Psychophysiology, 48 (2011), 908–922. Wiley Periodicals, Inc. Printed in the USA.

DOI: 10.1111/j.1469-8986.2010.01170.x

An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions

VITALIY KOLODYAZHNIY,[a] SYLVIA D. KREIBIG,[b] JAMES J. GROSS,[b] WALTON T. ROTH,[b,c] AND FRANK H. WILHELM[a]

[a]Department of Clinical Psychology, Psychotherapy, and Health Psychology, University of Salzburg, Salzburg, Austria
[b]Department of Psychology, Stanford University, Palo Alto, California, USA
[c]Department of Veterans Affairs Health Care System, Palo Alto, California, USA

Abstract

The hypothesis of physiological emotion specificity has been tested using pattern classification analysis (PCA). To address limitations of prior research using PCA, we studied effects of feature selection (sequential forward selection, sequential backward selection), classifier type (linear and quadratic discriminant analysis, neural networks, k-nearest neighbors method), and cross-validation method (subject- and stimulus-(in)dependence). Analyses were run on a data set of 34 participants watching two sets of three 10-min film clips (fearful, sad, neutral) while autonomic, respiratory, and facial muscle activity were assessed. Results demonstrate that the three states can be classified with high accuracy by most classifiers, with the sparsest model having only five features, even for the most difficult task of identifying the emotion of an unknown subject in an unknown situation (77.5%). Implications for choosing PCA parameters are discussed.

Descriptors: Emotion, Pattern classification, Feature selection, Autonomic nervous system, Cardiovascular system, Respiration, Electrodermal system, Affective neuroscience, Affective computing

There exists a long tradition of psychophysiological research on differential physiological responding among emotions, originally based on James' (1884) peripheral perception theory and revived by basic emotion theory (e.g., Ekman, Levenson, & Friesen, 1983; see Friedman, 2010, for a review). These theories hold as a central tenet that basic human emotions have distinct physiological patterns. In an influential series of publications, Fridlund and colleagues (Fridlund & Izard, 1983; Fridlund, Schwartz, & Fowler, 1984) argued for the advantages of a formal pattern-classification approach for the study of such physiological patterns of emotion. This approach recognizes the interactive and configurative nature of physiological response systems and determines individual physiological response patterns that are most discriminative of emotional states. With the advancement of computational techniques, this analysis approach is now increasingly applied to study physiological emotion specificity (e.g., Christie & Friedman, 2004; Kreibig, Wilhelm, Roth, & Gross, 2007; Nyklicek, Thayer, & Van Doornen, 1997; Rainville, Bechara, Naqvi, & Damasio, 2006; Sinha & Parsons, 1996). However, although traditional statistical analysis packages often offer only a restricted functionality of pattern classification analysis, the affective computing movement promises a reviving cross-fertilization by applying a full-fledged automated pattern classification approach to such data sets. The present article thus aims to introduce a comprehensive automated pattern classification approach to the study of physiological emotion specificity.

We first review prior research on physiological emotion specificity using pattern classification analysis and identify three aspects of pattern classification analysis, where the application of automated computational methods promises important advancements. In an exemplary fashion, we next demonstrate the various steps of conducting an automated pattern classification analysis on a comprehensive physiological data set. Finally, we discuss implications of the results of the pattern classification analysis and point to important directions for future research. We

The first and second authors contributed equally to this work. This research was supported by the 6th Framework Programme Project EU-CLOCK (018741) funded by the European Commission (F.W., V.K.), the Basel Scientific Society (F.W.), and the National Center of Competence in Research Affective Sciences funded by the Swiss National Science Foundation (51NF40-104897) and hosted by the University of Geneva (S.K.) and the Swiss National Science Foundation (PBGEP1-125914; S.K.). Some of these data were presented at the annual meeting of the Society for Psychophysiological Research (October, 2007).

Address correspondence to: Frank Wilhelm, University of Salzburg, Institute of Psychology, Department of Clinical Psychology, Psychotherapy, and Health Psychology, Hellbrunnerstrasse 34, A-5020 Salzburg, Austria. E-mail: frank.wilhelm@sbg.ac.at

PSYCHOPHYSIOLOGY

introduce some terminology of automated pattern classification, as it will be used throughout the article, in the glossary in Table 1.

Emotion Classification Based on Physiological Signals

The pattern classification approach to physiological response patterns in emotion represents a complementary approach to the traditional group mean analysis. In fact, to test the discriminability of physiological response patterns, pattern classification has been suggested to be the better-suited analysis approach of these two (Fridlund et al., 1984). In pattern classification analysis, units consisting of multiple response measures are classified into well-defined groups (Huberty, 1994). An individual's emotion is thus predicted from a combination of physiological variables (Kreibig, Brosch, & Schaefer, 2010). Classification analysis has therefore been used to corroborate the differentiation of physiological responses between emotions (e.g., Christie & Friedman, 2004; Fridlund et al., 1984; Kreibig et al., 2007; Nyklicek et al., 1997; Rainville et al., 2006; Sinha & Parsons, 1996). This approach is also used in the field of affective computing, where it has inspired a number of applied research projects (e.g., Kim & André, 2008; Kim, Bang, & Kim, 2004; Lisetti & Nasoz, 2004; Nasoz, Alvarez, Lisetti, & Finkelstein, 2004; Picard, Vyzas, & Healey, 2001).

Previous research on emotion classification has documented a considerable degree of patterning among physiological response variables between different emotions (e.g., Friedman, 2010). Thus, it is instructive to review the classification approaches used and associated classification performance of previous studies on emotion classification. It should, however, be kept in mind that numeric results of such previous studies are not directly comparable if the number of emotions considered in the analysis varies between studies. This, in turn, affects the chance classification rate, against which such results are compared (e.g., chance correct classification is 50% if only two emotions are considered, 33.3% if three emotions are considered, and 25% if four emotions are considered).

Research published in psychophysiology journals typically relied on the linear classifier approach and has used discriminant function analysis (DFA)[1] on an ad hoc preselection of a large number of physiological variables (features). Whereas not all studies included cross-validation, if it was included, it was most often of the leave-one-out nature. These studies achieved average classification accuracy on data used for training of 37% to 84% (e.g., Christie & Friedman, 2004; Fridlund et al., 1984; Kreibig et al., 2007; Nyklicek et al., 1997; Rainville et al., 2006; Sinha & Parsons, 1996).

By applying methods of automated pattern classification developed within the affective computing research framework, this approach can be refined and improved in three respects. First, whereas ad hoc selected feature sets may not represent the optimal feature set, *automated feature selection* may achieve improved classification accuracy with a reduced feature subset, thus resulting in a sparser classification model. Picard and colleagues (2001), for example, explored feature selection and transformation with Sequential Floating Forward Search (SFFS), Fisher Projection (FP), and a

[1]Because the focus here is on predicting the outcome (emotion) based on a combination of physiological variables, we use DFA to refer to predictive discriminant analysis, as contrasted to descriptive discriminant analysis that characterizes the observed differences between groups (Huberty, 1994) or—in the present context—emotions.

Table 1. *Glossary of Terms for Automated Pattern Classification Analysis*

Terms	Definition
Chance classification rate	The *classification accuracy* achieved if the data were classified by random guessing.
Class	Discrete emotional state that needs to be discriminated, for example, fearful, sad, or neutral.
Class label	An identifier assigned to a *class*.
Classification accuracy or correct classification rate	The percentage of correct classifications achieved with a particular *classification model* computed as the ratio of correctly classified patterns to the total number of presented *patterns*.
Classification model or classifier	A mathematical model or an algorithm used to automatically assign *class labels* to incoming *patterns*, that is, to predict to which *class* the incoming *observation* belongs.
Cross-validation	Common approach for estimating the *classification accuracy* with unknown data (i.e., data outside of the *training set*). In this approach, the entire data set is divided into N nonoverlapping parts. *Training* and *validation* are performed repeatedly N times. At iteration k of the *cross-validation*, all parts of the data except for the kth part are used for *training*, and the kth part of the data is used for *validation*.
Feature	A variable involved in analysis; in the context of psychophysiological measurements, psychophysiological variables represent features.
Feature selection	Method for selecting a subset of *features* providing optimal *classification accuracy* of the *classification model*. Although it might be counterintuitive that fewer *features* result in better *classification accuracy*, *feature selection* usually results in improvement of *accuracy* by eliminating redundant or irrelevant *features* (Fridlund et al., 1984; Huberty, 1994). This is particularly true if not all variables in the classification analysis contribute substantially to the intergroup differences. By adding or substituting *features*, the variance overlap engendered under these conditions may alter the roles of some *features* from discriminative variables to moderator or suppressor variables, thus affecting the patterns of correlation structure, and by this, *classification accuracy*.
Observation	Data point representing one emotion for a particular subject.
Overfitting	Poor *classification accuracy* on the *validation data set*, while the *classification accuracy* on the *training set* approaches 100%, that is, the *model* loses its generalizability due to an excessive number of adjustable parameters in the model in relation to the number of *observations* in the *training data*.
Pattern	A vector comprising feature values for one *observation*.
Pattern classification	The process of assigning *class labels* to *observations* or the result of this process.
Training	Determining the parameters of a *classification model* based on a special representative data set called *training data*.
Training data	Representative data set for which the *class labels* are known and based on which the parameters of a *classification model* are determined.

