An Interface to Simplify Annotation of Emotional Behaviour

Shazia Afzal, Peter Robinson
Computer Laboratory, University of Cambridge
15 JJ Thompson Avenue, Cambridge, CB3 ODF, UK
E-mail: Shazia.Afzal@cl.cam.ac.uk, Peter.Robinson@cl.cam.ac.uk

Abstract

Research in affective computing is increasingly moving towards naturalistic data. Capturing and annotating such complex data is a massively challenging task. This paper describes a simple and efficient annotation scheme that promotes context-sensitive data labelling via an easy to use interface in an attempt to reduce reliance on expert or trained coders. Additionally, the same labelling interface can be used to obtain self-report of emotional behaviour from subjects. This annotation method has been designed to allow faster labelling of data with a minimal learning curve as part of our research in studying non-verbal expressivity of affect in computer based learning environments and is currently being evaluated. Design decisions are based on feedback from usage of initial prototype as well as relevance to our domain of interest. We anticipate that this can enable faster preparation of representative data in an effective manner for use in automatic analysis studies.

1. Introduction

As affect (or emotion) research gradually integrates with HCI studies and matures in application from mere prevention of usability problems to promoting richer user experiences, the need to capture 'pervasive emotion' (Cowie et al., 2005) and also its context of occurrence is becoming an increasing concern. Our research involves modelling affective aspects of learner experience in computer assisted learning environments. As such we are interested in studying how non-verbal behaviour from multiple-cues like facial expressions, eye-gaze and head posture can be used to infer a learner's affective state during interaction and learning with a computer tutor. The ultimate objective is to abstract this behaviour in terms of features that can enable automatic prediction and reliable computational modelling of different affect states. The need for representative data is therefore essential in order to carry out realistic analysis, to develop appropriate techniques and eventually perform validation of inferences.

Capturing naturalistic data - as it occurs and in all its complexity, is however a massively challenging task. Existing databases are often oriented to prototypical representations of a few emotional expressions, being mostly posed or recorded in scripted situations. Such extreme expressions of affect occur rarely, if at all, in HCI contexts. The applicability of such data therefore becomes severely limited because of observed deviation from real-life situations (Batliner et al., 2003) and for our purpose, their relevance to a learning situation like one-on-one interaction with a computer tutor. For developing systems that generalise to real world applications there is now an increasing shift from easier to obtain posed data to more realistic naturally occurring data in the target scenarios. Dealing with the complexity and ambiguity associated with natural data is however a significant problem. Automatic prediction using machine learning relies on extensive training data which in this case implies preparation of labelled representative data. This also serves as a ground-truth for validation of developed techniques and is therefore a crucial necessity. Non-verbal behaviour is rich, ambiguous and hard to validate making labelling of data a tedious, expensive and time-consuming exercise. In addition, lack of a consistent model of affect makes the abstraction of observed behaviour into appropriate labelling constructs very arbitrary. To achieve a compromise between descriptive detail and economy of annotation effort as in Kipp et al. (2007), this paper describes an annotation scheme tailored to our research but also applicable to similar areas. It is designed to map spontaneous interpretation of recorded behaviour onto different affect states and is currently being evaluated.

In Section 2 we describe the annotation method along some parameters that we deem to be important while considering an annotation scheme. Section 3 discusses some limitations and possible improvements to enhance the procedure while Section 4 concludes the paper by summarising the main idea.

2. Annotation Method

The annotation method that we describe evolved from various domain relevant decisions related to the choice of labelling constructs and modality, anticipated technical constraints in target scenario, relation to context and ease of interpretation. It is inspired by socially-based coding schemes; that is, observational systems that examine behaviour or messages that have more to do with social categories of interaction like smiling rather than with physiological elements of behaviour like amplitude (Manusov, 2005). Precisely, Bakeman & Gotham (1997) define a socially based scheme as one "that deal with behaviour whose very classification depends far more on the mind of the investigator (and others) than on the mechanisms of the body".

The scheme is designed to allow a split-screen viewing of a subjects' recorded behaviour along with the time synchronised interaction record obtained via screen capture. It is implemented in the form of an easy to use annotation interface that combines viewing, navigation and labelling of recorded data. Figure 1 below shows a snapshot of a labelling session using the interface.

