Multimodal Affect Recognition in Intelligent Tutoring Systems

Ntombikayise Banda and Peter Robinson

Computer Laboratory, University of Cambridge
15 JJ Thompson Avenue, Cambridge, CB3 0FD, UK

Abstract. This paper concerns the multimodal inference of complex mental states in the intelligent tutoring domain. The research aim is to provide intervention strategies in response to a detected mental state, with the goal being to keep the student in a positive affect realm to maximize learning potential. The research follows an ethnographic approach in the determination of affective states that naturally occur between students and computers. The multimodal inference component will be evaluated from video and audio recordings taken during classroom sessions. Further experiments will be conducted to evaluate the affect component and educational impact of the intelligent tutor.

Keywords: Affective Computing, Multimodal Analysis, Intelligent Tutoring Systems.

1 Introduction

In human-interaction, 55% of affective information is carried by the body whilst 38% by the voice tone and volume, and only 7% person by the words spoken [1]. Ekman [2] further suggests that non-verbal behaviours are the primary vehicles for expressing emotion. With the availability of computational power, and great advances in the fields of computer vision and speech recognition, it is now possible to create systems that can detect facial expressions, gestures and body postures from video and audio feed. Furthermore, systems that can integrate different modalities can offer powerful and much more pleasant computer experiences as they would be embracing users' natural behaviour.

The research is directed towards equipping intelligent tutoring systems with the ability to infer complex mental states from multiple modalities. Until recently, emotion has been a neglected dynamic in the design of intelligent tutors. Ingleton [3] argues that emotion is not merely the affective product of learning but is rather a constitutive of the activity of learning. She further states that emotions shape learning and teaching experiences for both students and teachers as they play a key role in the development of identity and self-esteem. The work therefore seeks to allow the tutoring system to maximize the learning potential of a student by detecting mental states such as frustration, interest and confusion from facial expressions, head gestures and speech prosody.

S. D´Mello et al. (Eds.): ACII 2011, Part II, LNCS 6975, pp. 200–207, 2011.

The core component of the research is related to the *affective response* of the intelligent tutor upon detection of a mental state. The aim of the system is to keep the student in a motivated state throughout the learning session. This requires the system to determine strategies that will trigger an *affect transition* from a negative to a positive state in an efficient and appropriate manner. We will therefore seek to answer questions such as “which affective states are conducive to learning and which ones are not?”, “which interventions are effective?”, “when should the tutor intervene?” and “how often should the tutor intervene?”.

2 Background

2.1 Affect-Sensitive Intelligent Tutors

Human tutors have been shown to be effective due to their ability to provide students with constant, timely and appropriate feedback, and the interactive manner in which they guide the student towards a solution. This in turn prevents the student from disengaging from the studies when they are unable to find solutions. According to Wolcott [4] teachers rely on nonverbal means such as eye contact, facial expressions and body language to determine the cognitive states of students, which indicate the degree of success in the instructional transaction. Embedding an affect-recognition component in an intelligent tutoring system will enhance its ability to provide the necessary guidance, and make the tutoring sessions more interactive and thus effective.

2.2 Intervention Strategies

Prior to the expansion of research in affective computing, feedback-oriented intelligent tutors used various cognitive and expert models which would trigger a response when the student behaviour diverges from that of the expert or cognitive model [5]. To take into account the difference in cognitive skills of learners, Ohlosson's study emphasized the need to model student patterns and learn their academic weaknesses and strengths, and consider this information when deciding on the appropriate teaching strategy. He further suggested that intelligent tutors have internal representation of the subject matter so that it can generate appropriate material specifically adapted for the learner. For example, it should be able to offer a definition, an explanation or a practise problem when the student responses indicate that the material has not been grasped [6]. Adding affect to the tutoring system builds on Ohlosson's ideas in that by tailoring material based on the emotion detected and other cognitive variables, the student will be at a better position to learn.

