Criteria and metrics for thresholded AU detection

Jeffrey M. Girard and Jeffrey F. Cohn
Department of Psychology, University of Pittsburgh
Pittsburgh, PA, 15260 USA
jmg174@pitt.edu, jeffcohn@pitt.edu

Abstract

Implementing a computerized facial expression analysis system for automatic coding requires that a threshold for the system's classifier outputs be selected. However, there are many potential ways to select a threshold. How do different criteria and metrics compare?

Manually FACS coded video of 45 clinical interviews (Spectrum dataset) were processed using person-specific active appearance models (AAM). Support vector machine (SVM) classifiers were trained using an independent dataset (RU-FACS). Spectrum sessions were randomly assigned to training (n=32) and testing sets (n=13). Six different threshold selection criteria were compared for automatic AU coding.

Three major findings emerged: 1) Thresholds that attempt to balance the confusion matrix (using kappa, F1, or MCC) performed significantly better on all metrics than thresholds that select arbitrary error or accuracy rates (such as TPR, FPR, or EER). 2) AU detection scores for kappa, F1, and MCC were highly intercorrelated; accuracy was uncorrelated with the others. And 3) Kappa, MCC, and F1 were all positively correlated with base rate. They increased with increases in AU base rates. Accuracy, by contrast, showed the opposite pattern. It was strongly negatively correlated with base rate.

These findings suggest that better automatic coding can be obtained by using threshold-selection criteria that balance the confusion matrix and benefit from increased AU base rates in the training data.

1. Introduction

In behavioral science, facial expression analysis has relied on human coders trained in observational measurement of facial actions and to a lesser extent on facial EMG from selected muscle regions [1]. Of the various approaches, the Facial Action Coding System (FACS) is the most comprehensive [2, 3]. FACS defines 30+ anatomically separable facial actions, referred to as action units (AUs), which may occur individually or in combinations to describe nearly all possible facial expressions. Because of its descriptive power, FACS has been widely used in behavioral science and has influenced efforts in computer vision and graphics (e.g. MPEG-4 facial animation parameters) [4].

Manual FACS coding, however, is time-consuming, costly, and restrictive. Computerized systems for automatic facial expression analysis (AFA) have been proposed to address these limitations [5, 6]. Initial efforts in AFA have shown promise in automatic measurement of pain [7, 8], emotion [9, 10], psychopathology [11, 12], parent-infant communication [13], and adult attachment [14] among others.

To evaluate the accuracy of AFA systems, investigators have often used receiver operating characteristic (ROC) curves or Precision-Recall curves, both of which are popular in signal detection and information retrieval. These curves and many popular performance metrics derived from them, such as area under the curve, are threshold-independent. That is, they allow for an evaluation of the system as a whole, given any possible cutoff point for its classifier output. However, to implement an AFA system for automated coding requires the selection of one specific threshold (i.e. a threshold is required to convert the classifier's continuous output into binary classifications). Thus, ROC and Precision-Recall curves (and the areas under them) tell us little about the performance of a given implementation/configuration of an AFA system. For this purpose, a threshold for the system's classifier outputs must be selected and a threshold-specific measure of performance utilized.

Various criteria have been used for threshold selection. Many involve selecting an *a priori* accuracy or error rate. For instance, setting the true positive rate (TPR) at 80% or the false positive rate (FPR) at 10% are two common criteria. Another is equal error rate (EER). By selecting the threshold that is equally likely to make false-positive and false-negative errors, this criterion seeks to minimize system bias. Other approaches seek to maximize performance metrics that represent the entire confusion matrix. These include the F1 score, the Matthews correlation coefficient (MCC; aka phi coefficient), and Cohen's kappa. (See the appendix for more information.)

