Interacting with Computers 24 (2012) 154–172

Contents lists available at SciVerse ScienceDirect

Interacting with Computers

journal homepage: www.elsevier.com/locate/intcom

A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing ☆

Domen Novak *, Matjaž Mihelj, Marko Munih

Faculty of Electrical Engineering, University of Ljubljana, Trzaska cesta 25, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 11 May 2011
Received in revised form 15 February 2012
Accepted 30 April 2012
Available online 9 May 2012

Keywords:
Affective computing
Physiological computing
Psychophysiology
Data fusion
Autonomic nervous system

ABSTRACT

Physiological computing represents a mode of human–computer interaction where the computer monitors, analyzes and responds to the user's psychophysiological activity in real-time. Within the field, autonomic nervous system responses have been studied extensively since they can be measured quickly and unobtrusively. However, despite a vast body of literature available on the subject, there is still no universally accepted set of rules that would translate physiological data to psychological states. This paper surveys the work performed on data fusion and system adaptation using autonomic nervous system responses in psychophysiology and physiological computing during the last ten years. First, five prerequisites for data fusion are examined: psychological model selection, training set preparation, feature extraction, normalization and dimension reduction. Then, different methods for either classification or estimation of psychological states from the extracted features are presented and compared. Finally, implementations of system adaptation are reviewed: changing the system that the user is interacting with in response to cognitive or affective information inferred from autonomic nervous system responses. The paper is aimed primarily at psychologists and computer scientists who have already recorded autonomic nervous system responses and now need to create algorithms to determine the subject's psychological state.

1. Introduction

Physiological computing represents a mode of human–computer interaction where the computer monitors, analyzes and responds to the user's psychophysiological activity in real-time (Fairclough, 2009). It can be divided into cognitive physiological computing, which aims to maximize user performance, and affective physiological computing, which aims to maximize user pleasure. Through analyzing psychophysiological measurements such as heart rate and brain activity, a new, subconscious channel of communication is established between the user and the machine (Hettinger et al., 2003).

Physiological computing has many potential applications. Cognitive psychophysiology can, for instance, be used to recognize periods of very high or very low workload and adapt to them by taking appropriate actions. Examples exist in simulated flight (Wilson and Russell, 2007), learning (Shen et al., 2009), biomedical applications (Novak et al., 2011) and other fields. Affective psychophysiology, on the other hand, can be used to recognize undesirable emotional states such as anger and guide the user to a more positive emotional state. This has been demonstrated in, for example, computer games (Mandryk and Atkins, 2007; Liu et al., 2009) and human-robot interaction (Rani et al., 2004). Of course, cognitive and affective physiological computing frequently overlap, with the overall goal of providing a pleasant environment for the user that will lead to improved performance and efficiency. In such settings, physiological measurements have the advantage that they provide an objective estimate of the user's psychological state that can be obtained unobtrusively without his or her active participation.

Physiological computing is heavily based on earlier experiments in psychophysiology, which extensively used physiological measurements to identify psychological states (e.g. Ekman et al., 1983; Cacioppo and Tassinary, 1990). It also partially overlaps with affective computing, the study of systems that can recognize and mimic human affect (Picard, 1997). Both psychophysiology and affective computing have explored many avenues of research, including speech, facial expressions, gestures, central nervous system responses and autonomic nervous system responses (Zeng et al., 2009; Calvo and D'Mello, 2010). Among these, autonomic nervous system (ANS) responses such as cardiorespiratory and electrodermal responses hold a great deal of promise in physiological computing since they can be measured more cheaply, quickly and unobtrusively than central nervous system responses.

Though measurement of ANS responses is not a difficult task, their interpretation in a psychophysiological context is much

☆ This paper has been recommended for acceptance by Timothy Bickmore.

* Corresponding author. Tel.: +386 14768373 (O), mobile: +386 41969753; fax: +386 14768239.

E-mail addresses: domen.novak@robo.fe.uni-lj.si (D. Novak), matjaz.mihelj@robo.fe.uni-lj.si (M. Mihelj), marko@robo.fe.uni-lj.si (M. Munih).

0953-5438/$ - see front matter © 2012 British Informatics Society Limited. All rights reserved.
http://dx.doi.org/10.1016/j.intcom.2012.04.003

more difficult. Perhaps the most influential paper on this topic was published by Rosalind Picard and covers psychophysiological data fusion: the extraction of various features from different physiological responses, the automated selection of the most appropriate features, and the classification of these features into different possible emotion classes (Picard et al., 2001). Since then, many machine learning approaches have been used to infer cognitive and affective information from ANS responses. However, despite a vast body of literature available on the subject, there is still no universally accepted set of data fusion rules that would translate physiological data to psychological states.

Nonetheless, though no universal set of rules currently exists, it is possible to look at all the work that has been done and identify some of the most promising strategies. We thus conducted a review of psychophysiological studies performed in the last ten years and examined all those dealing with the fusion of ANS responses, whether in physiological computing or elsewhere. Our goal was specifically to provide an overview of the different steps and methods necessary for data fusion. Thus, this review is aimed primarily at psychologists and computer scientists who have already measured ANS responses and now need to create algorithms to fuse these measurements into the cognitive or affective state of the subject. Such reviews have already been done for electroencephalography (Lotte et al., 2007) and speech (El Ayadi et al., 2011), but not for ANS responses, which come with their own challenges and applications. General reviews of ANS activity in psychophysiology (e.g. Kreibig, 2010) are not very useful for data fusion, though they are extremely valuable otherwise. Similarly, reviews focusing on issues such as experiment design, user modeling and ethical issues in physiological computing (e.g. Fairclough, 2009) are very useful in general, but do not provide enough information about data fusion methods. However, general overviews of methods for affect detection such as those by Zeng et al. (2009), Calvo and D'Mello (2010), or Gunes and Pantic (2010) could also be of interest to readers of this paper even if they do not cover ANS responses in detail.

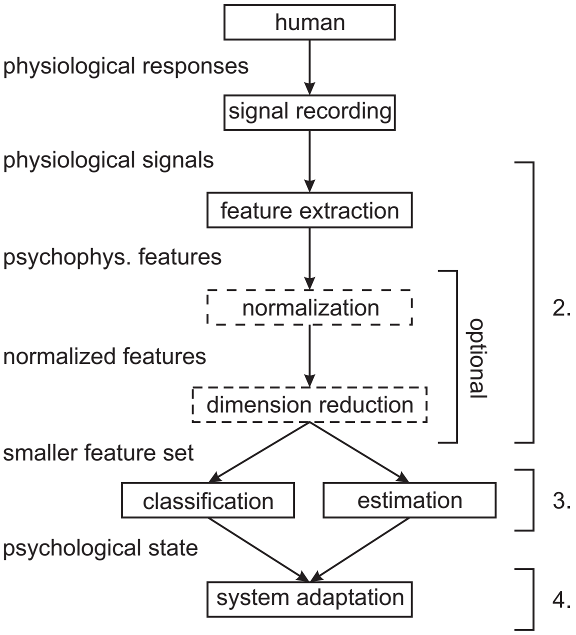

Fig. 1. The general process of measuring, interpreting and using autonomic nervous system responses in psychophysiology and physiological computing. The blocks contain the human (top) and different steps to be performed. The data used at each stage is written on the left, while the numbers on the right show which section of this paper covers the different steps. Classification and estimation are shown in parallel since they are two possible but mutually exclusive data fusion methods.

Although the most complex challenge in physiological computing has been the interpretation of ANS measurements, the final goal is still to adapt various systems in response to the detected cognitive or affective changes. This has already been done in several applications, from computer games to automatic pilots. We thus also attempt to cover these existing applications in our review. Unfortunately, due to space limitations, we have had to limit ourselves only to studies which present actual implementations; numerous papers provide interesting ideas and frameworks for physiological computing, but without implementation, it is difficult to gauge whether these ideas are feasible.

Fig. 1 shows the general process of measuring, interpreting and using ANS responses in physiological computing. This paper is likewise divided into three main sections corresponding to the three main steps. Within the context of this paper, we define 'data fusion' to correspond to strictly the classification or estimation of psychological states from multiple psychophysiologically relevant features obtained from different physiological signals.

Section 2 thus describes the prerequisites for data fusion (psychological model selection, training data set preparation, feature extraction, normalization and dimension reduction), Section 3 describes different data fusion methods (either classification or estimation), and Section 4 covers system adaptation in response to the inferred psychological state. Many of these steps are not specific to physiological or affective computing, but have been common in psychophysiological studies since the 1980s. For this reason, many of the procedures will be referred to as psychophysiological methods rather than physiological computing methods.

2. Prerequisites for data fusion

This section describes the steps taken to obtain a set of features relevant for physiological computing and suitable for use in data fusion. The first, selection of an appropriate psychological model, is crucial for all psychophysiological studies and affects all of the following steps. The second, preparation of a training data set, is required for supervised data fusion methods. Since the majority of data fusion methods used in physiological computing are supervised (with the exception of expert-defined rules described in Sections 3.2.1 and 3.2.2), a training data set is necessary in most cases. The third step, feature extraction, is also practically necessary since psychological information is difficult to glean from raw signals. The fourth, normalization, is not absolutely necessary but is commonly taken in order to improve data fusion. The fifth, dimension reduction, is also not strictly necessary, but is useful in settings where a large number of possibly irrelevant features are available.

Psychological model selection needs to be done before any measurements can even begin. Training data set preparation must be planned before the actual measurements and affects the measuring process. The remaining three steps are performed after the measurements have been collected.

2.1. Psychological model selection

One of the first steps of any psychophysiological study is to select an appropriate model that describes the user's psychological state. This is integral to data fusion, as it defines the states that can be identified from physiological measurements. A number of emotion models currently exist for emotion recognition in human–computer interaction (Cowie et al., 2001). Of those, the categorical model and the two-dimensional arousal-valence model are two of the most frequently used in psychophysiology.

The first model tries to classify psychophysiological measurements into one of several basic emotions (anger, sadness, fear, surprise, happiness...) (Ekman, 1992). The second posits that a person's psychological state is multidimensional and thus described with multiple variables. The most popular multidimensional model in psychophysiology is the arousal-valence model (Russell, 1980). Valence (sometimes also called pleasure) is defined as positive versus negative affective states (e.g. humiliation, disinterest, and anger at one end versus excitement, relaxation, and tranquility at the other) while arousal is defined in terms of mental alertness and physical activity (e.g. sleep, inactivity, boredom, and relaxation at the lower end versus wakefulness, tension, exercise, and concentration at the higher end) (Mehrabian, 1996). While valence and arousal are generally continuous variables, it is also possible to separate the arousal-valence space into quadrants: low arousal/positive valence, low arousal/negative valence, high arousal/positive valence and high arousal/negative valence. This is commonly done for classification problems (Section 3.1).

Both of the above models were originally developed for general psychology, not necessarily involving physiological measurements. However, they are not always appropriate in physiological computing. It is not always necessary to identify a large number of possible basic emotions, and valence is relatively difficult to detect using ANS responses alone (e.g. Peter and Herbon, 2006). Many studies thus use simpler, ad hoc models that include only the psychological variable of greatest interest to that particular application such as: stress, frustration or mental workload. These are especially preferable when system adaptation is desired. Since an adaptation action needs to be defined for each possible psychological state, a simpler model with fewer states requires fewer actions to be defined.

Researchers should be aware that, since the psychological model affects every part of a study, from the experiment design to the data analysis, it is practically impossible to change models once measurements have begun. While it is possible to, for example, convert basic emotions to arousal-valence quadrants or vice versa (e.g. Christie and Friedman, 2004), it can be problematic since a perfect conversion is difficult. For instance, fear and anger both occupy the same arousal-valence quadrant (high arousal, negative valence) despite being two very different emotions.

2.2. Training data set

Generally, a training data set is a set of measurements taken in known conditions. In physiological computing, it refers to a set of physiological measurements (electrocardiogram, skin conductance . . .) associated with induced psychological states (fear, anger, boredom, low stress . . .). A supervised data fusion method uses this data set to learn the associations between psychophysiological measurements and psychological states since both the inputs and outputs are known. The learned associations can then be applied to new measurements where the psychological state is not yet known.

The majority of data fusion methods in physiological computing are supervised techniques that are trained in advance and thus require a prerecorded training data set. Linear sums and fuzzy logic, described in Sections 3.2.1 and 3.2.2, do not require training data at all. A few physiological computing techniques (e.g. Liu et al., 2008; Novak et al., 2011) combine initial training with online supervised learning, but they are not yet widespread. The training data set is thus crucial to psychophysiological data fusion and must be properly constructed. While this is primarily a matter of study design, a few words should nonetheless be dedicated to it since it is such an important part of data fusion.

2.2.1. Psychological state induction

Since supervised learning depends on the training data set, the same psychological states that we wish to identify should also be represented in the training data set. These states are defined by the psychological model (Section 2.1) and should be properly induced (elicited) in the subjects so that we can be certain that the training data set actually contains useful information. Common induction techniques include actively performing tasks as well as more passive methods such as remembering past emotional experiences or viewing affective pictures and videos. A good review of induction techniques was performed by Kreibig (2010).

In laboratory experiments involving induction of psychological states using videos, pictures or memory recall, it is common to elicit each possible psychological state in each subject (e.g. Christie and Friedman, 2004; Rainville et al., 2006; Frantzidis et al., 2010). This can also be done in applied studies by, for example, exposing the subject to tasks that are obviously boring, stressful or frustrating (e.g. Scheirer et al., 2002; Wilson and Russell, 2003a). Another possibility in applied studies, however, is to simply let the subject perform the task and record the subject's experience, thus resulting in data sets where not every subject experiences every psychological state and some subjects experience the same state multiple times (e.g. Katsis et al., 2008; Woolf et al., 2009). The first approach is preferable from a theoretical viewpoint, as it provides a well-structured database with all psychological states equally represented (assuming they have been successfully elicited), while the second is more easily achievable in applied settings where some extreme states are rare and difficult to elicit. The choice thus depends primarily on the goal of the study.