The critical questions in the tutoring system's teaching strategy relates to the *timing*, *frequency* and the *nature* of the intervention. The help policy adopted by Project LISTEN's Reading Tutor [7] is a great example of timing of the tutor's intervention. In their work, help is offered when a student skips a word, misreads a word, gets stuck or clicks for help. Beck et al. [8] explored the possibility of offering pre-emptive assistance by letting the tutor pronounce a difficult word before an attempt by the student. The system was however found to be biased towards long words and a student's past mistakes. One could argue that students should be allowed to attempt to pronounce the word first, and the tutor can then confirm or correct the pronunciation. Allowing students to make mistakes leads them into a correction cycle which increases the time exposure to material, which may yield better recall rates (a theory supported by Herrington et al [9]).

The affective content of an intervention is another important factor when formulating strategies. Jennifer Robison [10] conducted a study on the application of parallel and reactive feedback, where the former relates to identifying with the student's emotion, and the latter concerns the display of emotions that aim to alter or enhance the observed affect. Her preliminary study suggests that the nature of affective feedback given could lead to either positive or negative consequences, and that due consideration should be given to the current affect state of the learner when selecting affective responses.

2.3 Evaluating Intelligent Tutoring Systems

The common goal of intelligent tutors is to impart knowledge as effectively and efficiently as possible. We will be focusing on the achievement and affect measures related to the educational impact of the system. Many studies perform pre- and post-tests to determine if the learning objectives were met. Since affective computing is a relatively new field, measures to evaluate the impact of affective feedback are still in their early stages of design. One measure that seems to be widely used is the *L* measure introduced by D'Mello [11] which is a probability function that analyses and maps out transitions between affective states making it possible to extrapolate the success or failure of the strategies employed.

2.4 Related Work

The recent focus towards emotionally-sensitive intelligent tutors has led to studies exploring the inference of academic-related emotions from various channels such as facial expressions, speech prosody and physiological signals. These tutors have the ultimate goal of keeping the student motivated or interested in their work by adapting their tutoring strategies based on the observed behaviour of the student. Amongst these are the ITSPOKE and AutoTutor intelligent tutoring systems.

The ITSPOKE (Intelligent Tutoring SPOKEn) dialogue system [12] guides a student through an essay-type qualitative physics problem by eliciting complete explanations and correcting misconceptions. It uses acoustic and prosodic features extracted from student speech to recognize three affect categories, namely, *positive* (encompassing affect states such as confidence and interest), *neutral* and *negative* (a blanket category for emotions such as uncertainty, confusion, boredom, frustration) states. It achieves an accuracy of 80.53% for the three-way classification.

While this is a great start towards adapting to students' emotional states, intelligent tutors need to have a more granular understanding of emotions as mental states such as boredom and confusion have distinct causes and should be addressed appropriately.

The AutoTutor [13] addresses this problem by detecting more affective states, namely, boredom, confusion, flow, frustration and neutral. It infers emotion from conversational cues, body posture and facial features and uses an embodied pedagogical agent to synthesize affective responses through animated facial expressions and modulated speech. The authors are still however investigating the fusion of these channels. The proposed work differs from the current approaches in that it explores the fusion of visual and audio cues to infer academic mental states (which according to our knowledge has not been addressed in the learning domain). Other research works in audio-visual emotion recognition systems have been centred on basic emotions which have distinct signatures thus making such systems easier to model. The various ways that one can express mental states such as thinking, compounded by the task of working with unrestricted natural behaviour of students, makes this a challenging problem.

3 Work to Date

We propose the framework depicted in Figure 1 for the recognition of emotion in an intelligent tutoring system. The emotion recognition components (which were developed in-house) provide automatic analysis of facial expressions, head gestures and speech prosody. The first research task was to adapt and re-train these two subsystems to meet the specifications of the project. This includes identifying new features and incorporating them into the system.

