

Exploring the Biocybernetic loop:
Classifying Psychophysiological Responses
to Cultural Artefacts using Physiological
Computing

Alexander John Karran

A Thesis submitted in partial fulfilment of the requirements of
Liverpool John Moores University for a degree of Doctor of
Philosophy

April 2014

Acknowledgements

I would like to thank my director of studies Professor Stephen Fairclough for providing the opportunity to complete the research and for mentoring me through its various stages

Secondary supervisor: Paul Fergus for advice and support

Project ARtSENSE (EU FP7 project No.270318) and Liverpool John Moores University for sponsoring this research

Heather Wake for being a constant when all else was change

I would also like to thank my office colleagues Ute Kreplin and Chris Burns, without whom life would have been dull

Abstract

Background

The aim of this research project was to provide a bio-sensing component for a real-time adaptive technology in the context of cultural heritage. The proposed system was designed to infer the interest or intention of the user and to augment elements of the cultural heritage experience interactively through implicit interaction. Implicit interaction in this context is the process whereby the system observes the user while they interact with artefacts; recording psychophysiological responses to cultural heritage artefacts or materials and acting upon these responses to drive adaptations in content in real-time.

Real-time biocybernetic control is the central component of physiological computing wherein physiological data are converted into a control input for a technological system. At its core the bio-sensing component is a biocybernetic control loop that utilises an inference of user interest as its primary driver. A biocybernetic loop is composed of four main stages: inference, classification, adaptation and interaction. The programme of research described in this thesis is concerned primarily with exploration of the inference and classification elements of the biocybernetic loop but also encompasses an element of adaptation and interaction. These elements are explored first through literature review and discussion (presented in chapters 1-5) and then through experimental studies (presented in chapters 7-11).

Experimental work

The goal of the experimental work was to explore the issues involved in constructing a biocybernetic loop and build a real-time biocybernetic control loop to work in a cultural heritage context. With the thesis was concerned with several key questions:

- How to define an appropriate psychological construct? I.e. exploring the concept of interest in an applied context.
- Which physiological measurements best capture interest? i.e. investigating measures of the autonomic and central nervous system using ambulatory sensors.
- What classification strategy is best-suited the purpose and requirements of the biocybernetic loop? I.e. examining subject independent and subject dependent approaches in conjunction with machine learning algorithms.
- How should classifiers be trained for use in a real-time application? I.e. exploring how best to aggregate classifier training data for use in a real-time system
- How do users perceive system accuracy? i.e. exploring the relationship between mathematical accuracy and a users' perception

A programme of five studies was completed to explore these issues in depth

- Study one: explored a psychophysiological inference (as autonomic activation) using a range of autonomic measures and classification algorithms under laboratory conditions. It was concluded that the support vector machine classifier was the more accurate of the classification algorithms tested and that indices of autonomic activation are best measured and classified using a subject dependent approach (see chapter 7).
- Study two: The aim of this study was to investigate cross-session classification of autonomic activation wherein a support vector machine classifier was trained on session one and applied to data from session two. It was concluded that indices of autonomic activation are best classified within the same session and that classifier training should occur on the same day for subject dependent classification (see chapter 8).
- Study three: This study was concerned with the classification of multiple psychophysiological measures recorded using ambulatory sensor apparatus in response to audio material in a cultural heritage setting. It was concluded that combining measures of autonomic and central nervous system activation resulted in a high classification accuracies of when inferring participant interest (see chapter 9)
- Study four: represents a replication of study three, using multiple sources of media (audio, video, still image and combinations thereof) in a cultural heritage setting. From the results it was concluded that subject dependent approach to classification and classifier training was more accurate when compared to subject independent. In addition, combining measures of autonomic and central nervous system activation provided the highest classification accuracies (see chapter 10).
- Study five: represents a culmination of the previous studies to create a real-time classification protocol to capture high or low interest in response to video material. The results of this study showed that classifying the inference of interest in real-time was stable across the experimental session and that user perception (due to human factors) of system accuracy varied across the session starting with a perception of high system accuracy, then perceiving a drop in accuracy, and by the end of the session perceiving system accuracy to be higher than the initial perception (see chapter 11).

Main Conclusions

The results from the experimental work are discussed (see chapter 12) in the context of a functioning biocybernetic control loop, identifying issues concerned with the psychological construct of interest, measuring psychophysiological responses and classifying psychophysiological states in real-time. The following guidelines represent a summary of the findings from the classification analyses performed in this thesis:

- Normalisation of psychophysiological data is not indicated for use in subject dependent systems and presents increased computational cost with no increased benefit to classification accuracies when compared to absolute values
- Classifiers perform poorly when tasked with classifying data across repeated sessions and repeated exposures to same stimuli
- In the case of subject dependent applications, classifiers should be trained for each session using a combination of psychophysiological data and subjective assessment for training data captured during that session
- When designing systems to integrate real-time machine learning classification into biocybernetic control loops, there is a trade-off between the time required for classifier training, accuracy of the resulting classifier and speed of deployment
- Classifiers can more accurately reflect a user's appraisal of psychophysiological state when trained repeatedly during the same session, resulting in more accurate classifications and potentiating an increase in user acceptance or trust towards the system
- Users are likely to overestimate the accuracy of the system

Table of Contents

Table of Contents.....	6
1. Introduction.....	12
2. The inference model: Psychological Constructs, Measurement and Inference.....	16
2.1. The Psychological Construct.....	16
2.2. Measuring the psychophysiological response.....	19
2.3. Creating the Psychophysiological inference.....	22
3. Classification.....	24
3.1. K-Nearest Neighbour.....	27
3.2. Decision Trees.....	29
3.3. Support Vector Machine.....	32
3.4. Artificial Neural Networks.....	34
4. Adaptation and interaction at the interface.....	37
5. Challenges for real-time psychophysiological state classification.....	41
6. Outline of experimental studies.....	43
7. Study One Classification of Psychophysiological Activation States using K Nearest Neighbour (KNN).....	45
7.1. Abstract.....	45
7.2. Introduction.....	46
7.3. Methods.....	47
7.3.1. Participants.....	47
7.3.2. Design.....	47
7.3.3. Apparatus.....	47
7.3.4. Experimental Measures.....	48
7.3.5. Procedure.....	49
7.4. Analysis Framework.....	51
7.5. Results.....	53
7.5.1. Comparing KNN and other classification algorithms.....	56
7.6. Conclusion.....	57
8. Study Two: Test Retest classification of autonomic activation using Support Vector Machine (SVM).....	59
8.1. Abstract.....	59

8.2.	Introduction	59
8.3.	Methods	61
8.3.1.	Participants	61
8.3.2.	Experimental design	61
8.3.3.	Experimental Measures	61
8.3.4.	Experimental Material	62
8.3.5.	Procedures	64
8.3.6.	Support vector machine parameterisation & accuracy estimation	65
8.4.	Analysis	67
8.4.1.	Feature Extraction	67
8.4.2.	Trial 1 Raw Physiological Feature Analysis using Survey Labels	68
8.4.3.	Trial 2 Normalised feature analysis using Survey Labels	69
8.4.4.	Trial 3 Dimension reduction using principal component analysis	69
8.4.5.	Trial 4 Raw data analysis with associated subjective labels	70
8.5.	Results	70
8.5.1.	Trial 1 Classification of raw physiological features using survey labels	71
8.5.2.	Trial 2 Classification of normalised features using survey labels	71
8.5.3.	Trial 3 Effects of dimension reduction using principal component analysis	72
8.5.4.	Trial 4 Classification of features using subjective ratings	72
8.5.5.	Trial 5 Generalised models	73
8.6.	Discussion	74
8.7.	Conclusion	77
9.	Study 3: A Virtual Heritage installation	79
9.1.	Abstract	79
9.2.	Introduction	80
9.2.1.	The cultural heritage experience	81
9.2.2.	Conceptual model of interest	82
9.2.3.	Operationalising the model	82
9.2.4.	Study Goals	84
9.3.	Methodology	84
9.3.1.	Participants	84
9.3.2.	Experimental Design	84
9.3.3.	Apparatus and Experimental Measures	85
9.3.4.	Task definition	85
9.3.5.	Procedure	87
9.4.	Analysis	88
9.4.1.	Feature Derivatives	88

9.5.	Results	89
9.6.	Discussion.....	92
9.7.	Conclusion.....	94
10.	Study Four: Liverpool FACT Study – Classification of multimodal cultural heritage material	95
10.1.	Abstract.....	95
10.2.	Introduction	96
10.2.1.	The Interest as Binary - Interest as State (IBIS) Framework.....	97
10.2.2.	Psychophysiological Signal Processing Pipeline:.....	99
10.2.3.	Study Goals.....	101
10.3.	Methods	101
10.3.1.	Participants.....	101
10.3.2.	Experimental Design.....	101
10.3.3.	Experimental Measures.....	102
10.3.4.	Materials	102
10.3.5.	Procedures.....	103
10.4.	Analysis	104
10.4.1.	Feature Extraction.....	104
10.4.2.	Classification Trials	105
10.4.3.	Deriving Binary Labels from Likert Scales	106
10.5.	Results	107
10.5.1.	Trial 1 Composite classification using “interest” labels	107
10.5.2.	Trial 2 Component classification with individual response labels.....	108
10.5.3.	Trial 3 Generalised model using both “composite” and “component” labels	108
10.6.	Discussion.....	109
10.7.	Conclusion.....	112
11.	Study 5: classifying the interest state in real-time	113
11.1.	Abstract.....	113
11.2.	Introduction	114
11.3.	Classifying the Interest Response in Real-time	117
11.3.1.	The Interest as Binary - Interest as State (IBIS) Model.....	117
11.3.2.	Applied Real-Time Interest Classification Framework (ARTIC).....	119
11.4.	Study Goals	124
11.5.	Methods	124
11.5.1.	Participants.....	124
11.5.2.	Experimental Design.....	124
11.5.3.	Experimental measures	125

11.5.4.	Materials	125
11.5.5.	Procedures.....	126
11.6.	Analysis	128
11.6.1.	Feature Extraction.....	128
11.7.	Results	129
11.8.	Discussion.....	133
11.9.	Conclusion.....	137
12.	Discussion	138
12.1.	The inference model	139
12.2.	Classification	143
12.3.	Interaction and Adaptation	151
12.4.	Limitations.....	155
12.5.	Future work	158
13.	Conclusion	160
14.	References.....	161

Table of Figures

Figure 1.1. Representation of the human-machine biocybernetic loop.....	14
Figure 2.1 The circumplex model of emotion showing both axes (Russell 1980).....	17
Figure 3.1 Example k-nearest neighbour query; 3 nearest neighbours.....	28
Figure 3.2 Simple regression decision tree for a biocybernetic loop.....	30
Figure 3.3 Example SVM separating hyperplane.....	32
Figure 3.4 Example Neural Network with one hidden layer.....	35
Figure 7.1 Experimental procedure timeline.....	49
Figure 7.2 Experimental Analysis Framework.....	52
Figure 7.3 KNN Classifier accuracies 3 classes vs. 2 classes all analyses.....	54
Figure 7.4 KNN classification showing no clear differentiation between classes.....	55
Figure 8.1 Sequence of testing for SVM experiment (a) and (b) examples of low and high activation images (left to right).....	64
Figure 8.2 The stimulus presentation timeline.....	64
Figure 9.1 The still image used in the experiment with highlighted sections that were linked to the audio.....	86
Figure 9.2 Experimental stimulus timeline.....	87
Figure 9.3 Mean Classifier Recall Accuracy comparison generalised model vs. subject dependent model.....	91
Figure 10.1 The Interest as Binary, Interest as Scale (IBIS) classification framework.....	98
Figure 10.2 Meta Process Pipeline: feature extraction processing.....	100
Figure 10.3 The procedure and stimulus timeline.....	103
Figure 11.1 The revised IBIS Framework.....	118
Figure 11.2 The display video and subjective feedback process.....	120
Figure 11.3 The train classifier process.....	121
Figure 11.4 The Applied Real-Time Interest Classification Framework (ARTIC).....	122
Figure 11.5 Experimental “Wizard of Oz” interaction procedure.....	127
Figure 11.6 ROC Build 1.....	130
Figure 11.7 ROC Build 2.....	130
Figure 11.8 ROC Build 3.....	131
Figure 11.9 ROC Build 4.....	131
Figure 11.10 Mean system accuracy per build.....	132
Figure 11.11 Mean perceived accuracy per build.....	132
Figure 11.12 Trend of true positive classifications.....	132
Figure 11.13 Trend of false positive classifications.....	132
Figure 11.14 Trend of true negative classifications.....	133

Figure 11.15 Trend of false negative classifications.....	133
Figure 12.1 A Cultural Heritage “digital curator” Framework	153

1. Introduction

A central feature of the early 21st century has been the acceleration of technological progress in the fields of ubiquitous computing (such as smart phones and tablet computers) and wearable technologies (such as smart watches). These innovations have set in motion a trend of human-technological integration which is set to accelerate in the coming decades, with the further integration of computing technology into most aspects of daily life (Kurzweil, 2005). This trend encompasses a range of fields including: self-health monitoring, entertainment, communication and biological enhancement.

