Empirical analysis of continuous affect

Peter Robinson & Tadas Baltrušaitis
Computer Laboratory
University of Cambridge
England
{peter.robinson,tadas.baltrusaitis}@cl.cam.ac.uk

***Abstract*—Automatic analysis of affect from facial expressions has been extensively studied, but most work has considered only a small set of discrete emotions, typically Ekman's six basic emotions, or a small number of continuous measures, typically valence and arousal. We have developed a system that accommodates a much larger vocabulary of discrete emotions and links them with continuous measures that have been aggregated over a few seconds. The approach has a sound theoretical basis in multi-dimensional statistics, making it both principled and robust, while a graphical presentation makes it easy to understand.**

Keywords*—*affect measurement; dimensional affect; discrete affect; variance.

I. Introduction

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 many years, using a variety of techniques – rule-based classifiers, neural networks, support vector machines, and Bayesian classifiers[1] – but often only considering Ekman's six basic emotions or a couple of continuous measures. Recognising the complex, cognitive mental states that arise in everyday life is more difficult, but probably more useful as part of general interaction with computer systems [2], [3]. We have been exploring the relationship between complex, cognitive mental states and continuous measures of valence and arousal.

Categorical descriptions of mental states are part of our everyday language, which gives them the advantage of being commonly understood and easy to interpret. On the other hand, continuous measures arise naturally when applying computational techniques to the analysis of affect, and have the advantage of precision. Resolving this tension between intuition and precision is difficult [4]. Our statistical analysis has revealed two problems with a naïve translation between categorical and continuous classifications of emotion. The continuous valence and arousal classifications for valid videos of a single emotion vary considerably, and the range of values for distinct emotions overlap considerably.

We have devised a new approach using statistical techniques. We consider continuous measures of affect as distributions in a multi-dimensional space. The multi-dimensional means and variances of continuous measurements taken at video frame rates provide a way of aggregating information about facial expressions over a few seconds, and comparing them with other aggregates for the same or different emotions.

TABLE I
The 24 high-level categories in Baron-Cohen's taxonomy [8]

Afraid	**Angry**	Bored	Bothered
Disbelieving	**Disgusted**	Excited	Fond
Happy	Hurt	Interested	Kind
Liked	Romantic	**Sad**	Sneaky
Sorry	Sure	**Surprised**	Thinking
Touched	Unfriendly	Unsure	Wanting

We believe that this gives a principled and practical approach to analysing continuous measures of affect.

The remainder of this paper gives a brief summary of these two models of affect. Then Section III present the mathematical basis of our statistical measures for continuous affect in two or more dimensions. Section IV presents the EU-Emotion Stimulus Set used for our empirical analysis and Section V presents an analysis using the new statistical measures. Finally Section VI discusses the implications of this approach.

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. Affective computing needs to consider the more common but subtler mental states experienced in everyday life.

A broader taxonomy of human emotions has been developed by Simon Baron-Cohen 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. Table I shows the 24 high-level categories, including Ekman's basic emotions in bold type. The 18 additional complex mental states seem to require a second or two of continuous observation to be recognised by humans, rather than the single image that suffices for basic emotions [9].

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

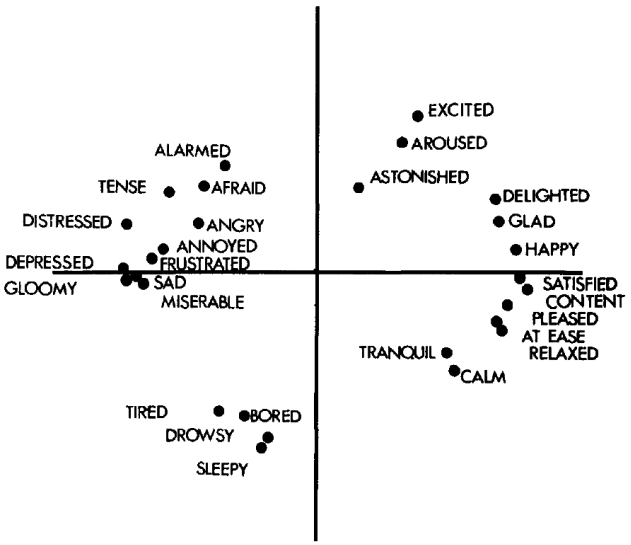

Fig. 1. Russell's circumplex showing two principal components of 28 affect words [10]

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 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. These measures are *continuous* in two senses. The coordinates are quantified on a continuous scale, often between -1 and $+1$, and they are measured continuously in time, or at least at a rate approximating to video frame rates.