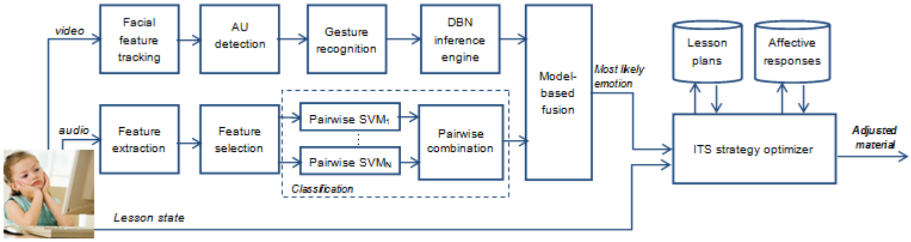

Fig. 1. A multimodal framework for the affect-sensitive intelligent tutor

3.1 Data

The initial system tests were based on the MindReading DVD due to the lack of audio-visual databases enacting discrete complex mental states [14]. The audio analysis section is evaluated on the extended MindReading audio DVD as the number of training files in the original DVD is insufficient to characterize emotional speech.

3.2 Facial Expression Analysis

The recognition of affect from the face remains a challenging task due to the variability of an expression amongst different people, and even within the same person as it is time and context-dependent. The first task in facial expression analysis is representing the face as abstract components for measurement. The most commonly used system for the coding and measurement of the face is the Facial Action Coding System developed by Ekman and Friesen [15]. FACS is an anatomically-based system which uses minimal units (called action units) to describe visible muscle activities that occur for all types of facial movements. The 44 defined action units include eye and head positions and movements.

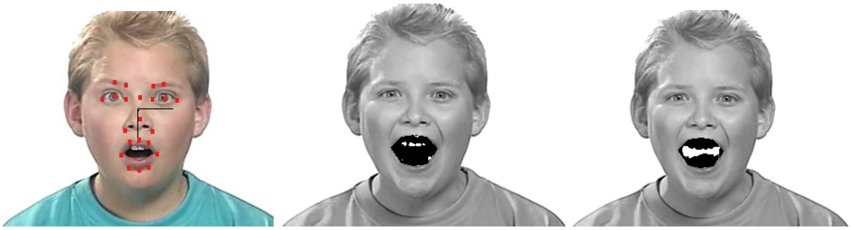

Fig. 2. Three image frames from the MindReading DVD of a child actor expressing the affective state of '*impressed*' within the *interest* mental state group at different time instances, with (a) showing the 22 feature points from the NevenVision tracker, (b) tracking of teeth to represent the *teeth present* gesture, (c) aperture tracking gesture indicating the *mouth open* gesture

The research builds on an in-house real-time complex mental state recognition system developed by el Kaliouby and Robinson [16]. The system abstracts the recognition process into three levels, namely, action unit detection, gesture recognition and mental state inference. The first level involves the tracking of the face from a video sequence using the FaceTracker[1] which outputs 22 feature points (see Figure 2(a)) and head orientation measures (pitch, yaw, roll). The FaceTacker uses Gabor wavelet image transformation and neural networks for its tracking. Action units are detected from the displacement of the features and through appearance-based features.

The second level encodes the detected action units into gestures to allow for complex movements such as head nods and head shakes to be represented. The sequence of action units representing a gesture are modelled using hidden Markov models (HMM) to capture the temporal nature of the gestures. During executing time, the HMMs output quantized probabilities of whether or not a gesture was observed from the sequence of detected action units. The system was extended to recognize gestures such as the presence of teeth and the tracking of an open mouth as seen in Figures 2(b) and 2(c), and to detect furrows.

The final level uses dynamic Bayesian networks (DBNs) to model the unfolding emotion based on the quantized probabilities from the gesture recognition component. Inference is carried out in real time with each emotion modelled as a separate DBN. The inference engine employs a sliding window technique which allows it to predict an emotion based on the history of six gesture observations. The probability scores from the DBNs are integrated over a time period (video length or turn basis) and the emotion with the highest score is selected.

3.3 Audio Analysis

Emotion recognition from audio is concerned with *how* speech is conveyed. The audio analysis component used in the research is an adaptation of the framework introduced by Sobol-Shikler [17] and enhanced by Pfister [18]. The OpenSMILE library extracts 6555 features which represent pitch, spectral envelope, and energy feature groups (amongst others); delta and acceleration information; and their

[1] FaceTracker is part of the NevenVision SDK licensed from Google Inc.

functionals (e.g. min, max, mean, percentiles and peaks). A correlation-based feature selection method is applied to reduce dimensionality. The selected features are used in the training and classification of emotions using pairwise support vector machines with radial basis function kernels. Table 1 shows the performance of the pairwise support vector machines.