These new technologies and supporting infrastructures are becoming increasingly complicated; however there has not been a concomitant increase in support or usability for the user of these systems. The lack of support for users of increasingly complex systems can lead to an asymmetry of communication or purpose between the user and the system, asymmetry in this context is an artifact of poor human computer interface (HCI) design that implies that the system has been designed with optimal usability and that user needs past the point of operation are for the most part are an irrelevancy. The more advanced the system, the greater the chance that communication between the user and the system will become more asymmetrical (Norman 2007), rather than a transparent communication where interaction and purpose is one and the same thing. Assimilating these technologies will require novel forms of human computer interaction (HCI), which are task appropriate and enhance the user experience. Most crucially, the actions and assessment from the system should be transparent to the user.

A significant issue is that the technology or system remains largely unaware of and unaffected by the affective “state” of the user, their goals or the environment in which they work, whereas the system can convey a disproportionate amount of information about its own state to the user. The majority of today’s computing systems utilise keyboards, mice and screens to enable a dialogue with the user and this human-computer interaction relationship which has remained largely unchanged since the 1970s. Under the traditional interaction model of mouse and keyboard, it is the user that adapts to the system, such that the computer issues commands (i.e. “to do this task - first complete this task” and so on.) In this example the system makes no attempt to adapt to the needs of the user. To counter this asymmetry, researchers and practitioners in HCI are seeking to develop new paradigms and techniques to enable greater interactivity between users and systems.

In the fields of physiological and affective computing, researchers have considered how systems might adapt to users based upon situational needs and psychophysiological state. The goal of these approaches is to devise computer systems that respond in a logical, considered and timely fashion

to real-time changes in a user's cognitive state (e.g. workload or inattention), affective state (e.g. frustration) or motivations, as represented by psychophysiology. A core element of this approach is to open a channel of implicit and symmetrical communication between a user and the computer system, by granting the system access to a representation of the psychological status of the user (Serbedzija & Fairclough, 2009), and this channel of communication is achieved by monitoring, analysing and responding to covert psychophysiological activity from the user in real-time (Fairclough, 2009). Systems that are designed in this way can potentially promote greater performance efficiency (by monitoring the cognitive workload of the user) or maximise learning and information retention (by monitoring the affective state or motivations of the user).

The mechanism of action which initiates the communication then manages and transforms the psychophysiological data into a control signal useable by systems, is known as a biocybernetic loop (Pope, Bogart, & Bartolome, 1995). Typically a biocybernetic loop is designed to monitor or promote specific psychological states in the user of computer or adaptive automated systems (in which the level of automation is dependent on the state of the operator). These states often represent positive or negative control loops. In a negative control loop (such as that designed for NASA by Pope et al, 1995) the goal of this control system is to avoid operator states that are detrimental to task performance, such as extreme cognitive workload. The goal of a positive control loop is to promote operator states to achieve a system or user assignment, such as improving knowledge transfer (in a learning environment) or increasing the entertainment value of an activity (in a leisure-focused environment). In order to influence the psychological state of the user effectively, biocybernetic loops must be designed with a degree of autonomy (Serbedzija & Fairclough, 2009).

The creation of a biocybernetic loop is a multidisciplinary task involving elements from psychophysiology, computer science and human factors/HCI and constructed in a series of stages (see Figure 1.1); each stage is reliant upon the other for data and pre-processing and when taken as a whole becomes an analysis framework which reinforces the behaviour and purpose of both the system and the user. Fairclough & Gilleade (2012) described the four stages that typically form a biocybernetic loop.

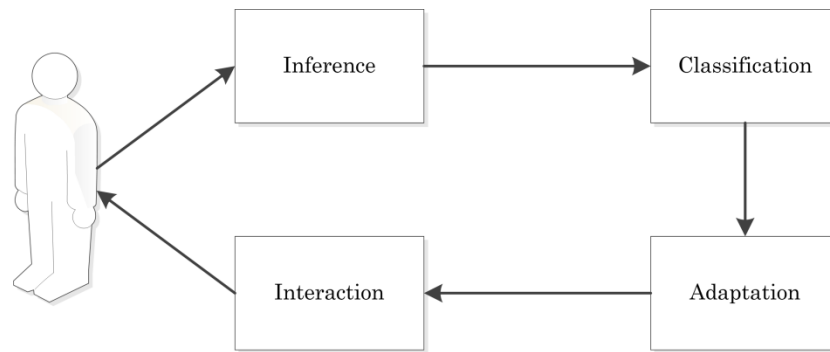


Figure 1.1. Representation of the human-machine biocybernetic loop

1. Inference

This stage is concerned with making a link between the target psychological state and a physiological measure. A psychophysiological construct is created that best represents the target psychological state (such as a state of high cognitive workload) and physiological measures are selected which provide the most valid operationalisation of that psychological state. The choice of sensor technology and signal processing techniques are crucial for this stage of the loop, which must be appropriate for application in the field and provide high signal fidelity. The selection of features of the inference model is central to the effectiveness of the loop. If the physiological measures do not capture the psychological construct with sufficient sensitivity and reliability then the inference model does not provide a clear link between the user state and system operation.

2. Classification

Classification concerns the identification of the psychophysiological state in real-time or near real-time. It is important that information passed from this stage be timely if the loop is to function dynamically. The choice of classification algorithm is crucial at this point. The classifier must be capable of processing and categorising information in both an accurate and timely manner. The cost of misclassification of user responses must be considered carefully as ultimately the classifier feeds forward judgements into the adaptation engine and thus shapes the efficacy of system adaptation in response to user behaviour.

3. Adaptation

At this stage the psychophysiological response has previously been measured and classified. The results from the classification are then used to inform what form of adaptation is to be used at the interface. Thus adaptation is concerned with employing the governing rule set or purpose of the loop, that is, what actions should be taken at the interface in response to classification judgements about the user's state.

4. Interaction

The process of adaptation is given form at the interface between the user and the system. The form of adaptation will shape user perceptions of system efficacy from the psychophysiological inference to classification and adaptation. The form of adaptation must be carefully designed to provide timely and relevant action or feedback at the interface in order to engender the trust of the user.

The research presented within the thesis is concerned with the construction of a biocybernetic loop for the purpose of enhancing cultural heritage experiences. Experimental work will be presented which details the stages completed to build the loop: creation of psychophysiological inference suitable for cultural heritage; classification of the psychophysiological response and output of the classification judgement for use in an adaptive system.

To examine the issues involved with creating a biocybernetic loop in more detail, the literature has been reviewed with a focus on the core processes involved in creating the biocybernetic loop. In the following text, the way in which psychological constructs, physiological measurement and machine learning classification as separate elements interact to form the basis of a biocybernetic control loop will be discussed. The theoretical psychological basis that informs the experimental work presented in this thesis will be discussed, principally describing how an inference of psychological state is created and operationalised using a psychological construct and physiological measurements and the scope in which the inference is valid. A section will describe machine learning classification in the context of physiological computing, the automatic detection of psychophysiological states and adaptive biocybernetic control. The issues surrounding adaptive biocybernetic control and how users interact with such systems will also be discussed. The challenges involved with applying adaptive biocybernetic control in a real-time context will be defined before outlining the planned research project to explore and investigate these issues.

2. The inference model: Psychological Constructs, Measurement and Inference

In this chapter how psychological theory and physiological methods converge to create a psychophysiological inference will be discussed, from initial theoretical foundation, to measuring the psychophysiological response and creating the inference.

2.1. The Psychological Construct

One of the main issues with the use of psychological constructs within a bio-cybernetic loop, involves the identification and definition of a psychological state (e.g. emotion, motivation, cognition) to drive the biocybernetic loop. When creating a loop for any task involving human operators (such as a cultural heritage experience), consideration must be given to the motivations or emotional state of the user as the majority of user interactions are goal driven and imply a degree of motivation and emotional engagement. In a recent survey of emotion recognition for affective computing literature Calvo and D’Mello (2010) identified no clear definition of what constitutes the basis of psychophysiological phenomena such as emotion. They detailed six theoretical perspectives on the issue. For the purposes of this thesis, three of these perspectives (embodiment, neuroscience and core-affect) can be seen to offer a firm but flexible epistemological footing, from which psychophysiological responses and motivations can be inferred within the boundaries of a biocybernetic loop.

Embodiment sometimes referred to as the James-Lange (James 1894) theory of emotion, emphasises emotional experience as being “embodied” within human physiology. This theory states that emotions are both “felt” physiologically and manifested as changes in the sympathetic nervous system (SNS) which is part of the autonomic nervous system (ANS). Thus emotional experience is embodied in peripheral physiology. This position contrasts with the Cannon-Bard (Cannon 1927) theory where responses to emotional stimuli or events occur simultaneously in the brain and body. This minor but crucial distinction, places equal focus on the brain, specifically the thalamus (the part of the brain that deals with sensory and motor processing) and the autonomic nervous system. This suggests that emotions result when the thalamus signals other areas of the brain in response to a stimulus, resulting in a top-down physiological reaction. The theory of embodiment is important to the construction of a biocybernetic loop as it provides the theoretical grounding that psychophysiological reactivity can be measured and recorded by placing sensors on the body. Indeed, the approach of using sensors placed around the body to capture emotion-autonomic nervous system responses has met with much success, as detailed in Kreibig (2010).

The field of affective neuroscience offers a unique perspective towards understanding the general organisation of the brain and its relation to those physiological and cognitive processes underpinning emotional experience. This area of research concentrates on the development of techniques and methods to uncover the basic principles and underlying circuitry involved in attention, language, emotion and consciousness (Dalgleish et al. 2009, Panksepp 2004, Damasio 2003). This is achieved by mapping psychophysiological states onto neural circuitry using indices of brain activation, measured by a variety of techniques such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG) etc.

Core affect is a theory posited by Russell (1980, 2003) which argues that the different theories of emotion all examine different forms of emotional phenomena which may be reconciled within a two-dimensional space. This theory has its basis in the concept of “core affect” expressed as a circumplex model (see Figure 2.1), which places emotional experience upon a point in two dimensional space described as valence (pleasure – displeasure) and arousal (sleepy – activated).

