one might describe a person's entire life experience as a pattern matching exercise. Nevertheless, it is a reasonable place to start. Recognizing sentic expression and using fuzzy logic mappings to make a crude 'best guess' at the subject's affective state is indeed a pattern matching exercise. To make a human style 'fine resolution' affective assessment of course, requires more than just observed sentic expression – people display emotions differently and in different situations. Therefore, a 'fine-resolution' assessment requires knowledge of the individual and their situation, which implicitly is an open-ended question. It is the AI problem of representing all knowledge in potentially every domain. Since we need to start somewhere, however, pattern matching sentic data is certainly a valid starting point.

Examples of sentic expression

Any sentic expression can be expressed digitally - it could be argued that if it cannot be expressed digitally then it is not a sentic expression. Examples of potentially useful sentic expression could originate from an audio stream and take the form of captured voice inflection, pitch, rhythm or phrasing. Alternatively the features could be extracted from a video stream and indicate movement, rhythms, posture and gestures. Sentic indicators other than audio and video might include, for example, heart rate, respiration and blood pressure, all of which are measurable and can be analyzed for meaningful patterns. Facial expressions are a valuable source of sentic data to humans but a little harder to collect and measure for machines. Most attempts at facial expression recognition systems to date are based on Ekman's (1977) Facial Action coding system which attempts to map facial muscles to emotion space or use an analysis of relative positions of facial features such as eyebrows and edges of the mouth (Huang, 1998).

Affect granularity

By observing such features as a subject's movement bursts and the voice patterns, we may have enough data to conclude that our subject, for example is in a state of *agitation*. *Agitation* is a vague emotion describing a broad chunk of the emotional spectrum. It implies a high state of ***arousal***, and evidence of erratic and frequent motion. For some affective computing applications this may be a sufficiently fine assessment of the affective state. If we wanted to go finer, the next step might be to determine whether the sentic data indicated a positive, neutral, or negative state – this is sometimes referred to as ***valence***. If the video data and voice patterns, for example, were consistent with a strong positive state we might describe the emotion as *excitement*, if not, then perhaps we might call it *distress*. What if we want to describe it with even finer resolution? We may need more than simply sentic data. This is the point where humans overlay the sentic analysis with knowledge of the subject and the situation. Cognitive processes analyze the known facts and look for data that will illuminate and help to further define the affective state. Let's say we know the subject is a young dancer desperate for a position in the New York Ballet; we know her agent has just called to say she has some big news - too big to tell her on the phone; the agent is on her way over to the dancer's house. We compare the known situation with the sentic data and we make a fine resolution conjecture that the affective state is *excited anticipation*. The point here is that although observing and mapping sentic expression can lead us to conclude affective states that may be accurate in terms of an acceptable granularity, the observations will always lack the fine-grain resolution that comes from knowledge of the subject and their situation. Another constraining factor to consider here is the tendency of human beings to simultaneously experience a range of sometimes-conflicting emotion. It should be recognized then that when an affect recognition system attempts to distill a single dominant affective state from the observed data it is not only reducing granularity but also complexity.

Representing affective states

We come now to consider systems for describing and representing affective states. Sometimes it is quite difficult to find an appropriate emotion-word that has a universally unambiguous meaning and describes the emotion with sufficiently fine resolution.