Table 1. (Contd.)

Terms	Definition
Validation	The estimation of *classification accuracy* of a *classification model* with a data set different from that used during *training*, called *validation data*.
Validation data	Representative data set for which the *class labels* are known and based on which the *classification accuracy* of a *classification model* is tested; the *validation data set* should be nonoverlapping with the *training data set*. A good *classification model* with properly chosen *training data* should have comparable *classification accuracy* with the *training and validation data sets*.

combination of both (SFFS-FP) and applied the *k*-nearest neighbors algorithm (KNN) and Maximum a Posteriori (MAP) for classification. Average classification accuracy was up to 81% with 2 to 19 features on eight emotion classes.

Second, whereas linear classifiers have been applied widely to the study of physiological emotion specificity, it remains unclear whether this is the *most suitable type of classifier*. A systematic comparison of the performance of different linear and nonlinear classifiers is therefore necessary. Linear discriminant functions are best suited for data with linear boundaries between different classes in the feature space classes and are inherently connected with the assumption of normality of data distribution in each class, such that the covariance matrices of the multivariate normal distributions for each class are equal (Hastie, Tibshirani, & Friedman, 2001). A *linear classifier* decides class membership by comparing a linear combination of the features to a threshold and is represented by a line in two dimensions or a hyperplane in higher dimensions. If the covariance matrices of different classes cannot be assumed to be equal while the assumption of normality is still valid, *quadratic discriminant analysis* can be applied (Hastie et al., 2001). If the class boundaries are nonlinear and the assumption of normality is not valid, methods that do not rely on the type of data distribution are worth attention. Among a great variety of such nonlinear methods, particularly the KNN algorithm (Cover & Hart, 1967) and artificial neural networks (ANN; Haykin, 1999; for a detailed description of these approaches, see below) have found application in research on emotion classification. Work by Nasoz, Lisetti, and colleagues, for example, analyzed classification of physiological emotional responding using DFA, KNN, and ANN (Lisetti & Nasoz, 2004; Nasoz et al., 2004; Nasoz, Ozyer, Lisetti, & Finkelstein, 2002). They found classification accuracies between 72% and 75% using DFA, 72% and 84% using KNN, and 84% using ANN. This suggests that nonlinear classifiers can achieve higher classification accuracy.

Third, whereas most of the emotion classification studies did not test their classifiers with a different data set than it was trained with, *cross-validation* is an essential step for evaluating the performance of a given classifier in case of unknown data. Moreover, as individual and situational specificity has long been acknowledged in psychophysiological research (Hinz, Seibt, Hueber, & Schreinicke, 2000; Lacey & Lacey, 1958; Marwitz & Stemmler, 1998), it is important that also emotion classification studies acknowledge the need for a differential treatment of subject and situation dependence in trained classifiers. Supporting the importance of considering the influence of such varying factors, generalizability theory (Webb & Shavelson, 2005) suggests that facets, across which generalization is sought, such as subject and stimulus characteristics should be taken into account.

Those studies on emotion recognition that employed cross-validation (e.g., Kim & André, 2008; Kreibig et al., 2007; Lisetti & Nasoz, 2004; Picard et al., 2001; Rainville et al., 2006) were based on the leave-one-out approach. In this approach, one observation representing one emotion for a particular subject is left out at a time, the rest of the data is used for training of the classifier, and the left-out observation is used for validation. This means that both a known subject and known stimulus material (except for one emotion for the classified subject) are used in the training of the emotion classifier. However, this approach does not reflect the accuracy of emotion classification when all measurements to be classified come from a subject not included in the training set or are induced by stimulus material (e.g., film clips) other than those used for training of the classifier. Only a few studies have implemented subject-independent (e.g., Bailenson et al., 2008) or context-independent cross-validation (e.g., Sinha & Parsons, 1996) in the context of psychophysiological emotion classification. Still, a comprehensive evaluation of subject- *and* context-independent cross-validation, has—to our knowledge—not yet been applied to this research question. For both theoretical questions of advancing our knowledge regarding the tenability of physiological emotion specificity and practical applications of human–computer interactions (HCI; Picard, 2000), subject-independent or stimulus-independent emotion discrimination is of central importance. On the theoretical level, generalizability theory (Webb & Shavelson, 2005) recognizes that measurements can be independently influenced by a number of error sources, including subject and stimulus characteristics. In an applied context, a subject- or stimulus-independent technique could be employed, for example, in an HCI system capable of emotion recognition without need of adaptation to an unknown user or unknown context.

Taken together, the current psychophysiological approach to emotion classification calls for improvement and refinement in three respects: (1) applying an objective method for selecting a minimal or optimal feature subset, rather than ad hoc selected features; (2) systematically evaluating and making an informed choice of the specific type of classifier rather than exclusive reliance on DFA; and (3) using a systematic approach to ensure both subject and stimulus independence of results rather than nonindependence of cross-validation. To realize improvements 1 and 2, various forms of feature selection methods and classifiers are explored in the present article. To address improvement 3, we propose the following schematic of different types of cross-validation, summarized in Table 2. The "standard technique" or "leave-one-out cross-validation" used in many other studies corresponds to the upper left cell of the table (*subject- and stimulus-dependent cross-validation*). The upper right cell represents *subject-independent cross-validation* or the "leave one subject out" approach, whereas the lower left cell represents *stimulus-independent cross-validation*. Finally, the lower right cell combines the two latter approaches to *subject- and stimulus-independent cross-validation*.

Linear and Nonlinear Classification Approaches

The classification models that we consider in this article fall into the following three categories:

1. *Discriminant analysis*. Two models fall within the category of discriminant analysis, namely, the classical linear discriminant

Table 2. *Types of Cross-validation*

	Known subject	Unknown subject
Known stimulus material	*Subject- and stimulus-dependent cross-validation* One sample representing one emotion for a particular participant is left out at a time; the rest of the data is used for training of the classifier and the left out sample for validation of the classifier.	*Subject-independent cross-validation* In each iteration, all measurements corresponding to a particular participant are removed from the training set and used for validation.
Unknown stimulus material	*Stimulus-independent cross-validation* All measurements from one of the two film sets are used for training and from the other film set for validation.	*Subject- and Stimulus-independent cross-validation* The data set is divided into two parts corresponding to the two film sets. Then, each of the two parts is used for a cross-validation, where all measurements corresponding to a particular participant are removed from the training set and validation is done with the respective measurements for the same participant but from the other film set. Results from the two runs are averaged.

analysis (LDA) and quadratic discriminant analysis (QDA; for details, see Hastie et al., 2001). LDA is a proven, simple, reliable, and therefore widely used classification technique. QDA is an extension of LDA that allows the covariance matrices of individual classes to differ from each other and the class boundaries to be quadratic rather than linear. The parameters of the algorithm estimated based on the data are the covariance matrices and class means in the feature space. The LDA estimates one covariance matrix, which is assumed equal for all the classes, whereas the QDA estimates a separate covariance matrix for each class.

2. *Artificial neural networks*. The most popular types of artificial neural networks used for pattern classification are multilayer perceptrons (MLP; Haykin, 1999) and radial basis function networks (RBFN; Moody & Darken, 1989) models. The advantages of neural networks compared to the conventional techniques, like LDA, is the ability to cope with data with arbitrary nonlinear input–output relations and non-Gaussian distributions. In the context of pattern classification, this also means that the boundary between classes in the feature space can be different from a linear one and are not restricted to a particular type of nonlinearity. An ANN consists of a number of elementary processing units called artificial neurons, which work in parallel in a way inspired by the parallel information processing in living brains. The desired input–output mapping of an ANN is realized via the so-called *training algorithm* based on input–output examples (input patterns and desired responses) from a representative *training data set*. In such a way, an ANN learns to *approximate* the distribution of data in the training set and to *generalize* for unknown data outside of the training set. The RBFN model is closely related to Gaussian mixtures (Yiu, Mak, & Li, 1999) and approximates non-Gaussian distributions with a weighted combination of Gaussian basis functions. The MLP model is constructed of two or more layers of artificial neurons, each corresponding to a logistic regression model (Hastie et al., 2001) capable of linear classification. By combining such elementary classifiers in layers, an arbitrary nonlinear boundary between classes can be constructed. In ANNs, all layers of neurons except the output layer are called "hidden layers." An RBFN always has one hidden layer of basis functions, whereas an MLP can have a different number of hidden layers, usually one or two, rarely more. We used an MLP model with one hidden layer because of the relatively small size of our data set. Using only one layer reduces the number of model parameters and thus guards against *overfitting* that may result in poor generalization to unknown data.
3. *Memory-based classification*. The KNN algorithm is *memory-based* in the sense that it has no model to fit and the classification result is determined as a majority vote of the k nearest neighbors of the observation that is being classified (Cover & Hart, 1967). The KNN model does not involve any training other than assigning a "training" data set to the "memory" of the classifier, from which the nearest neighbors are being selected in the classification phase based on the distances from the classified observation. Despite its simplicity, the KNN algorithm is one of the most successful classification techniques (Hastie et al., 2001). Like ANN, the KNN algorithm can classify data with complex nonlinear boundaries between classes in the feature space.