The present paper asks how these different threshold-selection criteria compare. To inform about threshold selection and to encourage emerging standards in choice

978-1-4673-0063-6/11/$26.00 ©2011 IEEE

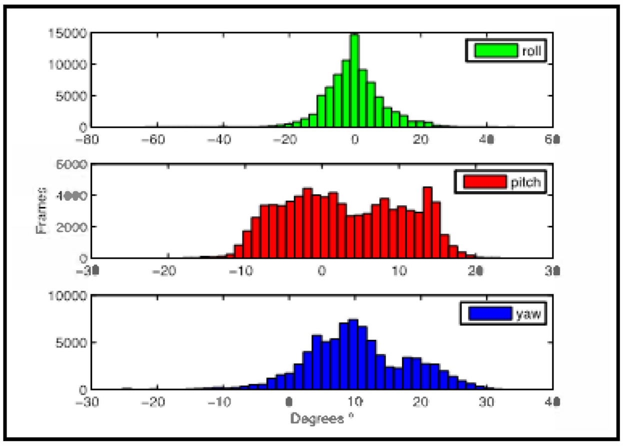

Figure 1: Head rotation angle estimation from Spectrum database.

of performance metrics, we evaluated six different threshold selection criteria on four different threshold-specific performance metrics. For evaluation, we included accuracy and the F1 score (the harmonic mean of precision and recall) due to their widespread use in machine learning. Cohen's kappa and the Matthews Correlation Coefficient (MCC, also called the phi coefficient) were also included as metrics that attempt to control for chance agreement.

We test the hypothesis that better performance can be achieved by using more data-driven criteria for threshold-selection. Rather than choosing an arbitrary accuracy or error rate, thresholds may be selected that maximize the performance metrics that attempt to balance the entire confusion matrix. Thus, we compare "error/accuracy rate" criteria (TPR80, FPR10, and EER) to "confusion matrix" criteria that maximize the F1 score (maxF1), kappa (maxKappa), or MCC (maxMCC) in the training set. Classifiers were first trained on an independent FACS-coded dataset (RU-FACS) [15]. We then applied each of the threshold criteria to a training set from the FACS-coded Spectrum dataset. In a test set from Spectrum, we compared the inter-correlation between each of the four metrics; we then evaluated six thresholds on these metrics. In addition, we evaluated the correlation of each metric with the base rate of AUs. To our knowledge, this is the first effort to compare threshold-selection criteria and performance metrics for automated AU detection.

2. Methods

2.1. Image data and FACS coding

Image data were from the Spectrum database [11]. Twenty-nine men and women (69% female, 86% Caucasian, average age 43 years) with major depressive disorder [16] were interviewed with the Hamilton Rating Scale for Depression [17]. They were interviewed on one or more occasions to assess symptom severity of depression over the course of treatment. The interviews were recorded using four hardware-synchronized analogue cameras and digitized into 640x480 pixel arrays with 24-bit resolution. Video from one of the four cameras located about 15 degrees to the participant's right was used in the current study. Non-frontal pose and moderate head motion, estimated using a structure from motion algorithm described below, were common (Figure 1).

We selected for analysis seven AUs from the first 45 FACS-coded sessions. The AUs were ones theoretically related to smiles or negative affect: AU 4, AU 6, AU 10, AU 12, AU 14, and AU 17 [18]. AU 4 (corrugator) occurs in concentration and negative affect; AU 6 (orbicularis oculi) in both positive and negative affect; AU 10 (levator labii superioris) in disgust; AU 12 (zygomatic major) in positive affect; AU 14 (buccinnator) in contempt; AU 15 (triangularis) in sadness and smile control; and AU 17 (mentalis) in anger, sadness, and smile control. With the exception of AU 6 and AU 15, the base rates were 18% or higher. Base rates for the latter two were only 11% and 4%, respectively. Intensity for all of the AU tended to be low relative to that in RU-FACS [19].

2.2. Automatic facial image analysis

Automatic facial image analysis included three steps. These were 1) extract the face shape and appearance using a person-specific active appearance model (AAM) [20]; 2) normalize shape and appearance to control for variation due to rigid head motion (e.g., turning toward or away from other participants); and 3) detect FACS action units.

