ELSEVIER

Journal of Neuroscience Methods 163 (2007) 350–361

JOURNAL OF NEUROSCIENCE METHODS

www.elsevier.com/locate/jneumeth

Computerized measurement of facial expression of emotions in schizophrenia

Christopher Alvino [a], Christian Kohler [b,1], Frederick Barrett [b,2], Raquel E. Gur [b,3], Ruben C. Gur [b,4], Ragini Verma [a,*]

[a] *Section of Biomedical Image Analysis, Department of Radiology, University of Pennsylvania, 3600 Market, Suite 380, Philadelphia, PA 19104, United States*

[b] *Brain Behavior Center, Department of Psychiatry, University of Pennsylvania Medical Center, Philadelphia, PA 19104, United States*

Received 21 January 2007; received in revised form 6 March 2007; accepted 6 March 2007

Abstract

Deficits in the ability to express emotions characterize several neuropsychiatric disorders and are a hallmark of schizophrenia, and there is need for a method of quantifying expression, which is currently done by clinical ratings. This paper presents the development and validation of a computational framework for quantifying emotional expression differences between patients with schizophrenia and healthy controls. Each face is modeled as a combination of elastic regions, and expression changes are modeled as a deformation between a neutral face and an expressive face. Functions of these deformations, known as the regional volumetric difference (RVD) functions, form distinctive quantitative profiles of expressions. Employing pattern classification techniques, we have designed expression classifiers for the four universal emotions of happiness, sadness, anger and fear by training on RVD functions of expression changes. The classifiers were cross-validated and then applied to facial expression images of patients with schizophrenia and healthy controls. The classification score for each image reflects the extent to which the expressed emotion matches the intended emotion. Group-wise statistical analysis revealed this score to be significantly different between healthy controls and patients, especially in the case of anger. This score correlated with clinical severity of flat affect. These results encourage the use of such deformation based expression quantification measures for research in clinical applications that require the automated measurement of facial affect.

Keywords: Facial expressions quantification; Elastic deformations; Regional volumetric differences; Support vector machines; Pattern classification

1. Introduction

Facial expressions play an important role in the clinical manifestation of several neuropsychiatric disorders, such as schizophrenia, where perception (Berenbaum and Oltmann, 1992; Edwards et al., 2001; Heimberg et al., 1992; Mandal et al., 1998; Morrison et al., 1988) and expression of emotion are impaired (Kring and Neale, 1996; Salem et al., 1996; Sison et al., 1996; Tremeau et al., 2005). Indeed, facial expressions are an important window for assessing the level of affective impairment in schizophrenia, with respect to expression of emotions and their recognition. Distinguishing the degree of expressivity both between healthy and impaired people and within patients to establish presence and severity of flat affect, is a challenging problem because of normal variation in expressivity and subtlety of difference in expressions. Much of the research in this area focuses on the emotional recognition abilities of patients when compared with healthy controls (Edwards et al., 2001; Heimberg

* Corresponding author. Tel.: +1 215 662 7471; fax: +1 215 614 0266.
E-mail addresses: calvino.ece05@gtalumni.org (C. Alvino), kohler@bbl.med.upenn.edu (C. Kohler), fbarrett@bbl.med.upenn.edu (F. Barrett), raquel@bbl.med.upenn.edu (R.E. Gur), gur@bbl.med.upenn.edu (R.C. Gur), Ragini.Verma@uphs.upenn.edu (R. Verma).

[1] Present address: Hospital of the University of Pennsylvania, 3400 Spruce Street, 10th Floor, Gates Building Philadelphia, PA 19104, United States. Tel.: +1 215 662 2826; fax: +1 215 662 7903.

[2] Present address: Hospital of the University of Pennsylvania, 3400 Spruce Street, 10th Floor, Gates Building Philadelphia, PA 19104, United States. Tel.: +1 215 662 7389; fax: +1 215 662 7903.

[3] Present address: Hospital of the University of Pennsylvania, 3400 Spruce Street, 10th Floor, Gates Building Philadelphia, PA 19104, United States. Tel.: +1 215 662 2915; fax: +1 215 662 7903.

[4] Present address: Hospital of the University of Pennsylvania, 3400 Spruce Street, 10th Floor, Gates Building Philadelphia, PA 19104, United States. Tel.: +1 215 615 3604; fax: +1 215 662 7903.

0165-0270/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jneumeth.2007.03.002

et al., 1992; Kohler et al., 2003; Mandal et al., 1998; Morrison et al., 1988; Salem et al., 1996), self-ratings and psychophysiological measures of emotional experience (Berenbaum and Oltmann, 1992; Kring et al., 1993; Schneider et al., 1995) or rating of emotional expressions using the facial action coding system (FACS) (Berenbaum and Oltmann, 1992; Tremeau et al., 2005). Clinical rating scales, such as the scale for assessment of negative symptoms (SANS, Andreasen, 1984) and positive and negative symptoms scale (PANSS, Kay et al., 1987), include assessment for affective flattening and remain the most widely used instruments to examine facial emotion expression in schizophrenia. However, there is no quantitative objective measurement that both indexes the ability of a person to express emotion and also correlates this with affective impairment in patients compared to healthy controls. This underlines the need for an objective quantitative measure of emotional expression that can identify and quantify subtle changes in affect and hence help in a group based comparative analysis between patients and healthy controls, thereby enabling the assessment of treatment efficacy and progression of the disease.

Quantitative analysis of facial emotional expression is a challenging research problem. Work on automated facial expression recognition by the computer vision community (Black and Yacoob, 1997; Cohn and Kanade, 2006; Cohn et al., 1999; Essa and Pentland, 1997; Lien et al., 2000; Littlewort et al., 2006; Lucey et al., 2006; Terzopoulos and Waters, 1990; Tian et al., 2001; Zhang, 1999) has enabled extraction of features that represent a given expression. Unfortunately, these measures are either too sparse to capture the elastic changes within facial regions during the course of an expression change, or are global and lack the specificity required to capture fine-grained changes that may occur within facial regions. Therefore, these features are inadequate in quantifying the expression changes. Indeed, some of these features have been incorporated in facial expression recognition systems, however these perform well only on expressions of extreme intensity (Pantic and Rothkrantz, 2000). Since patients show more subtle and pervasive impairment in affect, these methods are not able to capture the subtle changes. In FACS (Ekman and Friesen, 1978), each expression is characterized as a combination of muscular movements of the fundamental unit of appearance change called an action unit (AU). Existing methods for studying expression differences, such as FACS (Ekman and Friesen, 1978) are time and labor intensive, are prone to inter-rater variability due to differences in experience based assessment, and do not produce a quantitative measure of expression change. There are automated versions of FACS that are based on extracting action units from facial video sequences (Bartlett et al., 1999; Cohn et al., 1999; Donato et al., 1999; Essa and Pentland, 1997; Lien et al., 2000; Tian et al., 2001). However, FACS, like its automated versions, is unable to quantify the intensity and degree of difference between emotion expressions and hence provides no quantification of expression change.