The horizontal axis is usually referred to as *valence* and the vertical axis as *arousal*. Further axes have been given names like *dominance, expectation, intensity* and *tendency*. This has led to a popular belief that emotions can be measured precisely by coordinates in a suitably high-dimensional space. Unfortunately, our experiments show that analysis is more complicated in practice. One obvious reason is that the expression of an emotion will move through phases from neutral through onset to apex, and then back through offset to neutral again. Indeed, spontaneous expressions may move between appearances without returning to neutral. It makes more sense to consider a trajectory traced over time in the multi-dimensional space, or simply a set of coordinates following some statistical distribution.

III. Variance in Dimensional Affect

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

$$(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \leq \chi^2_k(p) \quad (1)$$

where $\chi^2_k(p)$ 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 a k-dimensional ellipsoid 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. If these ellipses are scaled so that the axes have a length equal to twice the square root of the corresponding eigenvalues, they will extend two standard deviations from their means and encompass about 86% of the probability mass. This gives a useful visualisation of the two-dimensional mean and variance in $(valence, arousal)$ space.

IV. The EU-Emotion Stimulus Set

Our analysis has used a database collected and validated for the European ASC-Inclusion project [11], [12]. The overall project built and evaluated an internet-based game platform, intended for children with Autism Spectrum Conditions (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.

A major component in the ASC-Inclusion game is a collection of model depictions of the mental states being taught. 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 this content. 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

TABLE II
EMOTION CATEGORIES IN THE EU-EMOTION STIMULUS SET

Category	# videos	# ratings	# accepted	Duration/s
Afraid	17	2 113	17	139
Angry	17	1 997	10	88
Ashamed (Sorry)	8	829	5	46
Bored	8	895	7	67
Disappointed (Sad)	10	1 260	6	55
Disgusted	18	2 025	14	111
Excited	9	985	9	75
Frustrated (Angry)	12	2 017	11	89
Happy	14	1 498	11	84
Hurt	10	1 106	8	72
Interested	11	1 360	8	63
Jealous (Wanting)	7	774	0	
Joking (Happy)	9	1 083	9	80
Kind	9	969	0	
Neutral	17	1 927	17	95
Proud (Happy)	11	1 348	7	54
Sad	14	1 506	13	103
Sneaky	11	1 221	8	71
Surprised	18	2 249	16	71
Unfriendly	9	1 156	0	
Worried (Afraid)	8	759	5	46
Total	**247**	**29 077**	**181**	**1 407**

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.

A subset of the videos has been published as the EU-Emotion Stimulus Set (EESS) [13]. This includes 181 videos encompassing 18 of the emotions, with a total duration of almost 24 minutes. Some of these correspond directly to high-level categories in the Baron-Cohen taxonomy, and some consider only subsets. Table II shows the number of videos considered as candidates for each emotion and also the number actually included in the corpus. The parent categories of subsets are shown in parentheses. The final column shows the total duration of the validated videos (in seconds).

V. ANALYSIS OF VARIANCE IN EESS

The Cambridge Facetracker [14], [15] was used to determine continuous ratings of valence and arousal for the validated videos in the EU-Emotion Stimulus Set. This was trained using the Denver Intensity of Spontaneous Facial Action (DISFA) database [16] to infer action units, and then the SEMAINE database [17] to infer valence and arousal. The videos were analysed at their original 30fps frame rate, giving a total of 42 216 classifications.

