Modelling emotions in an on-line educational game

Peter Robinson

Computer Laboratory
University of Cambridge
Cambridge CB3 0FD England
peter.robinson@cl.cam.ac.uk

***Abstract*—Emotional expression and recognition are important in social interactions between people. This is particularly evident in communications between a teacher and a pupil when facial expressions signal levels of understanding and enjoyment will change the teacher's explanation to a pupil, and effective e-learning systems must adapt in the same way if they are provide the social interactions that are necessary for effective pedagogy. Affective computing can equip computer systems with the ability to process social signals and respond accordingly. However, social signals are inherently ambiguous and confusion will result if the approach to processing them is too mechanistic. This paper presents and analyses empirical evidence for this ambiguity, and proposes a possible solution. The techniques are applicable in a wide variety of applications where continuous measures of performance are being assessed.**

***Index Terms*—Affective computing, educational games, emotions, human-computer interaction, social signal processing.**

I. Introduction

The ability to display and recognise emotions is an important aspect of social interaction between humans. We monitor each other's facial expressions, vocal nuances and body postures and gestures, and use them to make inferences about other people's mental states. People who are unable to do this are at a social disadvantage. This is particularly important in an educational context where a teacher has to read the expressions of pupil's faces, make inferences about whether they are confused or understanding, interested or bored, and adapt the lesson accordingly. Effective e-learning systems need to adapt dynamically in the same way [1], [2].

Autism Spectrum Conditions (ASCs) are neurodevelopmental conditions characterized by social communication difficulties and restricted and repetitive behaviour patterns. The European ASC-Inclusion project [3], [4] aims to create and evaluate the effectiveness of an internet-based game platform, intended for children with ASCs and their carers. The platform combines several state-of-the art technologies in one comprehensive virtual world providing training through games, and including feedback from analysis of the player's gestures, facial and vocal expressions using a standard web-cam and microphone. The game also includes text communication with peers and smart agents, animation, video and audio clips.

One component of the game monitors the player's face while he or she is acting a particular emotion. Computer vision and machine learning are then used to infer the emotion depicted and report back, both assessing the player's performance and also suggesting changes to make it resemble a canonical performance more closely. This is an extremely challenging test for automatic analysis of emotions and provides useful information for the more general use of affective feedback to guide social interactions in adaptive e-learning systems. In particular, care must be taken in the choice of system adopted for modelling emotions.

978-1-4799-6773-5/14/$31.00

Validation of videos for use in the ASC-Inclusion game and pilot trials of the game itself have revealed considerable ambiguity in both categorical labels of emotions and dimensional measurements of human feelings. This paper presents and analyses empirical evidence for this ambiguity, and proposes a possible solution to allow affective monitoring to be used in adaptive e-learning systems.

The remainder of the paper is in three parts. Section II presents a summary of the two main models of emotion that are used in affective computing. Section III presents an analysis of data collected for the ASC-Inclusion project that gives some insights into the problems that can arise if these models are used too naïvely. Section IV proposes a more measured approach to the use of affective feedback in computer applications and Section V explains how this can be applied in practice. Finally, Section VI concludes the paper by exploring the broader implications of this work.

II. Models of affect

Charles Darwin considered seven categories of emotion in his seminal work on *The expression of the emotions in man and animals* [5]. A century later, Paul Ekman refined this into a classification of six basic emotions – *anger*, *disgust*, *fe*ar, *joy*, *sadness* and *surprise* [6]. The six basic emotions and Ekman's Facial Action Coding System (FACS) [7] have been widely used in the study of emotions over the past 35 years, and particularly for work on affective computing in the past 15 years. However, they are not particularly representative of people's everyday experiences.

More recently, Simon Baron-Cohen has devised a new taxonomy of human emotions based on a linguistic analysis [8]. 412 distinct emotion concepts are identified and grouped into 24 disjoint categories. These include Ekman's six basic emotions and a further 18 further groups that cover complex mental states reflecting cognitive activity. They also require a few seconds of continuous observation to be recognised by humans, rather than the single image that suffices for basic emotions [9].