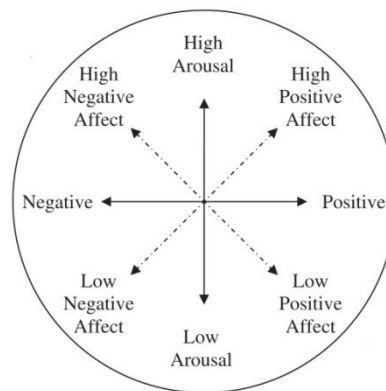


Figure 2.1 The circumplex model of emotion showing both axes (Russell 1980)

A key component of Russell’s theory emphasises the importance of context (e.g. the exact event that elicits the response) when separating emotional episodes, such as separating a particular episode of happiness from the base emotion category. While this appears to be a concept specific to emotion recognition, the same holds true for all psychophysiological states, in that the context of the event that elicits the response is a key component when attempting to measure and differentiate a psychophysiological state from background physiological activity; for example, measuring indices of brain activity to determine cognitive overload during a demanding mental task, while performing routine manual operations. Russell’s circumplex metaphor becomes a useful tool when constructing a biocybernetic loop, the model allows the heterogeneous elements that comprise psychological phenomena to be unified as a set of independent yet loosely coupled components, such as appraisals, physiological responses, expressions etc. and visualised upon a scale within n

dimensional space, providing a basis from which ad-hoc models of psychophysiological phenomena can be created for specific contexts, such as measuring cognitive workload or a level of vigilance.

These three approaches become a road map for conceptualising and building a psychological construct which forms the basis around which a biocybernetic loop can be assembled. This roadmap provides a number of testable assumptions to be made:

1. Psychological phenomena are embodied physiologically
2. Psychological phenomena have neurological and autonomic nervous system correlates
3. Psychological phenomena can be placed upon a uni-dimensional or two-dimensional scale, such that magnitude changes in physiology can be measured and classified into scalar or binary states.

Evidence for assumptions 1 and 2 has been gathered by empirical work completed over 120 years, demonstrating an association between psychological states and physiological responses (see Cacioppo et al. (2007) for in-depth review). Measuring the psychophysiological response will be discussed in section 3.2. The third assumption refers to a conceptual psychological space and there is an issue to be addressed concerning the role of subjective judgements during the development of a biocybernetic loop and during the systems interaction with the user.

To build a psychological construct that captures the cultural heritage experience, an assumption of “interest” on behalf of the heritage visitor could be made to describe a cultural heritage experience. The concept of interest was described by Berlyne (1960) as an exploratory drive, defining “interest” as a psychophysiological state that fosters curiosity and the drive to explore the object or situation at hand. Furthermore, Berlyne posited that interest is experienced through increased arousal and sensation seeking, i.e. objects that inspire curiosity via novelty and emotional conflict. Silvia (2005; 2010; 2008) expanded upon this concept to incorporate a cognitive dimension, whereby interest is driven by stimulus complexity. However, these two theoretical approaches do not take into account the influence of emotional resonances within the cultural heritage experience. An object may create interest as a result of its novelty or complexity, but the emotional states that accompany curiosity or interest also represent an important component of the process. For example, a visitor may be repulsed by a painting which leads to a sudden interest in the object.

The experience of interest is followed by a sense of positive emotion derived from intellectual engagement; positive emotions experienced as a result of interest can therefore occur even during engagement with negative material (Hidi & Renninger, 2006).

In forming a psychological construct to describe a cultural heritage experience the assumptions made of a user's psychophysiological state may be:

- A level of cognitive engagement based on the novelty and complexity of an artefact,
- A level of physiological stimulation while engaged with the artefact,
- An emotional response (positive or negative) towards the artefact

2.2. Measuring the psychophysiological response

Measuring psychophysiological responses is typically achieved by placing sensor hardware around the body. Sensor types include (but are not limited to) the Electroencephalogram (EEG) which measures voltage fluctuations resulting from brain activity at the surface of the cortex; Electrocardiogram (ECG) used to for recording measures of heart muscle activity over time; Skin conductance, also known as galvanic skin response (GSR) or electrodermal activity which is a method of measuring changes in skin electro-conductivity due to variations in eccrine secretions; respiration (RSP) as measured using a thoracic band to capture the depth and rate of chest cavity expansion/contraction and facial electromyography (fEMG) commonly measured from the face which measures electrical activity produced by skeletal muscle.

There is a current trend in industry, driven by the quantified self-movement, e-health and mobile computing industry, to miniaturise sensor technologies and output the data gathered to handheld devices for self-monitoring. These devices include the iHealth (iHealth, 2014) suite of ambulatory products containing a pulse oximeter (to measure blood oxygen), an adhesive patch ECG (to measure heart rate) and a wearable blood pressure vest monitor pulse oximetry. Advances in EEG monitoring in the form of the "in-the-ear" EEG (Looney et al. 2012) for long term monitoring of brain activity preceding epileptic seizures, similarly the development of pulse rate measurement and classification (in an adaptive game) from the ear using pulse oximetry to help gamers maintain a level of calm and vigilance during gameplay (Matson, 2014) and a mass market heart rate monitoring watch (coupled with mobile phone for processing) produced by Samsung (Samsung Gear 2 2014.). However, research in the field normally utilises laboratory grade sensor equipment or high quality ambulatory sensors.

From the standpoint of the biocybernetic loop, choice of sensor is closely related to the rationale for the loop itself. In the case of cultural heritage, the rationale is to measure a level of "interest" consisting of indices of cognition (as attention), physiological arousal (as activation) and approach/avoidance (as valence). In this instance the most sensitive and robust measures appear to be, skin conductance (EDA) which is linearly correlated with levels of physiological arousal

(Dawson 2007), cardiovascular responses (ECG) such as heart rate (HR) which is also used as a metric for levels of physiological arousal (Anttonen & Surakka 2005) and measures of brain activity from electroencephalography (EEG) which have been used as indices of arousal, attention and cognitive work load in biocybernetic loops specifically for adaptive task automation (Scerbo et al. 2001).

Speaking purely in terms of a biocybernetic loop constructed around an inference of user “interest” towards cultural heritage artefacts, understanding the underlying neural pathways, and their connections to psychophysiological states, during cultural heritage experiences is therefore important to the development of a functional biocybernetic loop. The Inference of interest in this context concerns the creation of a one-to-many relationship in which two or more physiological elements or measures are associated with one psychological element or construct.

Cognitive engagement (as cognition) can be quantified using Electroencephalography (EEG), particularly using alpha waves which have been associated with changes in cognitive load, i.e. a higher cognitive load is indicative of greater cognitive engagement (Goldman et al, 2002). Furthermore, recent studies in the field of neuroaesthetics have used functional magnetic resonance imaging (fMRI), functional near infrared spectroscopy (fNIRS) and EEG to investigate the relationship between brain activity and cultural heritage experiences, in particular the perception of beauty and aesthetics (Nadal & Pearce, 2011). This research contends that the prefrontal cortex (PFC), in particular Brodman’s area (BA) 10 located in the dorsal PFC, plays an important part in the evaluation of artworks through attentional top-down feedback that is the interpretation of sensory processing through cognitive engagement with the stimuli (e.g. Cupchik et al, 2009; Vessel et al, 2012; see Hahn et al, 2006 for a review).

Moreover, it has also been noted that alpha activation in the PFC is reduced during aesthetic experiences¹, particularly during the judgment of beauty (Cela-Conde et al, 2011), making EEG an appropriate measure to encapsulate cognitive engagement in cultural heritage settings. Cognitive engagement can therefore be captured and quantified using spontaneous EEG measures of electrocortical activation in CH contexts. Additionally, the aspect of arousal or activation described by Berlyne (1960) and Russell (1980) can be captured through changes in the visitor’s psychophysiology. Thus, cognitive engagement can be quantified through changes in psychophysiology and brain activation. In addition it has been hypothesised that greater activation of the left hemisphere of the PFC is associated with positive emotions whereas greater activation of the right hemisphere is linked to negative emotions (see Coan & Allen, 2004 for a review), thus the

¹ EEG alpha activation has a converse relationship with brain activity (Goldman et al, 2002), i.e. higher alpha activity is associated with reduced brain activation.

emotional response (as valence) to cultural heritage artefacts could also be captured using spontaneous EEG measures of electrocortical activation.

The level of physiological stimulation (as activation) associated with the construct of interest can be captured via the level of skin conductance (SC) and supplemented by the measurement of heart rate (HR); SC is highly sensitive to sympathetic nervous system activity (Boucsein, 1992) and HR captures both sympathetic and parasympathetic components of the autonomic nervous system. Both SC and HR have been found to be appropriate measures to be used in CH environments (Tschacher et al, 2011).

Measuring the physiological response for the detection and categorisation of psychological states for use within biocybernetic loops presents a number of technical and ergonomic challenges. For example, for ethical or ergonomic reasons, methods of signal acquisition must be as non-intrusive and transparent to the user as possible (Sakr et al. 2010). This would impact on the availability of physiological signals and restrict the range of sensor hardware that can be applied for signal acquisition used to monitor signals from the central (CNS) and autonomic nervous system (ANS). Furthermore, Fairclough (2009) identifies sensitivity and diagnosticity as fundamental criteria when choosing measures of psychophysiological responses and the sensor hardware with which to capture them. These criteria arise from the rationale of the biocybernetic loop e.g. ascertain a level of interest in a cultural heritage exhibit. The success and effectiveness of the loop is dependent on the assumption that the psychophysiological measure (or array of measures) is an accurate and sensitive representation of the relevant psychological element or dimension.

Sensitivity refers to attributes within the physiological measure that have high temporal resolution and capability of differentiating multiple levels along a psychological dimension. Features such as noise and interference and the filters required to detect and reduce them can affect the sensitivity of physiological signal data. For example EDA and ECG signals are highly susceptible to movement artefacts; skin conductance level (a measure of EDA) is dependent on continuous contact with the skin; loose sensor contacts result in loss of signal and data. Diagnosticity refers to the ability of the measure to target the specifics of the psychological construct, while at the same time remaining unaffected by related influences. Thus, diagnosticity can be seen as a function of context, in that physiological measures may have many psychological effectors, for example blood pressure as an indicator of frustration (while learning) or a state of positive challenge (while playing games) (Fairclough 2009). In these instances, it is the context that provides the association between measure and psychological state and allows the scope of the link to be reduced and an inference to be defined.

The intrusiveness, sensitivity and diagnosticity of measures is inextricably linked to the form factor and type of sensor hardware used to measure the physiological response. These concerns have implications when making any inference between psychological state and physiological response, for the given context. Therefore, the biggest challenge when constructing a biocybernetic loop be it for physiological or affective computing applications is finding stable relationships (one-to-one or many to one) between physiological response and psychological state across different contexts and users.

The definition of what constitutes a valid psychophysiological inference should be treated with care. Psychophysiological inference does not provide a literal, isomorphic representation of a given thought, intention or emotion; but rather represents an operationalisation of internal states, the quality of which may vary from measure to measure, and between different states (Fairclough 2009). Therefore, the caveat that psychophysiology provides a less-than-perfect representation of internal states must be considered explicitly during the design and construction of a biocybernetic loop, and addressed by asking and answering the question: is the psychophysiological inference between the psychological states and physiological responses sufficiently sensitive and diagnostic to realise the query criteria that is the basis of the loop?

2.3. Creating the Psychophysiological inference

Creating psychophysiological inferences has its basis in reductionism. Reductionism holds that any complex system may be understood as the sum of its parts and the examination of its individual parts increases our understanding of the complex system. For example, a complex psychological state such as interest can be measured with reference to one component of its manifestation in the body, e.g. heart rate.