A review of the existing methods of facial expression analysis underlines the need for a framework that can identify group differences between patients and controls by capturing subtleties of expression change and provide a measure that can be correlated with a clinical scale of affect impairment. We model the faces as a combination of elastic regions and a facial expression change as an elastic transformation that transforms a neutral face to a face with expression and produces an expression quantification map (Verma et al., 2005). We use this expression quantification map to train automated facial expression classifiers for the four universal emotions of happiness, sadness, anger, and fear, using images of actors in varying degrees of emotion expression. These form profiles for each of these emotions. These classifiers, when applied to patients and controls provide scores of emotion expressivity that quantify the extent to which an intended emotion has been expressed. These scores, when statistically analyzed, show significant group difference between patients and healthy controls. The classifier-based framework that we propose for determining subtle expression changes is general and applicable to group-wise analysis of all affect-related disorders, against healthy controls.

2. Methods

2.1. Design of expression quantification framework

In the subsequent sections, we present the details of our framework for computerized expression quantification using expression classifiers that provide a measure of the likelihood of the emotion being expressed. As in the design of any classifier, we follow a sequence of steps: (1) extraction of expression features (Section 2.2) in which we discuss the creation of regional volumetric maps that identify the face as a combination of deformable regions and expressions as a combination of elastic changes in these regions. Wavelet features extracted from these volumetric maps are the features used to train the classifiers; (2) creation of expression classifiers (Section 2.3) using the features computed as part of step 1. Four classifiers are trained once for each expression of happy, sad, anger and fear using advanced pattern classification techniques. These classifiers, when applied to a facial expression image, produce a probability score, that indicate the probability of a face expressing a particular emotion (Section 2.3.1). The classifiers are created by training on an actors' database that is described in Section 2.4.1. These classifiers are cross-validated using a leave-one-out (or jackknife) paradigm on the same database, and are then applied to a patients-controls database described in Section 2.4.2 and classification scores are computed for each expression. These classification scores are then correlated with clinical measures described in Section 2.5. We now provide details of each of these aspects.

2.2. Extraction of expression features

Each facial image is modeled as a combination of elastic regions and their boundaries, with the regions chosen on the basis of their role in the expression and comprehension of an emotion. Facial changes, resulting from an expression change from a neutral face to a face with emotion, are modeled as a deformation of a combination of these regions. Fig. 1 shows facial regions selected in a neutral face (Fig. 1(b)) and in a face with expression (Fig. 1(c)), with each pre-selected deformable

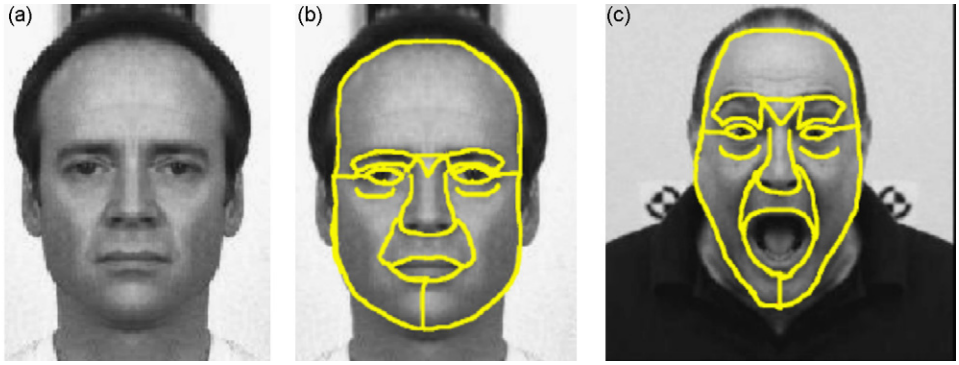

Fig. 1. (a) Predefined template image of neutral expression; (b) labeled template image; (c) labeled subject facial expression image. These boundary labels identify the regions in which we consider change to occur during expression formation. Elastic registration is performed constrained by the corresponding regions.

region demarcated by hand labeled boundaries. We selected several anatomically meaningful regions, such as the outline of the face, the lips, the eyes, the eyebrows, the nose, the mouth, etc. The boundaries were chosen to lie on naturally occurring lines on the face that contain strong edges or appear in an expression, in order to maintain consistency. Manual labeling adds flexibility for identifying subtle facial regions.

2.2.1. *Computation of the deformation changes*

The boundaries in this figure and landmark points on them guide the computation of the elastic transformation that maps the enclosed demarcated regions of the face from the template image to their counterparts in the subject facial expression image (for details see Verma et al., 2005). The resulting displacement field thus characterizes the shape change of the subject face with respect to the template face as shown in Fig. 1(a). Computing the displacement in the domain of the template allows for comparison of the displacement fields of regions across subjects. The specific choice of template image is immaterial and will not bias the results (Davatzikos et al., 1996).

2.2.2. *Removal of inter-subject variability—normalizing expression changes*

In order to remove variability in the displacement fields that results from subjects having differently shaped neutral faces, with variable features, we compute two displacement fields, one from the template to the subject's neutral image V_{TN}, and one from the template to the subject's expression image, V_{TE}. The desired *expression displacement* (see Fig. 2), is defined as,

$$V_{NE} = V_{TE} - V_{TN}$$

with the rationale that V_{NE} is the displacement that this expression would have produced given that the subject had a neutral face in the shape of the template face, thereby removing the inter-subject neutral shape variability. In addition, all images were scaled before analysis to remove camera zoom and head

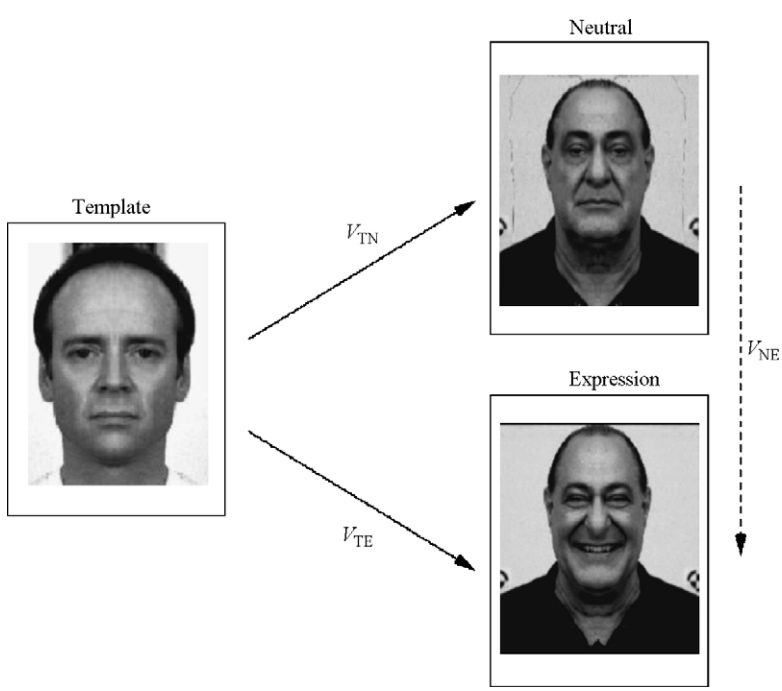

Fig. 2. Removal of inter-subject variability in the neutral face by computing deformations in the space of a common template. We use the information obtained from V_{NE} for all our analysis.