James Russell took a different approach by deriving a continuous, dimensional classification in his *Circumplex model of affect* [10]. This was formulated in the light of an experiment in which participants arranged 28 emotion words around a circle, with similar affects located close to each other and inverses on opposite sides of the circle. Principal Component

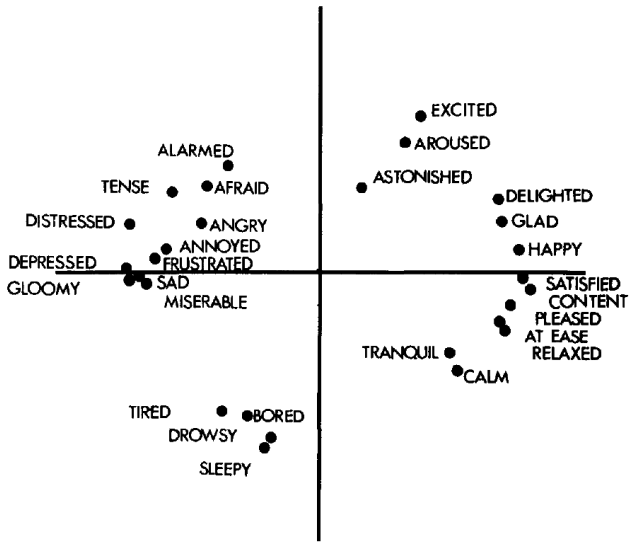

Figure 1. Two principal components of 28 affect words [10]

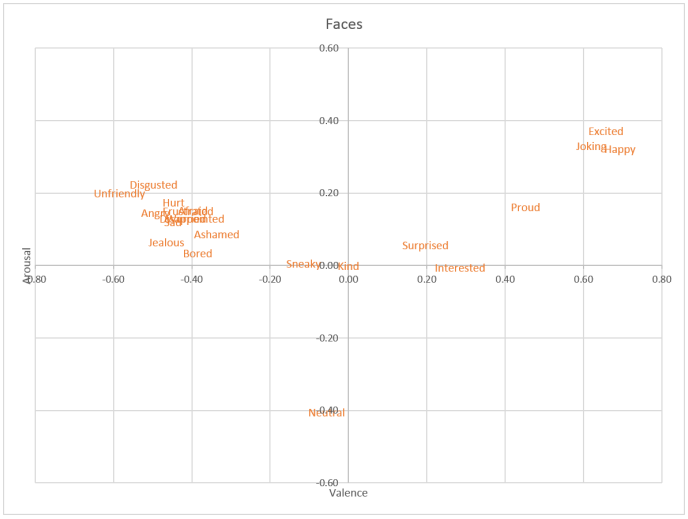

Figure 2. Average assessments of valence and arousal

Analysis was then used to identify various dimensions in the data. The first two components accounted for 46% of the total variance, and the next three only an additional 13%. The locations determined by the two principal components are shown in Figure 1.

The horizontal axis is usually referred to as *valence* and the vertical axis as *arousal*. The further axes have been given names like *intensity*, *expectancy* and *tendency*. This has led to a popular belief that emotions can be measured precisely by coordinates in a suitably high-dimensional space. Unfortunately, this is not true.

III. Measuring Mental States

The face is one of the clearest channels for communication of human emotion. People routinely express their mental states through their facial expressions. Inference of emotion from facial expressions has been studied for 10 years, using a variety of techniques – rule-based classifiers, neural networks, support vector machines, and Bayesian classifiers – but only considering Ekman's six basic emotions. Recognising the complex, cognitive mental states is more difficult, but probably more useful as part of general interaction with computer systems. We have developed a full automatic system requiring no human intervention which operates in real-time [11], [12]. Our Facial Affect Inference System uses a multi-level representation of the video input, combined in a Bayesian inference framework operating at four levels: facial feature points, FACS action units (AUs), gestures composed of several AUs, and mental states.