It is at this point in creating an inference between psychological state and physiological response where context becomes a key element. As previously discussed, the rationale of the biocybernetic loop provides the setting (e.g. to measure a level of interest and adapt information on this basis) and context provides the indication of state induction i.e. viewing a painting or artefact. Thus, psychophysiological inference takes place in the context of a specific psychological construct and a particular task - hence the process of making the inference is relative and must be grounded in an appropriate scenario that represents the operating conditions of the biocybernetic loop.

Cacioppo, Tassinary & Bernston (2007) proposed a general framework for psychophysiological inference, which includes rules of evidence and the limitations of psychophysiological inference. In this framework psychophysiological inference is separated into two domains; the psychological and

the physiological. Each domain is a separate entity consisting of the conceptual variables within the psychological domain (the psychological model) and empirical variables within the physiological domain (the physiological response). These domains are mapped onto each other forming a series of relationships, and it is the type of relationship that delineates the strength of the inference. Within the suggested framework five relationships are demarcated to represent the elements within each domain.

- A one-to-one relationship (i.e. one element in the psychological domain is associated with only one element within the physiological domain and vice versa).
- A many-to-one relationship (i.e. in which, two or more physiological elements are associated with one in the psychological domain).
- A one-to-many relationship (i.e. one element within the psychological domain is associated with many elements within the physiological domain).
- A many-to-many relationship (i.e. in which two or more psychological elements are associated with two or more elements within the physiological domain).
- A null relationship, in which, no association between elements within the psychological or physiological domains is possible.

This framework can be seen as a useful tool when creating experimental methodologies in which the relationship between psychological elements and physiological responses needs to be defined clearly as proof of concept, or in situations in which a relationship needs to be leveraged in an applied context, such as the dynamic basis of a biocybernetic loop. When establishing an inference between psychological states and physiological responses, the unique one-to-one relationship is clearly the gold standard for biocybernetic loops. However, instances of this form of isomorphic relationship are rare in the literature, relative to the other categories (Fairclough 2009). Such relationships are normally established within the laboratory and much research has been completed to define relationships between psychological states and physiological responses (see Kreibig 2010 for a review). In the case of a many-to-one relationship, a psychological state may be only be fully represented by a psychophysiological response pattern that incorporates several measures, and it is the aggregation of these responses that defines the strength of the inference. This pattern is inverted in a one-to-many relationship, in which one physiological response - e.g. systolic blood pressure - may increase in response to many psychological states – e.g. when a person is excited, frustrated or stressed (Cacioppo & Gardner, 1999). In the case of a many-to-many relationship, a mixture of psychological states may combine to exert multiple, overlapping paths of influence over many physiological responses.

The inference model has clear implications for the design and construction of biocybernetic loops regardless of application domain. The rationale for loop informs the creation of the psychological construct and this construct then becomes the framework into which physiological sensor technology is adapted for use. The type of sensor technology required should possess sufficient sensitivity (i.e. signal fidelity) and specificity (i.e. capture only what is under observation) to meet the requirements of the biocybernetic loop. The measures derived from the signal must be sufficiently diagnostic to allow a one-to-one or one-to-many inference link between psychological state and physiological response. The diagnosticity of measures can be maximised by using context to set the boundaries in which the psychophysiological inference is valid. A valid inference model is the fundament of the biocybernetic loop. Accurate classification of the psychophysiological response requires data that has a high degree of separation between psychological states. Without this separation, no adaptation will occur and the loop becomes ineffective. Thus, when constructing a biocybernetic loop an acceptable level of diagnosticity within the specific context of the task and the system must be established.

3. Classification

The process of classification plays a crucial role in the construction of a biocybernetic loop. Once the target psychophysiological construct has been operationalised and validated, classification provides the means by which physiological response data can be categorised as belonging to a specific inference class. This assessment can be subsequently made available to the adaptation component of the loop to be acted upon in real-time.

Currently the majority of research exploring methods for measuring and classifying psychophysiological states have been conducted in laboratory settings. These approaches combine multivariate physiological data, such as ECG, RSP, GSR, EEG fEMG etc. measured from the peripheral and central nervous system with statistical and machine learning classification algorithms (e.g. Picard 2003, Picard & Klein 2002, Regan & Atkins 2007, Wilhelm & Grossman 2010, Petrantonakis & Hadjileontiadis 2010). Examples include the application of fuzzy logic models to transform and transpose physiological signals into levels of arousal and valence (Mandryk & Atkins 2007), regression decision trees to determine affective states from ECG and GSR physiological data (Rani et al. 2005, Villon & Lisetti 2006). Picard and colleagues (2007) used K-means (nearest neighbour clustering algorithms) to partition and categorise physiological signals as affective states for use with adaptive computing systems. Other research utilises more advanced algorithmic techniques, such as Support Vector Machines, to partition incoming signals to detect levels of physiological activation as a measure of agitation transition for monitoring the

onset of epileptic episodes (Sakr et al. 2010) and Neural Networks to monitor the ANS for variance when presented with emotional stimuli (Lee et al. 2006).

Evaluating the performance of classifiers for physiological computing applications, involves gathering and aggregating physiological data (feature vectors) from individuals, in groups or singly, to create training datasets (feature sets). These datasets represent a matrix of “supervised” data in which the each row of physiological values represents an observation. Each observation then has a label associated, which denotes that observation’s (predictor) class. Classifiers are subsequently trained using these data and used as the basis for comparison against new instances of physiological data.

There are a number of issues to be addressed during the creation of training data for supervised learning algorithms. As supervised learning methods are completely dependent on the training dataset to create predictive models, it is imperative that the training dataset contains psychophysiological data that is completely representative of the psychological construct under examination. Therefore, the psychological construct under examination should be correctly induced as an exemplar and recorded before training of the classifier occurs or the predictive model will be ineffective. A second issue concerns the features of psychophysiological data itself, which exhibit intra and inter individual variability related to demographic factors such as age, state of health, gender and other factors (Novak et al. 2012). A subject dependent classifier is one that is trained using the psychophysiological data (both features and class labels) from a single individual, a subject independent classifier is one trained using psychophysiological data from n individuals associated with class labels from each individual or from some other source such as survey. The choice of subject-independent (nomothetic) or subject-dependent (idiodynamic) feature data also has implications for how the data is processed (nomothetic feature data may require normalisation to maintain similar value ranges) and ultimately for how a classifier is trained and applied i.e. to generalise predictions across a group of users or trained to specific individuals for specific contexts. Put simply, should the designer construct a biocybernetic loop that is trained to the individual or attempt to aggregate data from multiple users to create a generic training set that can generalise across a population?

In a recent survey of methods for data fusion, classification and adaptation using autonomic nervous system responses for use in physiological computing, Novak and colleagues (2012) present the results from 102 completed studies that applied machine learning algorithms to physiological data in three categories of task: classifying basic emotions, classifying responses in arousal valence space and classifying single specific responses (either as binary or multiple levels). The machine learning algorithms reported in the survey included: k-nearest neighbour (KNN); Naïve Bayes and Bayes networks (BN); linear discriminant analysis (LDA); support vector machine (SVM); regression decision trees (RDT); and artificial neural networks (NN).

The data shown in Table 3-1 displays the averaged accuracy for each classifier (averaged over all studies using highest reported accuracy for each study), the number of studies, the number of classes and classification methodology: Subject dependent (individual); subject independent (across a group); or unknown (where no methodology was reported).

Type	# Studies	Avg. # Classes	D	I	Unk.	Avg. Acc %
KNN	16	5	5	4	7	72.58
BN	10	4	5	1	4	77.04
LDA	25	3	10	8	7	70.53
SVM	22	3	10	6	6	80.33
RDT	12	3	5	1	6	84.85
NN	17	3	9	3	5	80.70
Totals	102		44	23	35	

Table 3-1 Averaged number of classes and accuracies across all studies: D = subject dependent; I = subject independent; Unk = unknown

In these results the data indicates that LDA has been used in the highest number of studies (25), assigned with classifying on average 3 classes of psychophysiological response ranging from basic emotion discrimination to workload level. The popularity of LDA is based upon its ease of use and transparency (the contribution of each physiological feature towards discrimination between the classes can be seen). However, LDA has the lowest average performance (accuracy 70.53%) of all the surveyed classifiers followed by KNN (72.58%) and BN (77.04%) shown in Table 3-1; however, the low number of studies reported (12) may have positively biased the average accuracy report in comparison with other classifiers which have a larger proportion of studies to providing a more balanced view of their accuracy. The two remaining classifiers SVM and ANN essentially score equally (80.33 and 80.70 respectively) when tasked with an average of 3 class discriminations and have a high number of studies supporting this high average accuracy. Of the classification methodologies used to complete the studies, 44 were subject-dependent, 23 subject-

independent and 35 unknown. From these data it can be seen that subject-dependant classification is the most applied technique and the most accurate on average (83.4% for subject-dependent versus 69.4% for subject independent). This observation has clear implications for the construction of a biocybernetic loop and makes an implicit recommendation that calibrating the loop for individuals appears to deliver the highest classification accuracy. Furthermore, it can be seen that RDT SVM and NN represent the top tier of accuracy for machine learning algorithms, able to deal effectively with the variance inherent in psychophysiological data and should be investigated for application within biocybernetic loops. However, KNN was used during a high number of studies and is also worthy of investigation due to a high degree of transparency.

The classification of psychophysiological data falls within the field of pattern recognition. In this field, there are two types of classification methods: numeric and non-numeric. Numeric methods consist of deterministic and statistical measures, which can be considered as measurements made on a geometric pattern space. Such methods are commonly referred to as clustering techniques or unsupervised classifiers, which seek to partition data and cluster data points around the boundary of each partition in order to make a determination of class membership. Non-numeric methods seek to objectify categorisation problems into symbolic abstract representations (such as algebraic form) that do not directly manipulate the numerical values they represent. This allows for simultaneous analysis and computation of numerous paths of class membership; techniques' that utilise this approach are commonly referred to as supervised learning classifiers and include Support Vector Machines (SVM) and Artificial Neural Networks (NN). The goal of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown (Kotsiantis et al. 2006).

In the following sections the machine learning algorithms identified from the survey data as potential candidates for application within a biocybernetic loop will be investigated further.

3.1. K-Nearest Neighbour

As identified from the survey data, KNN is a popular numeric method of classification (with 16 studies shown in Table 3-1), nearest neighbour techniques belong to the class of supervised learning algorithms. This algorithm classifies data based on the shortest distance of an item of test data to the neighbouring training data class sample(s), this makes KNN a two stage process, stage one gathers training data and stage two gathers test data for comparison. Test data is then assigned to whichever class it appears closest to, using a distance metric such as Euclidean distance for

comparison. Computing “k” involves determining the number of nearest neighbours (via majority vote) required to provide the most accurate prediction of which class an item of test data belongs to.

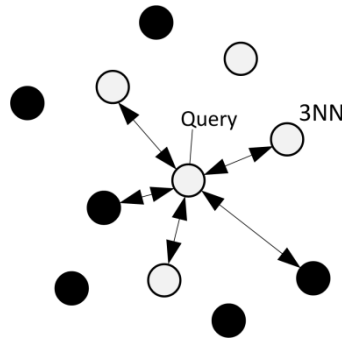


Figure 3.1 Example k-nearest neighbour query; 3 nearest neighbours

By taking several distance measures against many class samples (Figure 3.1), the effect of any noisy physiological measurement is likely to be averaged out over k-samples. However, reliable classification of psychophysiological data by numeric pattern recognition systems depends heavily on the use of noise-free features as input. Therefore, the KNN approach is best applied in situations where a high number of classes are to be classified and where physiological data is continuous and free of noise, such as within laboratory settings (Petrantonakis & Hadjileontiadis, 2010; Novak et al. 2012). KNN has been shown to have a high accuracy rate in this setting when applied to classifying basic emotions with an average of 5 classes (Picard et al. 2001; Lisseti et al. 2003; Wagner et al. 2005).