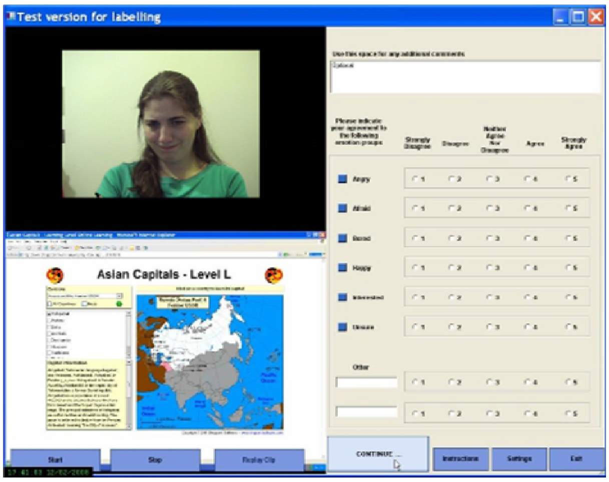

Figure 1: Snapshot of the annotation interface

The annotation scheme and different features are described here along the following parameters.

Labelling Constructs: Decisions related to the representation of affect permeate every subsequent step in the automatic analysis of non-verbal behaviour. Annotation schemes commonly employ either categorical, dimensional or appraisal based labelling approaches. In addition, free-response labelling may also be used for subjective descriptions. For a description and relative merits of each method the interested reader is referred to (Cowie et al., 2005; Douglas-Cowie et al., 2004).

We are using a variant of categorical labelling where coders are asked to rate their agreement on each of the pre-selected categories based on a Likert scale ranging from Strongly Agree to Strongly Disagree. The categories reflect the macro-classes of a taxonomy of complex mental states selected for their relevance to the domain of study and therefore include affect states that are considered pertinent in learning situations (Afzal & Robinson, 2007).

Getting agreement ratings on all affect descriptors on a single data segment allows a greater degree of freedom in inference tasks. To reduce the bias of forced choice on selected affect labels - an often listed drawback in categorical methods, the scheme allows the coder to define his/her own category or label if the perceived state is not represented by the categories. Additionally, there is provision for a free-form description should the coder wish to include comments or other observations not captured via categorical listing. The flexibility in labelling is provided consciously in order to characterise mixed emotions that are known to occur frequently in realistic settings.

Level of measurement: Observational assessment can be done along two different frames of reference – at a macro level to capture the social meaning of behaviour or at a micro level to analyse specific cues or displays in behaviour (Manusov, 2005). Our purpose is of capturing the affective component in behaviour - which is influenced by social meaning, rather than coding of individual displays like smiles, head gestures, eye-gaze, etc.

Context Information: Expressivity and context interact in complex ways as behaviour is always interpreted within a certain context. To emphasise the significance of context in the perception of meaning Russel (1997) cites the example of an experiment where three silent film strips each ending with the same footage of a deliberately deadpan face of an actor were created. In each strip the face was preceded by a different picture - a bowl of soup, a dead woman in a coffin and a young girl playing with a teddy bear. The result was an illusion – audiences saw emotions expressed in the actors' deliberately posed expressionless face (Russel, 1997). Such varying interpretations based on varying context information indicates the danger of forming judgements in isolation from the context of occurrence. Ignoring context can thus dangerously introduce a relative bias in inferences made solely from non-verbal behaviour (Russell J.A., Fernandez-Dols, 1997; Jinni et al., 2005).

In order to represent context information we explicitly try to recreate the interaction by making available both the activity and the user views so that the coder does not need to spend additional time in 'creating' a context. This coupling of information recreates the evolution of behaviour on task and primes the coder into making context-sensitive judgement. The idea is contextualisation of meaning by combined assessment.

Coding Unit: The coding unit refers to the decisions regarding when to code within the interaction and the length of time the observation should last (Manusov, 2005). It has two broad variants - event based and interval based. The choice of the coding unit depends upon the research view and the level of accuracy required, complexity of the coding scheme and the frequency of behaviour occurrence (Bakeman & Gotham, 1997). Our method implements the interval based coding unit through fixed time slots. Also known as systematic observation, this has the advantage of allowing behaviour to be observed consistently throughout the interaction allowing a more accurate reflection of how it is represented in the data (Manusov, 2005).