The Present Study

In the present study, we tested the applicability of feature selection, linear and nonlinear classifiers, as well as systematic cross-validation (see Table 2) to automated emotion classification. To this aim, we reanalyzed psychophysiological data from an emotion elicitation study using film clips (Kreibig et al., 2007). This data set is of particular interest for the present research question as it consists of a test and retest within the same experimental session with different stimulus material. In the previous study, Kreibig et al. (2007) used a predictive discriminant analysis based on 14 features. This approach achieved an average classification accuracy of 84.5% with training data and 69.0% with validation data using the leave-one-out approach, that is, when one observation at a time was removed from the training data set and was used for testing. The data set consisted of 84 data points altogether, collected from 28 subjects (recordings from another 6 subjects had missing values and were not included in that analysis). The data were averaged between two sets of three films each, thus giving three data points for each subject. For the present study, the reanalyzed data set contained 204 data points coming from six film clips (three clips from two separate film sets) for all 34 subjects, with missing values interpolated.

Method

Participants

Thirty-four healthy student volunteers (19 women), age 18 to 26 years, participated in the experiment. All participants provided

written informed consent prior to the experiment and were paid US$30.

Materials

Six 10-min film clips (two frightening, two sad, and two neutral) were used to elicit fear, sadness, and a neutral emotional state. Film clips were edited from the movies *I Know What You Did Last Summer* and *I Still Know What You Did Last Summer* (for fear), *Steel Magnolias* and *John Q*. (for sadness), and *Alaska's Wild Denali* (two different clips for a neutral emotional state). Each clip was preceded by a 30-s audio introduction. For a detailed description of the content of the film clips, see Kreibig et al. (2007).

Startle was elicited with brief (50 ms) bursts of 95 dB(A) white noise (20–20,000 Hz), 0.1-ms rise/fall time (Lang, 1995), administered binaurally by headphones, with a 60-s intertrial interval (unpredictably varying). There were a total of 10 stimuli per film and three stimuli per pre-film rest period.

Procedure

Data recording was conducted individually for each participant, who was seated in a chair. After attachment of physiological sensors and amplifier calibration, participants viewed the series of six film clips, presented in a Latin square design between participants. Films were preceded by 3-min rest periods, during which participants were instructed to sit quietly. Immediately after each film, participants' retrospective feeling self-report to the film clip were assessed (self-report results have been reported in Kreibig et al., 2007). Finally, participants were disconnected from monitoring devices, debriefed, reimbursed, and dismissed.

Physiological Measures

An SA Instruments (San Diego, CA) 12-channel bioamplifier was interfaced with a Data Translation (Marlborough, MA) DT3001 PCI 12-bit 16-channel A/D conversion board to a computer, which sampled physiological channels at 400 Hz, except for electromyography (EMG), which was sampled at 1000 Hz, and impedance cardiography, which was sampled at 1200 Hz. Recorded signals were analyzed and averaged for each film period using an integrated set of biosignal analysis programs written in MATLAB (Mathworks, Inc., Natick, MA; Wilhelm, Grossman, & Roth, 1999; Blechert, Lajtman, Michael, Margraf, & Wilhelm, 2006; see SPR Software Repository, http://www.spr-web.org/). Table 3 provides an overview of the assessed physiological measures. They represent a broad range of organismic responding, including responses from the cardiovascular, electrodermal, respiratory, and motor systems. Details of parameter quantification are reported in Kreibig et al. (2007).

Data Reduction

Artifactual epochs were edited manually for each channel. Film periods were subdivided into three intervals of equal length, averaging 180 s, for each of which a mean score was calculated. Likewise, the average of the 180-s baseline preceding each film clip was calculated.

Missing data points were interpolated following the procedure described by Stemmler (1989). Percentage of missing data for individual physiological variables ranged from 9 to 33 of 816 conditions or 1.10% to 4.04% of total conditions recorded (1.93% on average).

Reactivity scores were calculated by subjecting the data to a modified Z-standardization within participants based solely on the pooled within-conditions sums of squares (Stemmler, 1987). Maximal reactivity from baseline was determined for each emotion induction, resulting in one score per film clip. By quantifying responses as maximal reactivity from baseline over three consecutive 3-min averages over the 10-min film clip, we strove to combine advantages of increased reliability from averaging over relatively long measurement intervals with increased sensitivity to short-term reactivity changes (Ax, 1953). Determination of maximal reactivity was guided by empirically based hypotheses about direction of change (Kreibig et al., 2007) and confirmed by inspection of waveforms. Briefly, the following algorithm was used: For heart rate (HR) and systolic blood pressure (SBP), reactivity was defined as maximum increase from baseline for the fear condition, whereas maximum decrease was used for the sad and neutral conditions. Conversely, for preejection period (PEP), end-tidal carbon dioxide partial pressure (pCO_2), and finger temperature (FT), reactivity was defined as maximum decrease from baseline for the fear condition and as maximum increase for the sad and neutral conditions. Reactivity was defined as maximum increase from baseline for skin conductance level (SCL), rate of skin conductance fluctuations (SRR), respiratory rate (RR), general somatic activity (ACT), startle response magnitude (SM), musculus corrugator supercilii (CS), and musculus zygomatic major (ZM) and as maximum decrease from baseline for respiratory sinus arrhythmia (RSA) and tidal volume (Vt) for all conditions. By subtracting the prior rest period from the film mean score, standardized difference scores were determined (additional detail is available in Kreibig, 2004). For each of 34

Table 3. *Assessed Physiological Channels, Computed Features, Abbreviations, and Units*

Channels	Features	No.	Abbr.	Unit
Electrocardiography	Heart rate	1	HR	beats per min
Electrocardiography	Respiratory sinus arrhythmia	2	RSA	ms^2
Impedance cardiography	Preejection period	3	PEP	ms
Blood pressure monitor	Systolic blood pressure	4	SBP	mmHg
Electrodermal activity	Skin conductance level	5	SCL	μS
Electrodermal activity	Rate of skin conductance fluctuations	6	SRR	fluctuations per min
Respiratory inductive plethysmography	Respiratory rate	7	RR	cycles per min
Respiratory inductive plethysmography	Tidal volume	8	Vt	ml
Capnography	End-tidal carbon dioxide partial pressure	9	pCO_2	mmHg
Temperature sensor	Finger temperature	10	FT	°C
Piezo-electric sensor	General somatic activity	11	ACT	g
Eyeblink EMG	Startle response magnitude	12	SM	mV
Facial expression EMG	Musculus corrugator supercilii	13	CS	mV
Facial expression EMG	Musculus zygomatic major	14	ZM	mV

participants, six scores corresponding to each of the six film clips were computed, giving 204 data points altogether.

Classification Analysis

The problem to be solved was to classify this data set into three classes corresponding to the emotion that a particular film induced (fear, sadness, and neutral). Our classification analysis was performed for the five classification models introduced above (LDA, QDA, MLP, RBFN, and KNN; see also subsection "Classification models" below for implementation details) with the four different cross-validation types (see Table 2). The analysis consisted of the two main steps of feature selection and cross-validation:

1. *Feature selection for four different types of cross-validation*. At this step, features providing the best classification accuracy were selected based on the best performance of the respective classification models at each of the cross-validation iterations with the respective training data. The feature selection algorithms were run inside the respective cross-validation loops every time with different training data corresponding to the current cross-validation iteration, that is, N times where N is the number of cross-validation iterations ($N = 2$, 34, or 204, depending on cross-validation type, for details see the subsection Cross-validation below). Thus, separate subsets of selected features varying between the cross-validation iterations were produced. Where applicable, additional structure optimization was performed for the respective classification models. The whole procedure was repeated for five models and four different cross-validation types, that is, 20 times. Afterward, 20 subsets of most frequently selected features were determined from the results of feature selection, one subset for each combination of the five models and four cross-validation types.
2. *Cross-validation with preselected 20 subsets of most frequent features obtained from Step 1* with additional structure optimization for classification models (where applicable) and best model selection for the respective types of cross-validation. Additional structure optimization was performed, because feature subsets at this step remained constant for the entire cross-validation, in contrast to Step 1. Estimates of classification accuracy for five classification models with four cross-validation types were obtained, and the best model for each of the four cross-validation types was determined.