It is also recommended to use a second method to validate that the psychological states were successfully induced, as the training data may otherwise be unreliable. By far the most common methods for this are self-report techniques (e.g. Christie and Friedman, 2004; Lisetti and Nasoz, 2004; Haarmann et al., 2009). Their greatest advantage is that they can be administered easily and cheaply. If a self-report questionnaire has been validated in advance, its specific strengths and weaknesses are also known. Two examples of commonly used and well-validated self-report methods in psychophysiology are the Self-Assessment Manikin (Bradley and Lang, 1994) and the NASA-TLX (Hart and Staveland, 1988). However, it is often not possible to use a standardized questionnaire, so ad hoc self-report measures are used instead. Researchers should then be aware that several studies have found that subjects are sometimes unaware of their own emotions, are unable to report them, or are simply unwilling to report them (e.g. Schwerdtfeger, 2004; Liu et al., 2008; Koenig et al., 2011). In such cases, alternative measures should be used to determine what psychological state was induced. These include other physiological measurements (e.g. facial electromyography in Kreibig et al., 2007) or, more commonly, observation of the subject by others (e.g. Schwerdtfeger, 2004; Healey and Picard, 2005; Katsis et al., 2008; Liu et al., 2008; Koenig et al., 2011).

2.2.2. Crossvalidation

Since the training data set is recorded in advance, it is common to test different data fusion methods on it in order to compare their effectiveness. For psychophysiological studies that do not involve system adaptation, this is in fact the primary result of the study. In order to ensure that data fusion is not trained and tested on the very same data set (which could lead to unreliable results), it is common to use the technique of crossvalidation. The prerecorded data set is first divided into multiple parts. One part is designated as the validation data set and all the others are designated as the training data set. The rules for data fusion are constructed using data from the training set and tested on the validation set. This process is repeated as many times as there are parts, with each part serving as the validation set exactly once. The most common approaches to this are tenfold crossvalidation, where the data is divided into ten roughly equally sized parts, and leave-one-out crossvalidation, where the training set contains all but one sample and the validation set contains that remaining sample.

During crossvalidation, however, it is necessary to decide whether data fusion should be subject-dependent or subject-independent. Completely subject-independent data fusion means that the training data set and the validation data set contain entirely different subjects. Subject-dependent data fusion means that the training data and the validation data partially share subjects. An extreme version of subject-dependent data fusion is within-subject data fusion, where both the training and validation sets contain data from only a single subject. Within-subject fusion is obviously done in studies that involve long-term recordings of a single subject (e.g. Picard et al., 2001; Leon et al., 2004). It is also often seen in studies with long-term recordings of a small number of subjects: the data fusion is handled for each subject independently of the others (e.g. Wilson and Russell, 2003a; Liu et al., 2008). This is primarily done to avoid intersubject differences in physiology. Within-subject data fusion thus often yields higher accuracies than subject-independent fusion (as shown by, for example, Kim, 2007; Bailenson et al., 2008; Kim and Andre, 2008), but requires preparation of an extensive training data set for each subject. Subject-independent data fusion, on the other hand, requires a smaller amount of data from each subject. In practice, within-subject data fusion would be preferred for applications where few regular users are expected (e.g. pilot monitoring during flight) while subject-independent data fusion would be preferred for applications with a large amount of casual users (e.g. entertainment technologies).

Direct comparison of subject-dependent and subject-independent data fusion is difficult. Any study involving data fusion should thus report whether the data fusion procedure is completely subject-independent, partially subject-dependent, or completely within-subject. Though the vast majority of psychophysiological studies use crossvalidation, they often do not report whether, for instance, leave-one-out crossvalidation refers to leaving out one subject or one specific recording from that subject. While this is not problematic for studies with a large number of subjects (since the amount of data from that subject in the training data set is relatively small and thus unlikely to affect accuracy), it can be confusing in studies with less than ten subjects where including data from the validation subject in the training group can have a noticeable effect on accuracy.

2.2.3. Sample size

The choice of subject-dependent or subject-independent validation also affects the number of subjects that need to be included in the study. In within-subject studies, the number of subjects is unimportant since data fusion is done for each subject separately; it is only necessary to ensure that enough data is recorded for each subject. For this reason, within-subject studies often include a single subject (e.g. Picard et al., 2001; Leon et al., 2004) or a small number of subjects. They are especially preferable when focusing on a specific, small group of the population with a limited amount of available subjects (e.g. six autistic children in Liu et al. (2008), seven air traffic controllers in Wilson and Russell (2003a)).

The majority of subject-independent studies include 20 or more subjects (e.g. 20 in Setz et al. (2009), 20 in Chanel et al. (2011), 24 in Kapoor et al. (2007), 33 in Setz et al. (2010), 34 in Nasoz et al. (2010), 41 in Bailenson et al. (2008), 59 in Arroyo-Palacios and Romano (2010), 75 in Tognetti et al. (2010)). Smaller numbers are again possible in applied studies where the amount of available subjects is low (e.g. 11 patients undergoing motor rehabilitation in Novak et al. (2011)). However, in such cases, multiple recordings from each subject are generally required. Dimension reduction is also useful with small sample sizes, as the performance of a supervised data fusion method is strongly dependent on both sample size and the number of input features (Hua et al., 2005).

2.3. Feature extraction

In psychophysiology, feature extraction refers to extracting a number of psychophysiologically relevant features from raw physiological signals. The electrocardiogram, for example, is a raw physiological signal from which a number of features such as mean heart rate or various measures of heart rate variability can be extracted. The complexity of feature extraction depends on the raw signals involved. Extraction of features from skin temperature, for example, is a relatively simple process that generally only involves the mean, standard deviation and mean absolute derivative over a certain time period, while extraction of heart rate variability from the electrocardiogram involves careful filtering, peak detection, interpolation and power spectral density calculation (Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996).

The final result of feature extraction is a vector of multiple psychophysiological features calculated from different raw signals over a certain time period. In subsequent sections, this vector will be simply referred to as a ‘feature vector’. A matrix consisting of multiple feature vectors from different subjects or different time periods will be referred to as a ‘data set’. By far the majority of data fusion and system adaptation schemes in physiological computing are static: they are trained in advance and then use a single feature vector as input without online learning. Dynamic data fusion methods, which learn online or take the history of signals and features into account, exist in physiological computing, but are rare (e.g. Hidden Markov Models, mentioned in Section 3.1.7, and reinforcement learning, mentioned in Section 4.2).

Though there is no total agreement as to which features should be extracted from each physiological signal, some features have become fairly common. The mean and standard values of a signal over a certain time period, for instance, are commonly accepted psychophysiological features. Minimum and maximum values as well as mean absolute derivatives over a time period are also frequently seen. Additionally, certain signals have signal-specific features. For instance, heart rate is generally characterized by a number of time- and frequency-domain features of heart rate variability that have been defined by the Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996). Skin conductance, on the other hand, is often characterized by the amplitude and frequency of skin conductance responses (also called electrodermal reactions), transient increases in skin conductance that can occur in response to discrete psychological stimuli, but also occur spontaneously with no external trigger. A few examples of studies involving skin conductance responses in data fusion are works by Katsis et al. (2008), Kim and Andre (2008), and Frantzidis et al. (2010). It is also possible for features to be calculated using multiple raw signals. Pulse transit time, for instance, is defined as the time needed for the pulse pressure waveform to propagate through a length of the arterial tree, and is usually calculated using the electrocardiogram and the plethysmogram as the time between peaks in the two signals (as done by, for example, Liu et al. (2009)).

Due to the many different ANS responses that can be recorded, not all possible psychophysiological features are listed here. However, excellent lists of these features are available in Kreibig et al. (2007) and Kreibig (2010). Similarly, computational methods for the calculation of signal features are not described here, but are available in works such as Kim et al. (2004) (electrocardiography, skin conductance and skin temperature), Rani et al. (2004) (electrocardiography, skin conductance and electromyography) and Cacioppo et al. (2000) (general). All features can be calculated online over a sliding window, though some require larger windows. For instance, due to theoretical limitations, some features of heart rate variability cannot be calculated over a window shorter than two minutes (Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996).

A note: in other pattern recognition contexts, feature extraction often also includes dimension reduction methods such as principal component analysis. In this paper, dimension reduction methods are covered in Section 2.5.

2.4. Normalization

Psychophysiological features exhibit high intra- and intersubject variability as a result of age, gender, time of day and other factors. Normalization is primarily an attempt to reduce the effect of this variability prior to data fusion. For instance, if a training data set contains measurements from several subjects, some subjects may exhibit much larger responses than others or have different resting values for psychophysiological features (resting heart rate, for instance, can easily be anywhere between 60 and 80 beats per minute). This needs to be taken into account prior to data fusion. Furthermore, since different features are measured in different units, some features have much larger numerical ranges than others, which can be problematic for some data fusion methods (such as nearest-neighbor classification, described in Section 3.1.1). Normalization also attempts to reduce this effect. Three normalization approaches are commonly used, though it should be noted that not all psychophysiological studies use normalization and that not all studies report whether or not it was used.

The first approach is to record psychophysiological responses in a neutral or 'baseline' conditions where the subject is not exposed to stimuli or is only exposed to basic, relaxing stimuli. Psychophysiological features from other conditions (where the subject is performing a task or exposed to affective stimuli) are then normalized by either subtracting the baseline value (e.g. Christie and Friedman, 2004; Kim et al., 2004; Stephens et al., 2010), dividing by the baseline value (e.g. Zhai and Barreto, 2006; Arroyo-Palacios and Romano, 2010), subtracting the baseline value and dividing the result by the baseline value (e.g. Nasoz et al., 2004; Lisetti and Nasoz, 2004; Mohammad and Nishida, 2010), or a combination of these, with different options used for different features (e.g. Novak et al., 2010; Setz et al., 2010). Subtraction of the baseline value is aimed at reducing intersubject variability due to different baseline values while division is also partially aimed at reducing variability due to different response sizes. This approach can easily be used online, though it does require a baseline condition to be recorded first.

The second approach also begins by recording psychophysiological responses in a baseline condition. However, instead of subtracting or dividing the data from the 'task' or 'affective' conditions with the baseline data, the baseline features are added to the feature space as independent features: thus doubling the dimension of the feature space. This approach (called the 'baseline matrix') was used by, for example, Picard et al. (2001) and van den Broek et al. (2010). Like the previous approach, it can be easily done online, though it requires a baseline condition to be recorded first.

The third approach includes no baseline recordings, but simply involves normalizing the data from each subject separately or across all subjects to a certain range (e.g. from 0 to 1 or from -1 to 1). This is done for each feature separately by, for instance, subtracting the mean value of all feature vectors and dividing the result by the standard deviation of all feature vectors. If done for each subject separately, the goal is generally to reduce intersubject variability by scaling each person's features to a difference between their maximum and minimum value. If done across all subjects, the goal is simply to ensure that each psychophysiological feature has the same numerical range. This approach was used by, for example, Haag et al. (2004), Mandryk and Atkins (2007), Kulić and Croft (2007), Yannakakis et al. (2008) and Sakr et al. (2010). In online data fusion, normalizing features without a baseline recording can be done by calculating the maximum and minimum value of each feature across the entire training data set, then scaling features online between that maximum and minimum value.

The degree to which normalization improves data fusion is currently uncertain. Though improved correlations with self-report questionnaires and better discrimination between emotions as a result of normalization were established decades ago (e.g. Lykken et al., 1966; Ben-Shakhar, 1985), the exact effects of normalization are difficult to gauge. Some types of normalization (e.g. normalizing the data across all subjects to a certain range) are simple numerical rescaling and should thus have no effect on data fusion except in the case of algorithms such as k-nearest neighbors, which require normalization. However, normalization where the baseline values are subtracted from the values of the 'task' conditions is more than simple rescaling. Since each subject has a different baseline value, data from different subjects are scaled differently, thus 'warping' the dataset in a way.

While numerous studies have found improvements due to normalization, some have found certain normalization methods to be more effective and some have reported best results without any normalization. For instance, van den Broek et al. (2010) reported only a minimal improvement when using the baseline matrix approach, though the same study did find substantial improvement when normalizing by subtracting and dividing the baseline. Setz et al. (2010) compared data fusion with nonnormalized and normalized features (dividing by the baseline) and found better performance with nonnormalized features, though their application is fairly nonstandard (discriminating stress and cognitive load) and thus the results may not apply to the general case. Since normalization is only one step of the whole interpretation process, it is difficult to gauge its overall effect. A study that would analyze several different data sets and explore the effects of normalization when used with different other methods would be a useful way to clear up this uncertainty.

2.5. Dimension reduction

When multiple physiological responses are measured, it is possible that a very large number (20+) of features will be extracted from them. In data fusion, this can lead to the 'curse of dimensionality'. As the number of features (dimensionality) increases, the feature space grows in volume so quickly that it becomes difficult to find patterns and similarities in data. For machine learning techniques such as classifiers and estimators, the size of the required training data set can thus grow exponentially with the number of features. If the training data set is too small, overfitting can occur – data fusion rules trained on a small data set may not generalize well to new data. It can thus be beneficial to reduce the number of features prior to data fusion. Some data fusion methods already incorporate dimension reduction (e.g. classification tree pruning, random forests – Section 3.1.5), but several general dimension reduction techniques also exist. Two excellent introductory texts on dimension reduction are Guyon and Elisseeff (2003) as well as Liu and Motoda (2007). They are recommended for readers who are unfamiliar with the principles of dimension reduction and would like to learn about them in a broader context.

Dimension reduction can also help decrease the cost and complexity of a physiological computing application. If we know that some physiological signals do not provide any useful information, we can remove the sensors and feature extraction algorithms from the system entirely. One way to do this is as follows. First, record a training data set in different conditions with all potentially useful signals, and extract all potentially relevant features. Then, use dimension reduction to identify the most relevant features. If one of the measured physiological signals has no relevant features associated with it, it can be omitted in the future.

The techniques specifically used in psychophysiology can be roughly divided into three types: selection of individual features that ignores correlations between different features (Section 2.5.1), projection of the feature space onto a lower-dimensional space (Section 2.5.2) and selection of individual features that takes correlations between different features into account (Section 2.5.3). While the first and third type are mutually exclusive (with the third type having been found superior to the first as described in Section 2.5.3), the second can be used together with either of the other two (as described in Section 2.5.3). Other dimension reduction techniques that have seen only limited use in psychophysiology are mentioned in Section 2.5.4. All dimension reduction techniques depend on a training set and can be used both offline and online. For online use, either the best features are selected in advance or the projection of the feature space is calculated in advance. Online data fusion then either only uses the selected features or transforms the features online using the precalculated projection rule.