The initial prototype of the tool implementing the annotation method operated in two modes: manual and timed. In manual mode, the choice of determining when to label was left to the coder while in the timed mode the coder was prompted for noting the annotation after preset durations like every 5 seconds, 10 seconds or 20 seconds. It was observed that non-expert coders preferred the timed mode over the manual one as being much easier and convenient. Manual operation requires controlled navigation while maintaining the reference to context. Coders felt that it distracted them from observing the sequential behaviour. By not having to define segments of emotional episodes they could focus solely on the observation and hence labelling of behaviour. As such the manual mode was disabled in the current implementation of the annotation scheme and only the fixed interval coding unit was retained.

Dynamic Interpretation: Instead of pre-segmenting video clips for labelling, the method forces labelling in temporal sequence. In this way it retains the natural evolution of the behaviour and preserves the dynamics of expression and interaction.

Level of Expertise & Ease of Use: Selection of coders or raters is important for the labelling process as they should be able to discern meaning from behaviour and make judgements. Reliance on expert or trained coders makes the labelling task very time-consuming and expensive. Since the effectiveness of coders depends hugely on the nature and complexity of the coding system applied (Manusov, 2005), the design of the interface and coding scheme was simplified in an attempt to include non-experts as coders. Annotation tools like FEELTRACE (Cowie et al., 2000) and ANVIL (Kipp, 2001) require considerable training before use and restrict access to expert coders owing to the associated learning curve. Our proposed annotation method can on the other hand be used by diverse people without prior experience in labelling. To ensure quality of observation however, the coders can be pre-tested on their nonverbal decoding ability. Initial evaluations show that users are able to perform labelling smoothly soon after being familiarised with the interface and labelling procedure.

Self-Reporting: Inter-coder agreement scales like Cohen's Kappa are used for validation of annotation but are highly sensitive to the affect decoding skills and gender of individual coders (Abrilian, 2005). Obtaining self-report from subjects is an effective strategy of cross validation and interpretation of behaviour. Usage of standard self-report instruments like SAM (Lang, 1980) and EmoCards (Desmet et al., 2001) depends on specific research setups and factors like type of data sought, resources available, situation and users (Isomursu et al., 2007). Our method allows ease in comparison since the same interface used for labelling by external coders can also be used to obtain self-report. Verbal feedback from subjects using this method for self-reporting verified the utility of providing context knowledge and also the ease in usage. Of interest here is that even while self-reporting affect judgements, subjects preferred to work in the fixed interval timed mode rather than event based mode.

Optimisation of annotation effort: The method economises annotation effort by eliminating the need to iterate over data for hierarchical labelling as proposed Abrilian et al. (2005). The structure of the labelling format implicitly incorporates the elements of multi-step or hierarchical annotation as recommended.

Output Format: Each labelling session produces annotations in exportable *csv* or *xml* files. This allows seamless integration with data analysis tools and hence faster interpretations.

3. Limitations & Possible Extensions

Use of pre-selected categorical labels is an unavoidable limitation and has been done to cater to our domain of study. Also, dimensional constructs like valence have a relative meaning. Confusion, for instance, is considered a negatively valenced emotion and but has been found to have a positive effect on learning (Craig et al., 2004). So if a coder has to label the valence of a specific behaviour it will be difficult to establish whether the valence represents the objective view per se or is to be understood in relation to the current task.

Another drawback of our approach is that it will fail to account for emotional transitions occurring at the periphery of the fixed time intervals for observation. Depending on the frequency of such occurrences this can be easily overcome by repeating the annotation on a different time-scale. Interpreting results on the same source labelled on different time scales is trivial as the larger time grain can always be defined in terms of the smaller time segments and thus easily compared.

Further extensions to improve the annotation mechanism involve inclusion of context attributes like theme, degree of implication, target of emotion, communicative goal and the cause of emotion (Abrilian, 2005). Additions of more labelling attributes will however increase the complexity and difficulty of the labelling process. Online availability of the annotation tool to facilitate access and coordinate the labelling process is also proposed.