With reference to

Table 3-1, it can be seen that KNN is the second least accurate classification algorithm scoring an average 72.58% accuracy over 16 studies; KNNs popularity is based on its simplicity and transparency, in that the algorithm is easily understood and can be rapidly developed and deployed. However, in order to be an effective classifier, input data must be rescaled (i.e. normalised between 0 and 1) so all input features weigh equally due to the algorithm’s reliance on distance calculations to complete classifications. Furthermore, the selection of physiological features must be optimised, as high dimensional data increases computational load and decreases accuracy as some features may be irrelevant but weigh equally in the distance calculation (Novak et al. 2012).

No. Studies	No. classes	Classifying	D	I	Unk.	Avg. Accuracy
2	2	O		1	1	71.80
5	3	BE/O	1	1	3	72.13
5	4	BE/AV	4	2		70.51
4	>4	BE		1	2	72.52

Table 3-2 K-Nearest Neighbour number of studies, number of classes, Averaged accuracy per number of classes: D = subject dependent; I = subject independent; Unk = unknown

Looking closer at the studies which utilised KNN as a classifier, Table 3-2 shows that KNN has been applied to the classification of basic emotions (BE), activation and valence states (AV), and other psychophysiological states (O) such as stress and anxiety; of these applications classifying basic emotions with three to four classes is most common. For example, Kolodyzhniy and colleagues (2011) completed a study testing the applicability of feature selection, linear and non-linear classifiers and crossvalidation procedures to automated emotion classification. They reanalysed data from a previously completed study involving emotion elicitation using film clips, presented over two sessions to create a test-retest scenario. Thirty four participants took part in the video study. Kolodyzhniy et al, used six 10 minute long video clips pre-classified as frightening, sad and neutral to elicit fear, sadness and neutral emotional states, recording and classifying measures of the autonomic nervous system such as electrocardiogram (ECG), electrodermal activity (EDA) and respiration etc. To create the datasets for classification they first created an average of a 180 second baseline period which preceded each video clip was displayed, maximal reactivity from baseline was then calculated for each psychophysiological measure for each emotion induction and each participant. These data were then aggregated as test and retest data to create the training and testing datasets used for classification. Data was then classified in two ways; as subject - stimulus dependent and subject - stimulus independent, they achieved best results classifying subject-stimulus dependent data using KNN with 17 nearest neighbours reaching 81.9% \pm 17.1 averaged crossvalidation accuracy over the three emotion elicitation conditions, classified separately they achieved 80.9% for fear, 80.9% and 83.8% for the sadness and neutral conditions respectively. This subject dependent classification result compares favourably against a 77.5% \pm 20.1 averaged accuracy, fear 77.9%, sadness 70.6% and neutral 83.8% classification of subject independent data. These results indicate that it is possible to apply the KNN algorithm to classify and detect affective states from physiological signals with moderate accuracy. Furthermore, the computational simplicity of the algorithm makes it a primary candidate when considering applications involving real-time measurement and classification. However, the sensitivity of the algorithm to noisy features within psychophysiological data requires careful consideration.

3.2. Decision Trees

Regression decision trees (also known as decision or classification trees) work by creating a tree structure that maps observations about a value to assumptions about target class of that value. Within the tree structure, a leaf represents a class label and a branch represents a binary decision that denotes the class label. The algorithm works by progressing through several branching IF–

THEN logical rules, this branching logic structure is why they are referred to as trees. The tree shown in Figure 3.2, is a simple example of a decision tree classifying skin conductance response (SC) to trigger an adaptation component; IF no response is detected nothing is classified, IF the skin conductance response is greater than 0 then branching logic classifies it into one of 3 THEN states, in this case high medium or low arousal (using SC to infer a level of arousal) based on the level of response, once classified the judgement is passed to the adaptation component.

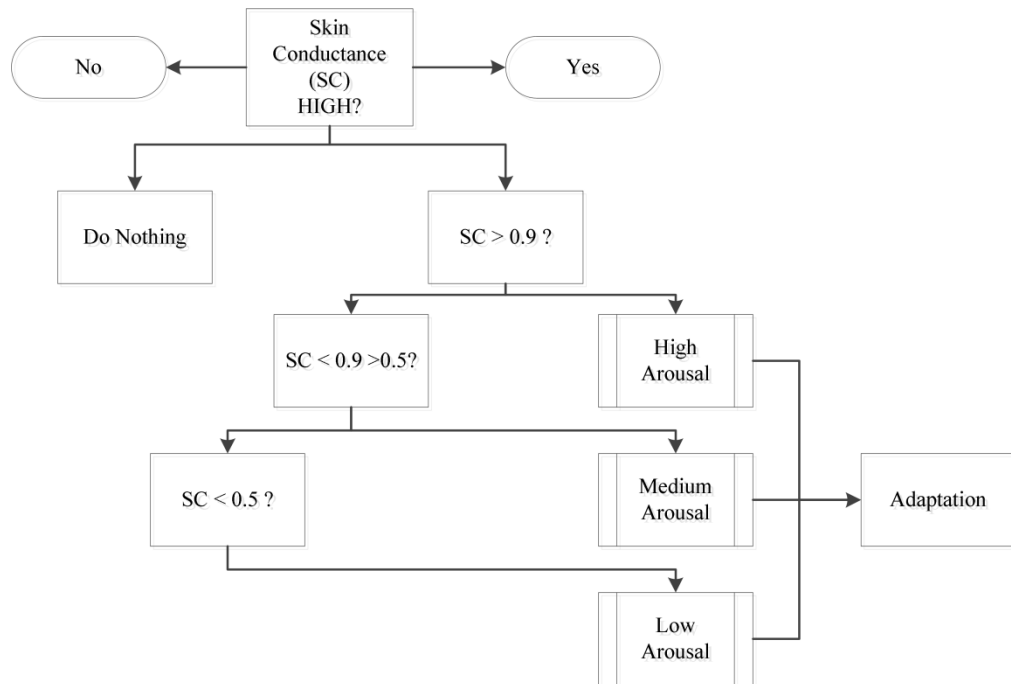


Figure 3.2 Simple regression decision tree for a biocybernetic loop

While the rules in this example are simple and could arguably have been set by an expert, decision tree rule-sets are not defined manually. Several different algorithms exist which learn and output the rules based on training data at each new node of the tree (see Kothari & Dong 2000 for a review), these algorithms select the feature that best discriminates between classes after all the previous decisions made in the tree have been taken into account. The efficacy of decision trees for psychophysiological data classification can be seen in

Table 3-1 by its high average accuracy of 84.85% over 12 studies. Decision trees offer a high degree of transparency, in that the tree structure is easy to comprehend and visualise. In some cases algorithms allow manual pruning of the “tree”, which reduces the complexity of the logic structures, preventing over fitting of the data, and can also act as a form of dimensionality reduction, increasing the strength of the inference. Decision trees perform best when applied to problems with a low number of classes but high number of levels within each class, for example: anxiety levels (Rani et al. 2007, 83.5%); (Liu et al. 2009, 88.5%), levels of stress and fatigue (Rigas et al. 2011, stress 76%;fatigue 81%), and stress levels alone (Plarre et al. 2011, 90.2%).

No. Studies	No. classes	Classifying	D	I	Unk.	Avg. Accuracy	Stdev
5	2	BE/O	1	2	4	81.68	5.57
5	3	BE/O	2		3	79.84	9.05
1	4	AV	1			77.00	0.00
1	>4	BE	1			89.20	0.00

Table 3-3 Decision Trees number of studies, number of classes, Averaged accuracy per number of classes: D = subject dependent; I = subject independent; Unk = unknown

The data in Table 3-3., shows that out of the 12 studies reported, decision trees are most frequently used to in the classification of basic emotions or other states, in this instance decision trees are used predominantly in the classification of other states such as stress, anxiety and amusement. For example, in a study investigating the characterisation of game players' experience using physiological signals and a decision tree classifier; Levillain and colleagues (2010) aimed to find the optimal balance between amusement and challenge then apply this information to drive changes inside of gaming environments. For this study 25 participants were required to play a first person shooter game on a gaming console while measures of autonomic nervous system activity (ECG, EDA and Respiration) were recorded. The game was split into 4 "game sequences" ranging from simple non-challenging to complex most-challenging gameplay. To derive class labels for classifier training, participants were asked to rate each game sequence by answering four questions posed as binary choices:

- Which sequence is most amusing?
- Which sequence is least amusing?
- Which sequence is most challenging?
- Which sequence is least challenging?

These labels were then associated with the features derived from the physiological measures to create classifier training datasets and classified using a decision tree algorithm. Their main objective was to extract and classify those physiological features that best characterise a player's level of enjoyment (as amusement); in this regard they achieved a modest level of classification accuracy of 80.40% classifying the least amusing game sequence and 71.30% for the most amusing sequence. This result shows that it is possible to identify and classify a key psychological state using measures of physiology with a moderate to high degree of accuracy and use this output inside possible biocybernetic loops with gaming applications to achieve an optimal state of satisfaction when playing a game. Furthermore, the computational simplicity and transparency of the decision tree algorithm makes it a prime candidate for real-time applications providing the algorithm has the ability to handle noisy features inherent to psychophysiological feature data.

3.3. Support Vector Machine

A considerable amount of research for pattern recognition within affective / physiological computing has concentrated on two non-numeric machine learning techniques the support vector machine (SVM) (22 studies) and artificial neural networks (NN) (17 studies). These techniques have been used primarily to classify psychophysiological data from multiple modalities (ANS, EEG, fEMG etc.), to discriminate between basic emotions, level of emotional response or physiological activation. A SVM is a set of related supervised machine learning algorithms that allow the analysis and categorisation of data and the recognition of patterns within that data. The SVM (Vapnik & Cortes, 1995) is a maximum margin, two stage non-probabilistic binary classifier that works within high dimensional or infinite feature space. Physiological input data is mapped onto this feature space in the form of feature vectors, using linear or nonlinear mapping and separated into two classes by a hyperplane (Figure 3.3). Values lying close to the intersection of the hyperplane become support vectors and distance based geometry (similar to KNN) is then used to determine the class of a new data point by: (1) calculating the distance from the hyperplane relative to its position (above or below) and (2) by its distance from a support vector if the value falls within the threshold boundary of the hyperplane intersection. In addition the SVM can perform non-linear classification by applying various “kernel tricks”, implicitly mapping inputs into higher dimensional feature space.