2.5.1. Feature ranking

A simple way to select the most appropriate features for data fusion is to rank the features according to a criterion of how much information each individual feature provides. Then, the best features (either a preselected number of features or all those who exceed a certain predefined threshold) are selected for data fusion. In psychophysiology, the most common way to rank individual features has been through analyses of variance, correlations and chi-square tests: statistical methods that find differences between different conditions (e.g. between 'sad' and 'angry' emotions) or connections between different variables (e.g. between arousal and heart rate). Only features that show statistically significant differences between conditions or significant connections between different variables can then be used in data fusion.

Analysis of variance (ANOVA) was used for feature selection by Wagner et al. (2005), where psychophysiological features were ranked according to their *p*-value and a preselected number of most significant features were selected. Analysis of variance was also used by van den Broek et al. (2010), where features with a *p*-value below 0.001 were selected, and by Chanel et al. (2011), where features with a *p*-value below 0.1 were selected. The chi-square test was used for feature selection by Pour et al. (2010), where the ten most significant features were chosen. Correlations with self-reported psychological variables were used for feature selection by Rani et al. (2007) and Liu et al. (2009), where only physiological features that had an absolute correlation coefficient of at least 0.3 were chosen. Correlations were also used to identify the most relevant physiological features by Bailenson et al. (2008), though this information was not later used in data fusion.

The weakness of this approach is that it ignores correlations between different physiological features. For example, if two features correlate highly with self-reported arousal, they may also correlate highly with each other. In this case, it may make sense to only include one of the two features in data fusion since the other one would not provide enough additional information to justify its inclusion.

2.5.2. Principal component analysis and Fisher's projection

Principal component analysis (PCA) is a method that transforms a large number of features into a smaller number of uncorrelated features (called principal components) that explain as much of the variability in the data as possible. Since it ensures that the principal components are uncorrelated with each other, PCA has an advantage over methods from the previous section which ignore correlations between different features. It has been used for dimension reduction in several psychophysiological studies, including Wagner et al. (2005), Rainville et al. (2006), Rigas et al. (2007), and van den Broek et al. (2010). However, it does have one important weakness: while the principal components explain as much of the variability in the data as possible, there is no guarantee that they are better-correlated with psychological states than the original features. If we have a training data set where each feature vector is labeled with a specific class (e.g. anger, fear...), it would be useful to take the labels into account during feature selection in order to ensure that the selected features discriminate between different classes. PCA, however, ignores any data labels.

The above weakness of PCA is addressed by Fisher's projection, which can be thought of as a supervised alternative to PCA. While PCA projects the original features onto a lower-dimensional space in such a way as to explain as much of the variability in the data as possible, Fisher's projection projects the original features onto a lower-dimensional space where between-class scatter is maximized and within-class scatter is minimized. In other words, it projects the original data into a lower-dimensional space where different classes (e.g. anger, fear...) are easier to linearly separate. Fisher's projection is essentially a version of linear discriminant analysis (Section 3.1.3), except used for dimension reduction rather than classification. It has been used by Picard et al. (2001), Healey and Picard (2005), Bonarini et al. (2008), and Gu et al. (2010). One weakness of Fisher's projection should be noted, however: since it transforms the original feature space into a space where different classes are linearly separable, it is less suitable for use with nonlinear data fusion techniques such as some variants of support vector machines and neural networks.

2.5.3. Sequential feature selection

Unlike PCA and Fisher's projection, which linearly transform the feature space, sequential feature selection methods (also known as stepwise methods) are methods that sequentially select individual features from the feature space. Unlike the approaches presented in Section 2.5.1, however, sequential feature selection methods do not ignore connections between different features.

Perhaps the most common sequential feature selection method is sequential forward selection, which works as follows. In the first step of the sequence, no features are included in the selection. The method evaluates all features to determine which one best discriminates between classes in the training data set (using criteria such as the *F*-value of each feature). That feature is included in the selection. In the next steps, all remaining features are evaluated to determine which one best discriminates between classes after the contributions of all previously selected features have already been taken into account. This process continues until no remaining feature contributes enough additional information to warrant its inclusion (for instance, the *F*-value of all remaining features is lower than a certain value). Sequential forward selection has been used in several psychophysiological studies, including Alpers et al. (2005), Wagner et al. (2005), Yannakakis et al. (2008), Yannakakis and Hallam (2008), Tognetti et al. (2010), Kolodyazhniy et al. (2011) and Wu et al. (2011). It has been shown to outperform feature ranking (Section 2.5.1) in Yannakakis and Hallam (2008), Yannakakis et al. (2008), and Yannakakis et al. (2010).

A very similar method to sequential forward selection is sequential backward selection. The difference is that, while forward selection begins with no features in the selection and sequentially adds features, backward selection begins with all features in the selection and sequentially removes features according to which one contributes the least to discrimination between classes. The process continues until the contribution of all remaining features exceeds a given threshold (for instance, the *F*-value of all remaining features is higher than a certain value). Sequential backward selection has been used in Kim (2007), Kim and Andre (2008), Giakoumis et al. (2011) and Kolodyazhniy et al. (2011). Kim and Andre (2008) reported that sequential backward selection outperformed sequential forward selection, though quantitative results were not reported for forward selection.

A combination of the above two methods is sequential floating forward selection, sometimes called sequential forward–backward selection. Starting with no features included in the selection, it sequentially adds features like sequential forward selection, but at each step it also evaluates whether any of currently included features can be removed. The most common criteria for inclusion or exclusion are *F*-value thresholds: a higher one for inclusion and a lower one for exclusion. Sequential floating forward selection has been used in several psychophysiological studies, including Picard et al. (2001), Gu et al. (2010), Chanel et al. (2011) as well as Wilson and Russell (2003a), where it is used in conjunction with discriminant analysis (Section 3.1.3) and thus called stepwise discriminant analysis.

It should finally be noted that Fisher's projection and sequential feature selection are not mutually exclusive. Instead of providing Fisher's projection with all possible features, it is possible to first select a subset of features using sequential feature selection and use Fisher's projection on this subset. This was first done by Picard et al. (2001), where the combination of the two approaches outperformed both of the two approaches used individually. Wagner et al. (2005) also found improved performance when using both approaches, though not for all classification methods. Finally, a combination of the two approaches was used by Gu et al. (2010), though it was not compared with using either approach individually. In principle, principal component analysis could also be used with sequential feature selection, and both Fisher's projection and principal component analysis could be used with selection of individually best features (Section 2.5.1). However, this has not been done in psychophysiology and there is no strong rationale for it since sequential feature selection has been shown to outperform feature ranking and since Fisher's projection takes class labels into account while principal component analysis does not.

2.5.4. Other

The aforementioned dimension reduction methods are of course not the only ones; they are simply the most prevalent in psychophysiology. Other methods include, for instance, Davies–Bouldin clustering (Leon et al., 2004; Leon et al., 2007), the Simba algorithm (Rigas et al., 2007) and genetic algorithms (Tognetti et al., 2010). However, these methods have not yet seen much use in psychophysiology, and additional studies will be required before their suitability for use in physiological computing can be properly assessed. Furthermore, a number of dimension reduction methods have not yet been evaluated with ANS responses in psychophysiology and should also be considered in future studies. Most are well-summarized by Guyen and Elisseeff (2003).

3. Data fusion

This section describes several possible methods for psychophysiological data fusion: a process which takes a physiological feature vector (consisting of several features extracted from multiple physiological signals) as input and assigns a psychological label to it. This psychological label can be categorical, in which case the feature vector is assigned to one of possible classes (e.g. 'angry', 'sad', 'low stress', 'high stress') and the process is called classification. Alternatively, the label can be a continuous value (e.g. an arousal of 9.2 on a scale between 0 and 10), in which case the process is called estimation. Classification and estimation are thus alternatives to each other, and externally differ mainly in whether the output is categorical or continuous. Since different methods are generally used for the two approaches, they are described in different sections. They are not, however, equally popular; in physiological computing, classification has been used far more than estimation.

Section 3.1 describes different classification methods while Section 3.2 describes different estimation methods. A comparison is then made in Section 3.3. All of these methods can be used both offline and online (real-time), though some are more computationally intensive and thus perhaps less suitable for online use (as discussed in Section 3.3.3). For each specific method, a brief description is provided followed by examples of use with ANS responses in psychophysiology and physiological computing. The goal is not to teach details about machine learning to the audience; rather, we provide a brief description so that readers unfamiliar with the concept of a particular method can obtain some initial information while readers already familiar with the method can use the examples to gauge its performance in specific psychophysiological applications. Those wishing more detailed descriptions of each method should refer to one of many available graduate-level pattern recognition textbooks such as the one by Bishop (2006).

3.1. Classification

There are many examples of classification in psychophysiology and physiological computing. However, their effectiveness is almost always gauged by their accuracy. Thus, each of the major methods described in this section has a corresponding table listing examples of the method's use. These tables have the following fields: the reference in which the study was published, the classification goal, the number of classes, the number of subjects, whether classification is subject-dependent or subject-independent, the accuracy rate (the percentage of feature vectors assigned to the correct class) and any measurements used in classification other than those of the ANS (e.g. electroencephalography – EEG, electromyography – EMG, electrooculography – EOG). However, a few other things should be noted regarding the tables:

- Direct comparison of accuracy rates between studies is difficult since accuracy depends on many factors, including the type of extracted features, normalization method and dimension reduction method.
- If multiple classification problems are performed in a study (e.g. different emotions classified separately), the results are averaged to obtain the overall accuracy if possible. If accurate averaging is not possible, an accuracy range is given.
- 'Levels of basic emotions' means that the emotion type is already known and the classification problem is only to determine the level of that emotion, even if multiple emotions are involved in the study.
- If multiple normalization or dimension reduction methods are used, the accuracy stated in the table is for the most accurate normalization or dimension reduction method.
- If a study is listed as subject-dependent, this does not necessarily mean that only data from a particular subject was used to classify that subject. It only means that the training data set included some measurements from the subject on which validation was performed (in extreme cases, the training data set can consist of 20 or more subjects in addition to the subject on whom validation was performed).
- If measurements other than those of the ANS are also used, the accuracy is specified for classification of all measurements unless otherwise stated.

3.1.1. Nearest neighbors

The k-nearest neighbor (kNN) algorithm is one of the simplest classification algorithms. When a new feature vector needs to be classified, the algorithm computes the (usually Euclidean or Mahalanobis) distance to each feature vector in the training data set. The training vectors are then ranked according to their distance to the new sample, and the k (where $k \geqslant 1$) nearest training vectors (neighbors) are used to classify the new feature vector using a majority vote: the sample is assigned to the class that is most common among the k nearest neighbors. The simplest version of this is the 1-nearest neighbor rule, where a new sample is assigned to the same class as the nearest vector in the training data set. Before calculating distances, it is usually necessary to scale the different data features (e.g. normalizing each feature to [0,1]) so that all features contribute equally to the distance calculation. Dimension reduction is also usually necessary, since the algorithm otherwise weighs all features equally even though some may not be relevant.

The k-nearest neighbor algorithm's simplicity is most likely the reason it has become popular in psychophysiology. Table 1 shows the studies that have used it.

An algorithm extremely similar to the k-nearest neighbors algorithm is the nearest class center. It differs in that, instead of distances being calculated for all feature vectors, each class is represented by the center of the feature vectors for that class (e.g. the mean and covariance of the data). A new feature vector is then simply assigned to the class with the nearest center, which is less computationally intensive than computing distances to every training vector. This approach was used by Frantzidis et al. (2010) and Setz et al. (2010).

3.1.2. Naïve Bayes classifier and Bayesian networks

A Bayesian network is, in essence, a probabilistic model of random variables and their conditional dependencies. Its simplest possible form is the naïve Bayes classifier, which assumes that all variables are independent of each other. Given the training data, it creates a probability model which estimates the probability that a feature vector belongs to a certain class. It then uses a decision rule to assign a class to the feature vector based on the probability model. Perhaps the most common rule is the 'maximum a posteriori' rule, which classifies the vector as coming from the class with the highest posterior probability.

Like the k-nearest neighbor algorithm, the naïve Bayes classifier has proven surprisingly effective despite its simplicity (e.g. Hand and Yu, 2001). One important advantage is that, by assuming independence between features, it requires a smaller training data set than other, more complex methods. Since the sample size in physiological computing is often limited, the naïve Bayes classifier could be an attractive option. However, several psychophysiological studies have also used more complex Bayesian networks, which do not assume that features are independent of each other. Examples of both the naïve classifier and more complex networks are given in Table 2.

3.1.3. Discriminant analysis

Discriminant analysis is a well-known classification method which finds a linear (linear discriminant analysis – LDA, also known as Fisher's linear discriminant) or quadratic (quadratic discriminant analysis – QDA) combination of input features which best separate feature vectors into two or more classes. This combination of input features is essentially a hyperplane in n-dimensional space (where n is the number of input features) that separates feature vectors of different classes. For a two-class problem, a linear discriminant function thus takes the form

$$D(\boldsymbol{x}) = \boldsymbol{w}\boldsymbol{x} + b \tag{1}$$

where $D(\boldsymbol{x})$ is the discriminant function, $\boldsymbol{x}$ is the vector of input features, $\boldsymbol{w}$ are the weights of the function and b is the intercept. $\boldsymbol{x}$ is then assigned to one class if $D(\boldsymbol{x})$ is positive and to the other class if $D(\boldsymbol{x})$ is negative. Both $\boldsymbol{w}$ and n are computed from training data.

Since each input feature has its own weight assigned to it, it is easy to determine how important it is to discrimination between classes. Though originally used for two-class problems, discriminant analysis can also be extended to multiclass situations. Its greatest limitation is that it only allows linear or quadratic relations between input and output; if strongly nonlinear relations are expected in the data, other methods may be preferable.

Because it is easy to use and transparently shows the contribution of each feature to discrimination between classes, discriminant analysis has been a popular data fusion method in psychophysiology. Recently, advanced extended versions of LDA have also been used, among them pseudoinverse LDA, which avoids singularity problems that can appear in classical LDA (Kim and Andre, 2008), and Kalman adaptive LDA, which can adjust the weights $\boldsymbol{w}$ online and thus gradually adapt to a particular subject (Novak et al., 2011; Koenig et al., 2011). Examples of all these are shown in Table 3.

3.1.4. Support vector machines

Similarly to discriminant analysis, support vector machines (SVMs) are a method of generating hyperplanes in n-dimensional space (where n is the number of input features) that separate feature vectors of different classes. The principal difference between the two is the criterion used to calculate these hyperplanes. While LDA maximizes a discriminative projection, SVMs are a maximum

Table 1
Psychophysiological studies that use the k-nearest neighbors algorithm. N = number of subjects, I = subject-independent classification, D = subject-dependent classification.