2.2.1. Active Appearance Model. AAMs decouple shape and appearance of a face image. Given a pre-defined linear shape model with linear appearance variation, AAMs align the shape model to an unseen image containing the face and facial expression of interest. To train an AAM for each participant, approximately 3% of keyframes were manually labeled during a training phase. The remaining frames were automatically aligned using a gradient-descent AAM fit described in [21, 22].

The *shape* $\mathbf{s}$ of an AAM is described by a 2D triangulated mesh. In particular, the coordinates of the mesh vertices define the shape $\mathbf{s}$ [6]. These vertex locations correspond to a source appearance image, from which the shape is aligned. Since AAMs allow linear shape variation, the shape $\mathbf{s}$ can be expressed as a base shape $\mathbf{s}_0$ plus a linear combination of m shape vectors

$$\mathbf{s}_{i} \quad \mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} p_i \mathbf{s}_i$$

where the coefficients $\mathbf{p} = (p_1, \ldots, p_m)^{T}$ are the shape parameters (See Figure 2). Additionally, a global normalizing transformation (in this case, a geometric similarity transform) is applied to $\mathbf{s}$ to remove variation due to rigid motion (i.e. translation, rotation, and scale). The parameters p_i are the residual parameters representing variations associated with the actual object shape (e.g. mouth opening and eye closing). Given a set of training shapes, Procrustes alignment is employed to normalize these shapes and estimate the base shape $\mathbf{s}_0$, and Principal Component Analysis (PCA) is then used to obtain the shape and appearance basis eigenvectors $\mathbf{s}_i$ [20]. A non-rigid structure from motion algorithm is used to estimate head pose parameters (e.g. pitch, etc.) [22, 23].

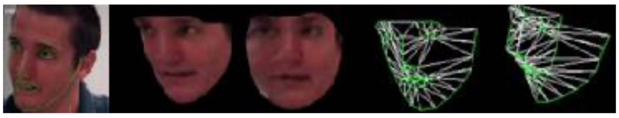

Figure 2: From left to right, an example of source video, 2D similarity and piece-wise normalized appearance, and 2D and 3D normalizations of face shape.

2.2.2. AAM features. Although “person-specifc” AAM models were used for tracking, a global model of the shape variation across all sessions was built to obtain the shape basis vectors and corresponding similarity normalized coefficients p_i. A model common to all subjects is necessary to ensure that the meaning of each of the coefficients is comparable across sessions. 95% of the energy was retained in the PCA dimensionality reduction step, resulting in 10 principal components or shape eigenvectors.

2.3. Action unit detection

Action units were detected using support vector machine classifiers (SVM) [21]. SVMs attempt to find the hyper-plane that maximizes the margin between positive and negative observations for a specified class. For AAM shape and appearance coefficients, they seek to maximize the boundary between each action unit (e.g., AU 6) and all instances of other action units including neutral faces (i.e., AU 0 in FACS). Both shape (66 tracked landmarks) and appearance were used for AU detection.

The appearance features were based on recent work that used fixed-scale-and-orientation SIFT descriptors [24]. Intuitively, the histogram of gradient orientations calculated in SIFT has the potential to capture much of the information that is described in FACS (e.g., the markedness of the naso-labial furrows, the direction and distribution of wrinkles, the slope of the eyebrows). At the same time, the SIFT descriptor has been shown to be robust to certain illumination changes and small errors in localization [25]. As noted above, an affine texture transformation was applied to each image so as to warp the texture into this canonical reference frame to provide some robustness to the effects of head motion. Once the texture was warped into this fixed reference, SIFT descriptors were computed around the outer outline of the mouth (11 points for lower face AU) and on the eyebrows (5 for upper face AU). The size of each side of the descriptor’s box was 4x15 pixels in the 400x400 reference frame.