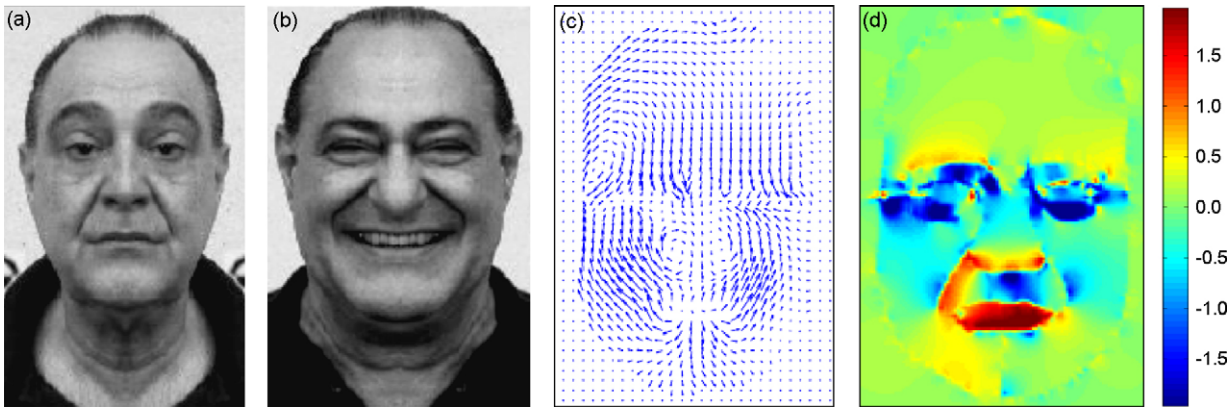

Fig. 3. Expression displacement from neutral to happiness. The left two images are neutral (a) and happy (b) facial images. The expression displacement field (c) characterizes the motion of facial regions using vectors. The log RVD map (d) is a quantification of this displacement field and represents a spatial profile of the expansion and contraction of the facial regions, as part of the expression change.

size variability by ensuring that the ear-to-ear distance was the same for each image. Ear-to-ear distance was chosen since it remains static during facial expression. An example of a normalized expression displacement field is shown in Fig. 3(c) for happiness.

Computation of deformation maps: The expression displacement, V_{NE}, quantifies the deformation from the neutral face to the expressive face. To locally measure the amount of contraction and expansion in the vector field, we compute the determinant of the Jacobian matrix for the corresponding deformation field and call it the RVD (Regional Volumetric Deformation) map as in (Davatzikos et al., 1996; Verma et al., 2005). The RVD value at each pixel represents the amount of local area change that the corresponding pixel in the neutral face undergoes when deforming to the expressive face. In order to treat area contraction on a comparable scale as area expansion, we use the log of the RVD values, with a zero value indicating no change, a negative value indicating a contraction and a positive value an expansion (the right of Fig. 3(d) shows a log RVD map corresponding to expression change). Each log RVD map is truncated to be in the interval $[-2, 2]$ (indicating an expansion or contraction of twice or half the current size) since numerical errors outside of this range can occur due to not solving the displacement field accurately and since most of the pertinent deformation information lies in this range.

2.2.3. Feature extraction

The log deformation map is then processed by a discrete wavelet decomposition using a 2D Daubechies wavelet (Daubechies, 1990). The wavelet approximation coefficients and detail coefficients, each filtered and down sampled three times and corresponding to mid-frequency components in the RVD maps, are compiled as a vector that serves as a finite dimensional representation of the deformation present from neutral to expression. Each of these coefficients corresponds to a region of size 8×8 in the original RVD map. In all, for each expression, we obtained 268 wavelet coefficients from the 109×156 region of the log RVD map where the template face is contained (Daubechies, 1990; Strang and Nguyen, 1996). Each of these four two-dimensional filtered and sub-sampled images is stacked into a vector. The feature vector can be summarized as the concatenation of the level four approximation coefficients, and the level four horizontal, vertical and diagonal detail coefficients. Level four was chosen because it represented the captured information at the scale size most appropriate for classifying emotional features. Including coefficients at too many scales only served to decrease the generalization performance of the classifier during validation. These features are used for training classifiers for each of the expressions.

2.3. Creation of expression classifiers

We developed a multi-class emotion classifier based on a pattern classification technique called support vector machines (SVM's) (Vapnik, 1998) that is trained on wavelet features computed from images of expressions from a professional actors' database (see Section 2.4.1 below). Then, given the wavelet feature vector corresponding to an expression of a new subject, the classifier identifies it as one of the four universal emotions of happiness, sadness, anger, or fear. SVM is a commonly applied pattern classification technique used widely in medical imaging and computer vision (LaConte et al., 2005; Schölkopf and Smola, 2001; Zhan and Shen, 2004; Zhang et al., 2002).

When used for classification, SVM's attempt to find classifiers that divides the feature space (in our case, the space of possible deformations) into regions that are separable on the basis of inherent underlying properties of the data. These expression classifiers (one for each expression) when applied to a facial expression image produces a real number that describes its estimated membership to one of the emotion classes. A classifier is trained for each of the four universal emotions. For example, to create an expression classifier for anger, we train on two classes of expressions, one containing only anger expressions denoted by A and the other containing all expressions other than anger, denoted by $\tilde{A}$. Training on these classes produces a classifier $f_{A\tilde{A}}$ that, when applied to an expression image feature, F, produces a positive value if the feature was from an anger expression and a negative value otherwise. Thus, the output of a given SVM signifies the membership of the test expression to that emotion. The outputs of the individual SVM's are numbers that typically

range from −1 to +1. We map these values, using a sigmoid function, $s = (x) = 1/(1 + e^{-x})$, to the range (0, 1) thereby associating a pseudo-probability score with each expression. Since four classifiers are applied to each test expression, we obtain a four-tuple of values for each expression feature vector, F,

$$t(F) = (s(f_{\mathrm{H},\tilde{\mathrm{H}}}(F)), s(f_{\mathrm{S},\tilde{\mathrm{S}}}(F)), s(f_{\mathrm{A},\tilde{\mathrm{A}}}(F)), s(f_{\mathrm{F},\tilde{\mathrm{F}}}(F)))$$

where H, S, A, and F correspond to happiness, sadness, anger and fear, respectively. We normalize this four-tuple, $t(F)$, by dividing by the sum of its elements so that we can interpret the normalized four-tuple, $t'(F)$, as an analysis of the composition of that expression in terms of the four universal emotions. Each value provides a measure of an expression belonging to a particular class. The expression corresponding to the highest value in the four-tuple signifies the most likely emotion being expressed according to the classifier. Values being nearly equal may indicate a mixture of emotions being expressed.