Table 1. Pairwise classification results of eight complex mental states using RBF SVMs. The number in brackets refers to the number of features selected for each pairwise class.

	excited	**interested**	**joyful**	**opposed**	**stressed**	**sure**	**thinking**	**unsure**
absorbed	84.2 [80]	86.3 [62]	86.5 [107]	81.7 [82]	83.3 [85]	83.2 [70]	80.7 [66]	78.2 [43]
excited		82.4 [76]	78.5 [38]	83.8 [40]	75.8 [22]	75.9 [68]	86.1 [134]	79.2 [102]
interested			86.8 [89]	85.0 [61]	87.0 [80]	91.2 [79]	80.1 [77]	75.0 [43]
joyful				81.8 [51]	79.7 [65]	85.3 [113]	84.7 [139]	83.1 [109]
opposed					85.3 [82]	76.6 [43]	88.2 [97]	86.9 [54]
stressed						85.0 [89]	86.2 [124]	76.4 [83]
sure							87.0 [115]	87.5 [78]
thinking								72.2 [75]

The pairwise comparisons are combined to calculate the average output probabilities and the count of pairwise wins for each emotion.

3.4 Multimodal Fusion

In Sharma's extensive introduction to fusion of multiple sensors [19], three distinct levels of integrating data are highlighted, namely, data, feature and decision fusion methods. Data fusion is automatically excluded from the consideration as it applies to observations of the same type (for example, two video camera recordings taken at different angles). Feature fusion is applied when the raw observations have been transformed into feature representations and is ideal for synchronized feeds. Decision fusion, also called late fusion, deals with the fusion of decisions computed independently by the respective components. Synchronizing the video and audio channels and aligning emotional segments is a challenging task, especially with the complexity of mental states investigated which need temporal and spatial information to be captured. We have chosen decision fusion given its robust architecture and resistance to sensor failure. The approach however loses information of mutual correlation between the audio and video modalities [20].

The probability scores from the facial expression and audio analysis subsystems are fed into an SVM for training and classification. Due to lack of corresponding training and test data, the multimodal component could not be reliably evaluated. This will however be remedied by a data collection exercise discussed in the next section.

4 Future Work

4.1 Ethnography Study and Data Collection

The next step involves spending time in primary schools with learners between the ages of ten and twelve to conduct a behavioural analysis to determine conditions that trigger negative affect, the immediate impact the affect has on the student's task and strategies that tutors and teachers employ to reverse the negative affect. This will involve recording students in their interactions with computers and tutors, and during traditional class sessions. Such a study will allow us to identify the common mental states experienced in a variety of learning activities, and to decide upon the mental states that our refined recognition system will detect. This will also serve as a great opportunity to collect recordings of natural (and elicited) data that will be used for training and testing the multimodal system. A crowdsourcing approach for the affect annotation of video and audio recordings will be followed.

4.2 Multimodal Recognition for Natural Data

Once the data has been collected, the multimodal recognition system will be configured to work with natural data. We will investigate the use of other sensors, such as an eye tracker for gaze detection, to increase the accuracy of the system.

4.3 Intervention Strategies

The main contribution of the research will stem from the task of developing computer-based affect-related intervention strategies to maintain interest and an overall positive affect during learning sessions. A study will be conducted to determine which emotions can provide the transition from a negative mental state to a positive one. The study will involve inducing positive affect in students, then subsequently inducing a negative affect and applying a random selection to reverse the affect transition. The findings of the study will be formulated into strategies and incorporated into the tutoring system. We will also investigate the timing and frequency of the interventions.

4.4 Evaluation

The multimodal recognition system will be evaluated on the natural emotions collected from the ethnographic study. The intervention component of the intelligent tutor will be evaluated through simulations and self-report, and the results will be analysed using the L probability function for successful affective transitions. The educational impact of the system will be evaluated through a control experiment involving two groups of students. One group will be subjected to learn unfamiliar material through self-study whilst the other through the intelligent tutor. Pre- and post-tests will be conducted to determine knowledge gained, and a subsequent test at a later date to test retention of information.