Study	Classification of	Classes	N	Indep.	Accuracy (%)	Other meas.
Picard et al. (2001)	Basic emotions	8	1	D	65	EMG
Lisetti et al. (2003)	Basic emotions	5	10	?	70–90	None
Nasoz et al. (2004) and Lisetti and Nasoz (2004)	Basic emotions	6	29	?	72	None
Wagner et al. (2005)	Basic emotions	4	1	D	90.9	EMG
Rigas et al. (2007)	Basic emotions	3	9	?	62.7	EMG
Nasoz et al. (2010)	Basic emotions	4	34	I	64.9	None
Kolodyazhniy et al. (2011)	Basic emotions	3	34	D, I	79.4 (D), 75.0 (I)	EMG
Rani et al. (2006)	Levels of 5 basic emotions	3 Per emotion	15	?	75.2	EMG
Shen et al. (2009)	Arousal-valence	4	1	D	60.3 ANS, 75.2 ANS + EEG	EEG
Gu et al. (2010)	Arousal-valence	4	28	D, I	90.7 (D), 50.3 (I)	EMG
van den Broek et al. (2010)	Arousal-valence	4	21	?	61.3	EMG
Bonarini et al. (2008)	Stress level	5	6	D	88.1	EMG
Liu et al. (2009)	Anxiety level	3	15	?	80.4	EMG
Kapoor et al. (2007)	Frustration level	2	24	I	66.7	Many
Tognetti et al. (2010)	Enjoyment level	3	75	I	57.0	None
Levillain et al. (2010)	Amusement level	2	25	?	76.9	None

Table 2
Psychophysiological studies that use the naïve Bayes classifier or a more complex Bayesian network.

Study	Classification of	Classes	*N*	Indep.	Accuracy (%)	Other meas.
Picard et al. (2001)[a]	Basic emotions	8	1	D	81.3	EMG
Calvo et al. (2009)[a]	Basic emotions	8	3	D	66.3 (1 session), 43.6 (all sessions)	EMG
Müller (2006)[a]	Arousal-valence	4	1	D	86	EMG
Zhai and Barreto (2006)[a]	Stress level	2	32	?	78.7	Eyes
Rigas et al. (2011)[a]	Stress and fatigue levels	2 Stress, 3 fatigue	1	D	66 stress (ANS), 74 fatigue (ANS)	Video
Calvo et al. (2009)[b]	Basic emotions	8	3	D	81.3 (1 session), 64.3 (all sessions)	EMG
Rani et al. (2006)[b]	Levels of 5 basic emotions	3 Per emotion	15	?	74.0	EMG
Conati (2002)[b]	Arousal-valence	Theoretical	0	N/A	N/A	N/A
Liu et al. (2009)[b]	Anxiety level	3	15	?	80.6	EMG
Kapoor et al. (2007)[b]	Frustration level	2	24	I	79.2	Many

N = number of subjects, *I* = subject-independent classification, *D* = subject-dependent classification.
[a] Naïve Bayes classifier.
[b] Bayesian network.

Table 3
Psychophysiological studies that use discriminant analysis.

Study	Classification of	Classes	*N*	Indep.	Accuracy (%)	Other meas.
Lisetti et al. (2003)[L] Nasoz et al. (2004)	Basic emotions	5	10	?	70–90	None
Lisetti and Nasoz (2004)[L]	Basic emotions	6	29	?	75	None
Christie and Friedman (2004)[L]	Basic emotions	7	34	?	37.4	None
Wagner et al. (2005) [L]	Basic emotions	4	1	D	92.1	EMG
Rainville et al. (2006) [L]	Basic emotions	4	43	?	49.0	None
Kreibig et al. (2007) [L]	Basic emotions	3	28	?	69.0	None
Setz et al. (2009) [L]	Basic emotions	4 or 5	20	I	58.8 for 4, 47 for 5 classes	EMG, EOG
Kolodyazhniy et al. (2011) [L]	Basic emotions	3	34	D, I	77.0 (D), 73.5 (I)	EMG
Kim (2007) [L]	Arousal-valence	4	3	D, I	71 (ANS, D), 51 (ANS, I), 79 (all, D), 54 (all, I)	Speech
Chanel et al. (2009) [L]	Arousal-valence	3	10	D	51 (ANS)	EEG
Wilson and Russell (2003a) [L]	Workload level	2	8	D	95	EEG, EOG
Healey and Picard (2005) [L]	Stress level	3	9	D	97.4	EMG
Chanel et al. (2011) [L]	Difficulty level	3	20	I	58 (ANS)	EEG
Giakoumis et al. (2011) [L]	Boredom level	2	19	D, I	94.7 (D), 89.4 (I)	None
Setz et al. (2010) [L]	Stress or cognitive load	2	33	I	82.8%	None
Alpers et al. (2005) [L]	Phobic and non-phobic subj.	2	38	?	95%	None
Blechert et al. (2006) [L]	Anxious and non-anxious subj.	2	42	?	83.3%	Movement
Chanel et al. (2009)[Q]	Arousal-valence	3	10	D	45 (ANS)	EEG
Setz et al. (2009)[Q]	Basic emotions	4 or 5	20	I	56 (4-class) 49 (5-class)	EMG, EOG
Kolodyazhniy et al. (2011)[Q]	Basic emotions	3	34	D, I	74.5 (D), 66.2 (I)	EMG
Chanel et al. (2011)[Q]	Difficulty level	3	20	I	59 (ANS)	EEG
Kim and Andre (2008)[P]	Arousal-valence	4	3	D, I	95 (D), 70 (I)	EMG
Novak et al. (2011)[K]	Difficulty level	2	24	I	76.4 (ANS), 84.7 (all)	Many
Koenig et al. (2011)[K]	Workload level	4	9	I	38 (ANS), 85 (all)	Many

N = number of subjects, *I* = subject-independent classification, *D* = subject-dependent classification. [L] = LDA, [Q] = quadratic discriminant analysis, [P] = pseudoinverse LDA, [K] = Kalman adaptive LDA.

margin classifier: it creates the hyperplane so that the distance (margin) between the hyperplane and the closest feature vectors on each side is maximized.

Basic SVMs thus have similar advantages and disadvantages as discriminant analysis. They are transparent and it is easy to determine the contribution of each input feature; but on the other hand, they are a linear classifier. To avoid the limitation of linearity, SVMs are commonly expanded using so-called kernels. A good explanation of kernels is provided by Schölkopf and Smola (2001), but in essence the training data is transformed into a higher-dimensional space and a hyperplane is generated in this space. While the hyperplane is linear in the new transformed space, it may be nonlinear in the original feature space, resulting in a nonlinear classifier.

The good performance and nonlinearity of SVMs has led to their frequent use in psychophysiology and physiological computing. Table 4 shows examples of their use.

3.1.5. Classification trees

Classification trees assign a class to a feature vector by progressing through several branching IF–THEN logical rules. This branching structure is the reason why they are called trees. An example of a psychophysiological classification tree would be “If skin conductance response frequency is below five per minute, the subject is bored. Otherwise, if skin temperature is below 33 °C, the subject is frustrated. Otherwise, the subject is entertained.” While not an accurate set of rules, this serves as a simple illustration of a classification tree. The rules are not defined manually; several different algorithms exist to learn the rules from training data. At each new node of the tree, these algorithms select the feature that best discriminates between classes after all the previous decisions made in the tree have been taken into account. Features are selected using criteria such as information gain.

Classification trees offer a very transparent way of classifying physiological data. The decision process can be easily followed by researchers and can be visualized graphically, making the trees a very ‘white-box’ approach. The tree building process acts as a form of dimension reduction, and many tree-building algorithms also incorporate tree pruning, which prevents the tree from becoming too complex and overfitting the data.

Table 5 shows examples of classification trees in psychophysiology and physiological computing. It also includes advanced

Table 4
Psychophysiological studies that use support vector machines.

Study	Classification of	Classes	*N*	Indep.	Accuracy (%)	Other meas.
Kim et al. (2004)	Basic emotions	3 or 4	50	?	78.4 (3-class), 61.8 (4-class)	None
Katsis et al. (2006)	Basic emotions	5	4	D	86	EMG
Katsis et al. (2008)	Basic emotions	4	10	?	79.3	EMG
Pastor-Sanz et al. (2008)	Basic emotions	6	24	?	63–83	EMG
Calvo et al. (2009)	Basic emotions	8	3	D	95.8 (1 session), 85.7 (all sessions)	EMG
Pour et al. (2010)	Basic emotions	2	16	D, I	42–84 (D), 52.6 (I)	EMG
Rani et al. (2006)	Levels of 5 basic emotions	3 Per emotion	15	?	85.8	EMG
Bailenson et al. (2008)	Levels of 2 basic emotions	2 Per emotion	41	D, I	80–95 (ANS, D), 30–90 (ANS, I), 94–99 (all, D), 50–95 (all, I)	Face
Liu et al. (2008)	Levels of 3 basic emotions	2 Per emotion	6	D	80–85	EMG
Müller (2006)	Arousal-valence	4	1	D	82	EMG
Chanel et al. (2009)	Arousal-valence	3	10	D	49 (ANS)	EEG
Shen et al. (2009)	Arousal-valence	4	1	D	68.1 ANS, 86.3 all	EEG
van den Broek et al. (2010)	Arousal-valence	4	21	?	60.7	EMG
Zhai and Barreto (2006)	Stress level	2	32	?	90.1%	Eyes
Kapoor et al. (2007)	Frustration level	2	24	I	70.8	Many
Liu et al. (2009)	Anxiety level	3	15	?	88.9	EMG
Chanel et al. (2011)	Difficulty level	3	20	I	56 (ANS)	EEG
Plarre et al. (2011)	Stress level	2	21	?	89.2	None
Rigas et al. (2011)	Stress and fatigue levels	2 Stress, 3 fatigue	1	D	78 stress (ANS), 85 fatigue (ANS)	Video
Sakr et al. (2010)	Agitation level	2	58	I	91.4	None
Wu et al. (2011)	Arousal level	3	18	D, I	96.5 (D), 36.9 (I)	EEG
Mohammad and Nishida (2010)	Behavior naturalness	2	44	?	81	None
Setz et al. (2010)	Stress or cognitive load	2	33	I	81	None

N = number of subjects, *I* = subject-independent classification, *D* = subject-dependent classification.

variants of classification trees. Fuzzy trees (Levillain et al., 2010) combine the hierarchical structure of classification trees with fuzzy logic (described in Section 3.2.2). Ensemble methods such as random forests (Rigas et al., 2007) and boosted decision trees (Bailenson et al., 2008; Plarre et al., 2011) produce sets of many trees whose outputs are combined to produce the final classification.

3.1.6. Artificial neural networks

Inspired by biological systems, artificial neural networks (ANNs) consist of a large number of simple, interconnected components ('neurons') operating in parallel. Each neuron receives a number of inputs and uses them to calculate the 'activation' of the neuron. Perhaps the simplest way to calculate this activation is to calculate a weighted sum of the inputs, then set the output as 1 if the weighted sum exceeds a certain threshold and 0 if the weighted sum does not exceed the threshold. This output is then fed to the next layer of neurons and so on until the final output is determined. Such a layered network with weighted sums and threshold is called a multilayer perceptron. Multilayer perceptrons can model functions of very high complexity if enough layers and neurons are used. However, other types of ANNs that incorporate more complex elements also exist (e.g. radial basis function networks). Complexity can be especially increased by allowing outputs of one layer of neurons to be used as inputs to both preceding and succeeding layers. This type of network is called a feedback network.

Table 5
Psychophysiological studies that use classification trees.

Study	Classification of	Classes	N	Indep.	Accuracy (%)	Other meas.
Rigas et al. (2007)	Basic emotions	3	9	?	62.4	EMG
Calvo et al. (2009)	Basic emotions	8	3	D	89.2 (1 session), 88.9 (all sessions)	EMG
Rani et al. (2006)	Levels of 5 basic emotions	3 Per emotion	15	?	83.5	EMG
Bailenson et al. (2008)[b]	Levels of 2 basic emotions	2 Per emotion	41	D, I	40–90 (ANS, I), 80–95 (ANS, D); 50–95 (all, I), 94–99 (all, D)	Face
Müller (2006)	Arousal-valence	4	1	D	77	EMG
Zhai and Barreto (2006)	Stress level	2	32	?	88.0	Eyes
Rani et al. (2007)	Anxiety level	3	4	D	83.8	EMG
Liu et al. (2009)	Anxiety level	3	15	?	88.5	EMG
Levillain et al. (2010)[a]	Amusement level	2	25	?	75.9	None
Plarre et al. (2011)[b]	Stress level	2	21	?	90.2	None
Rigas et al. (2011)[b]	Stress and fatigue levels	2 Stress, 3 fatigue	1	D	76 stress (ANS), 81 fatigue (ANS)	Video
Mohammad and Nishida (2010)	Behavior naturalness	2	44	?	79	None

N = number of subjects, *I* = subject-independent classification, *D* = subject-dependent classification.

[a] Fuzzy trees.

[b] Random forests or boosted trees.

Table 6
Psychophysiological studies that use artificial neural networks.

Study	Classification of	Classes	N	Indep.	Accuracy (%)	Other meas.
Nasoz et al. (2004) and Lisetti and Nasoz (2004)	Basic emotions	6	29	?	84	None
Wagner et al. (2005)	Basic emotions	4	1	D	88.6	EMG
Calvo et al. (2009)	Basic emotions	8	3	D	97.1 (1 session), 97.8 (all sessions)	EMG
Nasoz et al. (2010)	Basic emotions	4	34	I	73.3	None
Kolodyazhniy et al. (2011)	Basic emotions	3	34	D, I	77.9 (D), 76.5 (I)	EMG
Yannakakis et al. (2010)	Levels of 7 basic emotions	2 Per emotion	36	?	79	None
Müller (2006)	Arousal-valence	4	1	D	81–86	EMG
Arroyo-Palacios and Romano (2010)	Arousal-valence	4	59	I	78.4	None
van den Broek et al. (2010)	Arousal-valence	4	21	?	56.2	EMG
Leon et al. (2007)	Neutral/positive/negative	3	8	D	71.40	None
Wilson and Russell (2003a)	Workload level	2	8	D	99	EEG, EOG
Wilson and Russell (2003b)	Workload level	3	7	D	55.9 (ANS), 88.0 (all)	EEG
Wilson and Russell (2007)	Workload level	2	10	D	83.5	EEG, EOG
Yannakakis et al. (2008)	Entertainment preferences	2	18	?	76	None
Yannakakis and Hallam (2008)	Entertainment preferences	2	72	?	79.8	None
Leon et al. (2004)	Neutral or non-neutral state	2	1	D	Not given	EMG

N = number of subjects, *I* = subject-independent classification, *D* = subject-dependent classification.