2.3.1. C-SAFE scores

We wish to determine the level at which the expression classifier predicts how well a given expression matched that of the emotion that the subject intended. For this purpose, we consider selecting the specific score in the normalized four-tuple, $t'(F)$, that corresponds to the intended emotion as a figure of merit for how well the subject was able to express that emotion, and call it the *Computerized Scale for Assessing Facial Expressivity (C-SAFE)* score. For instance, if $t'(F) = (0.9, 0.05, 0.0, 0.05)$, and the intended emotion was happiness, then the score we derive from the classifier is 0.9. Thus, it is considered that the subject was able to express happiness well. However, if the intended emotion were anger, then, with a value of 0.0, it is considered that the subject did not express this emotion well. We will denote the classifier score for emotion E, and for the facial expression image whose intended emotion was I as $t'^{\mathrm{I}}_{\mathrm{E}}$. Thus $t'^{\mathrm{A}}_{\mathrm{A}}$ is the *C-SAFE Anger* score for a given subject expression when the intended expression was anger. Furthermore, we will refer to the *Average C-SAFE* score for a given subject as

$$\tilde{t}' = \frac{1}{4} \sum_{i \in \{\mathrm{H,S,A,F}\}} t'^{i}_{i}$$

which is the average of the C-SAFE scores for a given subject averaged over all four emotions in the current expression task. We will analyze differences between the patient and control groups using the Average C-SAFE score, as well as the C-SAFE for each of the emotions.

2.4. Databases used for validation and testing

2.4.1. Normative training database

We use a database of facial images of professional actors to serve as a set of widely recognized emotional expressions and as a training set for the four classifiers. This database contains posed and evoked expressions and has been validated (Gur et al., 2002). While posed databases have been used in the past for many expression studies, there is evidence that evoked emotions are more accurately perceived than posed expressions (Gur et al., 2002) and therefore we only use the evoked expressions in the present study.

A total of 32 professional actors were initially photographed with a neutral face. They were then guided by professional theater directors through enactments of each of the four universal emotions of happiness, sadness, anger, and fear based on the evoked emotions procedure (Gur et al., 2002). Images were acquired while these actors were expressing emotion at three levels of intensity: mild, moderate, and peak. This procedure produced 181 facial expression images, 74 females and 107 males. Validation of the database has been performed (Gur et al., 2002; Indersmitten and Gur, 2003). Select examples are shown in Fig. 4. We used this database because the experimental conditions were controlled and identical to those used for the patient/control database described below. Furthermore, few databases exist that contain true evoked expressions. For this study we used the peak evoked expressions from 20 male and 10 female actors ages 18–72 with an average age of 39 years.

The log RVD maps are generated for each of these images as explained in Section 2.2, quantifying the deformation from the neutral face to the expressive face. Wavelet features that serve as a more concise, lower dimensional numerical signature of this expression are then computed from the log RVD maps, and are used for training the four classifiers, one pertaining to each emotion.

2.4.2. Database of patients and controls

A database of facial expression images consisting of 12 patients with schizophrenia, diagnosed by DSM-IV criteria, and

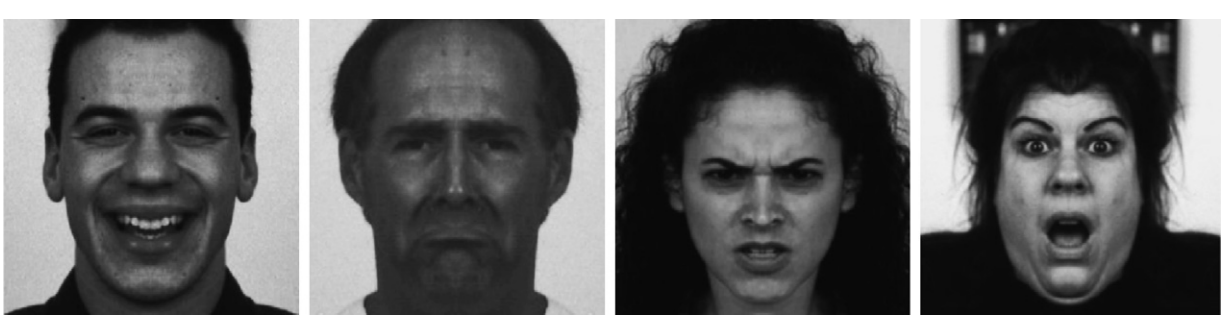

Fig. 4. Emotional expressions from the professional actors' database. This database is used as the "normative standard" of emotional expression to train SVM based classifiers.

Fig. 5. Facial images of healthy controls (top row) and schizophrenia patients (bottom row) expressing the four emotions of happiness, sadness, anger and fear.

12 healthy controls was acquired under the supervision of a psychiatrist (C. Kohler). Patients and controls were chosen in pairs matched for age (mean age 31.6 years, standard deviation 8.0 years), ethnicity (12 Caucasian, 12 African-American), and gender (14 Males and 10 Females). All participants underwent diagnostic assessment, including psychiatric interview, physical examination and blood work to rule out present or past DSM-IV Axis I disorders in healthy controls or a history in their family members, and any medical or neurological disorders, which may affect brain function. In addition, all schizophrenia participants met criteria for clinical stability, including living independently or with family members, no hospitalization or medication change within the past 3 months, and none were treated with first generation antipsychotics, which are associated with extrapyramidal symptoms and may affect facial expressions. The control-patient pairings were used in pair-wise statistical analysis to control for age, ethnicity, and gender. In order to elicit evoked expression, patients and controls were individually guided through vignettes provided by them. Each vignette described a situation in their life pertaining to each emotion. It was recounted back to them by C. Kohler, who guided them through the three levels of intensity: mild, moderate, and peak. For each expression, facial photographs were acquired at each of the three levels of intensity, of which we use only the facial images at peak expression levels. Participants provided written consent that included permission to use their photographs in scientific communications. Fig. 5 shows facial images of healthy controls and schizophrenia patients expressing each emotion. Note the impairment in the ability of the patients (bottom row) to express emotions, even in the evoked expression task, as compared with the controls (top row), although controls were less expressive compared to the actors.

2.5. *Ratings for comparison*

In order to validate the quantitative results of the classifier, we obtained (1) human ratings of expressions in the actors' database, (2) clinical ratings of the schizophrenia patients, and (3) ratings on videos of expression changes of patients and controls.

(1) *Human ratings of actors' database*: Forty-two students (13 males and 29 females), with average age 20.2, were recruited from undergraduate and graduate courses in psychology at Drexel University in Philadelphia, Pennsylvania. They rated each face for identity of expression (happy, sad, angry, fearful, or neutral) and for intensity.
(2) *Clinical ratings for patients' database*: The schizophrenia patients underwent clinical ratings within a 4-week period of when the facial expression images were recorded. In this study we used the SANS (Andreasen, 1984) items relevant for facial expression: unchanging facial expression, poor eye contact, affective non-responsivity, and inappropriate affect, (SANS items 1, 4, 5 and 6, respectively). Each score is an integer from 0 to 5, where 0 corresponds to no deficit and 5 corresponds to severe deficit.
(3) *Human ratings of videos (video-SANS)*: As SANS scores were generated for these subjects on a day different from the day of acquisition, videos acquired during image acquisition were rated independently by two raters on the SANS scale in the categories of *flat affect* and *inappropriate affect*. The first rater was C. Kohler, M.D., who is trained and reliable in clinical SANS rating. The second rater was R. Verma, a non-clinician trained by Dr. Kohler. A total of 21 subjects were rated with video-SANS (11 patients and 10 controls). The inter-class correlation (ICC) coefficients between the

two raters was $r = 0.89$ for flat affect and $r = 0.84$ for inappropriate affect. Due to this very high degree of reliability, all results are reported using the average of the two raters.