Acknowledgements. The financial assistance of the South African National Research Fund (NRF), Bradlow Foundation Scholarship and Google Anita Borg Memorial Scholarship towards this research is hereby acknowledged.

References

1. Paleari, M., Lisetti, C.: Toward Multimodal Fusion of Affective Cues. In: Proceedings of the1st ACM International Workshop on Human-Centered Multimedia (2006)
2. Ekman, P., Friesen, W.: Nonverbal behaviour in pschotherapy research. Research in Pschotherapy 3, 179–216 (1968)
3. Ingleton, C.: Emotion in learning: a neglected dynamic. Cornerstones of Higher Education 22, 86–99 (2000)
4. Wolcott, L.: The distance teacher as reflective practitioner. Educational Technology 1, 39–43 (1995)
5. Murray, T.: Authoring Intelligent Tutoring Systems: An Analysis of the State of the Art. Internal Journal of Artificial Intelligence in Education 10, 98–129 (1999)
6. Ohlsson, S.: Some Principles of Intelligent Tutoring. Instructional Science 14, 293–326 (1986)
7. Mostow, J., Aist, G.: Giving help and praise in a reading tutor with imperfect listening - because automated speech recognition means never being able to say you're certain. CALICO Journal 16, 407–424 (1999)
8. Beck, J.E., Jia, P., Sison, J., Mostow, J.: Predicting student help-request behavior in an intelligent tutor for reading. In: Proceedings of the 9th International Conference on User Modeling (2003)
9. Herrington, J., Oliver, R., Reeves, T.C.: Patterns of engagement in authentic online learning environments. Australian Journal of Educational Technology 19, 59–71 (2003)
10. Robison, J., McQuiggan, S., Lester, J.: Evaluating the consequences of affective feedback in intelligent tutoring systems, pp. 1–6 (2009)
11. D'Mello, S., Taylor, R.S., Graesser, A.: Monitoring Affective Trajectories during Complex Learning. In: Proceedings of the 29th Annual Meeting of the Cognitive Science Society, Austin, TX, pp. 203–208 (2007)
12. Litman, D., Forbes, K.: Recognizing Emotions from Student Speech in Tutoring Dialogues. In: Proceedings of the ASRU 2003 (2003)
13. D'Mello, S., Jackson, T., Craig, S., Morgan, B., Chipman, P., White, H., Person, N., Kort, B., el Kaliouby, R., Picard, R.W., Graesser, A.: AutoTutor Detects and Responds to Learners Affective and Cognitive States. In: Workshop on Emotional and Cognitive Issues at the International Conference of Intelligent Tutoring Systems (2008)
14. Baron-Cohen, S., Golan, O., Wheelwright, S., Hill, J.J.: Mind Reading: The Interactive Guide to Emotions. Jessica Kingsley Publishers, London (2004)
15. Ekman, P., Friesen, W.V.: Facial Action Coding System: a technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto (1978)
16. el Kaliouby, R., Robinson, P.: Real-time Inference of Complex Mental States from Facial Expressions and Head Gestures. In: Real-Time Vision for Human-Computer Interaction, pp. 181–200. Springer, Heidelberg (2005)
17. Sobol-Shikler, T., Robinson, P.: Classification of complex information: inference of co-occurring affective states from their expressions in speech. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1284–1297 (2010)
18. Pfister, T., Robinson, P.: Speech emotion classification and public speaking skill assessment. In: Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A. (eds.) HBU 2010. LNCS, vol. 6219, pp. 151–162. Springer, Heidelberg (2010)
19. Sharma, R., Pavlovic, V.I., Huang, T.S.: Toward a multi- modal human computer interface. In: Beun, R.-J. (ed.) Multimodal Cooperative Communication, pp. 89–112. Springer, Heidelberg (2001)
20. Zeng, Z., Pantic, M., Huang, T.S.: Emotion Recognition Based on Multimodal Information. In: Affective Information Processing (2008)