ANNs are taught to perform a particular function using a training data set by adjusting the weights of the connections between different neurons. They are either linear or nonlinear tools capable of modeling very complex relationships between variables, which can be very useful in physiological computing. Since many different types of neural networks exist, readers unfamiliar with the fundamentals are encouraged to refer to the book by Bishop (1996), which remains an excellent introduction to the topic.

To train an ANN as a classifier, it simply needs to be provided with a training data set where the inputs are physiological features and the outputs are numbers corresponding to different classes (e.g. 1 – 'angry', 2 – 'sad'). However, ANNs have one important disadvantage. Once trained, it is difficult to determine how different input variables contribute to the output. ANNs thus provide users with little information about the underlying system. Despite this lack of transparency, they have been frequently used with psychophysiological data. Examples are given in Table 6.

3.1.7. Other

Though the previous subsections describe the classification methods commonly used in physiological computing, there are also several other, less-often used methods that bear mentioning. Some of these are:

- *Fuzzy logic:* More properly an estimation technique, it has also been used for psychophysiological classification in Rani et al. (2007) and Katsis et al. (2008) by simply assigning classes to different values of the output variable. Notable for not requiring a training data set, it is described in more detail in Section 3.2.2.
- *Hidden Markov models (HMMs):* Actually a type of dynamic Bayesian network (Section 3.1.2), HMMs are notable because they allow the classification of temporal sequences. Though popular in research fields such as speech recognition and activity recognition, HMMs have seen little use in psychophysiology where the preferred approach is to calculate features from a temporal sequence and then classify those features instead. Three examples of HMMs in psychophysiological data fusion are Scheirer et al. (2002), Kulić and Croft (2007), and Plarre et al. (2011).
- *Relevance vector machines:* Functionally similar to support vector machines (Section 3.1.4), relevance vector machines (RVMs) are embedded in a Bayesian framework. They have been shown to provide results similar to SVMs, but with sparser solutions. They were used for psychophysiological data fusion by Chanel et al. (2009) and may represent an emerging alternative to the commonly used classifiers in psychophysiology.
- *Large margin algorithm (LMA):* A simpler version of support vector machines (Section 3.1.4), LMA makes certain assumptions about the data in order to reduce computational complexity. It was used by Yannakakis and Hallam (2008), but is unlikely to see wider use in physiological computing where computational complexity is generally not a problem.

3.1.8. Ensemble classification

Ensemble classification (also referred to as multiple classifier systems) refers to the practice of combining several classifiers (of the same type or different types) to obtain a final result. This approach has several advantages over using a single classifier, and has proven to be very effective at dealing with numerous types of classification problems. It can, for instance, reduce the risk of creating a classifier that does not generalize well to new data or allow different classifiers to deal with specific types of data. An excellent review of ensemble classification in general machine learning was published by Polikar (2006).

Though ensemble classification is not especially widespread in psychophysiology, the structure of the input data or the psychological model used may lend themselves naturally to such an approach. For instance, if other data modalities are used in addition to ANS responses, it is possible to obtain a classification result using data from each modality separately and then combine these unimodal results to obtain a final result. This is often called decision-level fusion and was performed using speech and physiology by Kim (2007) as well as with central and ANS responses by Chanel et al. (2009, 2011). On the level of sensors rather than modalities, Setz et al. (2009) used the same approach to obtain a separate classification result from each physiological sensor (one result for all features extracted from the electrocardiogram, one result for all features extracted from skin conductance etc.) and then fuse them together.

The above approach features several classifiers working in parallel. An alternate option is to have several classifiers in series. First, a classifier performs a rough separation into two broader classes. Then, separate classifiers are used on the broader classes to determine the final, specific class. Three examples of this exist in psychophysiology. In the first two examples, the first classifier classifies a feature vector into one half of the arousal-valence space while the second classifier classifies the feature vector into one of the two remaining possible quadrants (Kim and Andre, 2008; Frantzidis et al., 2010). In the third example by Sakr et al. the first classifier classifies the feature vector as either 'easy to classify' or 'difficult to classify', and different classifiers are used for these two possibilities in order to determine whether the subject is agitated or not.

Finally, a third approach to ensemble classification has been used in psychophysiology: so-called bagging or boosting. Though there are large differences in the two approaches, both work by randomly dividing the data into smaller subsets, training a classifier on each subset, and combining the outputs of all classifiers via majority voting. This approach thus splits the data randomly rather than according to the data modalities or the psychological model. Psychophysiological examples include random forests (Rigas et al., 2007) and boosted decision trees (Bailenson et al., 2008; Plarre et al., 2011). However, it is the opinion of the authors that such methods are still underutilized in psychophysiology and could result in improved classification.

3.2. Estimation

As previously stated, estimation is, like classification, a method of taking a physiological feature vector (consisting of several features extracted from multiple physiological signals) as input and assigning a psychological label to it. It differs from classification primarily in that the label is continuous (e.g. anywhere between 0 and 10) rather than categorical.

Estimation has been used far less frequently in psychophysiology and physiological computing than classification. When comparing different studies, it also has an important weakness: there is no commonly accepted way in psychophysiology to gauge the effectiveness of estimation (unlike the accuracy rate of classifiers). While it should be possible to determine the mean squared error, variance, or bias of an estimator, this is rarely done in psychophysiological studies. In fact, several of the examples listed in this section describe the results of the implemented methods only qualitatively. Thus, while classification examples were listed in tables due to the large number of studies and the ability to compare accuracy rates, estimation examples are described in text.

3.2.1. Linear sums and linear regression

A simple way to estimate a psychological quantity from physiological features is to define it as a weighted sum of (usually normalized) physiological features:

$$y(\boldsymbol{x}) = \boldsymbol{wx} + b \qquad (2)$$

where y is a psychological quantity (e.g. arousal), $\boldsymbol{x}$ are physiological features (e.g. mean heart rate, skin conductance response frequency), $\boldsymbol{w}$ are the weights assigned to the different features and b is the intercept. $\boldsymbol{w}$ and b can be defined manually (e.g. Toups et al., 2006; Grigore et al., 2008), but a more optimal approach is to perform linear regression on the training data set. Given a data set with known y and $\boldsymbol{x}$, linear regression usually estimates $\boldsymbol{w}$ and b using the least squares method, though other methods are also possible. It has been used for estimation of distress, worry and task engagement by Fairclough and Venables (2006), estimation of amusement and sadness by Bailenson et al. (2008) and estimation of arousal by Grundlehner et al. (2009).

3.2.2. Fuzzy logic

Fuzzy logic is an extension of classical binary logic. Fuzzy statements do not have to be absolutely true or false, but have "degrees" of truth. There are thus also no hard boundaries between categories or exclusive memberships. Perhaps the most famous example of fuzzy logic involves temperature control, described with the statements: "If the room is cold, the heating should be set to maximum. If the room is hot, the heating should be off." In fuzzy logic, the room can be both cold and hot to some degree (e.g. 0.8 cold, 0.2 hot), and the heating is thus also set to some intermediate value. An example from physiological computing would be "if heart rate is high and skin conductance is high, arousal is high". Ranges for each variable are defined using membership functions and can overlap.

Fuzzy logic is appropriate for situations where a precise mathematical model does not exist, but experts can identify general rules underlying the system – as in physiological computing. It is also appropriate for systems with a high level of noise, which is also common in physiological computing due to the intra- and intersubject variability. Expert-defined fuzzy rules have been used to estimate stress and anxiety by Rani et al. (2002) and Rani et al. (2004) as well as arousal and valence by Mandryk and Atkins (2007) and Mihelj et al. (2009), with the latter also estimating the level of physical activity. Expert-defined fuzzy rules are especially noteworthy because, unlike most of the methods described in this paper, they do not explicitly require training data.

If the underlying behavior of the system cannot be described by experts, machine learning approaches also exist to identify the parameters of a fuzzy logic system using training data. Examples of fuzzy system identification for the purpose of user state assessment from ANS responses are presented in Kumar et al. (2007), Katsis et al. (2008) and Ting et al. (2010).

3.2.3. Artificial neural networks

Previously described in Section 3.1.6, artificial neural networks (ANNs) consist of a large number of simple, interconnected components ('neurons') operating in parallel. They are taught to perform a particular function (which can be simple or very complex) using a training data set by adjusting the weights of the connections between different neurons. While mostly used in psychophysiology for classification, they do not necessarily have to output a categorical value (e.g. 1 – 'angry', 2 – 'sad'); they can easily be trained to output continuous values and thus estimate the level of a particular psychological variable. One classic example is by Haag et al. (2004), who use neural networks to estimate the level of arousal and valence. A more recent example is a study by Bailenson et al. (2008), where ANNs are used to estimate the level of amusement and sadness.

3.3. Comparisons of data fusion methods

Having reviewed several data fusion methods, it is only natural to ask ourselves "which method is the best?" The answer, of course, is not simple and depends critically on the properties of the data, the goal of the overall system and the desires of the researcher.

3.3.1. Classification or estimation?

Classification and estimation are two data fusion approaches that assign a psychological label to a psychophysiological feature vector. Both have their uses, and the choice between them depends primarily on the problem researchers are trying to solve and the corresponding study design. For instance, if an experiment is built around basic emotions (anger, sadness, fear, surprise, happiness ...) (Ekman, 1992), the psychological state can be described as one of several discrete classes, naturally creating a classification problem. If the psychological state is described in terms of arousal and valence, estimation is more useful since arousal and valence are both continuous quantities (Russell, 1980).

If we wish to act on the inferred psychological state, the choice of classification also depends on the properties of the physiological computing system we wish to use. If the ultimate goal of the system is to select one of several possible discrete actions (e.g. assist user or do not assist user), classification is the obvious choice. If the goal of the system is, however, to adapt a continuous value (for instance, to change the speed of an enemy in a computer game, with the speed being any value between a preset minimum and maximum), estimation is the better choice.

In many cases, though, the general problem or hypothesis does not strictly specify the appropriate approach. For instance, if the problem is to identify the subject's arousal, this can be framed either as a classification problem (with low/medium/high arousal as the three classes) or as an estimation problem (an arousal value between 0 and 10). This choice then guides the study design, particularly the selection of psychological state induction and validation method, and it is difficult to change to the other option once the measurements have been completed. Thus, in such a case we would recommend that researchers consider their options in advance. As previously mentioned, classification has been used in psychophysiology and physiological computing far more often than estimation. Even if the goal is to identify the level of a psychological variable, it can be difficult to induce a great number of arousal, valence, stress or anxiety levels. Researchers thus often settle for splitting psychological states into arousal-valence quadrants (e.g. Chanel et al., 2009; Frantzidis et al., 2010) or discrete levels of a psychological variable (e.g. Wilson and Russell, 2003b; Zhai and Barreto, 2006; Kapoor et al., 2007), creating a classification problem again. We would recommend that researchers who are unfamiliar with psychophysiology and/or data fusion choose classification over estimation if the problem they are trying to solve does not specifically require estimation. At this point, the literature on classification is much more extensive, preparation of training data is easier, and the success of a system can be easily gauged via classification accuracy.

3.3.2. Estimation methods

If the goal of data fusion is estimation, choosing the appropriate method is not difficult since only a few are used in physiological computing. If the training data set is small and a simple, transparent model is desired, linear regression is appropriate. If the training data set is larger and a more complex, nonlinear but less transparent method is desired, artificial neural networks are preferable. Finally, fuzzy logic should be used if the training data set is limited but the researcher is sufficiently familiar with both general psychophysiology and the specific application to accurately define the necessary rules. Automated fuzzy system identification is not yet well-established in psychophysiology and is not recommended for researchers without good prior knowledge of both fuzzy logic and psychophysiology.

3.3.3. Classification methods

If the goal of data fusion is classification, choosing the appropriate method is harder since so many classification algorithms are available and widely used in psychophysiology. Perhaps the most important quality of a classifier is its accuracy – how well it can classify feature vectors. To evaluate accuracy, we can first turn to large-scale classifier comparisons from other fields. One comprehensive nonpsychophysiological comparison of classifiers on different real-world data sets was made in the 1990s by King et al. (1995). Other classifier comparisons with nonpsychophysiological data include Harper (2005) (medical data), Hua et al. (2005) (with a special focus on classifier accuracy as a result of sample size), Caruana and Niculescu-Mizil (2006) and Caruana et al. (2008). Some general conclusions can be drawn from these comparisons that should also apply to psychophysiological data. However, the most relevant information can be obtained directly from psychophysiological data.

Table 7 lists a number of psychophysiological and physiological computing studies that have compared different classifiers on their data. Unfortunately, those looking for a quick response to the "which classifier is the best" question are likely to be disappointed again. Different studies report results that may at first glance be contradictory. For instance, Nasoz et al. (2004) and Nasoz et al. (2010) find ANNs to perform much better than kNN, but van den Broek et al. (2010) report higher classification accuracy with kNN than with ANNs. Similarly, Zhai and Barreto (2006) find SVMs to be much more accurate than the naïve Bayes classifier, but Müller (2006) reports similar accuracy for both methods.

It is important to once again realize that the best method critically depends on many different factors such as the input features

Table 7
Psychophysiological studies that compare different classifiers.