3. Results

We have constructed classifiers for each of the expressions using images from an actors' database (Section 2.4.1), with separate expression classifiers for each gender, since there is evidence for sex differences in emotion processing (Gur et al., 1992). Indeed, classification performance was better when the classifiers were separated on the basis of gender, i.e., when male patients and controls were tested on the classifier trained on only the male actors and similarly with females, but a larger sample size is required to confirm this finding. Prior to applying these classifiers to a database of patients and controls (Section 2.4.2), we cross-validated these expression classifiers and compared them against human ratings, to ensure that the classifiers have been able to learn and hence classify the emotions comparable to human raters.

3.1. Cross-validation on actor data

For the purpose of validation, we used leave-one-out analysis, i.e., we trained classifiers using all actors except one. This left out actor was classified using the expression classifiers. The confusion matrices in Table 1 show the frequency with which each expression was classified as the emotion it was intended to be. The columns represent the emotion that was intended by the subject and the rows represent the emotion determined by the classifier. The total number of images that are to be classified for each intended emotion is shown on the last row of the table. For example, in the case of the female actors, of the 20 expressions that were intended as happiness, 17 were classified as happiness, 0 were classified as sadness, 2 as anger, and 1 as fear. Hence, for female actors the overall percentages of correctly classified emotions were: 85.0% for intended happiness, 77.8% for sadness, 72.2% for anger, and 83.3% for fear. For male actors the overall percentages of correctly classified emotions were: 82.6% for intended happiness, 53.3% for sadness 87.5% for anger, and 75.0% for fear. The confusion matrices showed adequate cross-validation performance of the classifiers on the actors' database.

Table 1
Confusion matrices of "leave subject out" cross validation for multi-class expression classifier when tested on (a) female actors and (b) male actors

Classified emotion	Intended emotion			
	H	S	A	F
(a) Female actors				
H	17	3	1	2
S	0	14	1	0
A	2	0	13	1
F	1	1	3	15
Total	20	18	18	18
(b) Male actors				
H	19	3	1	0
S	4	16	3	2
A	0	10	28	3
F	0	1	0	15
Total	23	30	32	20

The confusion matrices show the frequency with which each expression was classified as a given emotion. The columns represent the emotion that was intended by the subject and the rows represent the emotion determined by the classifier. The total number of images corresponding to each intended emotion is shown on the last row of each table.

3.1.1. Correlation with human rated images

For all images used to design the classifier, we computed the percentage of raters that identified the expression correctly, irrespective of the intensity rating: 98% for happiness, 67% for sadness, 77% for anger, and 67% for fear. Fig. 6(a) shows a cluster plot highlighting the relationship between the C-SAFE score computed by the classifier and the percentage of human raters who correctly identified that expression and presents low but significant correlation at the level $p < 0.02$. In Fig. 6(b), we show a cluster plot highlighting the relationship between the C-SAFE score and the average intensity rating of the intended emotion provided by the human raters for that image. The linear correlation coefficient is $r = 0.24$ with $n = 145$, which represents low but significant correlation at the level $p < 0.004$.

The cross-validation results on the 181 images in the actors' database showed that the classifier was able to classify emotions well. Furthermore, the classifier performance matched human raters in that emotions identified well by the classifier were also better identified by human raters.

3.2. Classification of patient and control data

With emotion classifiers trained on the normative actors' database, we tested each subject in the patient/control data set. We tested the expressions of male subjects with a classifier trained only on the male actors' images, and similarly for the females. The C-SAFE scores resulting from this testing showed significant group-wise separation between patients and healthy controls in paired *t*-tests, both for the average over all emotions and specifically for anger. Importantly, these scores correlated with clinical ratings.

Table 2 shows confusion matrices of the patients and controls when classified using the classifier trained on actors' expressions. These confusion matrices represent the overall classification performance on the controls and patients database. The columns represent the emotion that was intended by the subject and the rows represent the emotion determined by the classifier. The total number of images that are to be classified for each intended emotion is shown on the last row of the table. We show these confusions matrices with gender combined since the classifier did not show significantly different performance for either gender. An expression is said to be correctly classified if it matches the intended emotion. In the group of healthy controls, 25 of the 48 expressed emotions were correctly classified for a correct classification rate of 52.0%. Notably, in the patients only 17 were correctly identified out of a possible 47

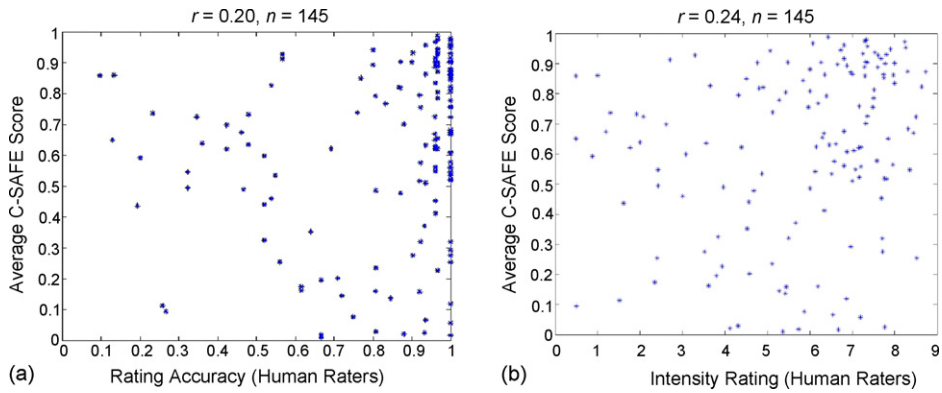

Fig. 6. For actors database, (a) cluster plot of C-SAFE score, vs. the percentage of the time the human raters correctly identified the emotion of expression and (b) cluster plot of C-SAFE emotion score, vs. the rated intensity of the intended emotion. Each point in this plot corresponds to an expression in the actors' database.

expressions, for a correct classification rate of 36%. As the classifier was validated (as above), this indicates that patients have lower expressivity than controls. While a χ^2-test (Hassard, 1991) between all correctly classified expressions and all incorrectly classified expressions for the patient and controls groups showed borderline significance, we would like to note that a χ^2-test within the same group, when only including expressions that were intended to be anger produced a significant ($p < 0.025$) difference. This can be clearly seen in that, for intended anger, controls were classified correctly in 9 out of 12 expressions, and patients were classified correctly in only 3 out of 12 expressions. No other intended emotions produced a significant difference.

To further underline the fact that patients and sometimes controls expressed emotions in atypical ways, Fig. 7 shows three illustrative examples where subjects intending to express one emotion produced a facial expression that subjectively appeared closer to another emotion. To qualitatively validate the emotion classifiers developed in this paper, we will show how the classifiers detected the subjectively apparent emotion rather than the incorrectly expressed intended emotion. On the left of Fig. 7 shows a schizophrenia patient intending to express happiness, however the facial expression appears to be neutral or flat. In this case the individual classifier scores demonstrate uncertainty about the expression. In the middle of Fig. 7 we show a healthy control intending to express fear however, the facial expression was rated as sadness by human raters and classified to be sadness by the classifiers. Finally, the right of Fig. 7 we show a schizophrenia patient intending to express anger however, the facial expression was rated as happiness by human raters as well as the classifiers. Such examples support the ability of the classifiers to correctly detect and identify the emotion apparent in facial expressions.