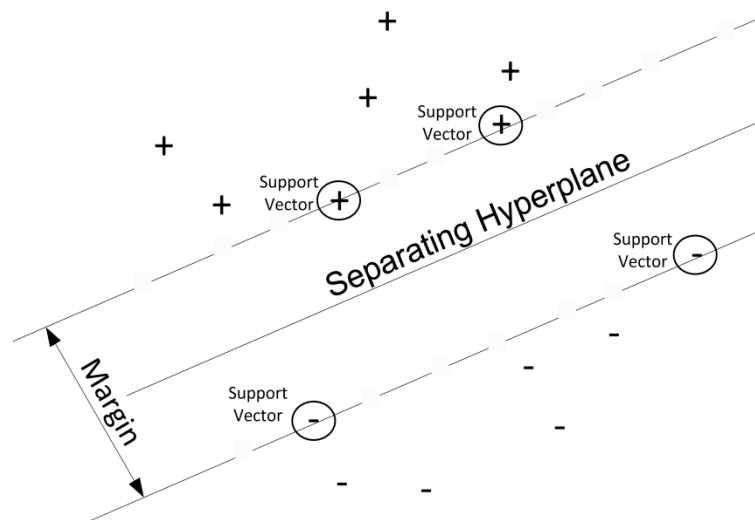


Figure 3.3 Example SVM separating hyperplane

No. Studies	No. classes	Classifying	D	I	Unk.	Avg. Accuracy	Stdev
10	2	O/BE/AV	2	3	5	84.61	6.51
6	3	AV/O	3	2	1	77.09	15.37
5	4	BE	2		3	71.78	9.58
1	>4	BE	1			95.80	0.00

Table 3-4 Support Vector Machine number of studies, number of classes, Averaged accuracy per number of classes: D = subject dependent; I = subject independent; Unk = unknown

As can be seen from Table 3-1, the SVM has been used in 22 studies and performs well when applied to psychophysiological data, with an average accuracy of 80.33 % over all of the studies reported. The data shown in Table 3-4 displays that, similar to KNN and decision trees, the classification of basic emotions and other states (such as stress) are the key classification tasks; and out of the 22 reported studies 10 apply the SVM to tasks involving two classes, six studies utilise it to identify three classes decreasing to five studies for four classes and only one study using the SVM to classify more than four classes. The high number of studies involving two class classification can be accounted for given the SVM's binary nature, meaning it lends itself to two class problems more easily than those involving more than two, which require careful methodological planning. On the whole the SVM is applied in subject dependent classification; however there are a high number of studies with unreported dependency methodologies.

The SVM has been employed within the fields of affective and physiological computing in a variety of ways, both singly and in conjunction with other analysis techniques. The bulk of these studies applied the SVM to discriminating between the "basic" emotional states: anger, sadness, joy, fear, disgust and surprise (Ekman, 1999) using a combination of ANS and fEMG (see Katsis et al. 2006, 2008; Pastor-Sanz et al. 2008; Calvo et al. 2009) or ANS with facial feature tracking (Bailenson et al. 2008). Exploring other affective dimensions, the SVM has been used to classify states in arousal-valence space using a combination of ANS and EEG features (Chanel et al. 2009; Shen et al. 2009). In research focused on healthcare and quality of life (Sakr et al, 2010), a multiple SVM architecture was used to detect levels of physiological agitation (an indication of stress) in healthy participants completing a stroop colour-word interference test in a laboratory environment. HR, GSR and skin temperature measures were used to create a model of agitation transition, for diagnosing the early onset of agitated state episodes related to dementia. SVMs show great promise in this area, with reported average detection rates of 91%, over 58 subjects.

Research in the field of performance enhancement used SVMs to capture physiological measures which infer cognitive performance and workload, in order to improve performance in a virtual

reality stroop colour interference task (Wu et al., 2010). Wu and colleagues measured indices of skin conductance, respiration, ECG and EEG activity continuously during task performance over three conditions, low threat (colour naming) high threat (word reading) and high threat (interference). Reported results from using the SVM approach were again very promising, with quoted accuracy ratings of 96.5% over 30 subjects tested against a training set of a further 120 subjects. However, the authors noted that a certain degree of data preparation was required before the SVM could be applied to subject test data, such as interval normalisation [0, 1], feature ranking to determine the optimal training set using sequential forward selection (SFS) (Guyon & Elisseeff 2003) and best 5 fold crossvalidation accuracy from each feature selected in SFS. In effect the raw data measures were processed for statistical significance and only those measures with the highest significance (per subject) were included in the SVM categorisation process.

The SVM is best applied in situations where a low number of classes are required or where high class problems can be reduced to a series of binary class discriminations. In these situations, the SVM has proven to be highly accurate. However, the high number of subject-dependent studies (10) versus subject-independent (6) may highlight issues regarding the ability of the classifier to generalise across a population of users and suggests that SVM should be considered for applications involving single users. There is some evidence to suggest that the SVM can be applied to raw psychophysiological data (data that has not been pre-processed). In a study completed by Rani and colleagues (2006) five affective states - engagement, anxiety, boredom, frustration, and anger - were invoked in participants while completing anagrams tasks and playing a game of pong. They reported an 85.81% accuracy rating (averaged over the five states) for the SVM using raw physiological data as input. If valid, this last point could have positive implications for the building of a biocybernetic loop which includes the SVM as the classification core, as this would indicate that the feature data captured from sensors would need very little preparation before classification, resulting in a more responsive system with reduced computation needs.

3.4. Artificial Neural Networks

Artificial neural networks (NN) are information processing constructs that attempt to mimic how biological systems process information. The design of these constructs is drawn from the components that make up the human brain and this biological approach is combined with numerical and statistical analysis elements. A neural network in general terms, is an interconnected network of a number of processing elements called “neurons” which is created for a specific purpose, such as signal processing or data classification. Figure 3.4 shows a simple representation of an example neural network. Here each circular node represents an artificial neuron and an arrow represents a connection from the output of one neuron to the input of another. Each node can contain one of a

number of processing techniques (such as propositional logic or statistical test). Each processing node (in the hidden layer) receives a number of inputs and uses them to calculate its activation based on the weighted sum of the inputs. This value is used as a threshold. If the value exceeds the threshold, the node outputs the class of the input data. This output is then fed to the next layer of neurons and so on, until the final output is determined. There is only one hidden layer in the example but there can be many. When layers are aggregated, they are referred to as a multilayer perceptron. This ability to aggregate processing units and the inherent interconnectedness give NNs the potential to process large volumes of complex or noisy data. NNs learn by example. They can be trained with known examples of problems in order to ‘acquire’ knowledge about them. Once appropriately trained, the network can then be applied to solve ‘unknown’ or ‘untrained’ instances of a problem.

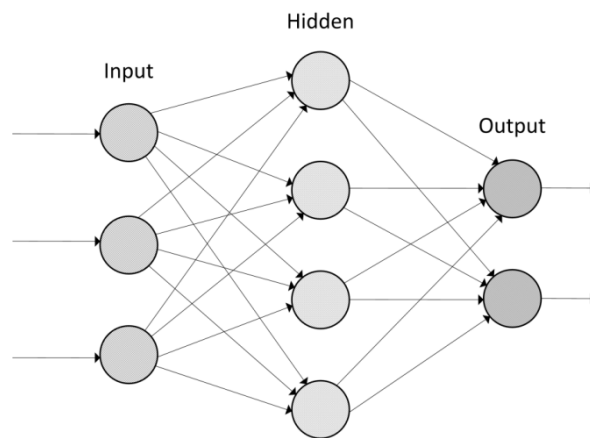


Figure 3.4 Example Neural Network with one hidden layer

No. Studies	No. classes	Classifying	D	I	Unk.	Avg. Accuracy	Stdev
6	2	O/BE	3		3	83.46	8.13
3	3	O/AV	3	1		73.94	10.51
5	4	BE	2	2	1	74.78	12.01
2	>4	BE	1		1	90.55	6.55

Table 3-5 Artificial Neural Networks number of studies, number of classes, Averaged accuracy per number of classes: D = subject dependent; I = subject independent; Unk = unknown

The use of neural networks within the fields of psychophysiology, physiological and affective computing is a relatively new technique. However, the data in

Table 3-1 Table 3-1 shows this technique is becoming increasingly popular in these fields with 17 studies reported to date. Table 3-5 shows that of these 17 studies, six applied NN’s to two class problems, classifying other psychophysiological states (such as entertainment preference and workload) and basic emotions and as the number of classes increases to five classes and beyond NN’s are exclusively applied to classifying basic emotions in a subject dependent manner. The

average accuracy of 80.70% shows neural networks to be in the same accuracy class as decision trees and support vector machines, NNs have been applied to a wide variety of data from basic emotion discrimination (Nasoz et al. 2004, 2010; Wagner et al. 2005), workload level (Wilson and Russell, 2003a/b, 2007) to entertainment preference (Yannakakis & Hallam, 2008). For example the study from Wilson & Russell (2007) used a NN to analyse psychophysiological signals (ANS and EEG measures) to aid in real-time adaptation of difficulty levels when subjects performed a complex aviation task (analysing radar data for potential military value), they found a significant increase in subject performance when comparing adaptation versus no adaptation in task difficulty. Using the NN to categorise mental workload they reported 89.7% and 80.1% accuracies for easy and hard condition adaptations respectively, they then compared this result against a new condition, that of subjects trained in the task versus non-trained and found that the NN technique scored significantly higher in terms of correct categorisations with 95.7% and 83.6% for easy and hard conditions respectively.

Further work by Kliensmith et al (2011) adopted the NN approach to recognize and categorise non-acted bodily gesture based indices of affect in subjects participating in a game task. They conducted an online posture evaluation survey with computer avatar stimuli, using a subset of postures (frustrated, concentrating, defeated and triumphant) to compare against NN categorisations for the same set of postures which used motion capture data as inputs for training. They then split these postures into ranks for arousal and valence and compared the results from human observation agreement and machine NN categorisation, the results from this comparison were very favourable with an 87.4% recognition rate for human observation and 87.2% for NN categorisation for arousal, and 84.3% and 83.9% for human and machine respectively for valence. Out of the 17 studies reported in the Novak et al. (2012) review, 9 used subject-dependent approaches and 3 used subject-independent (with 5 unreported), showing that overall this classification technique appears to perform optimally in situations that allow long calibration periods, calibrated to individuals and applied in real-time tasks; on average NNs were applied to tasks with 3 classes and in this type of scenario can output highly accurate classifications (for the given context). However, the long calibration periods (requiring a large dataset) coupled with the complexity of NNs once they are trained; give the NN a large computational requirement, in terms of both storage and processing. These requirements make NNs a classification technique best applied in contexts which do not require quick calibration and deployment.

The classification approaches discussed in the current section all share a common theme, they concentrate on the detection of affective states within the confines of the laboratory, where levels of signal noise and confounds (such as movement) can be controlled or kept to a minimum using state of the art sensor technologies. Furthermore, laboratory testing also allows data to be analysed

and classified post hoc, using statistical and data processing techniques, which can sometimes affect classification outputs significantly. As a consequence very few studies utilise these techniques inside a biocybernetic loop as part of a live real-time system. From the literature, KNN, RDT and SVM appear to be the best candidates for deployment in real-time systems as the classification component of a biocybernetic loop. These algorithms are fairly transparent and the results from the literature show they are able to classify subject-dependant psychophysiological data with high accuracy. The suitability of applying these algorithms to classify psychophysiological data inside a biocybernetic loop suitable for cultural heritage environments will be investigated in detail in the experimental studies presented in this thesis.

4. Adaptation and interaction at the interface

Adaptation is the final component of a biocybernetic loop. An adaptive component transposes classification output into adaptive control actions, in accordance with internal IF THEN rules to initiate interface adaptations or not. Adaptive systems that include a biocybernetic control loop can be divided into two categories: autonomous systems, where physiological input alone is used to drive adaptations; and hybrid decision support systems, where user decisions in conjunction with physiological input, are used to inform adaptations. The degree of automation provided by autonomous or hybrid user-driven adaptive systems can differ in both type and complexity. These automated responses range from the simple organisation and provision of information in response to physiological changes, to multi-layered response adaptations at key stages in a process in response to both user decisions and physiological changes or in extreme cases a hybrid system that relies on both decision based input and physiological changes but carries out a mission critical adaptation or decision automatically if certain response criteria are not met (Parasuraman et al, 2000).

An example of such a system could involve a biocybernetic loop used for information provision that monitors the level of interest of the user and makes information recommendations based on the classifications of this state. These recommendations can be implicit and fully automated; where information reaches the user in a pre-edited form, or there may be explicit recommendations e.g. you may be interested in x, which requires confirmation by the user.