Reference	Classes	*N*	Indep.	Classifiers compared (accuracy in %)
Wilson and Russell (2003a)	2	8	D	LDA (95), ANN (99)
Wagner et al. (2005)	4	1	D	kNN (90.9), LDA (92.1), ANN (88.6)
Müller (2006)	4	1	D	Naïve Bayes (86), SVM (82), ANN (81–86), trees (77)
Rani et al. (2006)	3 Per emotion	15	?	kNN (75.2), Bayesian network (74.0), SVM (85.8), trees (83.5)
Zhai and Barreto (2006)	2	32	?	Naïve Bayes (78.7), SVM (90.1), trees (88.0)
Kapoor et al. (2007)	2	24	I	kNN (66.7), Bayesian network (79.2), SVM (70.8)
Rani et al. (2007)	3	4	D	Trees (83.8), fuzzy logic (75.4)
Rigas et al. (2007)	3	9	?	kNN (62.7), trees (62.4)
Katsis et al. (2008)	4	10	?	SVM (79.3), fuzzy logic (76.7)
Pastor-Sanz et al. (2008)	6	24	?	SVM (63–83), kNN, naïve Bayes, trees (worse, no exact results given)
Yannakakis and Hallam (2008)	2	72	?	ANN (79.8), large margin algorithm (70.2)
Calvo et al. (2009) session-dependent	8	3	D	Naive Bayes (66.3), Bayesian network (81.3), SVM (95.8), ANN (97.1)
Calvo et al. (2009) session-independent	8	3	D	Naive Bayes (43.6), Bayesian network (64.3), SVM (85.7), ANN (97.8)
Chanel et al. (2009)	3	10	D	LDA (51), QDA (45), SVM (49), RVM (49)
Shen et al. (2009) – ANS only	4	1	D	kNN (60.3), SVM (68.1)
Shen et al. (2009) – ANS + EEG	4	1	D	kNN (75.2), SVM (86.3)
Mohammad and Nishida (2010)	2	44	?	SVM (81), trees (79)
Nasoz et al. (2010)	4	34	I	kNN (64.9), ANN (73.3)
Setz et al. (2010)	2	33	I	Nearest class center (78), LDA (83), SVM (81)
van den Broek et al. (2010)	4	21	?	kNN (61.3), SVM (60.7), ANN (56.2)
Chanel et al. (2011)	3	20	I	LDA (58), QDA (59), SVM (56)
Kolodyazhniy et al. (2011)	3	34	D, I	kNN (79.4 D, 75 I), LDA (77.0 D, 73.5 I), QDA (74.5 D, 66.2 I), ANN (77.9 D, 76.5 I)
Plarre et al. (2011)	2	21	?	SVM (89.2), trees (90.2)
Rigas et al. (2011) – ANS only	2 Stress, 3 fatigue	1	D	Naive Bayes (66–74), general Bayes (67–77), SVM (78–85), trees (76–81)
Rigas et al. (2011) – all measurements	2 Stress, 3 fatigue	1	D	Naive Bayes (76–79) general Bayes (79–81), SVM (86–88), trees (80–81)

N = number of subjects, *I* = subject-independent classification, *D* = subject-dependent classification.

and the possible classes used. For instance, Nasoz et al. (2004), Wagner et al. (2005) and Rani et al. (2006) report that certain classifiers are better at recognizing certain emotions, though it is uncertain whether or not this is a sampling fluke. Rani et al. (2007) compare classification trees and fuzzy logic on data sets of different qualities and find that while trees generally result in higher accuracy, fuzzy logic is more accurate if the data quality is low. Direct comparison of classification accuracies between studies is thus difficult due to all the factors that need to be taken into account. In any case, while some studies have found clearly better results with particular classifiers (e.g. Nasoz et al., 2004; Nasoz et al., 2010, where ANNs clearly outperform LDA and kNN, or Kolodyazhniy et al., 2011, where nonlinear classifiers outperform linear ones), several have reported similar results for different classifiers (e.g. Wagner et al., 2005; Müller, 2006; Rigas et al., 2007; Katsis et al., 2008; Chanel et al., 2009). It is our admittedly subjective opinion that the greatest limitation to high classification accuracy is in fact the nature of the data itself. Physiological responses are affected not only by psychological and affective stimuli, but by many confounding factors such as physical activity, environmental temperature and intersubject differences in physiology. Even the best classifier can obtain only a limited amount of information from physiological data, limiting the possible accuracy. Thus, while accuracy is certainly an important factor, it is not the only one that should be considered.

Another factor is the *robustness* of a classifier with regard to sample size and number of features, which strongly affect classification accuracy as shown for example by Hua et al. (2005) for non-psychophysiological data. This has unfortunately seen little study with affective ANS responses, but conclusions can also be drawn from other pattern analysis fields. In general, mathematically less complex methods usually require a smaller data set since fewer parameters need to be defined. There is also less danger of overfitting the data, making simple methods such as naïve Bayes classifiers and LDA suitable for smaller training data sets (as noted by e.g. Hand and Yu, 2001 for naïve Bayes classification). Complex nonlinear classifiers such as Bayesian networks and ANNs generally require a larger training data set to avoid overfitting (as noted by e.g. Kolodyazhniy et al. for QDA and ANNs). kNN algorithms are unsuitable for large numbers of features not only due to computational complexity, but also because they cannot easily handle irrelevant features.

The *speed* and computational cost of a classifier can also be important. While all classifiers can be used both offline and online, some are less suited for online use. Calvo et al. (2009), for example, found ANNs to be more accurate than SVMs, but also much slower than SVMs and thus less suitable for online use. However, it is important to differentiate between the time needed to train the classifier (which can be done in advance) and the time needed to apply the classifier to a new feature vector (which often needs to be done online). LDA, for instance, is simple to both train and apply. SVMs and ANNs can be time-consuming to train, but can be applied to new data much faster. On the other hand, kNN requires no advance training, but can be computationally intensive to apply to a new feature vector online since the distance to each feature vector in the training data set must be calculated in many dimensions. Although classification in physiological computing is generally not performed with a high frequency (most features are calculated over a range of 30 s to five minutes), it can nonetheless be preferable to employ a classifier which requires little time to classify a new feature vector.

The *compatibility* of a classifier with normalization and dimension reduction methods should also be considered. While all the described classifiers can in principle be used with all normalization and dimension reduction methods, some combinations are more or less suitable. kNN, for instance, practically requires all features to be normalized to the same range. If this is not done, not all features will contribute equally to the distance between feature vectors. kNN also generally requires dimension reduction since the algorithm otherwise weighs all features equally even though some may not be relevant. Fisher's projection is less suitable for use with nonlinear classifiers since it transforms the original feature space into a space where different classes are linearly separable. In such a space, nonlinear classifiers obviously lose a great deal of usefulness, though they may nonetheless be more accurate than linear ones since the transformation can never be perfect. Many classification tree generation algorithms already incorporate a form of dimension reduction similar to sequential feature selection, rendering additional dimension reduction less important.

A generally less crucial factor is the *transparency* of the classifier. Rather than the most accurate classifier, we might choose a slightly less accurate classifier whose classification procedure can be easily understood by humans. In this case, classification trees provide a very transparent method since their if–then reasoning can be easily followed. LDA is also fairly simple to understand while nonlinear methods such as neural networks are often looked down on despite attempts to do away with their reputation as a 'black box' (e.g. Benitez et al., 1997). Here, a consideration must be made whether the potential decrease in accuracy from using a transparent classifier is an acceptable sacrifice for increased transparency. Such a decision is fairly subjective and thus generally left to the researcher's preference.

Another problem that mainly depends on the individual researcher is the *ease of implementation* of the classifier. Not everyone involved in physiological computing has the time and knowledge needed to implement all of the aforementioned classifiers. The easiest two to implement are kNN and LDA. LDA requires only the mean value and covariance of each class to be calculated, and then a simple equation is used to classify the data. kNN only requires simple arithmetic operations, making it even simpler than LDA since it does not require covariance calculation. Bayesian networks and ANNs, on the other hand, can be very complex to properly understand and implement. Although most data analysis packages already provide classification options, programs such as SPSS (IBM Corporation) do not allow online classification. For use in a real-time application, it is thus necessary to use engineering software such as MATLAB (MathWorks) that includes classification or program the classifiers manually.

To summarize, a few recommendations can be given depending on what the researcher is interested in:

- *Transparency*: classification trees (best), linear discriminant analysis (second-best).
- *Small sample*: linear discriminant analysis, naïve Bayes classifier.
- *Ease of implementation*: linear discriminant analysis.
- *Nonlinearity, large sample*: support vector machines, artificial neural networks, Bayesian networks.

With regard to *accuracy*, perhaps the best (subjective) recommendation that can be given is this: Since most classifiers are not difficult to implement and since they all require training data, researchers should consider implementing several different classifiers (as well as dimension reduction) and comparing them using crossvalidation on the training data set in order to determine which one is most appropriate for their situation. However, since the overwhelming majority of existing studies use either only one classifier or only one data set, a very useful addition to the literature would be a study that would apply several different classifiers to several different sets of ANS responses in order to thoroughly analyze the benefits and drawbacks of each classifier in different conditions. Such a study may also consider analyzing the effects of ensemble learning, which remains underutilized in psychophysiology despite promising results in other fields.

4. System adaptation

While data fusion is primarily the process of interpreting physiological responses in a cognitive or affective context, system adaptation is the act of using the interpreted measurements in order to make changes to the system that the user is interacting with. These changes then again affect physiological responses and cause new adaptations. Such adaptation is not a new idea by any means; for instance, a review by Byrne and Parasuraman (1996) discusses several early attempts to use ANS responses to control the level of automation in a task.

Adaptive applications can broadly be divided into three categories. The first (described in Section 4.1) is adaptive automation: making a task easier for the user by providing automated assistance when necessary. The second (described in Section 4.2) is game difficulty adjustment: making a game easier or harder for the user in order to provide an appropriate challenge. The third (described in Section 4.3) is the adjustment of the audio or visual properties of an application that the user is interacting with in order to make it more pleasant and attractive for the user or to evoke a certain other mood. Finally, section 4.4 gives a brief summary of the different ways system adaptation has been achieved in physiological computing and gives some recommendations for future work.

4.1. Adaptive automation

Adaptive automation is the act of activating automated assistance systems (e.g. automatic pilots) in response to a detected high user workload. Though the majority of work on adaptive automation through physiology has focused on electroencephalography, some studies have also incorporated ANS responses, either by themselves or in combination with other measurements.

Prinzel et al. (2003), Liao et al. (2005) and Wilson and Russell (2007) all take a similar approach to adaptive automation: assistance is enabled when the user's level of stress or workload is high and disabled otherwise. A heart rate variability threshold is used to enable and disable assistance in Prinzel et al. (2003). A Bayesian network is used for fusion of ANS responses and video in Zhai and Barreto (2006) while ANNs are used with electroencephalography, electrooculography and ANS responses in Wilson and Russell (2007). The study by Ting et al. (2010) differs slightly from the previous three in that different automation levels are available; i.e. the control for automation is not only an on/off switch. The level of automation is determined by fusing features derived from the electrocardiogram and electroencephalogram using fuzzy logic.

Rani et al. (2004) suggest a system very similar to adaptive automation, except that automatic assistance activation is replaced by a simple query. A mobile robot performs tasks in the environment while monitoring a human's level of anxiety. The level of anxiety is calculated from heart rate, skin conductance and the electromyogram using fuzzy logic. If anxiety exceeds a certain threshold, the robot ceases its normal operations and queries the human whether he or she requires assistance. Shye et al. (2008) also suggest an interesting application (though of questionable practical value) that is similar to the idea of adaptive automation. A computer monitors the user's engagement level through a combination of skin conductance and nonphysiological signals. If the user is not focused on working with the computer, the computer decreases the microprocessor speed in order to save energy. This can be thought of as the opposite of adaptive automation: while adaptive automation has the computer take over some of the workload when the user is overworked, the application of Shye et al. decreases the computer's capabilities when the user is unlikely to need them.

4.2. Game difficulty adjustment

Multiple studies have used ANS responses to adjust the parameters of a computer game or similar system in order to make it easier or harder for the subject. The level of data fusion in these games differs strongly, from none at all to very complex.

Looking first at examples of game difficulty adjustment based on only one physiological measurement, Bersak et al. (2001) created a racing computer game where the speed of the car is inversely proportional to the value of the user's skin conductance: the lower the skin conductance, the faster the car. Nenonen et al. (2007) used heart rate to affect the difficulty of a biathlon computer game, though it is questionable whether changes in heart rate are caused by psychological factors. In their game, high heart rate results in fast skiing, but inaccurate shooting, and vice versa.

Moving onto studies combining multiple physiological measurements, Toups et al. (2006) used skin conductance and electromyography to increase or decrease the activity level of enemies in a computer game, though data fusion was simply performed as a linear sum of individual normalized features. Dekker and Champion (2007) changed the player's movement speed, visibility to enemies and the damage of his/her weapons in a first-person shooter game based on both heart rate and skin conductance. Similarly, Kuikkaniemi et al. (2010) and Nacke et al. (2011) controlled the player's movement speed, weapon strength and weapon accuracy in games using multiple physiological measurements, though no data fusion was performed. Haarmann et al. (2009) combined heart rate and skin conductance in a flight simulator. Manually set thresholds were used on the features to determine how aroused the subject was, and turbulence was turned on and off in the flight simulator depending on the level of arousal. Liu et al. (2009) used a classification tree on multiple physiological signals to estimate the level of anxiety and then used both task performance and anxiety to control the difficulty of a game of Pong. Novak et al. (2011) and Koenig et al. (2011) both used Kalman adaptive LDA on multiple physiological signals to control the difficulty of a game-like motor rehabilitation exercise.

A final, very interesting game-like scenario is a study by Liu et al. (2008) where children need to throw baskets through a basketball hoop controlled by a robotic arm. The hoop is constantly moved in different directions, with the speed and direction of movement changed to maximize the child's enjoyment of the game. The child's level of enjoyment during the game is determined by using SVMs to fuse multiple psychophysiological features. Furthermore, the robotic arm gradually adapts the system adaptation rules to the current subject. Since there is no guarantee that two users will respond to a particular action in the same way, the arm learns the subject's preferences through reinforcement learning, which learns by trying certain actions and noting the subject's response. Given enough time to try different actions, the system learns what action is likely to lead to a certain response for that subject.

4.3. Audiovisual adaptation

Unlike adaptive automation, which has been extensively applied to critical situations such as flight, audiovisual adaptation has primarily been explored within the context of multimedia applications, computer games and virtual reality. Here, the purpose is to have the environment reflect the user's current mood or to evoke a certain mood in the user using a feature of the environment.

Perhaps the first example of environments that try to match the subject's mood is described by Wang et al. (2004): an online chatting interface where the color and shape of the text changes to match the user's skin conductance. Similarly, Dekker and Champion (2007) and Groenegress et al. (2010) directly map physiological responses to audiovisual properties of a game or virtual environment, without any data fusion. In Arroyo-Palacios and Romano (2010), ANNs are used to classify the subject's mood, and appropriate wallpaper is displayed on the computer background.