Table 2
Confusion matrices of classifiers applied to (a) healthy controls and (b) schizophrenia patients

Classified emotion	Intended emotion			
	H	S	A	F
(a) Healthy controls				
H	8	1	0	0
S	3	4	3	3
A	1	5	9	5
F	0	2	0	4
Total	12	12	12	12
(b) Schizophrenia patients				
H	5	1	3	0
S	2	5	5	3
A	2	5	3	4
F	3	1	1	4
Total	12	12	12	11

These confusion matrices represent the overall classification performance on the controls and patients database. The columns represent the emotion that was intended by the subject and the rows represent the emotion determined by the classifier. We show these confusions matrices with gender combined since the classifier did not show significantly different performance for either gender. The total number of images corresponding to each intended emotion is shown on the last row of each table.

3.2.1. *Paired t-test on average C-SAFE scores*

Using the patient-control pairs matched for age, ethnicity, and gender, we performed a paired *t*-test on the Average C-SAFE score computed from the classifier and we obtained $t = 2.75$ with 11 degrees of freedom (12 valid pairs) for a significant $p < 0.05$.

3.2.2. *Paired t-test on C-SAFE Anger scores*

The C-SAFE Anger score was the individual emotion score that showed the highest difference between patients and controls of the four emotions in this study. This is consistent with the findings in the literature that people with schizophrenia with impaired affect are more impaired in threat-related expressions (Gur et al., 2006). Using the pairs matched for age, race, and gender, we performed a paired *t*-test on the C-SAFE Anger scores and obtained $t = 2.32$ with 11 degrees of freedom for a significant $p = 0.020$.

3.2.3. *Correlation with SANS*

Note that no controls have been rated with SANS, since they had no clinically detectable symptoms, and therefore controls were left out of this analysis. The correlation coefficients of the Average C-SAFE scores with the standard clinical SANS ratings were not significant. However, we also examined the correlation of the Average C-SAFE scores and intended emotion score for

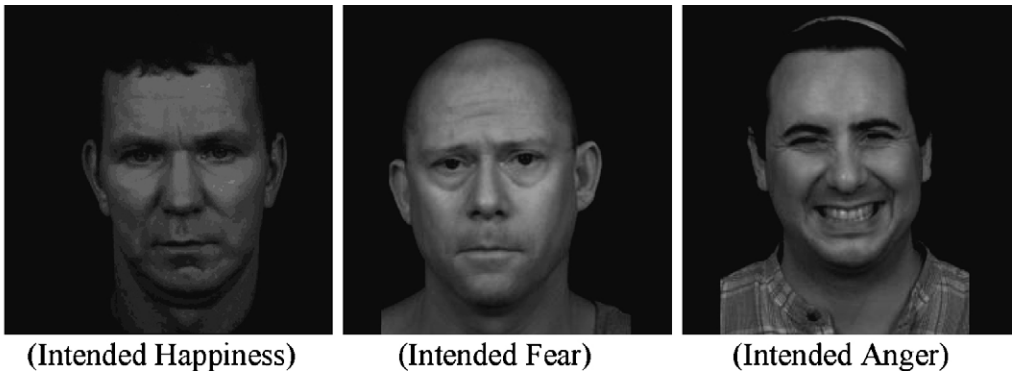

Fig. 7. Examples from the database of patients and controls where the intended emotion did not subjectively match the facial expression but where the classifier determined the expression based on the facial expression well. The first was classified as uncertain, the second as sad, and the third as happy.

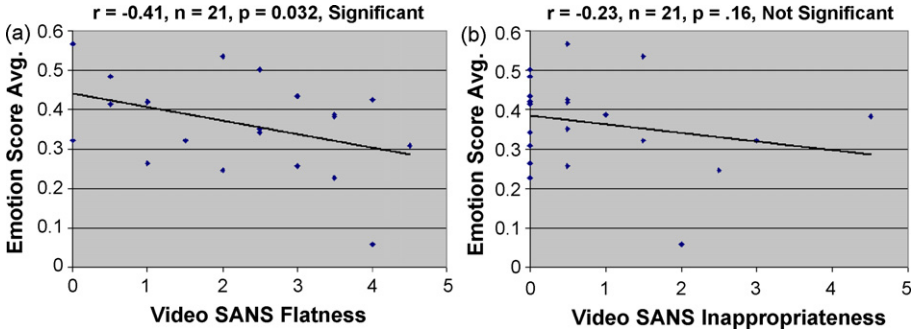

Fig. 8. (a) Average C-SAFE score vs. video SANS flatness and (b) average C-SAFE score vs. video SANS inappropriateness.

all emotions, with the video-SANS scores for "flat affect" and "inappropriate affect". Fig. 8 shows a cluster plot of the values of the Average C-SAFE scores for each subject versus the average of the video-SANS scores for each subject both for flat (Fig. 8(a)) and inappropriate affect (Fig. 8(b)). Each point in the plot corresponds to a subject. It shows that the Average C-SAFE score is negatively correlated (for a correlation coefficient of $r = -0.43$, which is significant at the $p < 0.05$ level) with the scores of flatness (Fig. 8(a)) and not significantly correlated with the scores of inappropriateness (Fig. 8(b)). Note that Spearman's rank order correlation for these values is $r = -0.40$ which is significant at $p < 0.04$ (Press et al., 1992). In repeating the analysis with only the C-SAFE score for each emotion separately, we found a strong negative correlation of the C-SAFE Anger score with "inappropriate affect" and of the happiness C-SAFE Happiness score with "flat affect" (Fig. 9).

4. Discussion and future work

We have developed a framework for identifying group differences between facial expressions of patients and controls based on computerized expression classifiers designed for each emotion. These classifiers were trained on region-based facial quantification maps and validated on a database of the evoked facial expressions of professional actors. It may be noted that while actors have more intense expressions than patients or controls, they are used as a normative baseline group, against whom the performance of the patients and controls are measured. The classifiers are able to differentiate the two groups, indicating that although they have been trained on more exaggerated expressions, the classifiers have been able to capture the nuances and subtleties of facial changes with change in expression. By identifying and quantifying these subtle expression differences, the classifiers showed a significant group separation between patients and controls, with healthy controls performing significantly better than schizophrenia patients. Patients showed more impairment than the controls, as demonstrated by a lower classification rate in patients, as well as a greater misclassification per expression, as indicated by the column scores of the classifier for each of the expressions in Table 2. This indicates that the classifiers are sufficiently sensitive to expressive appropriateness and flatness to establish that schizophrenia patients have difficulty expressing emotions.