Below, we explain the algorithms and models that we used in our analysis in Steps 1 and 2 for feature selection, classification, and cross-validation.

Feature selection. To determine optimal feature sets for each of the classification models, sequential feature selection algorithms were used: sequential forward selection (SFS) and sequential backward selection (SBS). With SFS, the algorithm starts with an empty set of features. A new feature is selected if the model performs better after the inclusion of this feature (i.e., if there are fewer classification errors). The algorithm stops when no improvement in model performance can be achieved or when a maximum number of features are selected.

With SBS, the algorithm starts with a full set of features and excludes them one by one if such exclusion results in improved or the same model performance. The algorithm stops when no performance improvement can be achieved or a minimum number of features remain in the feature set.

In our analysis, the feature selection procedures were included in the cross-validation loop as in Picard et al. (2001; for a substantiation of such an approach, see Zhang et al., 2006). According to this approach, feature selection in our computations was done with the training data only and repeated as many times as the number of iterations of cross-validation. For each model and cross-validation, the two feature selection procedures, SFS and SBS, were tried. Because the training data in each of the cross-validation iterations is different, feature subsets produced by feature selection procedures may also differ between cross-validation iterations. After the cross-validation with feature selection, we determined a subset of most frequently selected features between all cross-validation iterations as the overall result of the feature selection run: features selected by majority vote, that is, in more than 50% of cross-validation iterations (relative frequency $> 50\%$), were included in this subset. A similar feature selection procedure was performed for five classification models and for four different cross-validation types described in Table 2.

Classification models. Classification models that were used in our analysis are listed in Table 4. As in Kreibig et al. (2007), we used LDA for pattern classification, and additionally QDA

Table 4. *Classification Models*

Model	Standard abbreviation	Abbreviation in our analysis	Structure optimization	Notes
Linear Discriminant Analysis	LDA	L	N/A, one discriminant function per class	Linear class boundaries
Quadratic Discriminant Analysis	QDA	Q	N/A, one discriminant function per class	Quadratic class boundaries
Multilayer Perceptron	MLP	Mxx	Number of units (neurons) in hidden layers can vary	Arbitrary class boundaries can be approximated by hyperplanes defined by the parameters of artificial neurons
Radial Basis Function Network	RBFN	Rxx	Number of basis functions can vary	Arbitrary class boundaries, data distribution is approximated by a weighted combination of Gaussian basis functions
K-Nearest Neighbors	KNN	Kxx	Number of nearest neighbors can vary	Arbitrary class boundaries, no training is required, pattern classification is based on the voting scheme by the majority of the nearest patterns (neighbors)

Note: Models are specified as Mxx = MLP with xx hidden layer neurons, Rxx = RBFN with xx basis functions, Kxx = KNN with xx neighbors. For LDA and QDA models structure optimization is not applicable. Models MLP and RBFN belong to the class of artificial neural networks.

(Hastie et al., 2001), MLP (Haykin, 1999), RBFN (Moody & Darken, 1989), and the KNN algorithm (Cover & Hart, 1967). Particular models and reasons for choosing them are explained in the introduction above. Here we address some implementation-specific details to ensure replicability.

Both LDA and QDA models contain no additional parameters that need to be adjusted outside of the classification algorithm itself. The MLP, RBFN, and KNN models in our analyses have only one additional adjustable parameter that defines the complexity of their internal structure, subsequently called "the number of nodes." By "nodes" we mean hidden layer neurons in MLP, basis functions in RBFN, or "neighbors" in KNN. This parameter needs to be chosen carefully. For ANN, too few nodes result in poor performance on both the training and validation data set because the model is too simple to capture the relations between the input and output variables. On the other hand, too many nodes result in overfitting. For KNN, too few nodes result in too "noisy" classification whereas with too many nodes the boundaries of classes become too "smooth." In both cases the performance of the classifier is degraded. Because the number of nodes is an integer and has a limited range of possible values, it can be easily optimized by exhaustive search. More details on choosing the number of nodes are given at the end of this section under the heading "Structure optimization."

For the LDA, QDA, and KNN classifiers we used our own MATLAB implementation, whereas for the MLP and RBFN models we chose the free MATLAB neural network toolbox, NetLab (Nabney, 2003, 2004). The MLP model had the logistic activation function in the hidden layer and the softmax activation function in the output layer, which is a usual choice for classification problems. The RBFN had Gaussian basis functions. The MLP was trained with the scaled conjugate gradient algorithm, and for the RBFN the training algorithm was expectation maximization (Nabney, 2004). The number of training epochs for both the MLP and RBFN classifiers was 25. More information on ANNs can be found in Haykin (1999).

Cross-validation. Because our data contain recordings of emotions from a relatively large number of subjects ($N = 34$) and two different sets of film stimuli with three types of emotions each, we were able to simulate in our experiments all four possible combinations of cross-validation with known or unknown subjects and stimulus material. See Table 2 for a description of different types of cross-validation that were used to simulate these four combinations. These four different types of cross-validation were performed with all five classification models: LDA, QDA, MLP, RBFN, and KNN.

As we mentioned in the introduction, among the four presented types of cross-validation, only the leave-one-out cross-validation is usually encountered in the literature. In contrast to many other studies, we paid special attention to testing different classification approaches in settings with *subject independence*, *stimulus independence*, and *a combination of both*.

Our leave-one-out cross-validation used for comparison is similar to the usual approach, where one measurement is taken out of the data set and used to validate the classification model or algorithm trained with the rest of the data. This is done iteratively with all measurements. With our data, this was a 204-fold cross-validation, as the data set contained 204 measurements. Whereas the leave-one-out setting is the conventional approach, it is not very useful for real-world applications because it does not provide estimates of classification accuracy with unknown subjects or with unknown stimulus material (i.e., *subject- and stimulus-dependent cross-validation*). In order to simulate subject-independent, stimulus-independent, and the combination of both subject- and stimulus-independent classification, we did the following.

For *subject-independent cross-validation*, we carried out a 34-fold cross-validation removing *all* six measurements corresponding to a particular subject for *all* film clips (two film sets with three clips each) from the data set. The remaining 198 measurements from the other 33 subjects were used for training the classifiers, and the data from the removed subject was used for validation. Thus, this was *subject-independent cross-validation* because the data set used for training did *not* contain any data from any of the six film clips for the subject used for validation and, vice versa, the data used for validation were from the subject that was *completely* removed from the training data set. This procedure was repeated 34 times for each of the subjects in the sample, and results from the 34 iterations were averaged. For a substantiation of this approach, which is also called "leave-one-subject-out," see Esterman, Tamber-Rosenau, Chiu, and Yantis (2010). In other domains, this approach is the generally accepted choice for ensuring subject independence as opposed to the subject-dependent leave-one-out approach (see, e.g., Ho, Brown, & Serences, 2009; Howard, Plailly, Grueschow, Haynes, & Gottfried, 2009; Zhao & Lu, 2005).

For *stimulus-independent cross-validation*, no subject independence was assumed. We simulated the classification of unknown stimulus material for known subjects. In this setting, the data were divided into two halves, each containing 102 measurements from one of the two sets of film clips of the three types (fearful, sad, and neutral). The cross-validation was twofold: First, the classifiers were trained with 102 measurements of the first film set and validated against 102 measurements from the second film set. Then the procedure was repeated using the second film set for training and the first set for validation. Results from these two iterations were averaged.

For *subject- and stimulus-independent cross-validation*, we assumed *subject and stimulus independence at the same time*. The data set was divided as in the previous case in two parts containing 102 measurements each and corresponding to two film sets. Then we ran two 34-fold cross-validations in which three measurements corresponding to a particular participant (from 1 to 34) were removed from the training set, the training was based on the remaining 99 measurements, and the validation was carried out with the respective three measurements for the removed participant but from the other film set. The results from the two 34-fold cross-validations were averaged. Thereby, we ensured that the classification model or algorithm was validated using data from *both* a subject and a film set that were *not* used during training.

Training and validation data were restandardized in each cross-validation iteration to simulate a real situation where new observations with unknown mean and standard deviation need to be classified with a classification model based on the statistical characteristics of the known training data only. The data standardization for each of the cross-validation iterations was done as follows: First, means and standard deviations were computed for the respective variables in the training set. Then the means of the training data were subtracted from both the training and validation data, and the results were divided by the respective standard deviations of the training data.

Structure optimization. For MLP, RBFN, and KNN models, another external loop of structure optimization was executed in

order to find the number of nodes that yields the highest classification accuracy for the validation data set. Thus, the cross-validation for the respective classification models was iterated with an increasing number of nodes, and the optimal number of nodes was chosen based on classification accuracy. The minimum number of nodes was 2 for MLP and 3 for RBFN and KNN. The maximum number of nodes was 10 for MLP and 20 for RBFN and KNN. These numbers were chosen based on several trial runs. Further increasing of the indicated upper bounds on the number of nodes did not result in better classification models.