4. Summary & Conclusions

Labelling of data has a dual purpose. For computational analysis it serves as a ground-truth for evaluation and comparison of performance. More importantly, it serves as a key knowledge source to develop an understanding of affective behaviour that may occur in a learning situation and how it is perceived by humans. It is non-trivial in terms of the complexity associated with deciding the correct representation and descriptors of emotional behaviour as well as in the overall effort required for the task. Further, sensitivity of emotions to the form of measurement makes it more challenging to arrive at an optimal annotation format. Since quality of annotated data determines the efficiency of automatic prediction techniques, the choice of an annotation methodology is an important determinant of the true usefulness of collected video data.

This paper describes a simple and yet effective annotation method that can be easily administered to allow faster labelling of naturalistic data. The motivation to develop a simplified interface for annotation was to include non-experts in the coding process and utilise their general skills of decoding nonverbal behaviour. The annotation scheme is designed as part of our study of non-verbal behaviour in learning environments and is being evaluated.

5. Acknowledgements

This research is supported by the Gates Cambridge Trust and the Overseas Research Studentship of the University of Cambridge.

6. References

Abrilian, S., Devillers, L., Buisine, S., Martin, J-C (2005). “EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces”, *HCI International*.

Afzal S., Robinson P. (2007). A Study of Affect in Intelligent Tutoring, In *Proceedings of Workshop on Modelling and Scaffolding Affective Experiences to Impact Learning, International Conference on Artificial Intelligence in Education*, Los Angeles.

Bakeman R., Gothman J.M. (1997). Observing interaction: An introduction to sequential analysis, Cambridge University Press, UK.

Batliner A., Fischer K., Huber R., Spilker J., Noth E. (2003). How to Find Trouble in Communication, Speech Communication, 40 (1-2), pp. 117-143

Boener K., DePaula R., Sourish P., Sengers P. (2007). How emotion is made and measured, *Int. J. Human-Computer studies*, 65, pp. 275-291

Cowie, R., Douglas-Cowie E. & Cox C. (2005). Beyond emotion archetypes: Databases for emotion modelling using neural networks, *Neural Networks*, 18, 371-388

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). 'FEELTRACE’: An Instrument for Recording Perceived Emotion in Real Time. In *ISCA Workshop on Speech & Emotion*, (pp. 19-24). Northern Ireland.

Craig, S.D., Graesser A.C., Sullins, J. & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29, pp. 241-250.

Desmet P.M.A., Overbeeke P.J., Tax S.J.E.T. (2001). Designing products with added emotional value; development and application of an approach for research through design, *The Design Journal* **4** (1), pp. 32–47.

Devillers L., Vidrascu L., Lamel L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18, pp. 407-422.

Douglas-Cowie, E. et al. (2004). Deliverable D5c-Preliminary plans for exemplars: databases. *Project Deliverable of Humaine Network of Excellence*.

Isomursu M., Tahti M., Vainamo S., & Kuutti K. (2007). Experimental evaluation of five methods for collecting emotions in field settings with mobile application, *Int. J. Human-Computer Studies*, 65, pp. 404-418.

Jinni, A. Harrigan, J. A., Rosenthal, R., and Scherer, K. R. (2005). *The New Handbook of Methods in Nonverbal Behavior Research*. Oxford University Press.

Kipp, M. (2001). Anvil - A Generic Annotation Tool for Multimodal Dialogue. *Proceedings of Eurospeech'2001*.

Kipp M., Neff M., Albrecht I. (2007). An Annotation Scheme for Conversational Gestures: How to economically capture timing and form. *Journal on Language Resources and Evaluation-Special Issue on Multimodal Corpora*.

Lang P.J. (1980). Behavioral treatment and bio-behavioral assessment: computer applications. In: J.B. Sidowski, J.H. Johnson and T.A. Williams, Editors, *Technology in Mental Health Care Delivery Systems*, Albex, Norwood, NJ (1980), pp. 119–139.

Manusov V.L. (2005). *Sourcebook of Nonverbal Measures: Going Beyond Words.* Lawrence Erlbaum Associates.

Russell J.A., Fernandez-Dols J.M. (1997). *The psychology of facial expression*. Cambridge, MA: Cambridge UP.