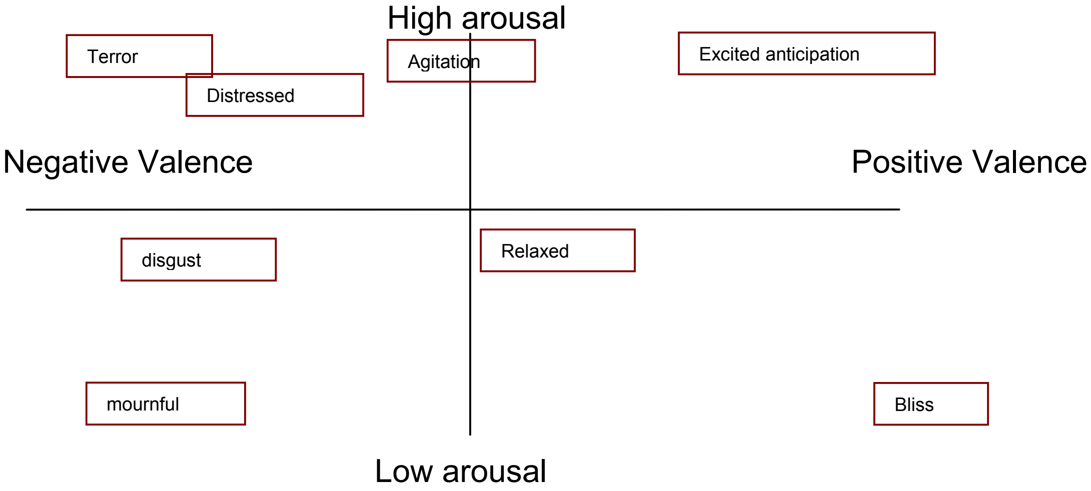

Figure 1. A two dimensional representation of affective state

For this reason researchers often prefer to represent affective states in terms of a number of defining properties rather than by a descriptive word. Figure 1 illustrates one of the most widely used models, a two dimensional representation of affective states giving values for each emotion's arousal and valence properties (Lang, 1995). More complex affect representation models use an increasing number of properties to capture the subject's intentions, effort and even states of mixed emotion (Baillie,2002). The strength of Lang's model, however, is its simplicity, representing an emotion similar to the way a small child will, that is without the need for cognitive understanding of the causes, or even the name of the emotion.

Affect recognition systems

This section describes a number of affect recognition systems, and demonstrates some examples of feature extraction from audio and video signals.

Controlling the test environment

In order to study the process of mapping sentic expression to affective state, researchers must conduct tests simulating emotional states on subjects under controlled conditions. This has in many cases, involved a number of subjects being given a set of emotion-inducing tests (visualizing a scene, listening to music, watching a film). Problems arise in trying to maintain consistency as individuals respond to emotion-inducing stimuli differently; and, express affective states differently.

The emotion-inducing tests themselves can also be a source of ambiguity. Attempting to artificially induce an emotional response from a subject in a laboratory is probably going to result in a quite different affective state than if the emotion is experienced as a part of the subject's normal life.

To overcome these difficulties some researchers have chosen to go down the path of wearable computing to improve the test environment by both monitoring a single subject over an extended period of time, and taking the subject outside the laboratory into real life situations.

Such a test is described in Picard (1998). The subject, an actress, was given a variety of emotion-inducing stimuli to help visualize 8 affective states over a period of 20 days. The physiological data collected from her in this study was electromyogram (EMG), blood volume pressure (BVP), galvanic skin response (GSR) and respiration. The affective states were no-emotion, anger, hate, grief, platonic love, joy, romantic love, and reverence. An analysis of the sentic data contrasting three of the 8 affective states resulted in around 87% recognition rates. Recognition rates dropped off significantly as the system analyzed the emotions with finer resolution, differentiating between all 8 states. Even leaving aside domain knowledge of the subject and their situation the question arises as to what are the key elements humans use to recognize affective states in others.

Bimodal feature extraction

Some work has been done in studying feature extraction for affect recognition by Huang et al (1998). In their study they used an earlier study by De Silva (1997) where subjects were asked to view video and audio of a person speaking and then to identify the affective state of the speaker as either happiness, sadness, anger, dislike, surprise, or fear. Subjects were tested with audio only, video only, and both. Huang used DeSilva's audio and video sets to perform bimodal computer-assisted analysis. In the audio processing, 5 features Huang found to be useful were average pitch, maximum pitch, standard deviation of pitch, average deviation of pitch and average rms energy envelope. In the video processing, Huang used a face-tracking algorithm developed by Tao et al (1998) where a tracking mesh deforms with the facial expressions (fig 2). Huang tracked horizontal and vertical positions of the eyebrows, cheek lifting, horizontal and vertical size of the mouth opening. The results showed the computer analysis had similar confusions to the humans. It also suggests that the indicators for certain affective states will lie in the audio features while others are best conveyed with the video features and that the combined modalities demonstrate an outstanding performance improvement. The reason for this improvement however is not attributable to simply an increase in the number of features being used. Huang reports that they tried using a larger number of features and the performance deteriorated. It would appear rather that the key is to select the most appropriate **complementary** audio and video features.

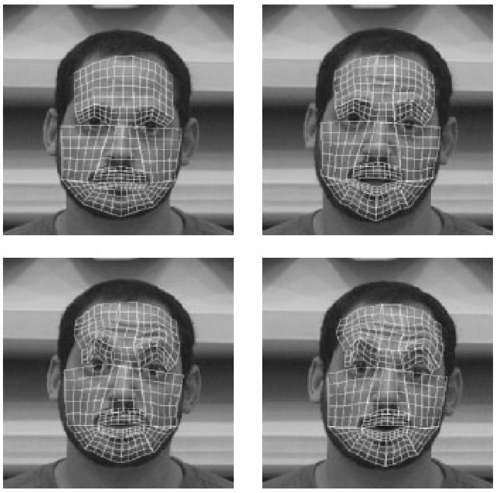

Figure 2. Tao's face tracking mesh (Huang 1998)