Information on Computational Complexity

The computations were performed in MATLAB v. 6.5 (The MathWorks, Inc., Natick, MA) under Windows XP (Microsoft, Redmond, WA) on a PC with 2 GB of RAM and an Intel Core Duo T2500 CPU running at 2.0 GHz. The total execution time was approximately 35 h for all combinations of five models, four types of cross-validation, two types of feature selection, and with structure optimization for MLP, RBFN, and KNN. On average, one feature selection procedure took about 4 s. For each of the LDA and QDA models, the feature selection procedures were run $2 \cdot (204+34+2+2 \cdot 34) = 616$ times. Taking into account structure optimization for the nonlinear models, the number of feature selection runs for MLP was $616 \cdot 9 = 5{,}544$, and $616 \cdot 18 = 11{,}088$ for each of the RBFN and KNN models. Thus, the total number of runs of the feature selection procedures was $2 \cdot 616+5{,}544+2 \cdot 11{,}088 = 28{,}952$. The total execution time of cross-validation with constant feature subsets was about 1 h including structure optimization for MLP, RBFN, and KNN.

Results

Feature Selection

First, we determined optimal subsets of features for the four different types of cross-validation shown in Table 2 using SFS and SBS algorithms. Outcomes of thus selected features are reported in Table 5. The quality of the feature selection was estimated based on the classification accuracies on the validation data sets, whereas the selection of features in the respective feature selection procedures was driven by the accuracy on the training set in each individual cross-validation run as described above. Classification accuracies in Table 5 were computed as averages of all cross-validation iterations for the respective models. In what follows, by "classification accuracy" we refer to that of the validation data unless otherwise indicated.

Classification accuracy ranged between 66.2% (based on 13 features) and 79.4% (based on 7 features), and was highest most often with the RBFN classifier (three times), followed by the KNN and MLP classifiers (both two times), which co-occurred with RBFN in a tie.

Feature subsets were determined separately in each of the cross-validation iterations because the feature selection procedures were run inside the cross-validation loop. The number of selected features ranged between 5 and 14 (i.e., the complete feature set). Features selected in more than 50% of cross-validation iterations are marked in Table 5 with a cross. Only the best results based on classification accuracy among the two feature selection procedures SFS and SBS (designated by "F" and "B," respectively) are shown. In some cases, the feature selection procedure did not improve the average classification accuracy for training data; therefore all 14 features were selected. For such cases, the best feature selection algorithm is designated by "N" (none).

Applying SBS most often resulted in the highest classification accuracy with the selected feature subset, followed by the count of N, that is, using the full feature set, and least often by applying SFS. Out of 20 analyses run, the full feature set resulted in highest classification accuracy nine times, a feature subset of equal to or more than 10 features resulted in highest classification accuracy four times, and a feature subset of equal to or less than 8 features resulted in highest classification accuracy seven times. These findings suggest that in more than half of the cases a considerable reduction of features used in classification is possible and achieves better classification outcomes than the full feature set, resulting in a less complex and thus sparser model.

Notably, in two cases, the same or similarly high classification accuracies were obtained with different classifier methods based on the full feature set or a significantly simplified feature subset: for the subject-independent cross-validation, we found a classification accuracy of 77.5% for MLP with 9 hidden layer neurons and 8 features and 77.9% for RBFN with 10 basis functions and 14 features. Similarly, for the stimulus-independent cross-validation, we found a classification accuracy of 77.9% for both RBFN with 9 basis functions and 14 features and KNN with 17 neighbors and 5 features. These results point to the importance of considering the application context and classifier used when trying to determine the appropriate number of features for classification.

It can also readily be seen that there are five common features that were selected in all runs for all models: PEP, SRR, pCO_2, CS, and ZM.

Cross-validation with Constant Feature Subsets

As was noted above, for practical applications it is reasonable to have a constant subset of features. Therefore we used the feature subsets selected by feature selection (as identified in Table 5), the full set of 14 features, and the common 5 features (PEP, SRR, pCO_2, CS, and ZM) to determine which feature subsets provide the best classification accuracy and with which models. Results are reported in Table 6. Table 6a contains results of cross-validation with the most frequent features (those encountered in more than 50% of feature selection results in Table 5) for each of the models separately. Table 6b contains results with the five common features for all models from Table 5 (PEP, SRR, pCO_2, CS, and ZM). Table 6c contains results with the full set of all 14 features. This information is used as a basis for comparison indicating the amount of improvement in classification accuracy that can be achieved with feature selection, given the added complexity in computation, on the one hand, but the outlook of a sparser model (and thus the need to assess fewer parameters) on the other hand.

In all the runs with constant feature subsets (results shown in Table 6), an external structure optimization loop was executed for the MLP, RBFN, and KNN models. This resulted in a different optimal number of nodes for these modes compared to those shown in Table 5, where the feature subsets varied among cross-validation iterations because a feature selection procedure was executed inside the respective cross-validation loops. As can readily be seen, there is little difference in accuracy between the previous analysis reported in Table 5 and the present one reported in Table 6, that is, with only five features it is possible to classify the three emotion states quite well.

Table 5. *Summary of Results of the Feature Selection Process*[a]

	Subject- and stimulus-dependentcross-validation					Subject-independent cross-validation				
Model	L	Q	M6	R10	K19	L	Q	M9	R10	K11
Feature selection	N	F	N	F	F	N	F	F	B[b]	B
Training accuracy (%)	81.3	86.6	96.1	80.9	82.8	81.4	86.2	96.8	79.6	83.9
Validation accuracy (%)	77.0	74.5	76.0	77.9	79.4	76.0	72.1	77.5	77.9	76.5
	Selected features					Selected features				
HR	x		x			x		x	x	x
RSA	x	x	x			x			x	
PEP	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**
SBP	x	x	x	x	x	x	x	x	x	x
SCL	x		x			x			x	x
SRR	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**
RR	x		x			x			x	x
Vt	x		x			x			x	
pCO_2	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**
FT	x		x	x	x	x		x	x	x
ACT	x	x	x			x			x	
SM	x		x			x			x	x
CS	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**
ZM	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**
No. of features	14	8	14	7	7	14	6	8	14	11
	Stimulus-independent cross-validation					Subject- and stimulus-independent cross-validation				
Model	L	Q	M7	R9	K17	L	Q	M3/4[c]	R13/11[c]	K5/8[c]
Feature selection	B	B	N	N	B	B	B	B/N[d]	N	N
Training accuracy (%)	85.3	98.0	97.1	77.5	81.4	85.0	98.4	94.9	80.6	75.1
Validation accuracy (%)	77.0	71.1	74.0	77.9	77.9	73.5	66.2	76.5	76.5	75
	Selected features					Selected features				
HR		x	x	x			x	x	x	x
RSA		x	x	x			x	x	x	x
PEP	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**
SBP	x	x	x	x		x	x	x	x	x
SCL		x	x	x			x	x	x	x
SRR	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**
RR			x	x		x	x	x	x	x
Vt			x	x				x	x	x
pCO_2	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**
FT	x	x	x	x		x	x	x	x	x
ACT	x	x	x	x		x	x	x	x	x
SM			x	x		x	x	x	x	x
CS	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**
ZM	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**	**x**
No. of features	8	11	14	14	5	10	13	14	14	14

[a]Marked with x are the most frequent features (selected in more than 50% of cross-validation iterations) for the different models. The best feature selection method is indicated (F = sequential forward selection, B = sequential backward selection, N = none). Classification accuracy based on idiosyncratic features is indicated for the training set and for the validation set. Five common features for all models are shown in bold. See Table 3 for abbreviations of features and Table 4 for abbreviations of models.
[b]Relative frequencies of selection for all 14 features were above 50%.
[c]Optimal number of nodes for the first/second film set.
[d]Backward feature selection for the first film set and no feature selection for the second film set.

Statistical Significance of Classification Results

Classification results were highly significantly better than chance for all the five models in all cross-validation runs; therefore the respective significance levels are not shown separately in Table 6. *T* tests with 33 degrees of freedom against the expected average classification accuracy of a random classifier (33.3%) resulted in *p* values smaller than 10^{-8} for all the models in all runs. Similar levels of significance of the classification results ($p < 10^{-9}$) with all the five models in all runs were indicated by *p* values computed with Huberty's (1994) method as in the work by Kreibig (Kreibig, 2004; Kreibig et al., 2007). For these tests, classification results were averaged for each of 34 subjects, giving 34 numbers corresponding to average classification accuracy with the respective classification models.