To maximize generalizability, we trained and tested the SVMs on independent data. For training, we used the RU-FACS [15] database. RU-FACS consists of digitized video and manual FACS coding of 34 young adults. They were recorded during an interview of approximately 2 minutes duration in which they lied or told the truth in response to an interviewer’s questions. Pose orientation was mostly frontal with small out-of-plane head motion. Image data from five subjects could not be analyzed due to image artifacts. Thus, image data from 29 subjects was used for training the classifiers. Classifier thresholds then were tested on the independent subjects from the current study.

2.4. Threshold Analysis

2.4.1. Threshold selection. Forty five sessions from the Spectrum dataset were randomly assigned to either the training set (n=32) or the testing set (n=13). Frame-level SVM output was compared to ground truth FACS codes in the training set to identify six thresholds of interest for each AU. First, the full range of SVM values across all included sessions was found for each AU. This range was then split into equally-spaced centiles. These 100 values then were used as thresholds for automatic coding in the training set; thus, any frame with an SVM value greater than or equal to that threshold would be coded as a positive instance of that AU. This automatic coding was then compared to ground truth FACS coding and thresholds were selected that granted (1) a true-positive rate of 80%, (2) a false-positive rate of 10%, (3) an equal error rate, (4) the maximum F1, (5) the maximum kappa, and (6) the maximum MCC. When the selection criteria were not perfectly met (e.g. when none of our centiles had a true-positive rate of exactly 80%), the thresholds were selected that came closest to filling them.

2.4.2. Threshold evaluation. To maximize generalizability, the six thresholds next were evaluated by using them to automatically code independent data, i.e. the videos in the testing set. These automatic codes were then compared to ground truth FACS codes using a number of threshold-specific reliability metrics including accuracy, F1, kappa, and MCC. For greater clarity, the cell sizes of

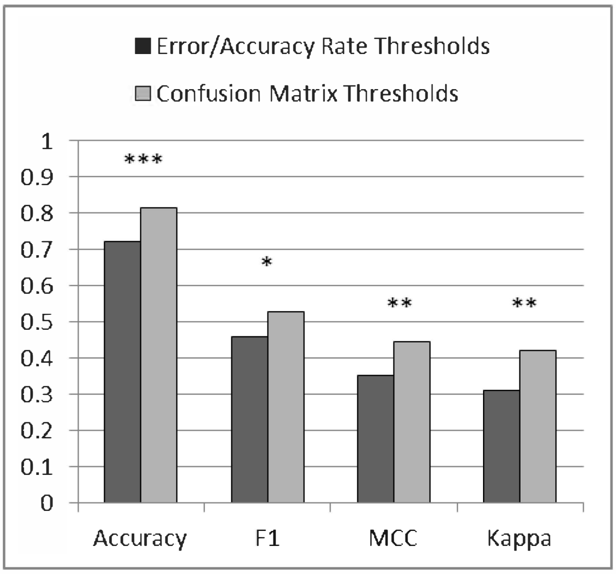

Figure 3: Performance per threshold-selection criteria type (for all Figures *p < .05, **p < .01, ***p < .001)