A great deal of data was necessary to determine the window sizes in the temporal abstraction and to train the statistical classifiers in the inference system. Baron-Cohen's Mind Reading DVD [8] proved ideal for this purpose. Our evaluation considered six conditions drawn from five of the 24 emotion groups and including 29 of the underlying mental state concepts, and chosen to be particularly relevant for human-computer interaction. For a mean false positive rate of 4.7%, the overall accuracy of the system is 77%. The system also generalises well to faces not included in the training data.

Subsequent work has led to a new facial tracker [13] and a version of the facial affect inference system that reports continuous values for valence and arousal [14].

Training in the ASC-Inclusion game introduces emotions using a categorical description based on Baron-Cohen's taxonomy, but provides feedback to players using a dimensional description based on valence and arousal. This requires translation between the two methods of classification.

The Autism Research Centre (ARC) and the Computer Laboratory at the University of Cambridge collected various media of actors displaying 20 different emotions plus neutral as part of the content creation for the ASC-Inclusion game. This is high quality material, carefully recorded, carefully validated and carefully labelled. It is a really useful resource and has proved valuable for the teaching aspects of the ASC-Inclusion game. However, it also indicates some limitations on the use of valence and arousal as indicators in feedback to game players.

The ARC recorded 496 videos of faces. These were then validated on-line, collecting a total of 54,097 assessments, an average of 109 for each video clip. The validation involved a six-way forced choice between the correct label, four foils and 'other'. Clips were deemed to be a reasonable representation of the emotion if at least 50% of labels are correct and no foil is chosen by more than 25% of the assessors. The latter condition turned out to be redundant – no video that achieved 50% correct labels had any foil rated more than 25%. 337 videos passed this qualification, just over two thirds of the total.

The assessors were also asked to rate valence, arousal and intensity for each clip on a five-point Likert scale. Averages were taken for each emotion and converted to the $[-1, +1]$ range used elsewhere in the project. The scatter plot in Figure 2 shows the distribution of the means for valence and arousal of the 21 conditions. This differs significantly from the pattern in Russell's circumplex.