Comparison of Classification Models

We used the Wilcoxon matched pairs test (Wilcoxon, 1945) to find significant differences between classification models using the same 34 values of average classification accuracy as in significance tests of classification results. These 34 numbers of the respective models were compared pairwise among different models (for the same cross-validation type) as dependent samples.

We also compared models of the same cross-validation type with different feature sets: the models with the most frequent features to the respective models with the five common features and to the models with the complete set of 14 features, as well as the models with the five common features to the models with 14 features.

As can be seen from the results in Table 6, the nonlinear models (MLP, RBFN, and KNN) were generally better than the LDA classifier. However, improvement was only marginally significant ($p \geq .063$) when nonlinear models with higher classification accuracy were compared to LDA. Similarly, differences reached only marginal significance ($p \geq .094$) when models with five common features were compared to the respective models with the most frequent features found with feature selection algorithms.

Best Classification Models

Details of the best classification results for the four different types of cross-validation with preselected feature sets across feature selection and classification methods are shown in Table 7. The KNN model was the best for both subject- and stimulus-dependent and subject- and stimulus-independent classification. The RBFN model was the best for subject-independent classification (the KNN model provided the same accuracy but with 11 features, whereas the RBFN model was sparser, being based on only 5 features), and the simplest LDA model was the best for stimulus-independent classification. For subject-independent and subject- and stimulus-independent classification, the same smallest subset of five common features provided the best results.

It is interesting to note that the correct classification rate for sadness was 81% for the subject- and stimulus-dependent cross-validation (upper left corner of Table 7) and 69%–70% for the other types of cross-validation. For fear, the highest classification accuracy was achieved with stimulus-independent cross-validation (88%), whereas other cross-validation types resulted in classification accuracy between 78% and 82%. For neutral emotional states, the difference in classification accuracy between different types of cross-validation was rather small.

Complementary Analysis without Data Interpolation for Missing Values

To check the impact of data interpolation for missing measurements, we did the same analyses with five classification models and four cross-validation settings but with data containing only 28 subjects without any missing values in the measured variables (as was done in Kreibig et al., 2007). This data set contained 168 data points instead of 204 for 34 subjects. Average classification accuracy for best models ranged from 73.2% (KNN, subject- and stimulus-independent cross-validation) to 80.4% (KNN, subject- and stimulus-dependent cross-validation), that is, it was very close to that achieved for the data set with 34 subjects and interpolated missing values. Interestingly enough, the LDA model was the best for stimulus-independent cross-validation, just as it was for the data set with 34 subjects. In the other three cross-validation settings, the nonlinear models provided better results than LDA. These additional results confirmed that neither the percentage of missing data in the data set with 34 subjects nor the data interpolation technique qualitatively impacted our main findings.

Discussion

In the present study, we tested the applicability of three improvements and refinements of pattern classification analysis: automated feature selection, various types of emotion classifiers, and systematic cross-validation. We demonstrated that three emotional states (fearful, sad, and neutral) can be successfully discriminated based on a remarkably small number of psychophysiological variables, by most classifiers, and independently of the stimulus material or of a particular person. For stimulus-independent classification, the best model used only eight variables and achieved a classification accuracy of 79%. For both the subject-independent as well as for the subject- and stimulus-independent classifications, only five variables were necessary to achieve a classification accuracy of 79% and 77%, respectively. Highest classification accuracy (82%) was achieved for the subject- and stimulus-dependent model, based on a considerably reduced feature set of only seven variables. The relatively small decrease in classification accuracy from subject- and stimulus-dependent cross-validation over subject-independent cross-validation and stimulus-independent cross-validation to subject- and stimulus-independent cross-validation is noteworthy. Thus, removing all information related to a particular person and a particular stimulus context from the data set still achieved comparable and strikingly high classification accuracy.

For all the four cross-validation types, the classification accuracy was higher than that reported in the previous study by Kreibig et al. (2007). Whereas in the previous study, sadness was the emotion condition less often correctly classified (64.3%), followed by fear (67.9%) and neutral (75.0%), in the present study, lowest classification accuracy was 69.1% up to 80.9% for sadness, 79.4% to 83.8% for neutral, and 77.9% to 88.2% for fear. This marks a considerable improvement of 4.4% to 8.8% for neutral, 4.8% to 16.6% for sadness, and 10% to 20.3% for fear in classification accuracy based on a rigorous feature selection and comprehensive cross-validation. Although different approaches for data reduction and response quantification were used in the present study (maximum reactivity of three 3-min means over the 10-min film clip) and the previous study by Kreibig et al. (2007; mean reactivity over the 10-min film clip),

Table 6. *Results of Cross-validation with Constant Preselected Feature Subsets and Optimized Number of Nodes*[a]

a: Most frequent features found with feature selection algorithms (relative frequency > 50%, see Table 5)

	Subject- and stimulus-dependent CV					Subject-independent CV				
Model	LS[b]	QS	M6S	R19S	K17S	LS	QS	M5S	R14S	K17S
No. Feat.	14	8	14	7	7	14	6	8	14	11
Acc. ± *SD* (%)	77.0 ± 15.4	74.5 ± 19.4	76.0 ± 15.5	80.9 ± 19.7	**81.9** ± 17.1	76.0 ± 18.0	77.9 ± 12.8	78.4 ± 15.6	77.9 ± 17.8	**78.9** ± 13.2
Wilc.		QA		R16A	QS, K19A					
	Stimulus-independent CV					Subject- and stimulus-independent CV				
Model	LS	QS	M7S	R9S	K5S[c]	LS	QS	M4S	R11S	K5S
No. Feat.	8	11	14	14	5	10	13	14	14	14
Acc. ± *SD* (%)	**78.9** ± 17.6	76.0 ± 18.0	74.0 ± 18.9	77.9 ± 16.8	77.9 ± 17.3	**76.5** ± 18.4	67.7 ± 20.5	76.0 ± 18.9	76.0 ± 19.3	73.5 ± 18.4
Wilc.	M7S	QA*				QS, LA	QA	QS		

b: Five features common to all models (PEP, SRR, pCO_2, CS, and ZM)

	Subject- and stimulus-dependent CV					Subject-independent CV				
Model	LC	QC	M9C	R11C	K5C	LC	QC	M3C	R12C	K20C
Acc. ± *SD* (%)	74.0 ± 17.5	73.5 ± 17.9	75.5 ± 16.0	**77.9** ± 19.1	76.5 ± 20.2	75.5 ± 14.9	75.0 ± 14.4	76.0 ± 17.5	**78.9** ± 13.8	77.0 ± 13.6
Wilc.										
	Stimulus-independent CV					Subject- and stimulus-independent CV				
Model	LC	QC	M2C	R9C	K5C	LC	QC	M2C	R10C	K7C
Acc. ± *SD* (%)	75.0 ± 15.5	77.5 ± 18.3	75.0 ± 16.5	77.5 ± 17.8	**77.9** ± 17.3	74.5 ± 19.8	77.0 ± 20.9	71.1 ± 19.4	76.5 ± 19.3	**77.5** ± 20.1
Wilc.		QA*					QA			

c: All 14 features

	Subject- and stimulus-dependent CV					*Subject-independent CV*				
Model	LA	QA	M6A	R16A	K19A	LA	QA	M4A	R14A	K18A
Acc. ± *SD* (%)	77.0 ± 15.4	69.1 ± 20.6	76.0 ± 15.5	**77.5** ± 14.7	74.5 ± 17.5	76.0 ± 18.0	69.1 ± 17.5	74.5 ± 17.5	**77.9** ± 17.8	75.5 ± 17.0
Wilc.	QA			QA					QA	
	Stimulus-independent CV					Subject- and stimulus-independent CV				
Model	LA	QA	M7A	R9A	K5A	LA	QA	M4A	R11A	K5A
Acc. ± *SD* (%)	75.0 ± 18.0	65.7 ± 16.9	74.0 ± 18.9	**77.9** ± 16.8	74.0 ± 16.0	73.0 ± 19.3	64.2 ± 21.8	**76.0** ± 18.9	**76.0** ± 19.3	73.5 ± 18.4
Wilc.	QA			QA*	QA	QA		QA		

[a]Best classification accuracy on the validation set (Acc.) for the respective cross-validation (CV) is shown in bold. Classification models are abbreviated as in Table 4 plus one letter denoting a feature subset (S = most frequent features found with feature selection algorithms, C = common 5 features, A = all 14 features). Wilc.: significant improvement over indicated models according to Wilcoxon signed-rank test ($p < 0.05$; see text under Comparison of Classification Models for details).
[b]Same as LA.
[c]Same as K5C.
*$p < .01$.