The first experiment calculated aggregate means and prediction intervals for each for the 247 videos taken separately. Figures 2, 3 and 4 show the means and 2σ prediction intervals in $(valence, arousal)$ space for six instances each of *Neutral, Happy* and *Sad.* The graphical presentation immediately reveals several interesting features. First, there is substantial variance within each single video, and this is mainly in the vertical, arousal axis. Arousal and intensity of expression are correlated, so this can be explained by the aggregation of measurements over the onset and offset of the expressions as

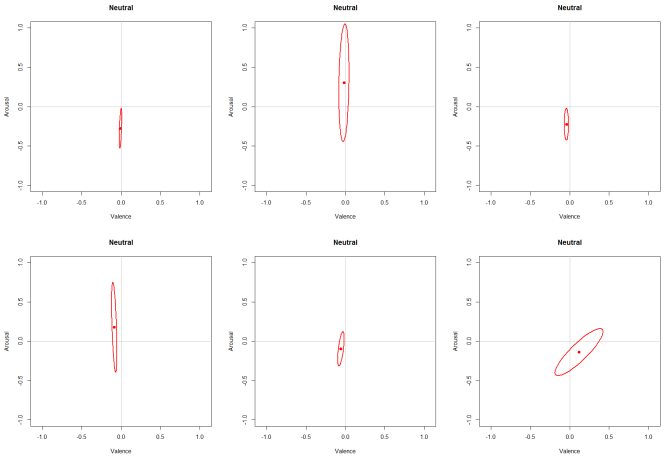

Fig. 2. 2σ prediction intervals for six instances of *Neutral*

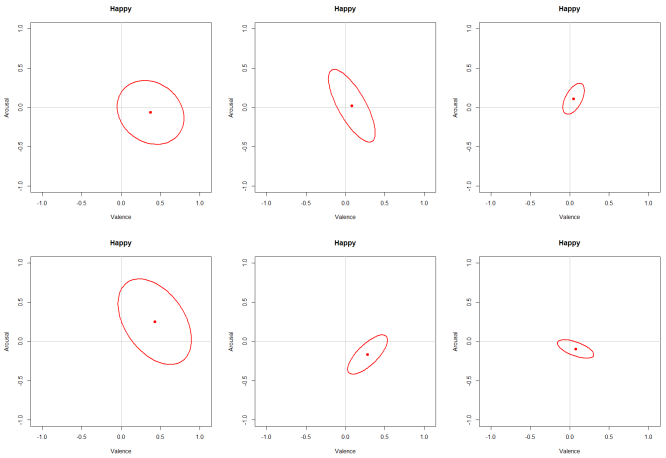

Fig. 3. 2σ prediction intervals for six instances of *Happy*

well as the peak intensity. It also confirms other observations that the face indicates valence while the voice indicates arousal [1].

Secondly, the analyses of videos for *Happy* show variation in valence as well as arousal. This can be explained by the brevity of the apex period for the distinctive smile indicating happiness.

The second experiment calculated aggregate means and prediction intervals for all the videos representing each emotion. Figure 5 shows the 2σ prediction intervals in $(valence, arousal)$ space for all 18 of the emotions in EESS, together with the means for each individual video. Again, the variance within each emotion is considerable, demonstrating that there is no simple mapping between discrete emotions and continuous measures.

Even the locations of the means differ considerably from Russell's original circumplex in Figure 1. This is confirmed by the mean $(valence, arousal)$ coordinates shown in Table III. Indeed, some of these are very different, although a lot of the difference is in the measurement of arousal. The classification