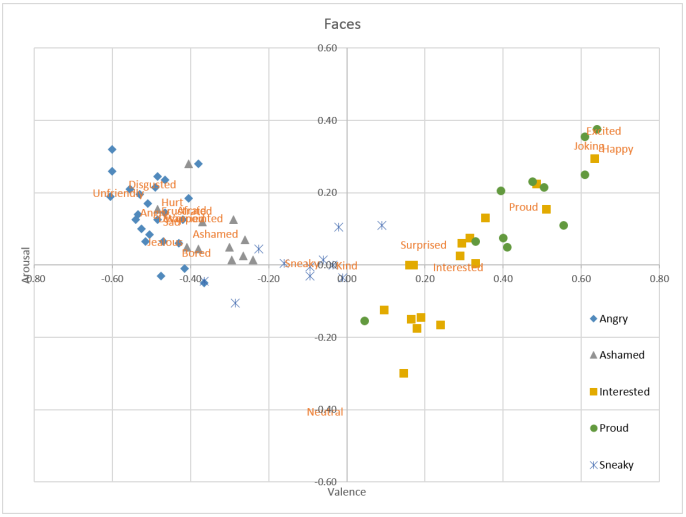

Figure 3. Scatter plots for assessments of individual videos

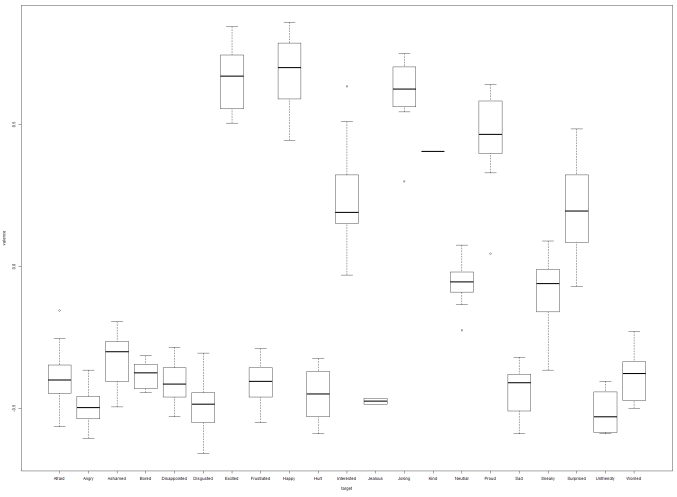

Figure 4. Box and whisker plots showing distributions of valence for the individual emotions

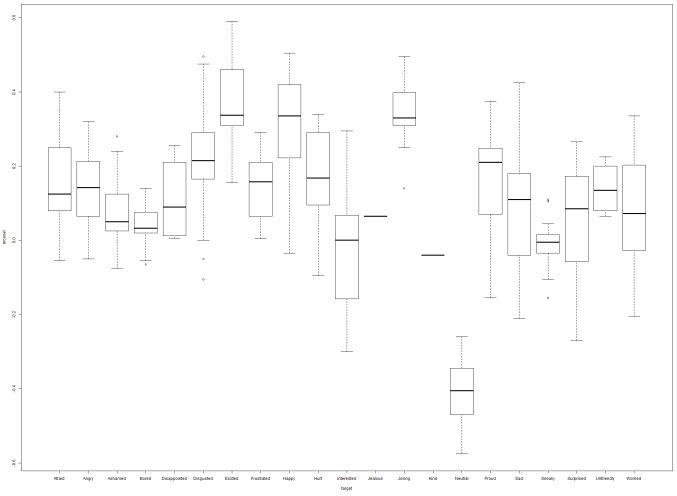

Figure 5. Box and whisker plots showing distributions of arousal for the individual emotions

There were no valid videos for *kind*, so it appears at the origin and can be ignored. *Neutral* averages $(-0.41, -0.06)$ rather than $(0, 0)$, which is curious. There are no other videos averaging negative arousal, which is possibly explained by the use of acted videos. *Sneaky* is fairly isolated. *Surprised*, *interested*, *proud*, *joking*, *happy* and *excited* all lie along a line where arousal and valence appear to be linearly correlated. Perhaps the assessors were not distinguishing arousal from valence particularly carefully, although that is also an effect that we observe in the continuous version of the facial affect inference system. *Bored*, *ashamed*, *jealous*, *sad*, *disappointed*, *worried*, *afraid*, *frustrated*, *angry*, *hurt*, *disgusted* and *unfriendly* are all grouped in a tight cluster in the slightly negative valence, slightly positive arousal area.

The valence and arousal values for valid videos for a single emotion vary considerably. The second scatter plot in Figure 3 superimposes the ranges of values for *angry*, *ashamed*, *interested*, *proud* and *sneaky*. There are significant overlaps for *angry* and *ashamed*, and for *interested* and *proud*. Again, the linear correlation between arousal and valence for the latter two is apparent.

The lack of distinction can be seen in the distributions of the valence and arousal assessments of videos for the various emotions. Figures 4 and 5 show box and whiskers plots for the 21 conditions. In each box, the solid horizontal bar show the median value, the box shows the upper and lower quartiles, and the whiskers show the extreme values except for any outliers that differ from the mean by more than one-and-a-half times the inter-quartile range, which are shown individually.

The bimodal distribution of valence assessments and the lack of distinction in the arousal assessments are apparent.