Table 1. Audio only recognition results

		Desired					
		Happiness	Sadness	Anger	Dislike	Surprise	Fear
Detected	Happiness	66.7	0	16.7	0	0	0
	Sadness	0	83.3	0	33.3	0	0
	Anger	33.3	0	66.7	0	0	0
	Dislike	0	16.7	0	66.7	0	0
	Surprise	0	0	16.7	0	83.3	16.7
	Fear	0	0	0	0	16.7	83.3

Table 2. Video only recognition results

		Desired					
		Happiness	Sadness	Anger	Dislike	Surprise	Fear
Detected	Happiness	83.3	16.7	0	0	0	0
	Sadness	0	33.3	0	0	0	16.7
	Anger	0	0	66.7	33.3	0	0
	Dislike	16.7	0	33.3	66.7	0	0
	Surprise	0	16.7	0	0	83.3	0
	Fear	0	33.3	0	0	16.7	83.3

Table 3. Bi-modal recognition results

		Desired					
		Happiness	Sadness	Anger	Dislike	Surprise	Fear
Detected	Happiness	100	0	0	0	0	0
	Sadness	0	100	0	0	0	0
	Anger	0	0	100	0	0	0
	Dislike	0	0	0	100	0	0
	Surprise	0	0	0	0	83.3	0
	Fear	0	0	0	0	16.7	100

From (Huang 1998)

Motion Analysis as a key indicator of affective state and intention

Some researchers have argued that there is indeed a set of cross-cultural, universal signals for emotion (Ekman, 1971). Grammar's (1989) observational approach to this question has revealed that pure emotion patterns do not map easily to facial signals. A smile can be modulated by other signals sent in parallel such as age, sex, and attitudes of dominance or submission. His conclusion is that it is not the 'type' of the non-verbal communication that should be analyzed but rather, it is the properties of the movement that actually holds the key information about the sender's intention/affective state.

Similarly, in subsequent studies he found no universal patterns, taken in isolation, of body postures that could be mapped to indicating attitudes of interest or rejection. But combined with laughter or with motion-energy the movements take on a new significance. Grammar has developed a system he calls Automatic Movie Analysis (AMA). Rather than building and installing expensive wearable devices AMA uses digital video of the subjects. Their movements are analyzed by calculating the difference between consecutive frames. After some noise and error filtering to allow for random faults in the video quality, the mean gray values for each frame can be plotted and the movement can be described in terms of the number of movement bursts, their duration, the intensity of the burst, the complexity of a burst (the number of different movement elements which contribute to a burst), and the speed of movement change (the intensity of the burst divided by its duration).

Perhaps the most important aspect of Grammer's work is that in the AMA he has demonstrated a powerful tool for studying visual sentic expression in a temporal context.

Comparing a number of significant studies

Since affect recognition is a relatively new field there are not a lot of studies available. Table 4 summarizes the feature extraction methods that were used in a number of recent prominent studies in the field. As video and audio modes are the least intrusive they will probably be the most useful for complex user interaction systems such as CBL environments. What emerges from this comparison is a lack of research into feature extraction methods for affect recognition, which incorporate audio and video input modes, with full body movement.

Table 4. A comparison of feature extraction methods from recent studies

Researcher	Data Mode	Subject Focus	Sentic Features	Hidden States
Picard (1998)	Wearable sensors	Actress with body sensors	• emg electromyogram • bvp blood volume pressure • gsr galvanic skin response • respiration	No emotion, anger, hate, grief, platonic love, joy, romantic love, reverence.
Marrin(1998)	Wearable sensors	Upper body	• muscle tension • heart rate • body temperature • respiration • skin conductance	Emotional and information communication from a conductor of an orchestra
Elliot(1997)	VR-headset	Field of vision	• Users field of vision relative to tutor's direction	boredom
Huang (1998)	Audio and Video	Talking Heads with some facial feature recognition	Audio: • average pitch • maximum pitch • standard deviation of pitch • average deviation of pitch • average rms energy envelope Video: • Eyebrow position • cheek lifting • mouth position • mouth opening	Happiness, sadness, anger, dislike, surprise, or fear
Rosenblum(1994)	Video Only	Faces	• mouth corners • eyebrows	Happiness, anger, surprise
Grammer (1997)	Video Only	Full Body movement	• motion intensity • motion duration • motion complexity	Sexual interest, attraction
Camuri (1999)	Video Only	Full Body movement with some human figure recognition	Relative and absolute velocities of • shoulder • elbow • hand • knee • ankle	Emotional content of dance movement
Vasconcelos(1998)	Video	Feature films	• close-up shots • pace of cutting	Romantic or action content of a film