Patients and controls are well separated by the classifiers ($p = 0.0055$), highlighting the ability of the C-SAFE scores in measuring differences between the two groups. However, when analyzed separately for each emotion, anger is the only emotion that shows significant group difference ($p = 0.013$). Possibly this is due to the small sample size. It may also be related to the fact that sadness and fear each have a high variability in expression, unlike happiness that is expressed in a very consistent way across people and across intensities. Indeed, happiness remains the most easily identifiable emotion by both humans and automated expression classifier. A direct comparison between the scores of the classifier and the human raters for each of the emotions reveals that happiness (with the highest score) is the easiest to identify and sadness (with the lowest score) is the most difficult, with anger and fear being of intermediate difficulty. When

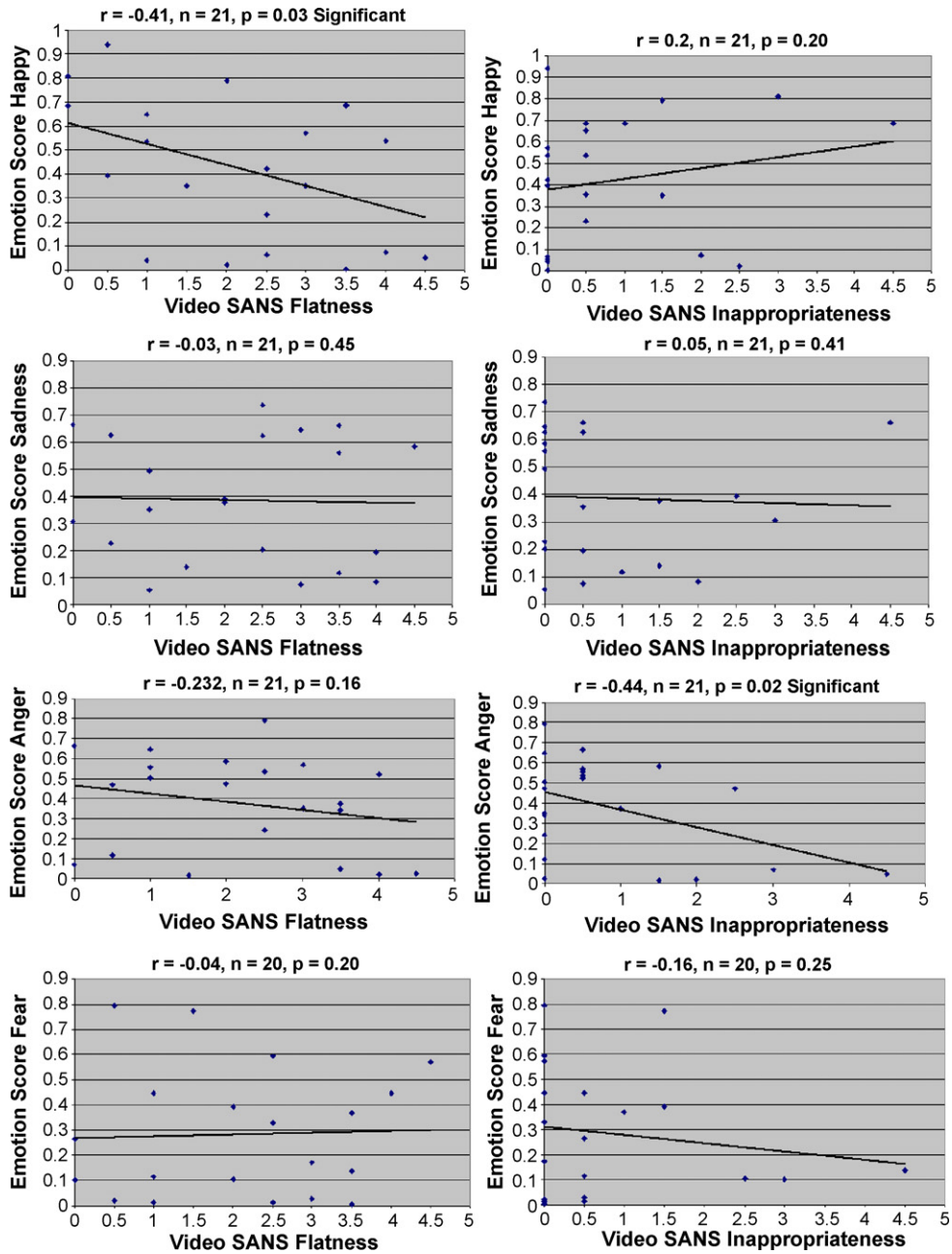

Fig. 9. Correlations by emotion and area of deficit. Each plot shows emotion specific C-SAFE scores vs. a specific video SANS scores. Anger is correlated with inappropriate affect and happiness is correlated with flat affect. All others are insignificant.

the output of the classifier for an intended emotion was compared against the percentage of human raters who were able to correctly identify a given emotion, as is shown in Fig. 6(a), we obtain a low but significant correlation coefficient, which shows that the C-SAFE score is higher for the images that were more easily identifiable by the human raters. The mismatches between the classifier ratings and the human ratings highlight the fact that humans and the computerized method use separate cues in understanding emotion expression. In addition, we found a small but significant correlation between the C-SAFE score and the average of intensity ratings provided by the raters (Fig. 6(b)). This indicates that expressions at a low intensity are more difficult to recognize for the classifier.

Finally, we found a correlation of these C-SAFE scores with video-SANS measures, designed to quantify the level of deficit in facial expressivity. There was significant correlation between the Average C-SAFE score and video-SANS score for flatness, the C-SAFE Anger score and video-SANS score for inappropriate affect, and the C-SAFE Happiness score and the video-SANS score for flatness of affect. This shows that the C-SAFE scores obtained from the classifiers are able to measure differences in expressivity between the two groups, and this group difference in inappropriate expression is dominated by anger. The high negative correlation of happiness with the flat affect score indicates that less flat affect is associated with a better expression of the emotion, especially in happy expressions. The fact that

video-SANS, unlike the clinical SANS ratings, correlates well with the computerized ratings may relate to the fact that SANS scores were obtained in a different setting than that of the image acquisition. Perhaps each patient responds differently to being in a clinical office setting compared to being in the experimental image acquisition setting, with several cameras capturing their facial expressions. This paves the way for video-SANS to be further investigated as an appropriate measure of affect deficit.

In the future, we propose to expand the set of training samples to add more expressions, including those at lower intensity. Using more training data can improve the classifier to achieve a level of accuracy that may match the emotion detection capability of humans. We expect that we will acquire more healthy controls, so that the classifiers can be constructed from the healthy controls, thereby providing a better normative standard for the patients. In addition, we will explore the use of 3D reconstructed faces (Gur et al., 2002) and video sequences in understanding small changes in expression. Impaired facial expressions of emotions in neuropsychiatric disorders with, in particular schizophrenia and autism, will exert considerable effects on interpersonal relatedness and social functioning. The construction of objective and quantitative instrumentation to assess emotional expressivity may have wide ranging clinical implications in patient populations, including application in expression training, assessment of treatment effects, and detection of persons at risk.

In conclusion, we have proposed a multi-class expression classifier that is able to determine the subtle expression differences between patients and controls. These classification scores correlate well with the human ratings of these expressions by non-experts, as well as expert determined clinical scores for expression change. The clinical implications of such a score are potentially far-reaching in that such an expression classification scheme is general and applicable to all disorders characterized by disturbed affect. Quantitative assessment afforded by this procedure can help in determining the diagnosis and progression of the disease, studying prognostic roles of affect deficit, and detecting subtle improvements due to treatment.

References

Andreasen, N. (1984). The Scale for the Assessment of Negative Symptoms (SANS).