In preliminary trials of the ASC-Inclusion game the clinical partners observed that participants found it hard to identify a facial expression that would steer their valence and arousal inferences into a target area. The same was also true of the vocal expressions and body gestures. Indeed, this was sufficiently difficult that it would be unhelpful to expect children to do it as part of the game. These plots help us understand why. Even well recorded, well validated videos exhibit such a wide range of valence and arousal values that it is virtually impossible to separate some mental states, still less to locate them accurately in a dimensional space. It simply is not clear what target coordinates should be considered as indicating a 'correct' facial expression.

IV. Affective Monitoring in Adaptive E-Learning

Affective monitoring is particularly challenging when trying to provide feedback in an adaptive e-learning system that is trying to teach emotions. However, there are general implications for all computer applications that feature social interactions. The analysis of the videos recorded for the ASC-Inclusion game presented in the previous section suggest a possible solution.

Problems arise if it is assumed that an emotion can be represented by a single point in valence-arousal space. Instead, it is necessary to accommodate the variation shown in Figure 3.

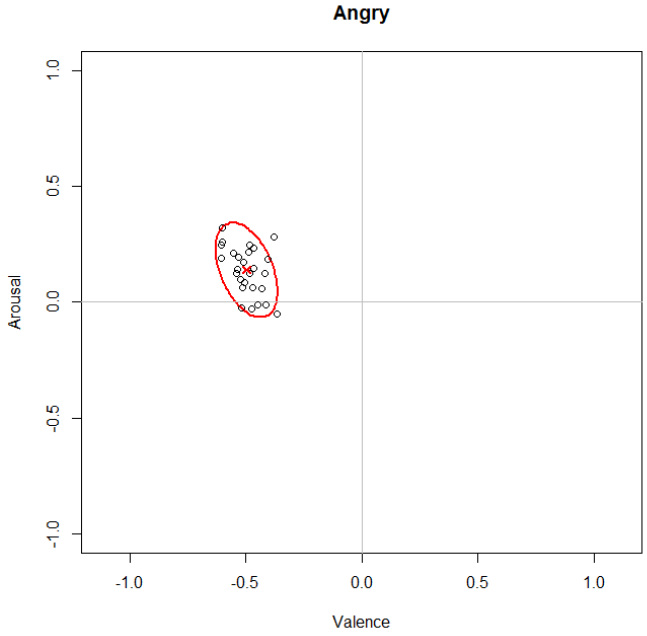

Figure 6. Prediction interval for angry.

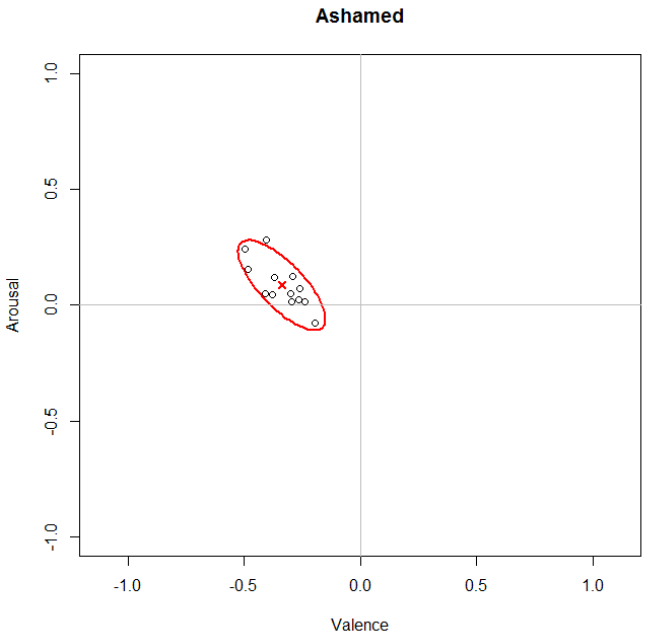

Figure 7. Prediction interval for ashamed.

The principled way to achieve this is to regard each emotion as a distribution in two (or more) dimensions.