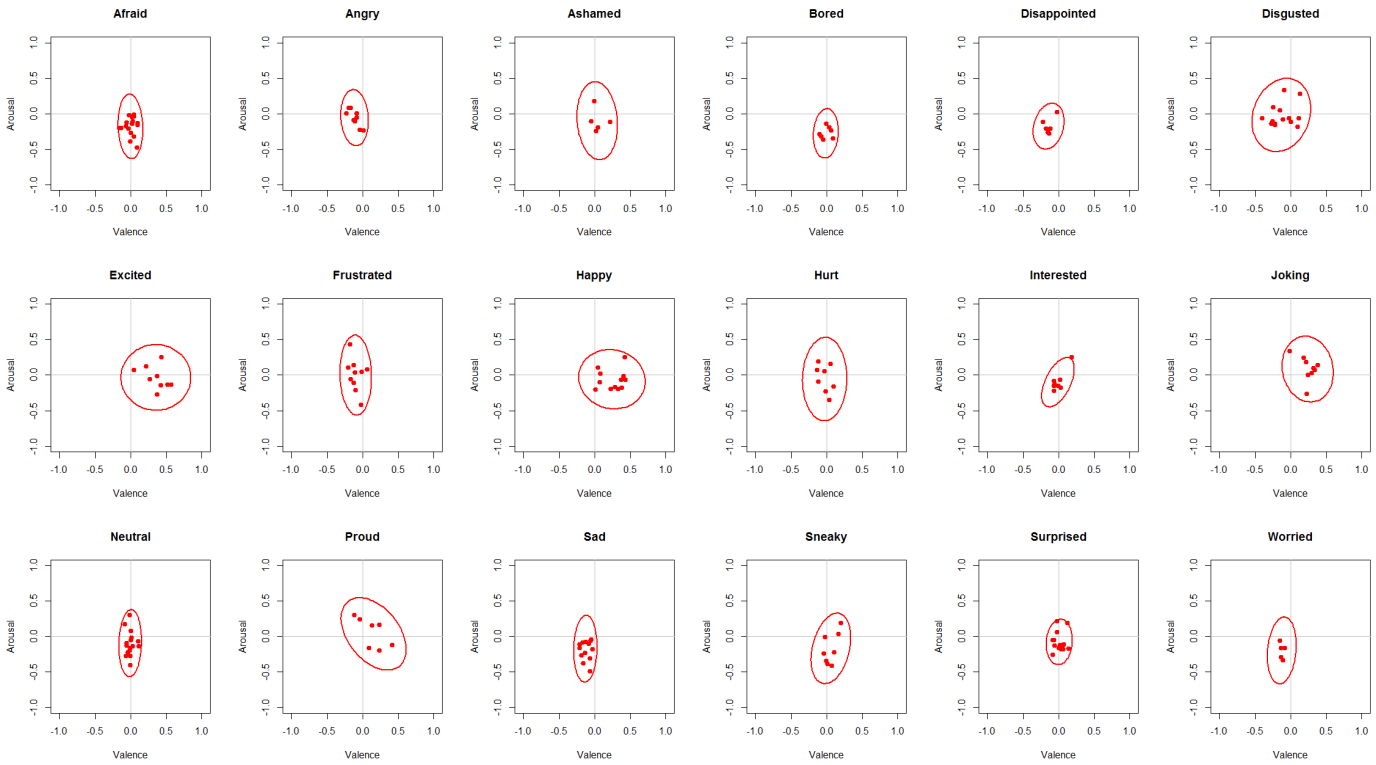

Fig. 5. 2σ prediction intervals in $(valence, arousal)$ space for 18 emotions

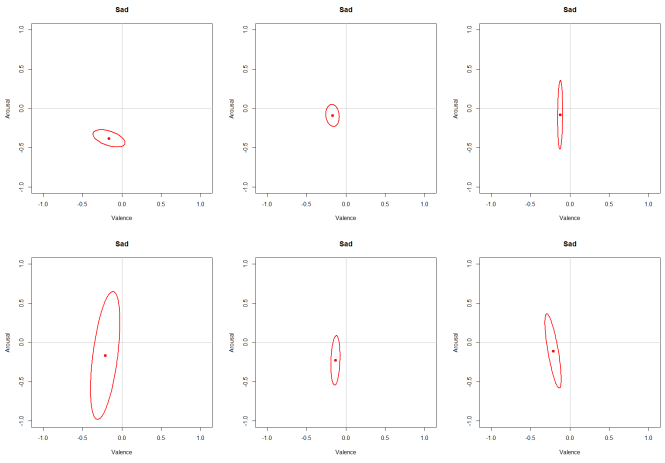

Fig. 4. 2σ prediction intervals for six instances of *Sad*

TABLE III
MEAN $(valence, arousal)$ COORDINATES FOR THE EU-EMOTION STIMULUS SET

Category	Valence Facetracker	Arousal Facetracker	Valence Validation	Arousal Validation
Afraid	0.00	-0.18	-0.40	0.15
Angry	-0.11	-0.06	-0.49	0.15
Ashamed	0.04	-0.10	-0.34	0.09
Bored	-0.01	-0.28	-0.39	0.03
Disappointed	-0.14	-0.18	-0.40	0.13
Disgusted	-0.13	-0.02	-0.50	0.22
Excited	0.36	-0.03	0.66	0.37
Frustrated	-0.10	0.00	-0.41	0.15
Happy	0.24	-0.06	0.69	0.32
Hurt	-0.03	-0.06	-0.45	0.17
Interested	-0.01	-0.10	0.29	-0.01
Joking	0.24	0.08	0.62	0.33
Neutral	0.00	-0.10	-0.06	-0.41
Proud	0.15	0.04	0.45	0.16
Sad	-0.13	-0.17	-0.45	0.12
Sneaky	0.06	-0.17	-0.11	0.01
Surprised	0.01	-0.08	0.20	0.06
Worried	-0.12	-0.20	-0.42	0.13

of the mean arousal for *Neutral* at -0.10 suggests that there may be a negative bias in the training data.

Assessors validating EESS were also asked to rate valence, arousal and intensity for each video on a five-point Likert scale. These were averaged for each emotion and converted to a $[-1, +1]$ range, and are shown in the final two columns of Table III and as distributions in Figure 6 [18]. These means also differ from both Russell's original circumplex and the means calculated by the automatic classification.

To some extent this is explained by scaling of the values. Figure 7 shows scatter diagrams for the automatic classifications of valence and arousal against the human assessments. The two sets of figures for valence have a Pearson R^2 correlation coefficient of 0.78, but those for arousal only 0.15. Again, this reflects the observation that the face is a clearer indicator of valence than arousal, and there is good