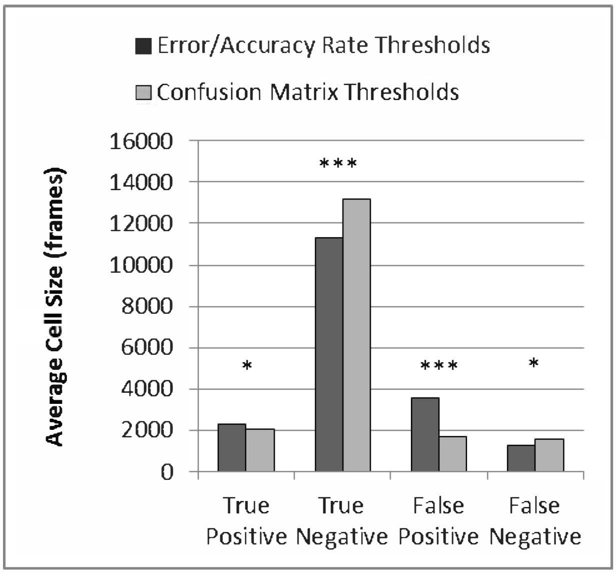

Figure 4: Average confusion matrix per threshold type

the confusion matrices were compared for each threshold. This process was then repeated using a temporal window of plus/minus 0.25 seconds, as is common practice when evaluating reliability amongst manual FACS coders. Finally, to enable comparison between inter-system and inter-observer performance, we computed all four performance metrics on a subset of videos (n=4) that had been coded by two manual FACS coders.

3. Results

3.1. Threshold selection

The primary analysis compared the two general types of threshold-selection criteria: *error/accuracy rate* thresholds (EER, FPR10, TPR80) and *confusion matrix* thresholds (maxKappa, maxF1, maxMCC). A series of paired t-tests found that confusion matrix thresholds yielded significantly higher performance than error/accuracy rate thresholds when performance was measured using accuracy (t = 6.11, p < .001), F1 (t = 2.96, p < .05), MCC (t = 3.75, p < .01), and kappa (t = 3.95, p < .01). Figure 3 shows the results for each threshold type on all four performance metrics. For each of the metrics, confusion matrix thresholds outperformed error/accuracy rate thresholds.

Different threshold selection criteria might have different effects on the cells of the confusion matrix. To evaluate this possibility, we compared error/accuracy rate and confusion matrix thresholds with respect to each cell in resultant truth tables; Figure 4 shows the results. Confusion matrix thresholds yielded significantly fewer false positive frames than error/accuracy rate thresholds (t = -7.817, p < .001) and significantly more true negative frames (t = 7.817, p < .001); they also yielded significantly fewer true positive frames (t = -2.253, p < .05) and significantly more false negatives (t = 2.253, p < .05).

3.2. Threshold evaluation

To evaluate the relation between them, we computed correlations between the different threshold-specific performance metrics that can be used to evaluate an implemented system. The performance metrics that attempt to balance the confusion matrix (kappa, F1, and MCC) had strong pairwise correlations (p < 0.01) while accuracy was only weakly correlated with the others (see Table 1).

	Kappa	F1	MCC
F1	.958**		
MCC	.986**	.945**	
Accuracy	.133	-.135	.131

Table 1: Performance metric inter-correlations

In many applications of automated facial expression detection, low base rates are common. To evaluate the influence of base rate on performance, we computed the correlations between the different reliability metrics and the prevalence of the different AUs. Kappa, F1, and MCC were all moderately positively correlated with AU base rate; as base rate increased, so did these measures. Accuracy was strongly negatively correlated with base rate; low base rates yielded high accuracy (see Table 2).

	Kappa	F1	MCC	Accuracy
Pearson r	.505	.694	.507	-.745

Table 2: Correlations with AU base rate

Finally, a second manual FACS coder was introduced to enable comparison between the inter-system performance of our AFA system and the inter-observer agreement of two manual FACS coders. Error in the ground truth effectively limits what can be achieved in inter-system reliability. For several AUs (AU 4, AU 10, and AU 12), intersystem reliability approached the level of agreement found between manual FACS coders (see Figure 5).