Given a set of videos representing an emotion, we can compute $(valence, arousal)$ coordinates either continuously at regular intervals through each video or as averages for each video. In general, we can compute k separate metrics X_i for $i = 1 \ldots k$ in this way, and treat them as k-dimensional samples from a multivariate normal distribution. We can then calculate the k-dimensional mean μ and covariance matrix $\Sigma = Cov(X_i, X_j)$. The *prediction interval* for the distribution is the set of vectors $\mathbf{x}$ satisfying

$$(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \leq \chi_k^2(p) \qquad (1)$$

where $\chi_k^2(p)$ is the quantile function for probability p of the chi-squared distribution with k degrees of freedom. This interval consists of points in the k-dimensional space that lie within the square of a given Mahalanobis distance of the mean.

The actual calculations can be implemented efficiently by deriving a Cholesky decomposition of the covariance matrix. This corresponds to a Principal Component Analysis with the eigenvectors giving the principal axes of the ellipse and the eigenvalues indicating the variance along them.

In the simple case where we are only considering valence and arousal, $k = 2$ and the prediction interval limits $\mathbf{x}$ to the interior of an ellipse. Figures 6-10 show the distributions for the five emotions discussed in the previous section together with their prediction intervals. The ellipses are scaled so that the axes have a length equal to twice the square root of the corresponding eigenvalues, so they extend two standard deviations from their means and the ellipses encompass about 86% of the probability mass.

In the higher-dimensional case where more than two metrics are calculated, the prediction interval is a k-dimensional

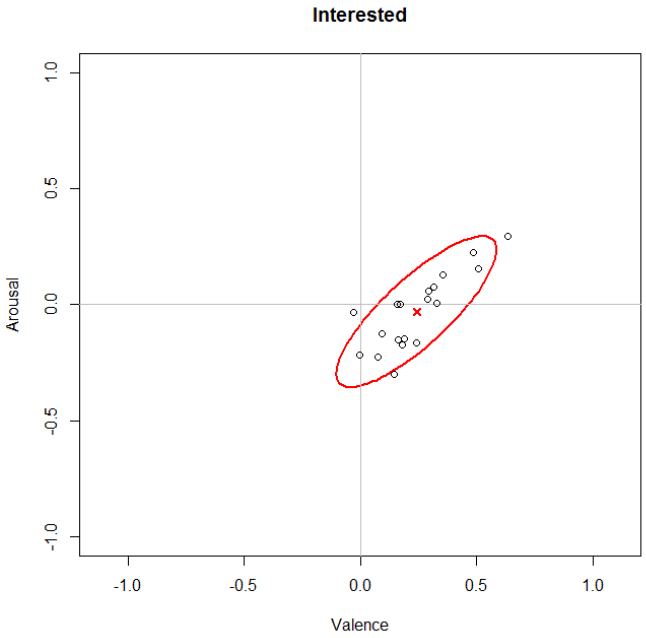

Figure 8. Prediction interval for interested.

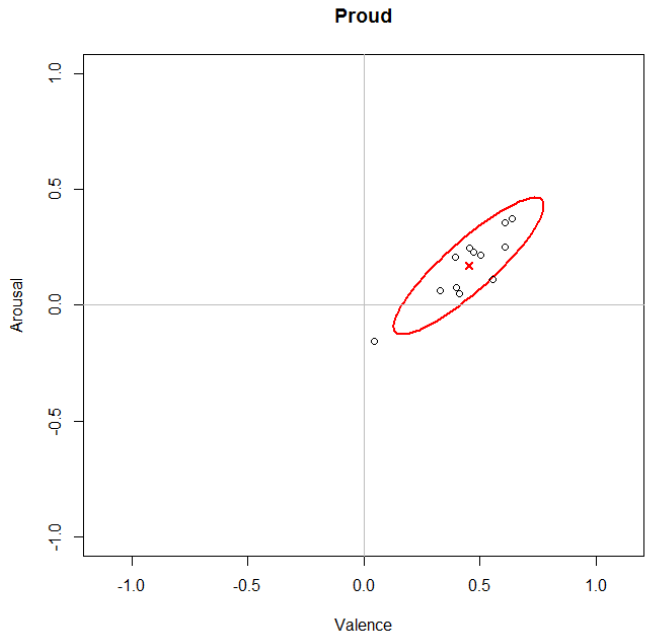

Figure 9. Prediction interval for proud.