Table 7. *Best Classification Results with Preselected Feature Subsets for the Respective Cross-Validation*[a]

	Subject- and stimulus-dependent cross-validation	Subject-independent cross-validation	Stimulus-independent cross-validation	Subject- and stimulus-independent cross-validation
Model	K17S	R12C	LS	K7C
Acc. ± *SD* (%)	81.9 ± 17.1	78.9 ± 13.8	78.9 ± 17.6	77.5 ± 20.1
Fear	80.9	82.4	88.3	77.9
Sadness	80.9	70.6	69.1	70.6
Neutral	83.8	83.8	79.4	83.8
	Selected features	Selected features	Selected features	Selected features
HR				
RSA				
PEP	**x**	**x**	**x**	**x**
SBP	x		x	
SCL				
SRR	**x**	**x**	**x**	**x**
RR				
Vt				
pCO_2	**x**	**x**	**x**	**x**
FT	x		x	
ACT			x	
SM				
CS	**x**	**x**	**x**	**x**
ZM	**x**	**x**	**x**	**x**
No. Feat.	7	5	8	5

[a]Five common features for all models are shown in bold.

still very similar results were obtained. This outcome demonstrates good generalizability of results based on different response quantification.

Linear versus Nonlinear Classifiers

It should be noted that our analyses showed only small improvements of the nonlinear classifiers (QDA, MLP, RBNF, and KNN) over the linear classifier (LDA). There are three possible reasons for this outcome:

1. *Ceiling effect*. The present data set may reflect a ceiling effect, because the base model already has very good classification accuracy, which may be due to the small number of different emotions, that is, three—fear, sadness, and neutral—in the present analysis.
2. *Suboptimal feature selection*. The analyses have been carried out on an already selected variable set, which was selected a priori to present an as diverse as possible feature set that is representative of different processes in autonomic and respiratory regulation. The feature selection performed might thus not have been as powerful as possible. Applied to a more complex feature set of many more—and possibly highly correlated—features, the feature selection approach may be even more informative. It is also known that simple sequential feature selection algorithms (SFS and SBS) effectively make univariate decisions on inclusion/exclusion of features for application in multivariate models, not taking into account the interdependencies of features (Pudil, Novovièová, & Kittler, 1994). Therefore, it might be necessary to use more sophisticated algorithms, such as sequential floating feature selection (Picard et al., 2001; Pudil et al., 1994), which might provide better feature subsets. However, these methods are computationally much more demanding and slower than simple sequential algorithms because they contain an additional nested loop of feature selection.
3. *Small data set*. Although our data set was collected from a larger number of subjects compared to most other studies, it still contained only 204 data points. Such a sample size might have been insufficient to reveal the advantages of such nonlinear classification models as neural networks in comparison to the conventional linear discriminant analysis, and was likely to result in at least a moderate degree of overfitting with the more complex models (i.e., QDA, MLP), whereas this did not seem to be a problem for the other models (i.e., LDA, RBFN, and KNN). This might have been most critical for stimulus-independent cross-validation and for the combination of both subject- and stimulus-independent cross-validation, where the training sets included only 102 and 99 data points, respectively. Therefore, the simplest models performed best (LDA with 8 features and KNN with $K = 7$ and 5 features). The overall poor performance of QDA was also most likely due to the overfitting effect because the QDA model contained three estimated covariance matrices instead of one with LDA (Hastie et al., 2001).

Nevertheless, for three of the four different cross-validation types, the best classification results were achieved with nonlinear models, and in only one case did the conventional LDA approach perform best, namely, for stimulus-independent cross-validation (see Table 7). Besides that, the nonlinear models were systematically better than the linear ones in all four cross-validation settings with five common features, PEP, SRR, pCO_2, CS, and ZM (Table 6b). That is, the nonlinear models were able to provide the same or a higher level of classification accuracy with a smaller number of input variables. Thus, a fearful, sad, and neutral state can be reliably discriminated using a remarkably small set of physiological measures of cardiac, electrodermal, respiratory, and facial muscular activity. For theoretical considerations, this finding provides further support to the notion of physiological emotion specificity. For real-world appli-

cations of HCI, this finding translates to a smaller number of sensors needed for discriminating these emotions and, consequently, lower cost, higher user comfort, lower subject reactivity to the measurement context, and thus higher reliability.

Taken together, these results suggest three central outcomes: First, feature selection can identify a considerably reduced subset of variables that performs extremely well on the emotion classification task. Second, compared to nonlinear classifiers, linear classifiers perform remarkably well on discriminating fearful, sad, and neutral states. It may thus be concluded that alternative nonlinear classification procedures are often not necessary for the specific task of discriminating a small subset of emotions. Third, either of the cross-validation strategies captures the magnitude of physiological emotion specificity well. Still, as a central tenet of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) holds that the facets across which generalizability is sought should be taken into account, classification accuracy decreases from subject- and stimulus-dependent cross-validation over subject-dependent or stimulus-dependent cross-validation to subject- and stimulus-independent cross-validation. Thus, by applying the three aspects outlined here of feature selection, classifier type, and cross validation to pattern classification analysis, important improvement and refinement of results can be expected.

Limitations and Future Directions

It is an important question how the results reported here will generalize. The present study is but a first conceptual step to systematically investigating the complex structure of psychophysiological emotion specificity. We thus had to make a number of limiting decisions and selections in the experimental design. First, the type and number of emotions considered in the present data set were restricted to three conditions—fear, sadness, and a neutral emotional state—with two of the conditions representing negative, withdrawal-related emotions. Future research should strive to apply these methods to data sets that include a larger number of emotion conditions and thus have a lower a priori chance classification rate of, for example, 25% (for four conditions) to 14% (for seven conditions). In addition, it would be desirable for future research to have emotion conditions varied on several dimensions, such as basic motive features of behavior (valence and arousal: Bradley & Lang, 2000; behavioral approach and avoidance: Carver, 2004) or appraisal processes (including intrinsic pleasantness, motive congruence, and coping potential; Frijda, 1986; Scherer, 2009).

Second, our analysis addressed subject- and stimulus-(in)dependence in cross-validation for classification analysis. Still, the sample size of our study was relatively small, and demonstration of replication of our results will be essential. As a result of the limited sample size, participants were drawn from a narrow age range; social backgrounds and cultural origin were also restricted. To evaluate whether a truly subject-independent system can be developed, it will be important for future studies to include participants from diverse social and cultural backgrounds as well as of different age groups, accounting for the diverse composition of our society.

Similar limitations are true for the stimulus material: Our selection of film clips was limited to two per emotion condition, and more would be desirable. Although film clips were selected to vary by emotion, they additionally differed in ways that we did not control, such as brightness, sound parameters, stimulus complexity, and thematic development. These additional aspects of between-emotion stimulus differences will need further examination.

It is, furthermore, noteworthy that emotional situations may vary on yet other features than the stimulus dimension considered here, such as the induction context (e.g., film, imagery, or music) or activity context (e.g., sitting, standing, moving). To what extent context independence of emotion classifiers can be found under varying stimulus conditions will be an important topic for future research.

Finally, whereas the present study identified the physiological parameters PEP, SRR, pCO_2, CS, and ZM to be important for the discrimination of fear, sadness, and a neutral state, these parameters do not need to be discriminative of other emotion contrasts. Stemmler (2004), for example, reported that HR, SBP, SRR, and FT were not discriminative for the emotion contrast of fear and anger. Rather, he found cardiac output, total peripheral resistance, and respiration rate to be fear specific and diastolic blood pressure, somatic motor activity, and facial temperature to be anger specific. Yet other physiological parameters, such as RSA, may be important for discriminating appetitive from defensive emotions. It will be an important task for future research to study such emotion-specific physiological discriminations in more detail.

We strongly encourage other researchers in this area to subject their data sets to these new methods to confirm and extend their utility. Applying the classification techniques described in the present article to other comprehensive data sets will more realistically model the real application context of these classifiers. With a larger and more diverse data set, it can also be expected that nonlinear classification techniques, such as neural networks, will provide significantly better results in comparison with the conventional discriminant analysis, first of all in terms of statistical significance of comparisons between the nonlinear models and LDA. Generalization of the results of the present study to a more varied data set, including ambulatory daily life settings (Rosenfield et al., 2010; Wilhelm & Grossman, 2010; Wilhelm, Pfaltz, & Grossman, 2006), is, moreover, of significant interest for practical implementation in user- and context-independent HCI systems with a minimum number of sensors and high accuracy in emotion recognition.

Different cross-validation settings should also be considered in future analyses, such as splitting the data set for subject-independent cross-validation in a different way, for example, a two- or a fivefold split. This would allow researchers to test the performance of the classification models with a larger number of unknown subjects whose data were not included in the training set of the classifier. However, such a split would require more data to avoid overfitting, and, second, it should be remembered that increasing N in an N-fold cross-validation theoretically leads to a higher variance of the estimated accuracy and to lower bias and vice versa (Hastie et al., 2001).