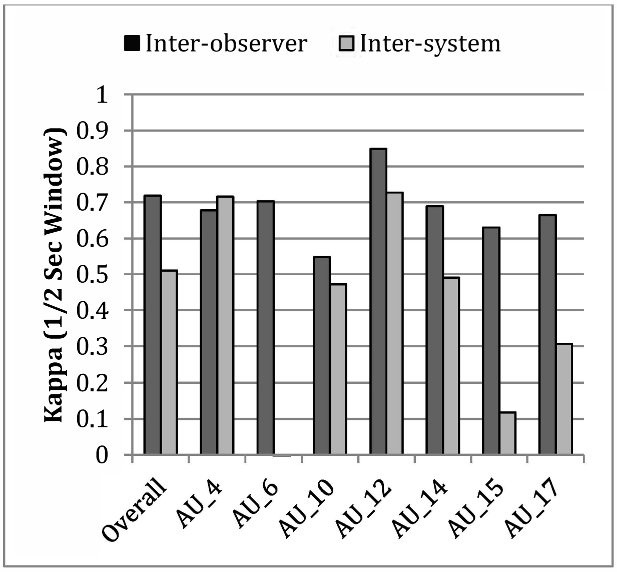

Figure 5: Inter-observer (two FACS coders) and Inter-system (FACS and maxKappa) performance

4. Discussion

4.1. Threshold selection

Results suggest that confusion matrix threshold-selection criteria provide significantly better performance than error/accuracy rate criteria. This result was found regardless of the performance metric used. This is likely due to the fact that error and accuracy rates tend to focus on only two cells of the confusion matrix [26]. The Equal Error Rate is an exception to this rule; by incorporating both error rates, it samples from all four cells of the confusion matrix. However, by requiring that the two rates be equal, the resultant thresholds often have much larger error rates than would normally be deemed acceptable. The confusion matrix thresholds, which attempt to balance three or four cells of the contingency table, showed higher performance on every metric. Thus, it is not simply the case that confusion matrix thresholds performed better on metrics that balance the confusion matrix; they also performed better in terms of accuracy. To explore this difference more deeply, we evaluated the average number of errors for each type of threshold across all videos in the sample. Basically, confusion matrix thresholds are more conservative than error/accuracy rate thresholds, slightly (but significantly) increasing the number of false negatives in order to greatly reduce the number of false positives.

The differences between the three "confusion matrix" threshold-selection criteria were minimal in this sample. As such, although confusion matrix thresholds are clearly preferable to error/accuracy rate thresholds, no single confusion matrix threshold stands above the others. However, it is possible that significant differences may become apparent between these criteria when applied to samples with different prevalence rates, rater biases, etc. Further work is needed to explore these possibilities.

4.2. Threshold evaluation

Results suggest that the four selected performance metrics fell into two separate groups. Kappa, F1, and MCC were highly intercorrelated, while accuracy was only weakly correlated with the others. However, although kappa, F1, and MCC were strongly correlated in our sample, they do have unique strengths and limitations. The F1 statistic does not include the number of true negative frames nor chance agreement. Cohen's kappa attempts to control for chance agreement, but suffers from potential problems with target prevalence and coder bias; for instance, with all else held constant, kappas are higher when codes are equiprobable and distributed similarly by the two observers [27]. In our sample, however, kappa and MCC were less correlated with AU base rate than were F1 and accuracy, suggesting that kappa may be promising in these circumstances nonetheless.

Finally, automatic coding using the optimal threshold configuration in this testing set (i.e. maxKappa) was compared to inter-observer reliability amongst manual FACS coders. Because the classifiers used by AFA systems must be trained on ground truth codes from manual FACS coders, inter-observer reliability sets a ceiling on what inter-system performance can be attained. Results suggest that for certain AUs, inter-system performance approaches the level of inter-observer agreement found between manual coders. Inter-system reliability for three of the seven AU compared favorably with that between manual FACS coders (Figure 5). For two of these (AU 4 and AU 12) inter-system reliability was within the acceptable range for use in behavioral research, which suggests the feasibility of beginning to employ automated FACS coding in behavioral studies.

For some AU, inter-system reliability was disappointing, which may be related to their low base rates (AU 6 and AU 15) and to the relatively low intensity of the AU in Spectrum. In Spectrum, AU intensity was lower than the intensity found in the RU-FACS on which the classifiers were trained. Consistent with this interpretation, a recent study found that RU-FACS-trained classifiers yielded high inter-system reliability for AU 6 and AU 12 (the two AU tested) in the Sayette Group Formation Task dataset [28]. In the Sayette dataset, AU intensity was as strong or stronger than in RU-FACS and much stronger than in Spectrum. Further research into the relation between intensity, base rate, and AU detection is needed.