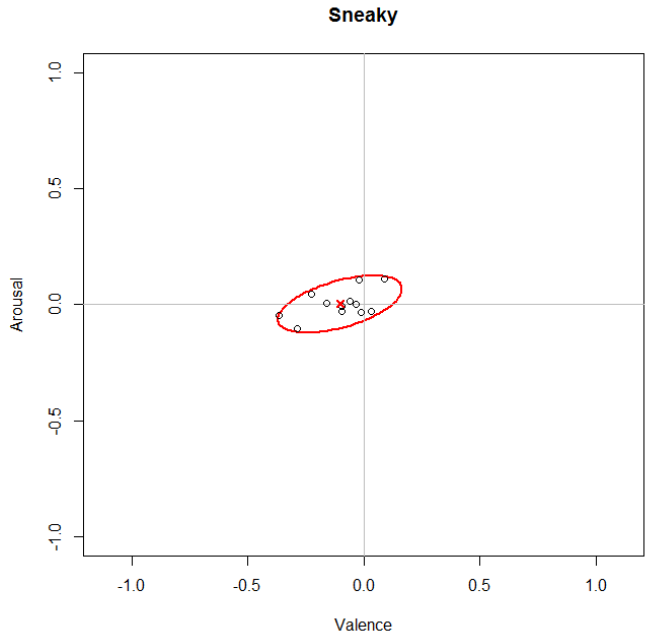

Figure 10. Prediction interval for sneaky.

ellipsoid.

V. Application

This suggests a straightforward mechanism for quantitative assessment of continuous answers in e-learning systems. Question with discrete answers are easy to mark; they are either right or wrong. However, continuous answers are more difficult to assess especially when they involve more than one dimension; we need to know if the value given is sufficiently close to the expected answer. Assessment is even more difficult when there is no 'correct' answer. As Figures 6-10 show, validated performances of emotions can given rise to a wide range of values in valence-arousal space. The solution is to assess a video by finding the smallest prediction interval encompassing it.

For example, Figure 6 shows 28 validated examples of *angry*. The outlier at the top right of the ellipse is on the 99% level of the distribution function, that on the bottom right lies on the 92% level, and there are a couple on the boundary near the 86% level. This distribution is to be expected given the underlying statistical model.

A sample video can now be assessed quantitatively. The video is processed to infer $(valence, arousal)$ coordinates and these are put into Equation 1 to calculate a level on the distribution function. A threshold can be picked to set the accuracy required in the answer. When assessing acted videos, a relatively large probability would be appropriate, perhaps around 85%, corresponding to the interval within two standard deviations of the mean.

However, the same technique could be applied in any exercise involving assessment of a multi-dimensional quantity. Greater accuracy might be expected in a more analytical example, and the probability might be set at 40%, within one standard deviation of the mean.

VI. Conclusions

Assessment in e-learning systems must handle ranges of values as acceptable answers to questions involving multi-dimensional continuous variables. This will typically involve a distribution of acceptable answers rather than simple ranges. The statistical analysis presented in this paper gives a principled approach to quantitative assessment of such answers.

The example of assessing acted facial expressions of emotions is a particularly challenging test. The expressions are ambiguous and a single acted video can reasonably represent several mental states. The approach presented here only tries to determine how good a representation the video is of a particular emotion, and does not exclude other possible interpretations.