Environments that try to evoke a specific emotion using purely audiovisual features usually make use of either music or lighting. An affective music player was suggested as early as 1998 by Healey et al. though the implementation is relatively basic since a single input feature (the average difference in skin conductance) is used. Oliver and Kreger-Stickles (2006) propose a music player that combines both physiology and body movement to suggest songs from a playlist, though this does not necessarily include psychological factors since ANS responses in their study are strongly affected by physical activity. Janssen et al. (2009) also suggest a music player that combines skin conductance and skin temperature using a Bayesian network in order to suggest songs. Liu et al. (2010) attempt to control heart rate around a certain threshold by playing appropriate music, though their heart rate sensor is embedded in a seat beneath the subject and is thus fairly nonstandard. A similar approach to these music recommendation systems is a content delivery system by Shen et al. (2009), which classifies the user's autonomic and central nervous system responses using *k*-nearest neighbors and SVMs. It then suggests content (different documents) that would be appropriate in that mood.

Moving onto lighting, Grigore et al. (2008) try to help the subject relax by adjusting the level of ambient light in a room. A simple weighted sum of different heart rate and skin conductance features is used to estimate the subject's current state. A final, unorthodox example is by Ritter (2011), who uses evolutionary algorithms to analyze skin conductance and dynamically adjust the shading of a user interface or the lighting of a room in order to improve the users' performance in a task. The evolutionary algorithms used are very different from the normal data fusion methods currently used in psychophysiology, but they are also fairly nontransparent, which may make it difficult for other authors to reproduce and validate the work.

4.4. Discussion and recommendations for physiological computing

Though adaptation in response to psychological information inferred from ANS responses has been used in many applications, most existing implementations fall into one of two categories. The first category includes systems without any explicit data fusion at all. In this case, either a variable of the system is proportional to a physiological feature (e.g. Bersak et al., 2001, where the speed of a car is inversely proportional to skin conductance) or different actions are taken depending on whether a feature is above or below a predefined threshold (e.g. Haarmann et al., 2009, where assistance is activated or deactivated depending on thresholds). The interpretation of physiological features can be considered as implicit in the adaptation rules. Though a very simple implementation, such adaptation can be useful when no training data is available, when only basic adaptation is required and when the relationship between psychophysiological features and psychological states is well-established. Skin conductance, for example, has been extensively documented as proportional to arousal. This approach, however, is not recommended in more complex applications since it cannot easily merge multiple psychophysiological features and thus cannot give a good approximation of the subject's psychological state.

The second category includes a classifier followed by a simple decision-making system that has a predefined action assigned to each possible class. An example from this category is the adaptive automation system by Wilson and Russell (2007), which uses an ANN with two possible classes: high and low workload. The decision-making system then enables task automation in the case of high workload and disables it in the case of low workload. A second example is the content delivery system by Shen et al. (2009), which uses SVM to classify the user's mood. It then suggests different documents in each possible mood. These systems can be thought of as the state of the art. We recommend that researchers interested in physiological computing implement such a system rather than one with no data fusion, despite the added requirements of training data preparation and classifier construction.

In these systems, an entirely practical question may then arise: how accurate must a classifier be before the system can be considered 'good enough' to make decisions? We believe that the required accuracy would depend on the application. For instance, in an adaptive automation or automated warning system (for instance, in a car or airplane) where the computer would need to either warn the user of potential danger or take over some of the workload, the required accuracy would be very high – certainly over 95%. With a lower accuracy, the computer would be annoying at best (warning the user in inappropriate moments) and dangerous at worst (ignoring actual cases of user stress). In such cases, within-subject classification would be necessary to achieve an acceptable accuracy. Since it is expected that a single car or airplane would not be used by a large number of people, it would be feasible to train a separate classifier for each subject separately, thus increasing the accuracy. In more casual applications where an incorrect decision cannot have serious consequences, the needed accuracy is lower. For instance, in a computer game where the difficulty is regularly adjusted accuracies of around 70% for a two-class problem (increase/decrease difficulty) may be acceptable since the general trend would lead the player toward the optimal difficulty given enough time. The classifier in such a case could be subject-independent, since users of entertainment technologies may not be willing to spend time building a training data set for a casual application. In both cases, psychophysiological features should also be combined with other, already available data (e.g. the speed of the vehicle or the player's score in the game) in order to increase accuracy.

Finally, a third category of physiological computing systems is emerging: systems that do not only use predefined classification and decision-making rules, but which gradually adapt to the user as they gain experience. One existing example is the work of Liu et al. (2008), where SVMs are used to classify the level of enjoyment in a 'game' and actions are then taken to increase enjoyment. However, while the SVMs are predefined and static, the decision-making system gradually learns what actions are likely to increase enjoyment for that particular user and adapts accordingly. We personally feel that such an approach is very promising and represents the next step in psychophysiological feedback. Just like the complexity of data fusion has increased from simple predefined thresholds to advanced classifiers such as neural networks, we hope that the complexity of decision-making will increase from simple predefined actions to dynamically adapting intelligent decision-making systems.

5. Concluding remarks

Having examined the different algorithms and methods for data fusion and system adaptation in physiological computing, some final remarks can be made for researchers interested in the topic. The majority of existing data fusion methods in physiological computing (with the exception of principal component analysis and fuzzy logic) is supervised, perhaps because connections between ANS responses and psychological states are still not yet precisely known. This is unlikely to change in the near future, and thus we recommend focusing on supervised data fusion methods. These methods require properly prepared training data that should be verified using nonphysiological measures such as self-report questionnaires or observers.

Data fusion should be preceded by feature extraction using features already well-established in the literature. Though the benefits of normalization are somewhat uncertain, some form of normalization should nonetheless be implemented and tested. Dimension reduction is also recommended, with the authors' subjective opinion being that sequential feature selection is perhaps the best of the three widely used approaches. Even if data fusion is planned with methods that are robust with regard to large numbers of features, dimension reduction can at least remove any irrelevant features.

The choice of optimal data fusion method depends on many factors. We recommend choosing classification over estimation, both because it is more prevalent in the literature and because discrete classes are easier to validate using questionnaires or independent observers than continuous values. Choosing a specific classifier is not easy since currently published studies do not identify any one as superior to the others when applied to psychophysiological data. Our subjective recommendations are as follows. With regard to accuracy, we recommend implementing several different classifiers and comparing them using crossvalidation on the training data set in order to determine which one is most appropriate for a given situation. Of course, only a few classifiers can be selected for implementation based on factors other than accuracy. If high transparency is desired, we recommend classification trees. If the available data set is limited and a simple algorithm is needed to avoid overfitting, we recommend either linear discriminant analysis or the naïve Bayes algorithm. Conversely, if the available data set is large and a complex nonlinear model is desired, we recommend either support vector machines, Bayesian networks or artificial neural networks.

Though the number of studies that use physiological measurements for system adaptation is increasing, many studies still use measured responses without any data fusion (and in many cases with only basic feature extraction). We expect that a major focus of future physiological computing research will be to effectively combine complex data fusion and decision-making methods. In this way, physiological computing should have a future both in serious applications such as adaptive automation and in light-hearted applications such as computer games.

References

Alpers, G.W., Wilhelm, F.H., Roth, W.T., 2005. Psychophysiological assessment during exposure in driving phobic patients. Journal of Abnormal Psychology 114, 126–139.

Arroyo-Palacios, J., Romano, D.M., 2010. Bio-affective computer interface for game interaction. International Journal of Gaming and Computer-Mediated Simulations 2 (4), 16–32.

Bailenson, J.N., Pontikakis, E.D., Mauss, I.B., Gross, J.J., Jabon, M.E., Hutcherson, C.A., 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.

Benitez, J.M., Castro, J.L., Requena, I., 1997. Are artificial neural networks black boxes? IEEE Transactions on Neural Networks 8, 1156–1164.

Ben-Shakhar, G., 1985. Standardization within individuals: a simple method to neutralize individual differences in skin conductance. Psychophysiology 22, 292–299.

Bersak, D., McDarby, G., Augenblick, N., McDarby, P., McDonnell, D., McDonald, B. et al., 2001. Intelligent biofeedback using an immersive competitive environment. In: Online Proceedings for the Designing Ubiquitous Computing Games Workshop.

Bishop, C.M., 1996. Neural Networks for Pattern Recognition. Oxford University Press.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

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.

Bonarini, A., Mainardi, L., Matteucci, M., Tognetti, S., Colombo, R., 2008. Stress recognition in a robotic rehabilitation task. In: Proceedings of "Robotic Helpers: User Interaction, Interfaces and Companions in Assistive and Therapy Robotics", A Workshop at ACM/IEEE HRI 2008, Amsterdam, Netherlands, pp. 41–48.

Bradley, M.M., Lang, P.J., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry 25, 49–59.

Byrne, E.A., Parasuraman, R., 1996. Psychophysiology and adaptive automation. Biological Psychology 42, 249–268.

Cacioppo, J.T., Tassinary, L.G., 1990. Inferring psychological significance from physiological signals. American Psychologist 45, 16–28.

Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (Eds.), 2000. Handbook of Psychophysiology, second ed. Cambridge University Press, Cambridge.

Calvo, R.A., Brown, I., Scheding, S., 2009. Effect of experimental factors on the recognition of affective mental states through physiological measures. In: Proceedings of 22nd Australasian Joint Conference on, Artificial Intelligence, pp. 62–70.

Calvo, R.A., D'Mello, S., 2010. Affect detection: an interdisciplinary review of models, methods and their applications. IEEE Transactions on Affective Computing 1, 18–37.

Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, pp. 161–168.

Caruana, R., Karampatziakis, N., Yessenalina, A., 2008. An empirical evaluation of supervised learning in high dimensions. In: Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, pp. 96–103.

Chanel, G., Kierkels, J.J., Soleymani, M., Pun, T., 2009. Short-term emotion assessment in a recall paradigm. International Journal of Human–Computer Studies 67, 607–627.

Chanel, G., Rebetez, C., Bétrancourt, M., Pun, T., 2011. Emotion assessment from physiological signals for adaptation of game difficulty. IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans 41, 1052–1063.

Christie, I.C., Friedman, B.H., 2004. Autonomic specificity of discrete emotion and dimensions of affective space. International Journal of Psychophysiology 51, 143–153.

Conati, C., 2002. Probabilistic assessment of user's emotions in educational games. Applied Artificial Intelligence 16, 555–575.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G., 2001. Emotion recognition in human–computer interaction. IEEE Signal Processing Magazine 18, 32–80.

Dekker, A., Champion, E., 2007. Please biofeed the zombies: enhancing the gameplay and display of a horror game using biofeedback. In: Proceedings of DiGRA 2007: Situated Play, Tokyo, Japan, pp. 550–558.

Ekman, P., 1992. An argument for basic emotions. Cognition and Emotion 6, 169–200.

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

El Ayadi, M., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: features, classification schemes and databases. Pattern Recognition 44, 572–587.

Fairclough, S.H., 2009. Fundamentals of physiological computing. Interacting with Computers 21, 133–145.

Fairclough, S.H., Venables, L., 2006. Prediction of subjective states from psychophysiology: a multivariate approach. Biological Psychology 71, 100–110.

Frantzidis, C.A., Bratsas, C., Klados, M.A., Konstantinidis, E., Lithari, C.D., Vivas, A.B., et al., 2010. On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications. IEEE Transactions on Information Technology in Biomedicine 14, 309–318.

Giakoumis, D., Tzovaras, D., Moustakas, K., Hassapis, G., 2011. Automatic recognition of boredom in video games using novel biosignal moment-based features. IEEE Transactions on Affective Computing 2, 119–133.

Grigore, O., Gavat, I., Cotescu, M., Grigore, C., 2008. Stochastic algorithms for adaptive lighting control using psycho-physiological features. International Journal of Biology and Biomedical Engineering 2, 9–18.

Groenegress, C., Spanlang, B., Slater, M., 2010. The physiological mirror: a system for unconscious control of a virtual environment through physiological activity. The Visual Computer 26, 649–657.

Grundlehner, B., Brown, L., Penders, J., Gyselinckx, B., 2009. The design and analysis of a real-time, continuous arousal monitor. In: 2009 Sixth International Workshop on Wearable and Implantable Body Sensor, Networks, pp. 156–161.

Gu, Y., Tan, S., Wong, K., Ho, M. R., Qu, L., 2010. A biometric signature based system for improved emotion recognition using physiological responses from multiple subjects. In: 2010 8th IEEE International Conference on Industrial Informatics, Osaka, Japan, pp. 61–66.

Gunes, H., Pantic, M., 2010. Automatic, dimensional and continuous emotion recognition. International Journal of Synthetic Emotions 1, 68–99.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182.

Haag, A., Goronzy, S., Schaich, P., Williams, J., 2004. Emotion recognition using biosensors: first steps towards an automatic system. In: Affective Dialogue Systems 2004. Springer-Verlag, Berlin, Heidelberg, pp. 36–48.

Haarmann, A., Boucsein, W., Schaefer, F., 2009. Combining electrodermal responses and cardiovascular measures for probing adaptive automation during simulated flight. Applied Ergonomics 40, 1026–1040.

Hand, D.J., Yu, K., 2001. Idiot's Bayes – not so stupid after all? International Statistical Review 69, 385–398.

Harper, P.R., 2005. A review and comparison of classification algorithms for medical decision making. Health Policy 71, 315–331.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock, P.A., Hancock, P.A., Meshkati, N. (Eds.), Human Mental Workload. Amsterdam, Netherlands, pp. 139–183.

Healey, J.A., Picard, R.W., Dabek, F., 1998. A new affect-perceiving interface and its application to personalized music selection. In: Proceedings of the 1998 Workshop on Perceptual User Interfaces, San Francisco, USA.

Healey, J.A., Picard, R.W., 2005. Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems 6, 156–166.

Hettinger, L.J., Branco, P., Encarnaco, L.M., Bonato, P., 2003. Neuroadaptive technologies: applying neuroergonomics to the design of advanced interfaces. Theoretical Issues in Ergonomic Science 4, 220–237.

Hua, J., Xioing, Z., Lowey, J., Suh, E., Dougherty, E.R., 2005. Optimal number of features as a function of sample size for various classification rules. Bioinformatics 21, 1509–1515.

Janssen, J.H., van den Broek, E.L., Westerink, J.H., 2009. Personalized affective music player. In: 3rd International Conference on Affective Computing and Intelligent Interaction. Amsterdam, Netherlands, pp. 1–6.