A striking exception to the positive correlation between AU detection and base rate was found when accuracy rather than F1, kappa, or MCC was the performance metric. Because accuracy is highly sensitive to low base rate, it may give deceptively high scores even when the classifier may be doing little more than guessing.

4.3. Future directions

In this study, thresholds were trained on 32 sessions and tested on 13. However, it is unclear how many training sessions are required to find adequate thresholds. A future study might explore this question by experimenting with different size training sets. Alternatively, a future study might utilize iterative cross-validation, such as k-fold. Further work is also needed to evaluate the generalizability of threshold selection criteria to other samples, AUs, and event-levels. In particular, another study might look at classifiers for peak events (as opposed to the frame-level occurrence explored here). Alternate reliability metrics and threshold selection criteria might also be explored, including "unbiased" reliability metrics (e.g. Bookmaker's Correlation) that take into account the prevalence of the target event [26]. Finally, future work might directly compare manual coders and automatic coding on the same video clips in terms of both reliability and time investment.

Acknowledgements

This research was supported in part by the U.S. National Institutes of Health grant MH 51435. The authors thank Tomas Simon, Fernando De la Torre, and Simon Lucey.

Appendix

A confusion matrix represents the pattern of agreement and disagreement between observers (e.g., manual FACS coders) or measurement approaches (e.g., manual FACS coding and an SVM classifier). If one coder or approach is considered the criterion ("gold standard"), one can distinguish true and false positives (TP, FP) and true and false negatives (TN, FN). For automated facial expression analysis, manual FACS coding typically is considered to be criterion even though manual FACS coders may agree only moderately between each other.

Agreement in confusion matrices has been quantified in numerous ways. *Accuracy* is the number of true positives and negatives in proportion to all cells of the confusion matrix [29]:

$$Accuracy = (TP + TN)/(TP + FP + TN + FN)$$

A problem is that when the number of true negatives is large (i.e. low base rate), accuracy can be spuriously high.

Statistics that attempt to 'balance the confusion matrix' seek to avoid this problem. We consider three: Matthews correlation coefficient, $F1$, and Cohen's kappa. Matthews correlation coefficient (MCC) [30] can be calculated by the formula:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

F1 seeks to balance the confusion matrix by taking the geometric mean of precision and recall:

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$F1 = 2\left(\frac{Precision * Recall}{Precision + Recall}\right)$$

Coefficient kappa is widely used in behavioral science to correct agreement due to chance, where chance is determined by base rates. Specifically,

$$Kappa = (p_o - p_e)/(1 - p_e)$$

where p_{o} is the proportion of observed agreement (i.e. accuracy), and p_{e} is the proportion of agreement expected by chance. This is found by summing over the agreement diagonals of the confusion matrix, the product of the proportions for the row and column for the cell [31, 32].

References

[1] J. F. Cohn and P. Ekman, "Measuring facial action by manual coding, facial EMG, and automatic facial image analysis," in *The new handbook of nonverbal behavior research*, J. A. Harrigan*, et al.*, Eds., ed New York: Oxford University Press, 2005, pp. 9-64.

[2] P. Ekman and W. V. Friesen, *Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto: Consulting Psychologists Press, 1978.

[3] P. Ekman*, et al.*, *Facial Action Coding System (FACS): A technique for the measurement of facial movement*. Salt Lake City, UT: Research Nexus, 2002.

[4] R. Koenen. (2002). *Overview of the MPEG-4 standard*. Available: http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm

[5] F. De la Torre and J. F. Cohn, "Facial expression analysis," in *Guide to visual analysis of humans: Looking at people*, T. B. Moeslund*, et al.*, Eds., ed New York: Springer, In press.

[6] Z. Zeng, *et al.*, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Trans Pattern Anal Mach Intell,* vol. 31, pp. 39-58, 2009.