It is worth noting that this analysis only provides summative assessment with very little formative guidance. Given an evaluation in $(valence, arousal)$ space that lies outside the prediction interval, the difference from the desired mean can be calculated and guidance in the form "Look less happy" or "Look more animated" given. However, this may be hard for the student to understand. We are investigating other approaches that use Parzen window estimation to produce

explanations in terms of more familiar attributes such as movements of the mouth or eyes.

In the wider context of on-line education, responsive systems should monitor the expressions of pupil's faces, make inferences about whether they are confused or understanding, interested or bored, and adapt the lesson accordingly. The statistical approach here allows the on-line system to set thresholds at which interventions might be made to adapt the lesson. This would allow such a system to cater individually for students with different analytical abilities, attention spans and general approaches to learning.

ACKNOWLEDGEMENTS

The work described here was undertaken by a large team in the Computer Laboratory at the University of Cambridge – notably Shazia Afzal, Tadas Baltrušaitis, Ntombi Banda, Ian Davies, Rana el Kaliouby and Marwa Mahmoud.

The videos and crowd-sourced validation used in this analysis were organised by Helen O'Reilly, Delia Pigat and Amandine Lassalle working with Simon Baron-Cohen in the Autism Research Centre at Cambridge.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 289021.

REFERENCES

[1] S. Afzal and P. Robinson, "Measuring affect in learning environments - motivation and methods," in *IEEE International Conference on Advanced Learning Technologies*, 2010.

[2] S. Afzal and P. Robinson, "Designing for automatic affect inference in learning environments," *Journal of Educational Technology and Society*, vol. 14, October 2011.

[3] B. W. Schuller, E. Marchi, S. Baron-Cohen, H. O'Reilly, P. Robinson, I. P. Davies, O. Golan, S. Fridenson, S. Tal, S. Newman, N. Meir-Goren, R. Shillo, A. Camurri, and S. Piana, "ASC-Inclusion: Interactive emotion games for social inclusion of children with Autism Spectrum Conditions," in *Intelligent Digital Games for Empowerment and Inclusion*, May 2013.

[4] S. Newman, O. Golan, S. Baron-Cohen, S. Bolte, A. Baranger, B. W. Schuller, P. Robinson, A. Camurri, N. Meir-Goren, M. Skurnik, S. Fridenson, S. Tal, E. Eshchar, H. O'Reilly, D. Pigat, S. Berggren, D. Lundqvist, N. Sullings, I. P. Davies, and S. Piana, "ASC-Inclusion — a virtual environment teaching children with ASC to understand and express emotions," in *International Meeting for Autism Research*, May 2014.

[5] C. Darwin, *The expression of the emotions in man and animals*. London: John Murray, 1872.

[6] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face*. New York: Pergamon Press, 1972.

[7] P. Ekman and W. V. Friesen, *Facial action coding system: a technique for the measurement of facial movement*. Palo Alto: Consulting Psychologists Press, 1978.

[8] S. Baron-Cohen, O. Golan, S. Wheelwright, and J. Hill, "Mind reading: the interactive guide to emotions." DVD, 2004.

[9] R. el Kaliouby, P. Robinson, and L. S. Keates, "Temporal context and the recognition of emotion from facial expression," in *International Conference on Human-Computer Interaction*, Lawrence Erlbaum Associates, 2003.

[10] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[11] R. el Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," in *Computer Society Conference on Computer Vision and Pattern Recognition*, (Washington, DC), p. 154, 2004.

[12] R. el Kaliouby and P. Robinson, "Generalization of a vision based computational model of mind-reading," in *Affective Computing and Intelligent Interaction*, (Beijing, China), pp. 582–589, 2005.

[13] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *International Conference on Computer Vision*, December 2013.

[14] T. Baltrušaitis, N. Banda, and P. Robinson, "Dimensional affect recognition using continuous conditional random fields," in *IEEE Conference on Automatic Face and Gesture Recognition*, April 2013.