Bartlett MR, Hager JC, Ekman P, Sejnowski T. Measuring facial expressions by computer image analysis. Psychophysiology 1999;36:253–63.

Berenbaum H, Oltmann TF. Emotional experience and expression in schizophrenia and depression. J Abnorm Psychiat Res 1992;101:37–44.

Black MJ, Yacoob Y. Recognizing facial expressions in image sequences using local parameterized models of image motion. Inter J Comput Vis 1997;25(1):23–48.

Cohn J, Kanade T. Use of Automated Facial Image Analysis for Measurement of Emotion Expression the Handbook of Emotion Elicitation and Assessment. Oxford University Press Series in Affective Science; 2006.

Cohn JF, Zlochower AJ, Lien JJ, Kanade T. Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. Psychophysiology 1999;36:35–43.

Daubechies I. The wavelet transform, time-frequency localisation and signal analysis. IEEE Trans Inform Theor 1990;36(5):910–6.

Davatzikos C, Vaillant M, Resnick SM, Prince JL, Letovsky S, Bryan R. A computerized approach for morphological analysis of the corpus callosum. J Comput Assist Tomogr 1996;20:88–97.

Donato G, Bartlett MR, Hager JC, Ekman P, Sejnowski T. Classifying facial actions. IEEE Trans Pattern Anal Mach Intell 1999;21(10):974–89.

Edwards J, Pattison PE, Jackson HJ, Wales RJ. Facial affect and affective prosody recognition in first-episode schizophrenia. Schizophren Res 2001;48(2/3):235–53.

Ekman P, Friesen WV. Facial Action Coding System. Consulting Psychologists Press; 1978.

Essa IA, Pentland AP. Coding, analysis, interpretation and recognition of facial expressions. IEEE Trans Pattern Anal Mach Intell 1997;19(7):757–63.

Gur RC, Erwin RJ, Gur RE, Zwil AS, Heimberg C, Kraemer HC. Facial emotion discrimination. II. Behavioral findings in depression. Psychiat Res 1992;42:241–51.

Gur RC, Sara R, Hagendoorn M, Marom O, Hughett P, Macy L, et al. A method for obtaining 3-dimensional facial expressions and its standardization for the use in neurocognitive studies. J Neurosci Meth 2002;115:137–43.

Gur RE, Kohler CG, Ragland JD, Siegel SJ, Lesko K, Bilker WB, et al. Flat affect in schizophrenia: relation to emotion processing and neurocognitive measures. Schizophren Bullet 2006;32(2):279–87.

Hassard TH. Understanding Biostatistics. Mosby Year Book; 1991.

Heimberg C, Gur RE, Erwin RJ, Shtasel DL, Gur RC. Facial emotion discrimination. III Behavioral findings in schizophrenia. Psychiat Res 1992;42: 253–65.

Indersmitten T, Gur RC. Emotion processing in chimeric faces: hemispheric asymmetries in expression and recognition of emotions. J Neurosci 2003;23:3820–5.

Kay S, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophren Bullet 1987;13(2):261–76.

Kohler CG, Turner TH, Bilker WB, Brensinger CM, Siegel SJ, Kanes SJ, et al. Facial emotion recognition in schizophrenia: intensity effects and error pattern. Am J Psychiat 2003;160:1768–74.

Kring AM, Kerr S, Smith DA, Neale JM. Flat affect in schizophrenia does not reflect diminished subjective experience of emotion. J Abnorm Psychol 1993;2:77–106.

Kring AM, Neale JM. Do schizophrenia patients show a disjunctive relationship among expressive, experiential, and psychophysiological components of emotion? J Abnorm Psychol 1996;105:249–57.

LaConte S, Strother S, Cherkassky V, Anderson J, Hu X. Support vector machines for temporal classification of block design fMRI data. Neuroimage 2005;26(2):317–29.

Lien J, Kanade T, Cohn JF, Li CC. Detection, tracking and classification of action units in facial expression. J Robot Autonom Syst 2000;31:131–46.

Littlewort G, Bartlett M, Fasel I, Susskind J, Movellan J. An automatic system for measuring facial expression in video. Image Vis Comput 2006;24(6):615–25.

Lucey S, Matthews I, Hu C, Ambadar Z, De la Torre Frade F, Cohn J. AAM derived face representations for robust facial action recognition. In: Proceedings of the Seventh IEEE International Conference on Automatic Face and Gesture Recognition; 2006. p. 155–60.

Mandal M, Pandey R, Prasad AB. Facial expressions of emotions and schizophrenia: a review. Schizophren Bullet 1998;24(3):399–412.

Morrison RL, Bellack AS, Mueser KT. Deficits in facial-affect recognition and schizophrenia. Schizophren Bullet 1988;14:67–83.

Pantic M, Rothkrantz LJM. Automatic analysis of facial expressions: the state of the art. IEEE Trans Pattern Anal Mach Intell 2000;22(12):1424–45.

Press W, Teukolsky S, Vetterling W, Flannery B. Numerical Recipes in C. Cambridge University Press; 1992.

Salem JE, Kring AM, Kerr SL. More evidence for generalized poor performance in facial emotion perception in schizophrenia. J Abnorm Psychol 1996;105:480–3.

Schneider F, Gur RC, Shtasel DL. Emotional processing in schizophrenia: neurobehavioral probes in relation to psychopathology. Schizophren Res 1995;17(1):67–75.

Schölkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization Optimization and Beyond (Adaptive Computation and Machine Learning). The MIT Press; 2001.

Sison CE, Alpert M, Fudge R, Stern RM. Constricted expressiveness and psychophysiological reactivity in schizophrenia. J Nerv Mental Dis 1996;184: 589–97.

Strang G, Nguyen T. Wavelets and Filter Banks. Wellesley-Cambridge Press; 1996.

Terzopoulos D, Waters K. Physically-based facial modeling, analysis, and animation. J Visual Comput Anim 1990;1:73–80.

Tian Y-L, Kanade T, Cohn JF. Recognizing action units for facial expression analysis. IEEE Trans Pattern Anal Mach Intell 2001;23:97–115.

Tremeau F, Malaspina D, Duval F, Correa H, Hager-Budney M, Coin-Bariou L, et al. Facial expressiveness in patients with schizophrenia compared to depressed patients and nonpatient comparison subjects. Am J Psychiat 2005;162(1):92–101.

Vapnik VN. Statistical Learning Theory. New York: Wiley; 1998.

Verma R, Davatzikos C, Loughead J, Indersmitten T, Hu R, Kohler C, et al. Quantification of facial expressions using high-dimensional shape transformations. J Neurosci Meth 2005;141(1):61–73.

Zhan Y, Shen D. An Efficient Method for Deformable Segmentation of 3D US Prostate Images. International Workshop on Medical Imaging and Augmented Reality (MIAR'04). Beijing, China: Springer-Verlag, GmbH; 2004.

Zhang J, Liu Y, Zhao T. SVM based feature screening applied to hierarchical cervical cancer detection. In: Proceedings of the International Conference on Diagnostic Imaging and Analysis (ICDIA 2002); 2002.

Zhang Z. Feature-based facial expression recognition: sensitivity analysis and experiments with a multilayer perceptron. Inter J Pattern Recogn Artif Intell 1999;13(6):898–911.