When designing an adaptive system that utilises biocybernetic control as its primary driver, the decision as to what level of automation is required is of paramount importance for the user experience and should be taken at the outset. The choice of autonomous or hybrid semi-autonomous system design can create adaptive system experiences ranging from supportive learning experiences to a deeper performance enhancement through reciprocal task allocation. Too

much automation creates a risk that users feel alienated with no control as information is pushed at them or tasks reassigned, too little and the adaptive system serves small purpose, as tasks or interactions require too much effort to complete. The potential exists should a fully symmetrical dialogue be created, to provide a beneficial synergy between the human and the system, analogous to the synergy that can exist between two or more functionally communicating human beings. Thus, the goal of adaptive systems is to explicitly open a symmetrical dialogue between the user and the system with respect to maximising efficiency of the task and interactions with the system, and then synergistically drive this communication using implicit and explicit input (or combinations thereof) and adaptations to user psychophysiological states.

Trust is a major concern in the communication between the user and the system during interaction with automated technology. Lee and See (2004) define trust as “the attitude that an agent will help achieve an individual’s goals in a situation characterised by uncertainty and vulnerability” (Lee and See, 2004, p. 51), Miller (2005) expands this definition into a series of “attitudes”:

1. Trust is a response to knowledge or belief about world states, but is not in itself those beliefs.
2. Trust is affective; trust and mistrust produce feelings about systems or agents they are directed at; future trust is in part a function of this affective response
3. Trust is egocentric and therefore based upon individual interpretations about a system or agent’s ability to achieve goals

Miller (2005) posits that these attitudes are dynamic and have greater and lesser importance dependant on the situation: greater if the situation is characterised by uncertainty and vulnerability; lesser if the situation is well understood and predictable. The fostering of trust between user and system is important to ensure the human-computer interaction is reciprocal and user understanding of the internal logic of the adaptive system is clear and unambiguous. Lee and See (2004) refer to this fostering as “tuning” and identify three routes to develop and tune user trust: analytic, analogic and affective methods:

- Analytic methods involve the detailed understanding and rational assessment of the mechanisms by which the system adapts to the user.
- Analogic methods involve the use of observable cues to infer system membership, such as an assumption of trust precisely because it was designed by members of the same group or endorsements based on the word of already trusted intermediaries.
- Affective methods are based on the affect generated by and towards the system, Lee and See (2004) base this on empirical findings that suggests that users tend to trust systems or devices that produce positive attitudes towards them

Lee and See (2004) also report that trust has a temporal element, meaning that it takes time for users to acquire trust in a system, whether through experience (analytic), training (analogic) or endorsement (analogic/affective). Furthermore, they point out that if a system receives a strong enough negative endorsement from an already trusted source, the likelihood that a user will trust the system is reduced even when presented with evidence to the contrary. From this analysis it can be determined that trust in adaptive systems is something that is engendered over time through use and continuous feedback. If feedback is positive the affective dimension of trust increases positive attitudes towards the system; if feedback is negative yet the reason for system failures is understood the analytic dimension is engaged, which again increases positive attitudes. If the system receives a positive endorsement from a trusted or authoritative source the analogic dimension is engaged and increasing the likelihood in the system outputs.

The majority of work in the field of adaptive systems has centred around one form of adaptive process; the modification of function allocation, in which the function modification acts as the interface between the user and the system, to determine which element performs which task i.e. the user or the system. Adaptations of function modification can be categorised within a 2x2 model by their target function level and immediacy; that is, adaptations can affect the semantic or syntactic levels of a system (Foley & van Dam, 1982) and be either immediate or future adaptations (Solevey et al, In Press). The semantic level refers to the internal values and parameters of the functions performed by the system; whereas the syntactic level refers to the input output operations performed to complete those functions without and reference to the values or parameters; immediate changes affect current interactive elements; whereas future changes adjust variable and elements which have not yet appeared as interactive elements. Thus, semantic adaptations are those that change the behaviour of the system and the goals and action of the user; and syntactic adaptations are those that take place at the level of the interface and do not modify the functional basis of the system (Jacob 2001).

Sheridan & Verplank (1978) defined adaptive systems for function allocation as a system with 10 levels of adaptation (LOA), ranging from the human operator completing the entire task, to the computer completing the entire task, with hybrid human-computer function allocations making up the intermediate stage. More recently Parasuraman, Sheridan and Wickens (2000) updated the 10 levels of adaptation model, to provide a framework consisting of four system function analogues of human information processing: information acquisition; information analysis; decision selection; and action implementation.

One of the research aims of this thesis is to utilise a real-time biocybernetic loop aimed at cultural heritage applications, in which users receive information and decisions adapted to psychophysiological responses automatically. In this use case scenario all function allocation and information provision is immediate, and users interact with the system implicitly via physiological sensors recording a psychophysiological state of interest and the system explicitly provides adaptation using this state information. Using the above function allocation mapping, the loop is opened when the system acquires psychophysiological data from the user in response to cultural heritage exhibits; the system then performs analysis in the form of classifications; the system decides what content to produce next; the system implements the decisions and content is displayed; this new content creates new responses to acquire, closing the loop.

5. Challenges for real-time psychophysiological state classification

Building adaptive systems for use in real-time contexts presents many challenges, such as: the choice of psychological model; type of sensor technology; measures of physiological activity to capture the psychological state; classifying the psychophysiological state; using classification output as part of an adaptation strategy; creating interface elements driven by the adaptation strategy and fostering trust in users of the system.

The choice of psychological model is one that is purely dependent upon the context in which the biocybernetic control loop is to be applied. However, care must be taken to ensure that the inference between psychological model and physiological response can be verified and operationalised experimentally before application in the field. Inferences should ideally be one-to-one or one-to-many (Cacioppo, Tassinari & Bernston, 2007) and model the psychophysiological state with sufficient diagnosticity, bounded by the rationale and task context of the system. Once the model is verified techniques should be applied to reduce the processing requirements and complexity of the system by reducing dimensionality of the physiological data (Novak et al. 2012). Reducing the dimensionality of the data also has the effect of reducing the complexity of classification protocols and the processing requirements for biocybernetic control systems. Specifically, a strong inference reduces in sensor requirements, measurements taken and data analysed; this simplifies the training of classifiers potentially leading to more accurate and timely user state classifications; more accurate classification output enables system adaptations that reflect the user state more effectively, allowing interface elements to change dynamically and accurately in response to the user state.

There are few examples in the literature of applied real-time adaptive systems (e.g. Chanel et al. 2011; Wilson & Russell, 2003; Pope et al. 1995; Scerbo et al. 2001), a review of the literature has shown that certain elements of the biocybernetic loop are receiving much interest (e.g. sensor technologies, psychophysiological measurement and classification); however, this research interest is focused on state detection within the confines of the laboratory and not in real-time field applications. The lack of lack of real-time data can be seen as a limiting factor in the research, development and deployment of biocybernetic control loops. Furthermore, this lack of real-time data leaves a gap in the knowledge base into which valuable contributions can be made.

From the literature it can be determined that reliable psychophysiological state classification using pattern recognition algorithms depends critically on the use of noise free and highly separable features as input. It remains to be determined whether training classifiers using data from groups of individuals, or single individuals presents with the best chance of success when building a biocybernetic loop for adaptive systems. It is acknowledged in the literature that physiological responses to affective stimuli are strongly characterized by individual differences (Kim, 2007; Krohne 2003). Therefore, in order to discriminate psychophysiological states among multiple individuals, each with their own personal psychophysiological traits, the choice of nomothetic (group) or idiodynamic (individual) classification methods and psychophysiological state induction paradigms requires in depth investigation and analysis. The results from the literature survey suggest that classifiers trained to individuals will prove the most accurate, if proven true this raises a number of issues such as: how to train classifiers in real-time for use with individuals; how can the classification output be validated i.e. can the classification output be trusted as reflecting the user state alone or will this output require combining with user judgements as a final stamp of validity and if so how can these judgements be included in the classification training process; finally, if user judgements are used to train a classifier what effect does this have on classification accuracy i.e. does it matter if the dataset is biased towards one class or another if it accurately reflects user judgements towards stimuli.

The literature contains little data concerning the test-retest reliability of psychophysiological measures and their effect on automatic psychophysiological state classification, especially when real-time systems are the consideration. This lack of data may become an issue of some import as computing applications including biocybernetic control loops gain popularity outside of controlled laboratory environments, and begin to be used in the field. Which raises a series of questions, can a system trained on one occasion using data from individual or multiple users be used on a separate occasion or multiple occasions? Will the system require re-training; if so when is it appropriate to do so and how long will it take to calibrate? Furthermore, in a real-time systems context users could reasonably be expected to be ambulatory, which raises questions about the sensitivity, specificity and diagnosticity of psychophysiological measures when recorded from non-stationary sources, how will non-stationarity affect the strength of the inference and classification accuracy.

6. Outline of experimental studies

The goal of this thesis is to develop a biocybernetic loop to adapt and personalise information for the individual in a cultural heritage setting. This will involve the design and development of a real-time data processing pipeline that will translate raw psychophysiological data into control input for adaptive information provision or media tagging. A psychological construct will be posited and operationalised as physiological measures of the autonomic and central nervous system to create an inference model for a state of interest. Machine learning algorithms will be investigated to determine the efficacy of psychophysiological classification in both offline and online contexts. A series of experiments will be conducted to explore the design and implementation issues within two components (inference model and classification) of the biocybernetic loop culminating in a framework that integrates each of the components into a real-time proof-of-concept application.

- Study one: explores a psychophysiological inference of participant interest (as autonomic activation) using a range of autonomic measures and compares the performance of the K-Nearest Neighbour, Decision Tree and Support Vector classification algorithms under laboratory conditions using both subject dependent and subject independent classification methods
- Study two: This study will explore and investigate cross-session classification of autonomic activation wherein a support vector machine classifier was trained on session one and applied to data from session two. The classification algorithm is applied to autonomic responses (heart rate, skin conductance and respiration) to art images (paintings), recorded in a laboratory setting
- Study three: The goal of this third study is to examine the cultural heritage experience, then posit a three dimensional psychological model of interest as a potential driver of this experience and operationalise the interest model as multiple measures of psychophysiological activation. The operationalised measures will then be used for the subject dependent classification of multiple psychophysiological measures recorded using ambulatory sensor apparatus in response to audio material in a virtual cultural heritage setting
- Study four: this study represents a replication in part of study three using multiple sources of media (audio, video, still image and combinations thereof) in a cultural heritage setting. A framework for a biocybernetic loop aimed at cultural heritage applications that utilises the interest model is proposed. The purpose of the framework is to take in psychophysiological measurement at one end and output classifications of user interest at the other. Two classification protocols are proposed and tested both subject dependently and independently

- Study five: represents a culmination of the previous studies in order to create a classification protocol and application framework to capture high or low interest in response to video material in real-time. The application utilises the input output framework from study four and proposes a new classification protocol to evolve the one proposed in the previous study. The support vector machine is tasked with classifying user interest as a binary condition (high or low) within the context of a user viewing video content over a number of training cycles. This experimental study investigates the concept of machine accuracy versus the perceived accuracy of the system by the user at runtime, the classifier is trained to classify user preference by the user while in operation over the course of the experiment
- Finally the results from each experimental study are discussed as a whole in the context of the biocybernetic loop i.e. inference, classification interaction and adaptation. Furthermore the limitations of the work presented are discussed and future research work posited

7. Study One Classification of Psychophysiological Activation States using K Nearest Neighbour (KNN)

7.1. Abstract

The following experimental study explores two stages of the biocybernetic loop: inference and classification. The first stage was investigated by ascertaining levels of psychophysiological activation towards still imagery as a three condition activation state (representing high, medium and low activation states) to create a psychophysiological inference (as autonomic activation), using a range of autonomic (heart rate and skin conductance) measures under laboratory conditions. The second stage was explored by determining which physiological measures provide the greatest contribution to classification accuracies and applying the k nearest neighbour (KNN) classification algorithm and comparing the classification results with those of a support vector machine (SVM) and regression decision tree (RDT). The classifiers were trained with data from 15 subjects using a subject independent approach, with either class labels provided by survey or by subjective assessment. The results showed that classifiers trained using subjective assessment were more accurate to those trained using survey class labels. Furthermore, the comparison of accuracies from each of the classification algorithms showed that in this instance the support vector machine was most accurate. From this the following conclusions were drawn; that the support vector machine classification algorithm is well placed for classifying psychophysiological responses in comparison to RDT or KNN; That classification accuracy increased when trained using subjective assessment when compared to survey labels; That subject dependent classification methods should be investigated as a means to increase classification accuracies further.