Affective states in learning

Learning agents capable of affect recognition have the potential to modify their behavior and tutoring methods in response to the learner. Apart from the rather superficial strategy of simply trying to keep the student in a state of perpetual happiness, it may be worthwhile engineering the tutor's behavior to maintain a balance between the user's satisfaction and frustration at an optimal level for learning. Picard (1995) notes that emotional state is a determining factor in activities demanding mental performance and Hebb (1966) demonstrates that mental performance is at it's lowest when the subject is just waking (low arousal) and when the subject is in an emotional disturbance (high arousal). Optimal mental performance occurs in a more average state of arousal. A tutoring agent, by deliberating pushing the user between satisfaction and frustration could aim to maintain the user's arousal levels for optimal learning.

Figure 4 demonstrates a repertoire of affective states that a tutoring agent could be designed to recognize, which could in turn trigger responsive behavior in the tutoring agent.

1. Learner performs routine learning, no discernable affective engagement (neutral)
2. Learner successfully performs a difficult learning task(pride)
3. Learner begins a new, and very engaging learning experience(fascination)
4. Learner repeatedly fails at a learning task (frustration/ anxiety)
5. Learner finds the learning environment tedious and irrelevant (boredom)
6. Learner becomes openly hostile toward the learning environment (aggression/anger)

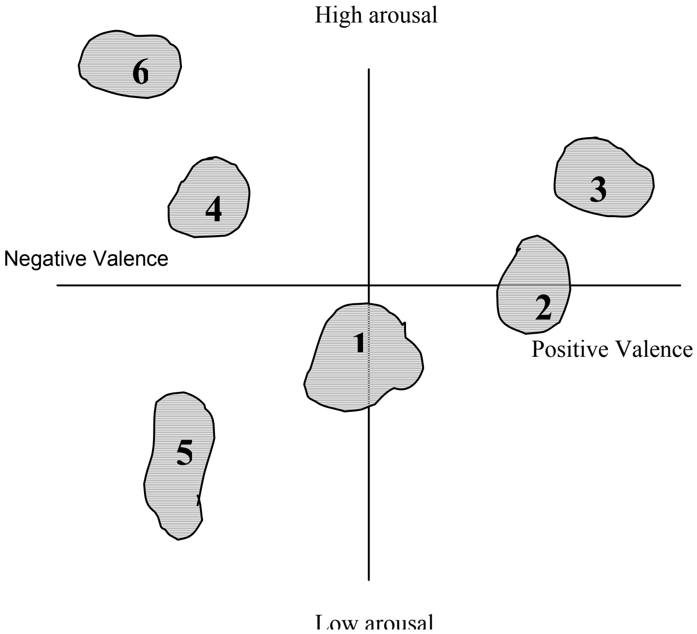

Figure 3. Representation of affective states in learning

Ethical questions of affect manipulation

This does raise questions about the ethics of designing a machine with the intention of deliberately manipulating the affective state of the user. On one hand this can be seen simply as an extension of the process of designing rich interactive experience for the user, making good use of the extra affective state information to maximize the learners' intrigue and minimize their anxiety in the same way a good human teacher would. On the other hand there is the potential for blatant and damaging misuse of the manipulation of individuals' emotions in a way comparable to the most excessive elements of the advertising industry. Perhaps the potential for malicious misuse is even greater from affect recognition agents than from the mass advertising since affect-aware agents have the potential to tailor the manipulation very specifically to an individual. Picard (1995) expresses the dilemma in this way, "Without emotion, computers are not likely to attain creative and intelligent behavior, but with too much emotion, we, their makers may be eliminated by our creation." The futuristic world-dominated-by-machines scenario has been rigorously explored in books and films from "Metropolis" to "The Matrix". The ethical questions of affect-recognition are not dissimilar to the ethical questions of the mass media and it may well be that a combination of legislation, self regulation, and industry codes of practice will grow up around the use of affect recognition applications as they have with the media.

CONCLUSIONS

From the outset computer-based learning has held an implicit promise of rich human machine interaction and to some extent this has been realized. There are numerous CBL applications designed to create an ‘illusion of intelligence’ (Newman, 2000), providing engaging interactive experiences for the user. In many computer-based learning application there are attempts at trying to track users emotional states, this is often done by tracking errant mouse and keyboard behaviour or by monitoring the number of times a user attempts a question, and the system then makes some general assumptions about the user’s level of engagement and level of satisfaction or frustration. With minimal inputs like these a system attempts to track complex affective state information and tries to create an Illusion of responsive intelligence to the user.