The present research has important implications for future studies on emotion classification. We here demonstrated the importance of a stringent feature selection, considering various forms of classifiers, and systematic cross-validation that acknowledges both subject and situation dependence of physiological emotion responses. Although feature selection can afford a classification model that uses a smaller number of measures, at the same or better classification accuracy, cross-validation will ensure generalization of classifiers to unknown individuals and unknown stimuli. Future research should aim to integrate these concepts into a programmatic analysis of emotion classification problems, using peripheral psychophysiological measures presented here and expanding this approach to electrocortical and functional magnetic resonance imaging data.

REFERENCES

Ax, A. F. (1953). The physiological differentiation between fear and anger in humans. *Psychosomatic Medicine*, *15*, 433–442.

Bailenson, J. N., Pontikakis, E. D., Mauss, I. B., Gross, J. J., Jabon, M. E., Hutcherson, C. A. C., et al. (2008). Real-time classification of evoked emotions using facial feature tracking and physiological responses. *International Journal of Human-Computer Studies*, *66*, 303–317.

Blechert, J., Lajtman, M., Michael, T., Margraf, J., & Wilhelm, F. H. (2006). Identifying anxiety states using broad sampling and advanced processing of peripheral physiological information. *Biomedical Sciences Instrumentation*, *42*, 136–141.

Bradley, M. M., & Lang, P. J. (2000). Measuring emotion: Behavior, feeling and physiology. In R. Lane & L. Nadel (Eds.), *Cognitive neuroscience of emotion* (pp. 242–276). New York: Oxford University Press.

Carver, C. S. (2004). Negative affects deriving from the behavioral approach system. *Emotion*, *4*, 3–22.

Christie, I., & Friedman, B. (2004). Autonomic specificity of discrete emotion and dimensions of affective space: A multivariate approach. *International Journal of Psychophysiology*, *51*, 143–153.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*, 21–27.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*, *221*, 1208–1210.

Esterman, M., Tamber-Rosenau, B. J., Chiu, Y.-C., & Yantis, S. (2010). Avoiding non-independence in fMRI data analysis: Leave one subject out. *NeuroImage*, *50*, 572–576.

Fridlund, A. J., & Izard, C. E. (1983). Electromyographic studies of facial expressions of emotions and patterns of emotions. In J. T. Cacioppo & R. E. Petty (Eds.), *Social psychophysiology* (pp. 243–280). New York: Guilford Press.

Fridlund, A. J., Schwartz, G. E., & Fowler, S. C. (1984). Pattern recognition of self-reported emotional state from multiple-site facial EMG activity during affective imagery. *Psychophysiology*, *21*, 622–637.

Friedman, B. H. (2010). Feelings and the body: The Jamesian perspective on autonomic specificity of emotion. *Biological Psychology*, *84*, 383–393.

Frijda, N. H. (1986). *The emotions*. Cambridge, UK: Cambridge University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining inference, and prediction*. Berlin: Springer.

Haykin, S. (1999). *Neural networks: A comprehensive foundation*. Upper Saddle River, NJ: Prentice Hall.

Hinz, A., Seibt, R., Hueber, B., & Schreinicke, G. (2000). Response specificity in psychophysiology: A comparison of different approaches. *Journal of Psychophysiology*, *14*, 115–122.

Ho, T. C., Brown, S., & Serences, J. T. (2009). Domain general mechanisms of perceptual decision making in human cortex. *Journal of Neuroscience*, *29*, 8675–8687.

Howard, J., Plailly, J., Grueschow, M., Haynes, J. D., & Gottfried, J. A. (2009). Odor quality coding and categorization in human posterior piriform cortex. *Nature Neuroscience*, *12*, 932–938.

Huberty, C. J. (1994). *Applied discriminant analysis*. New York: Wiley.

James, W. (1884). What is an emotion? *Mind*, *9*, 188–205.

Kim, J., & André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*, 2067–2083.

Kim, K. H., Bang, S. W., & Kim, S. R. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, *42*, 419–427.

Kreibig, S. D. (2004). *Situational and individual response specificity to emotional films: Effects on experiential, cardiovascular, electrodermal, respiratory and muscular responses*. Unpublished diploma thesis, University of Kiel, Kiel, Germany.

Kreibig, S. D., Brosch, T., & Schaefer, G. (2010). Psychophysiological response patterning in emotion: Implications for affective computing. In K. R. Scherer, T. Baenziger, & E. Roesch (Eds.), *A blueprint for an affectively competent agent: Cross-fertilization between emotion psychology, affective neuroscience, and affective computing* (pp. 105–130). Oxford, England: Oxford University Press.

Kreibig, S. D., Wilhelm, F. H., Roth, W. T., & Gross, J. J. (2007). Cardiovascular, electrodermal, and respiratory response patterns to fear and sadness-inducing films. *Psychophysiology*, *44*, 787–806.

Lacey, J. I., & Lacey, B. C. (1958). Verification and extension of the principle of autonomic response-stereotypy. *American Journal of Psychology*, *71*, 50–73.

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, *50*, 371–385.

Lisetti, C. L., & Nasoz, F. (2004). Using non-invasive wearable computers to recognize human emotions from physiological signals. *EURASIP Journal on Applied Signal Processing*, *2004*, 1672–1687.

Marwitz, M., & Stemmler, G. (1998). On the status of individual response specificity. *Psychophysiology*, *35*, 1–15.

Moody, J., & Darken, C. J. (1989). Fast learning in networks of locally tuned processing units. *Neural Computation*, *1*, 281–294.

Nabney, I. T. (2003). *Netlab neural network software*. Available at http://www.ncrg.aston.ac.uk/netlab/index.php

Nabney, I. T. (2004). *NETLAB. Algorithms for pattern recognition. Advances in pattern recognition* (3rd ed.). Berlin: Springer.

Nasoz, F., Alvarez, K., Lisetti, C., & Finkelstein, N. (2004). Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology & Work*, *6*, 4–14.

Nasoz, F., Ozyer, O., Lisetti, C. L., & Finkelstein, N. (2002). Multimodal affective driver interfaces for future cars. In *Proceedings of the Tenth ACM International Conference on Multimedia, France* (pp. 319–322). New York: ACM Press.

Nyklicek, I., Thayer, J. F., & van Doornen, L. J. P. (1997). Cardiorespiratory differentiation of musically-induced emotions. *Journal of Psychophysiology*, *11*, 304–321.

Picard, R. W. (2000). *Affective computing*. Cambridge, MA: MIT Press.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*, 1175–1191.

Pudil, P., Novovièová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, *15*, 1119–1125.

Rainville, P., Bechara, A., Naqvi, N., & Damasio, A. R. (2006). Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal of Psychophysiology*, *61*, 5–18.

Rosenfield, D., Zhou, E., Wilhelm, F. H., Conrad, A., Roth, W. T., & Meuret, A. E. (2010). Change point analysis for longitudinal physiological data: Detection of cardiorespiratory changes preceding panic attacks. *Biological Psychology*, *84*, 112–120.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, *23*, 1307–1351.

Sinha, R., & Parsons, O. A. (1996). Multivariate response patterning of fear and anger. *Cognition and Emotion*, *10*, 173–198.

Stemmler, G. (1987). Standardization within subjects: A critique of Ben-Shakhar's conclusions. *Psychophysiology*, *24*, 243–246.

Stemmler, G. (1989). The autonomic differentiation of emotions revisited: Convergent and discriminant validation. *Psychophysiology*, *26*, 617–632.

Stemmler, G. (2004). Physiological processes during emotion. In P. Philippot & R. S. Feldman (Eds.), *The regulation of emotion* (pp. 33–70). Mahwah, NJ: Erlbaum.

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: Overview. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (*Vol. 2*, pp. 717–719). Chichester, UK: John Wiley & Sons Ltd.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, *1*, 80–83.

Wilhelm, F. H., & Grossman, P. (2010). Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. *Biological Psychology*, *84*, 552–569.

Wilhelm, F. H., Grossman, P., & Roth, W. T. (1999). Analysis of cardiovascular regulation. *Biomedical Sciences Instrumentation*, *35*, 135–140.

Wilhelm, F. H., Pfaltz, M. C., & Grossman, P. (2006). Continuous electronic data capture of physiology, behavior and experience in real life: Towards ecological momentary assessment of emotion. *Interacting with Computers*, *18*(2), 171–186.

Yiu, K. K., Mak, M. W., & Li, C. K. (1999). Gaussian mixture models and probabilistic decision-based neural networks for pattern classification: A comparative study. *Neural Computing & Applications*, *3*, 235–245.

Zhang, X., Lu, X., Shi, Q., Xu, X., Leung, H. E., Harris, L. N., et al. (2006). Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*, *7*, 197.

Zhao, Q., & Lu, H. (2005). GA-driven LDA in KPCA space for facial expression recognition. In L. Wang, K. Chen, & Y. S. Ong (Eds.), *Lecture Notes in Computer Science* (*3611*, pp. 28–36). Berlin: Springer-Verlag.

(Received March 18, 2010; Accepted November 23, 2010)