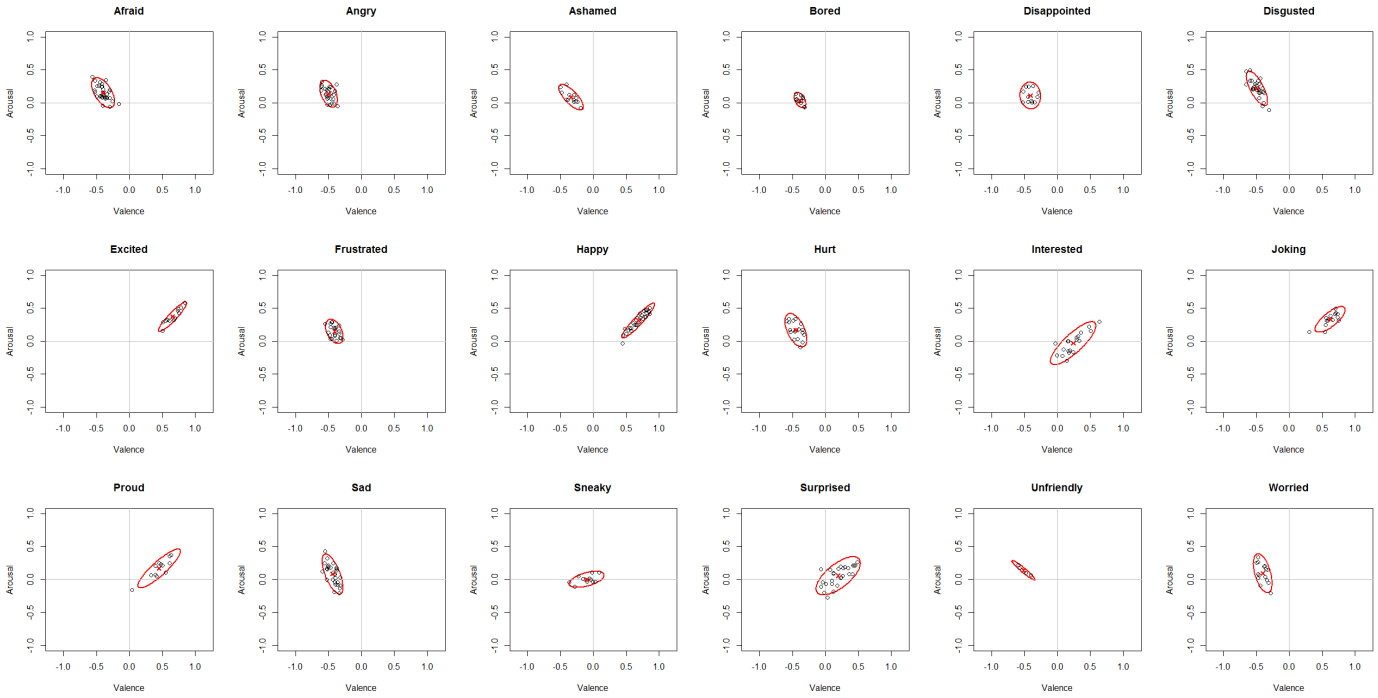

Fig. 6. 2σ prediction intervals in $(valence, arousal)$ space from manual assessment [18]

correspondence between the automatic and human assessments of arousal. The difference in magnitude can be attributed to human assessors recording the maximum valence during the expression's apex while the automatic classification is averaged over the whole video. A Pearson product-moment correlation test shows that the rank orderings of valence are strongly correlated ($p \ll 0.001$), but that those for arousal are not ($p > 0.05$).

This reveals a major problem with much research into affective computing. The results are only as good as the training data. The SEMAINE database used for training was carefully collected and validated, with multiple ratings achieving high levels of agreement, but still only achieving Pearson R^2 correlations around 0.5, although this can be attributed partly to timing errors in the annotation [19]. Any human translation from qualitative examination of videos into quantitative measurements in continuous time is fraught with difficulty.

VI. Discussion

Our statistical analysis has revealed two problems with a naïve translation between categorical and continuous classifications of emotion. The continuous valence and arousal classifications for valid videos of a single emotion vary considerably, and the range of values for distinct emotions overlap considerably. Figure 5 illustrates both of these difficulties.

These results indicate that attempts to make a universal "emotion meter" are unlikely to succeed. It is necessary to identify a particular application domain and then design a classifier that operates well in a single, specialised context.

For example, 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. This analysis of the EESS videos recorded for the ASC-Inclusion game suggests a possible solution.

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 has more general implications for the use of affective feedback to guide social interactions in adaptive e-learning systems.

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.

However, it is reasonable to assess an example piece of acting as acceptable if its prediction interval adequately overlaps that of the aggregate representatives of the intended emotion. We simply check that the two distributions are