With much more sophisticated methods of monitoring the users affective states how much more responsive can an application be? The implicit promise of future CBL agents and environments is less about creating clever illusion of intelligence and more about real interaction. We are now looking towards rich interactive experience with tutoring agents that not only exhibit believable behavior, but also recognize the affective state of the user and modify their behavior accordingly. This paper has identified constraining factors in affect recognition and discussed how features extracted from real-time data can be analyzed to make inferences about the subject’s affective state.

REFERENCES

Baillie P, Lukose D (2002), An Affective Decision Making Agent Architecture Using Emotion Appraisals, in PRICAI pp 581-590, Springer-Verlag Berlin.

Bates J (1992) The role of emotion in believable agents. Communications of ACM, 37(7):122-125.

Baum LE and Petrie T (1966) Statistical inference for probabilistic functions of finite state Markov chains, Ann.Math.Stat vol37 pp1554-1563.

Bradley E, Capps D, Rubin A (1999) Can computers learn to dance? In proceedings of International Dance and Technology IDAT, Tempe, Az, USA

Burgener R (2002)"The 20 Questions Site", http://q.20q.net/

Collins R et al (1999) A System for Video Surveillance and Monitoring, The Robotics Institue, Carnegie Mellon University, PA.

Damasio, A. (1994). Descartes' Error: Emotion, Reason and the Human Brain. New York, Gosset/Putnam Press

DeSilva L, Miyasato T, Nakatsu R, (1997) Facial Emotion Recognition Using Mutlimodal Information, in Proc IEEE Int Conf on Information, Communications and Signal Processing (ICICS97) Singapore, pp397-401.

Doleman, D (1995) Emotional Intelligence NY, Bantum Books

Ekman P, Friesen W (1971) Constants across cultures in the face of emotion, Consulting Psychologists Press

Elliott C, Rickel J, Lester J (1997) Integrating affective computing into animated tutoring agents, IJCAI Workshop on Animated Interface Agents, Nagoya, Japan.

El-Nasr M.S., Skubic M. (1998), A Fuzzy Emotional Agent for Decision Making in a Mobile Robot. A

& M University Press, Texas, US.

El-Nasr M.S. et al. (1999), Emotionally Expressive Agents. A & M University Press, Texas, US.

Grammer K, (1989) Human courtship: biological bases and cognitive processing. In Coalitions and alliances in humans and reproductive strategies, ed Rasa A & Vogel C, pp147-169 Chapman & Hall

Grammer K, (1997) The communication paradox and possible solutions, in New Aspect of Human Ethnology ed Schmitt et al, Plenum Press

Hebb D (1966) A textbook of Psychology, Philadelphia, WBSaunders and Co.

Houtsama A & Goldstein, J, (1972) The central origin of the pitch of complex tones: Evidence from musical interval recognition: The Journal of the Acoustical Society of America, vol 51, pp520-529.

Huanh T, Chen L, Tao H (1998) Bimodal Emotion Recognition by Man and Machine,

Lang, P (1995) The emotion probe: Studies of motivation and attention, American Psychologist vol 50, No 5, pp372-385

Marrin T, Picard R (1998) Analysis of Affective Musical Expression With the Conductor's Jacket, MIT Media Laboratory Perceptual Computing Section Technical Report No 475

Minsky M (1981) Music, Mind and Meaning, Computer Music Journal, Fall 1981, Vol 5 No3.

Newman, K (2000) The Illusion of Intelligence, in proceedings Intelligent Systems and Applications, ISA 2000, Wollongong, Australia.

Picard, R. (1995). Affective Computing. MIT Media Lab Perceptual Computing Section Technical Report No 321

Picard R (1998) Towards Agents that recognize emotions, MIT Media Laboratory Perceptual Computing Section Technical Report No 515

Prendinger H, Descamps S, Ishizuka M (2002), Scripting the bodies and minds of life-like characters, in PRICAI 2002, Springer Verlag, pp571-580

Rosenblum M, Yaser Y, Davis L, (1994) Human emotion recognition from motion using a radial basis function network architecture, IEEE Workshop on Motion of Non-rigid and articulated objects, Austin Tx.

Soonkyu Lee, DongSuk Yook (2002) Audio-to-Visual conversion using Hidden Markov Models, PRICAI 2002 pp 563-570 Springer-Verlag Berlin.

Tao H, Huang T (1998), “Connected vibrations: a modal analysis approach to non-rigid motion tracking” proc IEEE Compt Vision and Patt Recogn ‘98

Ken Newman
Griffith University
Computing & Information Technology
Nathan Campus, Brisbane, Qld 4111
Australia
E-mail : k.newman@griffith.edu.au