[7] G. C. Littlewort, *et al.*, "Automatic coding of facial expressions displayed during posed and genuine pain," *Image and Vision Computing,* vol. 27, pp. 1797-1803, 2009.

[8] P. Lucey, *et al.*, "Recognizing emotion with head pose variation: Identifying pain segments in video," *IEEE Transactions on Systems, Man, and Cybernetics - Part B,* In press.

[9] Z. Ambadar, *et al.*, "All Smiles are Not Created Equal: Morphology and Timing of Smiles Perceived as Amused, Polite, and Embarrassed/Nervous," *J Nonverbal Behav,* vol. 33, pp. 17-34, 2009.

[10] M. Hoque and R. W. Picard, "Acted vs. natural frustration and delight: Many people smile in natural frustration," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 354-359.

[11] J. F. Cohn, *et al.*, "Detecting depression from facial actions and vocal prosody," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, 2009, pp. 1-7.

[12] P. Wang, *et al.*, "Automated video-based facial expression analysis of neuropsychiatric disorders," *Journal of Neuroscience Methods,* vol. 168, pp. 224-238, 2008.

[13] D. S. Messinger, *et al.*, "Automated Measurement of Facial Expression in Infant–Mother Interaction: A Pilot Study," *Infancy,* vol. 14, pp. 285-305, 2009.

[14] Z. Zeng, *et al.*, "Audio-visual emotion recognition in adult attachment interview," presented at the Proceedings of the 8th international conference on Multimodal interfaces, Banff, Alberta, Canada, 2006.

[15] M. Frank, *et al.*, "RU-FACS-1 Database," Machine Perception Laboratory, Ed., ed. San Diego, Undated.

[16] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders (4th ed.)*. Washington, DC: Author, 1994.

[17] M. Hamilton, "Development of a rating scale for primary depressive illness," *Br J Soc Clin Psychol,* vol. 6, pp. 278-96, 1967.

[18] P. Ekman and E. L. Rosenberg, *What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS)*: Oxford University Press, 2005.

[19] T. K. Simon, "Action unit detection using SVMs based on AAM tracking: Tech report," Carnegie Mellon University, Pittsburgh, PA2010.

[20] I. Matthews and S. Baker, "Active Appearance Models Revisited," *International Journal of Computer Vision,* vol. 60, pp. 135-164, 2004.

[21] C. W. Hsu, *et al.*, "A practical guide to support vector classification," Department of Computer Science, National Taiwan University, 2003.

[22] I. Matthews, *et al.*, "2D vs. 3D Deformable Face Models: Representational Power, Construction, and Real-Time Fitting," *International Journal of Computer Vision,* vol. 75, pp. 93-113, 2007.

[23] J. Xiao, *et al.*, "Real-time combined 2D+3D active appearance models," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* vol. 2, pp. 535-542, 2004.

[24] M. Krystian, "A Performance Evaluation of Local Descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, pp. 1615-1630, 2005.

[25] P. Lucey, *et al.*, "Registration invariant representations for expression detection," *International Conference on Digital Image Computing: Techniques and Acpplications (DICTA),* 2010.

[26] D. M. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correaltion," School of Informatics and Engineering, Flinders University of South Australia, Adelaide 2007.

[27] J. Sim and C. C. Wright, "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements," *Physical Therapy,* vol. 85, pp. 257-268, 2005.

[28] J. Cohn and M. Sayette, "Spontaneous facial expression in a small group can be automatically measured: An initial demonstration," *Behavior Research Methods,* vol. 42, pp. 1079-1086, 2010.

[29] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters,* vol. 27, pp. 861-874, 2006.

[30] B. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta,* vol. 405, pp. 442-451, 1975.

[31] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement,* vol. 20, pp. 37-46, 1960.

[32] J. L. Fleiss, *Statistical methods for rates and proportions*, 2nd ed. New York: John Wiley & Sons, 1981.