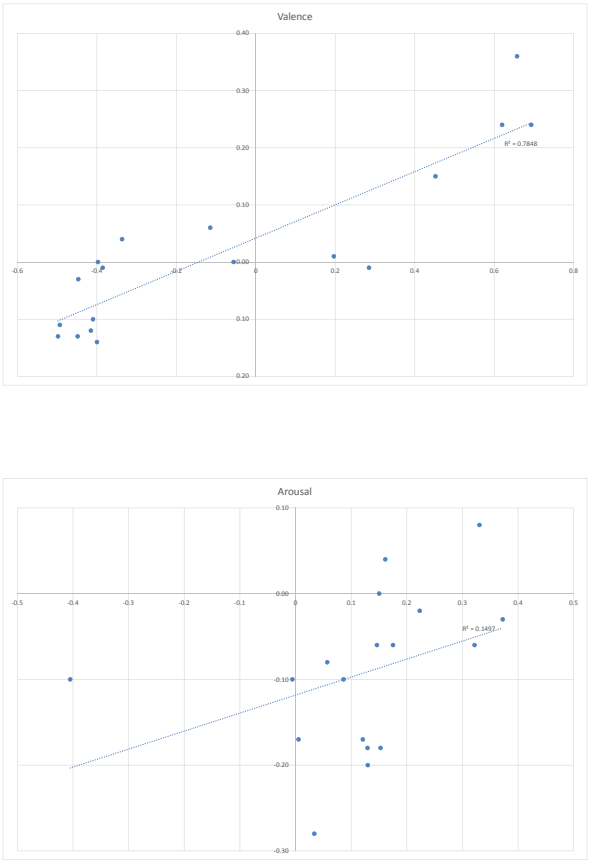

Fig. 7. Scatter diagrams of mean automatic classifications of valence and arousal plotted against human assessments.

sufficiently similar. Our statistical analysis provides the basis for a quantitative measurement of this overlap using statistical tests for change detection in multi-dimensional data [20].

VII. Conclusion

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 5. One principled way to achieve this is to regard each emotion as a distribution in two (or more) dimensions. We have developed a statistical measure that is principled and robust, but still easy to understand. This can form the basis of a system for interpreting classifications in a wide variety of applications as long as the context of each individual application is clearly established.

The work also poses interesting questions for future work. While basic emotions can be inferred from still images, complex mental states require continuous video evidence for a second or two to be understood [9]. Analysing these gives trajectories in a dimensional space which require temporal models with dynamics to achieve reasonable recognition. Previous work has used Dynamic Bayesian Networks for this [2], but more recent approaches such as recurrent neural networks or Continuous Conditional Neural Fields [14] look promising.

Acknowledgement

Many colleagues in the Computer Laboratory at the University of Cambridge have contributed to this work, while Simon Baron-Cohen and his colleagues in the Autism Research Centre have provided numerous technical insights. Parts of the work have been undertaken with support from the EPSRC, the European Commission, the Cambridge-MIT Institute, the Gates Cambridge Trust, Hanson Robotics, the Yousef Jameel Studentship, the Qualcomm Studentship, Thales Research & Technology, Toyota ITC, Vicon UK, the Neil Wiseman Fund and the University of Cambridge. ASC-Inclusion received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 289021

References

[1] J. F. Cohn and F. D. la Torre, *Handbook of affective computing*, ch. Automated face analysis for affective computing, pp. 131–150. Oxford University Press, 2015.

[2] 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.

[3] P. Robinson and R. el Kaliouby, "Computation of emotions in man and machines," *Phil. Trans. R. Soc. B*, vol. 364, pp. 3441–3448, December 2009.

[4] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input," *Image and Vision Computing*, vol. 31, pp. 120–136, 2013.

[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] 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.

[12] 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.

[13] H. O'Reilly, D. Pigat, S. Berggren, S. Fridenson, S. Tal, O. Golan, SvenBolte, S. Baron-Cohen, and D. Lundqvist, "The EU-Emotion Stimulus Set: A validation study," *Behavior Research Methods*, vol. (under review), 2015.

[14] T. Baltrušaitis, L.-P. Morency, and P. Robinson, "Continuous conditional neural fields for structured regression," in *European Conference on Computer Vision*, (Zurich, Switzerland), Sept. 2014.

[15] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *Facial Expression Recognition and Analysis Challenge*, (Ljubljana, Sovenia), May 2015.

[16] S. M. Mavadati, M. Mahoor, K. Bartlett, P. Trinh, and J. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, pp. 151–160, Apr.-June 2013.

[17] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, pp. 5–17, Jan.-Mar. 2011.

[18] P. Robinson, "Modelling emotions in an on-line educational game," in *International Conference on Control, Decision and Information Technologies*, (Metz, France), Nov. 2014.

[19] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. 6, pp. 97–108, April-June 2015.

[20] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Statistical change detection for multi-dimensional data," in *ACM Conference on Knowledge discovery and data mining*, pp. 667–676, 2007.