Kapoor, A., Burleson, W., Picard, R.W., 2007. Automatic prediction of frustration. International Journal of Human–Computer Studies 65, 724–736.

Katsis, C.D., Ganiatsas, G., Fotiadis, D.I., 2006. An integrated telemedicine platform for the assessment of affective physiological states. Diagnostic Pathology 1, 16.

Katsis, C.D., Katertsidis, N., Ganiatsas, G., Fotiadis, D.I., 2008. Toward emotion recognition in car-racing drivers: a biosignal processing approach. IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans 38, 502–512.

Kim, J., 2007. Bimodal emotion recognition using speech and physiological changes. In: Plutchik, R., Kellerman, H. (Eds.), Robust Speech Recognition and Understanding. I-Tech Education and Publishing, Vienna.

Kim, J., Andre, 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.

King, R.D., Feng, C., Shutherland, A., 1995. StatLog: comparison of classification algorithms on large real-world problems. Applied Artificial Intelligence 9, 259–287.

Koenig, A., Novak, D., Omlin, X., Pulfer, M., Perreault, E., Zimmerli, L., et al., 2011. Real-time closed-loop control of cognitive load in neurological patients during robot-assisted gait training. IEEE Transactions on Neural Systems and Rehabilitation Engineering 19, 453–464.

Kolodyazhniy, V., Kreibig, S.D., Gross, J.J., Roth, W.T., Wilhelm, F.H., 2011. An affective computing approach to physiological emotion specificity: toward subject-independent and stimulus-independent classification of film-induced emotions. Psychophysiology 48, 908–922.

Kreibig, S.D., 2010. Autonomic nervous system activity in emotion: a review. Biological Psychology 84, 394–421.

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.

Kuikkaniemi, K., Laitinen, T., Turpeinen, M., Saari, T., Kosunen, I., Ravaja, N., 2010. The influence of implicit and explicit biofeedback in first-person shooter games. In: Proceedings of the 28th ACM International Conference on Human Factors in Computing Systems (CHI 2010). ACM, New York.

Kulić, D., Croft, E.A., 2007. Affective state estimation for human–robot interaction. IEEE Transactions on Robotics 23, 991–1000.

Kumar, M., Weippert, M., Vilbrandt, R., Kreuzfeld, S., Stoll, R., 2007. Fuzzy evaluation of heart rate signals for mental stress assessment. IEEE Transactions on Fuzzy Systems 15, 791–808.

Leon, E., Clarke, G., Callaghan, V., Sepulveda, F., 2007. A user-independent real-time emotion recognition system for software agents in domestic environments. Engineering Applications of Artificial Intelligence 20, 337–345.

Leon, E., Clarke, G., Callaghan, V., Sepulveda, F., 2004. Real-time detection of emotional changes for inhabited environments. Computers & Graphics 28, 635–642.

Levillain, F., Orero, J. O., Rifqi, M., Bouchon-Meunier, B., 2010. Characterizing players experience from physiological signals using fuzzy decision trees. In: 2010 IEEE Conference on Computational Intelligence and Games, pp. 75–82.

Liao, W., Zhang, W., Zhu, Z., Ji, Q., 2005. A decision theoretic model for stress recognition and user assistance. In: Twentieth National Conference on Artificial Intelligence (AAAI), pp. 529–534.

Lisetti, C., Nasoz, F., LeRouge, C., Ozyer, O., Alvarez, K., 2003. Developing multimodal intelligent affective interfaces for tele-home health care. International Journal of Human–Computer Studies 59, 245–255.

Lisetti, C., Nasoz, F., 2004. Using noninvasive wearable computers to recognize human emotions from physiological signals. EURASIP Journal on Applied Signal Processing 11, 1672–1687.

Liu, C., Conn, K., Sarkar, N., Stone, W., 2008. Online affect detection and robot behavior adaptation for intervention of children with autism. IEEE Transactions on Robotics 24, 883–896.

Liu, C., Agrawal, P., Sarkar, N., Chen, S., 2009. Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. International Journal of Human–Computer Interaction 25, 506–529.

Liu, H., Motoda, H. (Eds.), 2007. Computational Methods of Feature Selection. Chapman and Hall/CRC.

Liu, H., Hu, J., Rauterberg, M., 2010. iHeartrate: a heart rate controlled in-flight music recommendation system. In: Proceedings of Measuring Behavior 2010, Eindhoven, Netherlands, pp. 265–268.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B., 2007. A review of classification algorithms for EEG-based brain–computer interfaces. Journal of Neural Engineering 4, R1–R13.

Lykken, D.T., Rose, R., Luther, B., Maley, M., 1966. Correcting psychophysiological measures for individual differences in range. Psychological Bulletin 66, 481–484.

Mandryk, R.L., Atkins, S.M., 2007. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. International Journal of Human–Computer Studies 65, 329–347.

Mehrabian, A., 1996. Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. Current Psychology 14, 261–292.

Mihelj, M., Novak, D., Munih, M., 2009. Emotion-aware system for upper extremity rehabilitation. In: Virtual Rehabilitation 2009, Haifa, Israel, pp. 173–178.

Mohammad, Y., Nishida, T., 2010. Using physiological signals to detect natural interactive behavior. Applied Intelligence 33, 79–92.

Müller, M.E., 2006. Why some emotional states are easier to be recognized than others: a thorough data analysis and a very accurate rough set classifier. In: 2006 IEEE International Conference on Systems, Man and Cybernetics, pp. 1624–1629.

Nacke, L.E., Kalyn, M., Lough, C., Mandryk, R. L., 2011. Biofeedback game design: using direct and indirect physiological control to enhance game interaction. In: Proceedings of the 29th ACM International Conference on Human Factors in Computing Systems (CHI 2011).

Nasoz, F., Alvarez, K., Lisetti, C.L., Finkelstein, N., 2004. Emotion recognition from physiological signals using wireless sensors for presence technologies. International Journal of Cognition, Technology and Work 6, 4–14.

Nasoz, F., Lisetti, C.L., Vasilakos, A.V., 2010. Affectively intelligent and adaptive car interfaces. Information Sciences 180, 3817–3836.

Nenonen, V., Lindblad, A., Häkkinen, V., Laitinen, T., Jouhtio, M., Hämäläinen, P., 2007. Using heart rate to control an interactive game. In: Proceedings of the 25th ACM International Conference on Human Factors in Computing Systems (CHI 2007), San Jose, CA, pp. 853–856.

Novak, D., Ziherl, J., Olenšek, A., Milavec, M., Podobnik, J., Mihelj, M., et al., 2010. Psychophysiological responses to robotic rehabilitation tasks in stroke. IEEE Transactions on Neural Systems and Rehabilitation Engineering 18, 351–361.

Novak, D., Mihelj, M., Ziherl, J., Olenšek, A., Munih, M., 2011. Psychophysiological measurements in a biocooperative feedback loop for upper extremity rehabilitation. IEEE Transactions on Neural Systems and Rehabilitation Engineering 19, 400–410.

Oliver, N., Kreger-Stickles, L., 2006. PAPA: physiology and purpose-aware automatic playlist generation. In: Proceedings of 7th International Conference on Music, Information Retrieval, pp. 250–253.

Pastor-Sanz, L., Vera-Munoz, C., Fico, G., Arredondo, M.T., 2008. Clinical validation of a wearable system for emotional recognition based on biosignals. Journal of Telemedicine and Telecare 14, 152–154.

Peter, C., Herbon, A., 2006. Emotion representation and physiology assignments in digital systems. Interacting with Computers 18, 139–170.

Picard, R.W., 1997. Affective Computing. MIT Press, Cambridge, MA.

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.

Plarre, K., Raij, A., Hossain, S. M., Ali, A. A., Nakajima, M., al'Absi, M. et al., 2011. Continuous inference of psychological stress from sensory measurements collected in the natural environment. In: Proceedings of the 10th International Conference on Information Processing in Sensor, Networks, pp. 97–108.

Polikar, R., 2006. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine 6, 21–45.

Pour, P.A., Hussain, M.S., AlZoubi, O., D'Mello, S., Calvo, R.A., 2010. The impact of system feedback on learners' affective and physiological states. Intelligent Tutoring Systems, 264–273.

Prinzel III, L.J., Parasuraman, R., Freeman, F.G., Scerbo, M.W., Mikulka, P.J., Pope, A.T., 2003. Three Experiments Examining the Use of Electroencephalogram, Event-Related Potentials, and Heart-Rate Variability for Real-Time Human-Centered Adaptive Automation Design. NASA Langley Research Center, Virginia.

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.

Rani, P., Liu, C., Sarkar, N., Vanman, E., 2006. An empirical study of machine learning techniques for affect recognition in human–robot interaction. Pattern Analysis and Applications 9, 58–69.

Rani, P., Sarkar, N., Adams, J., 2007. Anxiety-based affective communication for implicit human–machine interaction. Advanced Engineering Informatics 21, 323–334.

Rani, P., Sarkar, N., Smith, C.A., Kirby, L.D., 2004. Anxiety detecting robotic system – towards implicit human–robot collaboration. Robotica 22, 85–95.

Rani, P., Sims, J., Brackin, R., Sarkar, N., 2002. Online stress detection using psychophysiological signals for implicit human–robot cooperation. Robotica 20, 673–685.

Rigas, G., Katsis, C.D., Ganiatsas, G., Fotiadis, D.I., 2007. A user independent, biosignal based, emotion recognition method. In: User Modeling 2007. Springer-Verlag, Berlin, Heidelberg, pp. 314–318.

Rigas, G., Goletsis, Y., Bougia, P., Fotiadis, D.I., 2011. Towards driver's state recognition on real driving conditions. International Journal of Vehicular Technology Article id 617210.

Ritter, W., 2011. Benefits of subliminal feedback loops in human–computer interaction. Advances in Human–Computer Interaction (article ID 346492).

Russell, J.A., 1980. A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178.

Sakr, G.E., Elhajj, I.H., Hujier, H.A., 2010. Support vector machines to define and detect agitation transition. IEEE Transactions on Affective Computing 1, 98–108.

Scheirer, J., Fernandez, R., Klein, J., Picard, R.W., 2002. Frustrating the user on purpose: a step toward building an affective computer. Interacting with Computers 14, 93–118.

Schölkopf, B., Smola, A.J., 2001. Learning with Kernels. MIT Press.

Schwerdtfeger, A., 2004. Predicting autonomic reactivity to public speaking: don't get fixed on self-report data! International Journal of Psychophysiology 52, 217–224.

Setz, C., Schumm, J., Lorenz, C., Arnrich, B., Tröster, G., 2009. Combining worthless sensor data. In: Measuring Mobile Emotions Workshop at MobileHCI.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, Ehlert, U., 2010. Discriminating stress from cognitive load using a wearable EDA device. IEEE Transactions on Information Technology in Biomedicine 14, 410–417.

Shen, L., Wang, M., Shen, R., 2009. Affective e-learning: using "Emotional" data to improve learning in pervasive learning environment. Educational Technology & Society 12, 176–189.

Shye, A., Pan, Y., Scholbrock, B., Miller, J.S., Memik, G., Dinda, P.A. et al., 2008. Power to the people: leveraging human physiological traits to control microprocessor frequency. In: 41st IEEE/ACM International Symposium on Microarchitecture, pp. 188–199.

Stephens, C.L., Christie, I.C., Friedman, B.H., 2010. Autonomic specificity of basic emotions: evidence from pattern classification and cluster analysis. Biological Psychology 84, 463–473.

Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, 1996. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. European Heart Journal 17, 354–381.

Ting, C., Mahfouf, M., Nassef, A., Linkens, D.A., Panoutsos, G., Nickel, P., et al., 2010. Real-time adaptive automation system based on identification of operator functional state in simulated process control operations. IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans 40, 251–262.

Tognetti, S., Garbarino, M., Bonanno, A. T., Matteucci, M., Bonarini, A., 2010. Enjoyment recognition from physiological data in a car racing game. In: Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments (AFFINE '10), pp. 3–8.

Toups, Z.O., Graeber, R., Kerne, A., Tassinary, L., Berry, S., Overby, K. et al., 2006. A design for using physiological signals to affect team game play. In: Foundations of Augmented Cognition, pp. 134–139.

van den Broek, E.L., Lisy, V., Janssen, J.H., Westerink, J.H., Schut, M.H., Tuinenbreijer, K., 2010. Affective man–machine interface: unveiling human emotions through biosignals. In: Biomedical Engineering Systems and Technologies: BIOSTEC2009. Springer-Verlag, Berlin, Heidelberg, pp. 21–47.

Wagner, J., Kim, J., André, E., 2005. From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification. In: 2005 IEEE International Conference on Multimedia and Expo. Amsterdam, Netherlands, pp. 940–943.

Wang, H., Prendinger, H., Igarashi, T., 2004. Communicating emotions in online chat using physiological sensors and animated text. In: CHI 2004 Extended Abstracts on Human Factors. ACM, New York, pp. 1171–1174.

Wilson, G.F., Russell, C.A., 2003a. Operator functional state classification using multiple psychophysiological features in an air traffic control task. Human Factors 45, 381–389.

Wilson, G.F., Russell, C.A., 2003b. Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. Human Factors 45, 635–643.

Wilson, G.F., Russell, C.A., 2007. Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. Human Factors 49, 1005–1018.

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., Picard, R., 2009. Affect-aware tutors: recognising and responding to student affect. International Journal of Learning Technology 4, 129–164.

Wu, D., Courtney, C.G., Lance, B.J., Narayanan, S.S., Dawson, M.E., Oie, K.S., Parsons, T.D., 2011. Optimal arousal identification and classification for affective computing using physiological signals: virtual reality Stroop task. IEEE Transactions on Affective Computing 1, 109–118.

Yannakakis, G.N., Hallam, J., 2008. Entertainment modeling through physiology in physical play. International Journal of Human–Computer Studies 66, 741–755.

Yannakakis, G.N., Hallam, J., Lund, H.H., 2008. Entertainment capture through heart rate activity in physical interactive playgrounds. User Modeling and User-Adapted Interaction 18, 207–243.

Yannakakis, G.N., Martinez, H.P., Jhala, A., 2010. Towards affective camera control in games. User Modeling and User-Adapted Interaction 20, 313–340.

Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S., 2009. A survey of affect recognition methods: audio, visual and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39–58.

Zhai, J., Barreto, A., 2006. Stress detection in computer users through non-invasive monitoring of physiological signals. Biomedical Sciences Instrumentation 42, 495–500.