7.2. Introduction

This experiment represents an exploration of the first stage of the biocybernetic loop investigating a psychophysiological state induction methodology; sensor hardware for autonomic nervous system measures; signal and measure analysis techniques and finally considering the suitability of KNN for classifying psychophysiological data when compared to the accuracy of SVM and RDT. The experimental approach builds upon the laboratory work of research in the field of physiological and affective computing as discussed above (e.g. Picard, 2003, Picard & Klein, 2002, Regan & Atkins, 2007, Wilhelm & Grossman, 2010, Petrantonakis & Hadjileontiadis, 2010, Lang, *et al* 2008, Rani, *et al* 2006). These studies form a core of research concerned with the application of physiology within computing and the categorisation of activation or affective states from physiological signals. Therefore, the overarching goal is to create a methodology and analytical framework that builds upon this previous research and forms the basis for future experiments, to inform the creation of an *ad-hoc* top-down psychophysiological activation-interest model with its basis in the circumplex model of Russell (1980) and to derive psychophysiological measurement and classification methods for use within a biocybernetic control loop.

This experimental study aims to ascertain levels of psychophysiological activation as a three condition activation state (representing high, medium and low activation states in response to visual stimuli); to determine the accuracy of the k-nearest neighbour categorisation algorithm when compared to subjective response data; and to validate this response data against the predetermined arousal/valence space data provided by the IAPS image database (Lang et al. 2008). Furthermore, this experiment will attempt to reveal how psychophysiological activation - taken as Heart Rate and Skin Conductance - varies in subjects presented with visual stimuli with known properties. These measures reflect two components of autonomic regulation; a sympathetic component and a parasympathetic component that may reflect a possible top-down neural influence (Rainville et al 2006). The variance in autonomic regulation will be recorded and features extracted from the measures to be used as input data for evaluating the efficacy of the k-nearest neighbour (KNN) algorithm when applied to the categorisation of levels of arousal to represent the three condition activation states.

The experimental study was conceived with the following aims:

- Ascertain levels of psychophysiological activation as a three condition activation state (representing high, medium and low activation states) and determine which physiological measures provide the greatest contribution to classification accuracies
- Determine the efficacy of the KNN categorisation algorithm to distinguish between levels of physiological variance in response to High, Medium and Low activation image stimuli taken from the IAPS database
- To determine the accuracy of the KNN categorisation algorithm when compared to subjective response data; and to validate this response data against the predetermined arousal/valence space data provided by the IAPS image database (Lang et al. 2008).
- To compare the performance of KNN with SVM and regression decision trees (RDT).

7.3. Methods

7.3.1. *Participants*

Fifteen participants, 9 male and 5 female, aged 20 – 45 years, took part in the experiment. The experimental protocol conformed to the requirements of the University Research Ethics Committee lease of ethical approval. Subjects were required to provide notarised consent.

7.3.2. *Design*

The experiment was designed as a subject static laboratory three factor repeated measures experiment; i.e. participants were exposed to IAPS images whose activation scores were classified as high, medium and low (Shown in Table 7-1).

There were 10 images in each group.

Category of Image	High	Medium	Low
Mean Arousal Rating	6.388	4.065	2.859
Mean Valence Rating	3.682	3.588	4.325

Table 7-1 Mean activation and valence values for each category of image

7.3.3. *Apparatus*

To capture physiological signals for this experiment, the Biopac MP150 and MP35 sensor networks were used in conjunction with the signal analysis software Acqknowledge (Biopac Systems Inc.). Data for three subjects was obtained using the MP150 sensor hardware which later developed a fault, resulting in subsequent recording being performed using the MP35. The implementation of

the KNN algorithm used the city - block distance measure (which measures the distance between two points as the sum of absolute differences of their Cartesian coordinates), this is available in the bioinformatics module of Matlab (at2012) a mathematical analysis software application.

7.3.4. *Experimental Measures*

The features derived from the measures heart rate and electrodermal activity were used on the results of research reviews that focus on autonomic nervous system activity using affective stimuli, two such reviews by Kreibig (2010) and Rani (2007) present tables that detail the physiological measures (and sets of features) associated with an affective or psychophysiological state.

- For heart rate (in beats per minute) maximum, minimum, mean and standard deviation.
- For electrodermal activity (in micro Ohms); area, mean, maximum and minimum

For the electrocardiogram (heart rate) a sampling rate of 1 kHz was used with a band pass filter of 5 – 35 Hz to remove noise and baseline drift. A low pass filter with a sampling rate of 1 kHz (1 thousand samples per second) and fixed at 5Hz was applied to the EDA channel to remove high level and allow tonic changes in EDA to be observed.

7.3.5. Procedure

Upon arrival at the laboratory the participants were given an information pack detailing the experimental procedure. Once notarised consent was given, they were seated and electrodes were applied. The participants were then required to view the thirty images chosen from the IAPS database and record a subjective two dimensional measure of level of arousal and valence using the Self-Assessment Manikin (Lang, 1985) provided. Upon completion electrodes were removed and any questions the participant wished to ask were answered.

To present the stimulus images and record subjective responses an application was developed using the E-Prime experiment design environment (MacWhinney 2001, PSTinc 2011). The following categories of image were used as a stimulus pool for the experiment; images were counterbalanced within category (Van Ulzen et al., 2008) and presented to the participant randomly:

- Ten images categorised as High arousal neutral valence
- Ten images categorised as Medium arousal neutral valence
- Ten images categorised as Low arousal neutral valence

Three test images selected randomly from the Low arousal category of IAPS database were displayed at the start of the experiment, to establish signal fidelity and baseline sensor readings. The stimulus presentation timeline is illustrated in Figure 7.1.

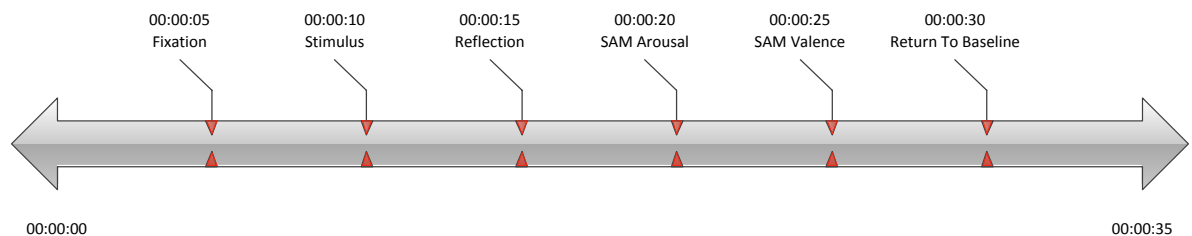


Figure 7.1 Experimental procedure timeline

The following tables (Tables 7 (2-4)) detail each image used, their IAPS code and valence/arousal ratings, and includes a representative sample of a stimulus image (one from each category) used for psychophysiological state induction.

High Activation Neutral Valence

Image	IAPS	Valence	Arousal
Snake	1022	4.26	6.02
Spider	1200	3.95	6.03
Pitbull	1300	3.55	6.79
Bear	1321	4.32	6.64
Shark	1930	3.79	6.42
War	2683	2.62	6.21
Openchest	3250	3.78	6.29
Lava	5940	4.23	6.29
Tornado	5971	3.49	6.65
Aimedgun	6250	2.83	6.54
Mean Arousal	6.388	Stdev	0.25
Mean Valence	3.682	Stdev	0.55

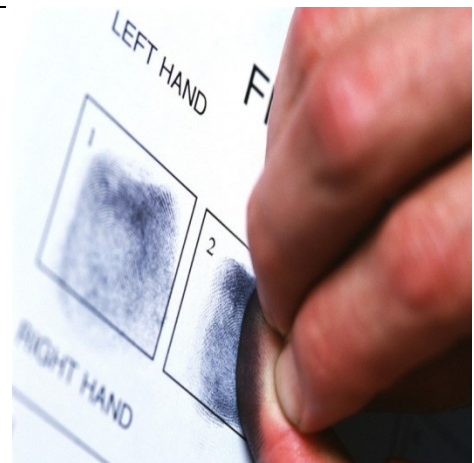


Sample Image High Category: Bear, Valence rating: 4.32, Arousal rating: 6.64

Table 7-2 Valence and Arousal values per IAPS image for high activation neutral valence category

Medium Activation Neutral Valence

Image	IAPS	Valence	Arousal
fingerprint	2206	4.06	3.71
Neutral face	2210	4.38	3.56
Boy	2280	4.22	3.77
Sad girls	2455	2.96	4.46
Police	2682	3.69	4.48
Alcoholic	2752	4.07	4.3
Jail	6010	3.73	3.95
Cemetery	9000	2.55	4.06
Cocaine	9101	3.62	4.02
Hung man	9265	2.6	3.34
Mean Arousal	4.065	Stdev	0.31
Mean Valence	3.588	Stdev	0.63



Sample Image Medium Category: Fingerprint, Valence rating: 4.45, Arousal rating: 2.81

Table 7-3 Valence and Arousal values per IAPS image for medium activation neutral valence category

Low Activation Neutral Valence

Image	IAPS	Valence	Arousal
Sick man	2491	4.14	3.41
Judge	2221	4.39	3.07
Pine needle	5120	4.39	3.07
Ironing board	7234	4.23	2.96
Office	7700	4.25	2.95
File cabinets	7224	4.45	2.81
Neutral girl	2440	4.49	2.63
Empty pool	9360	4.03	2.63
Trashcan	7060	4.43	2.55
Rocks	5130	4.45	2.51
Mean Arousal	2.859	Stdev	0.27
Mean Valence	4.325	Stdev	0.15



Sample Image Low Category: Filing Cabinets, Valence rating: 4.06, Arousal rating: 3.71

Table 7-4 Valence and Arousal values per IAPS image for low activation neutral valence category

7.4. Analysis Framework

A data analysis framework was developed to fit with the experimental study elements. This framework is outlined in Figure 7.2 and is detailed as follows:

- Capture physiological signals from participants using sensor hardware
- Select features from the captured signals
- Extract features to separate datasets
- Form Analysis instances
- Apply Principal Component Analysis to one instance of raw feature data
- Order one instance of feature data using IAPS image scoring
- Titrate one instance of feature data using participant subjective scores
- Process one instance of feature data as normalised change scores (z-scoring) to reduce the effect of individual difference within participant responses as in Mandryk & Atkins (2007), who used this approach to normalise electrodermal activity data.
 - Create two instances of normalised feature data and order one using IAPS scoring and the other using subjective titration

In this context the concept of titration is used to describe the process of re-ordering the feature data into categories based on the subjective scores from each participant for each image. In this way the

titrated dataset is labelled to better reflect the personalised experience and response of each participant.

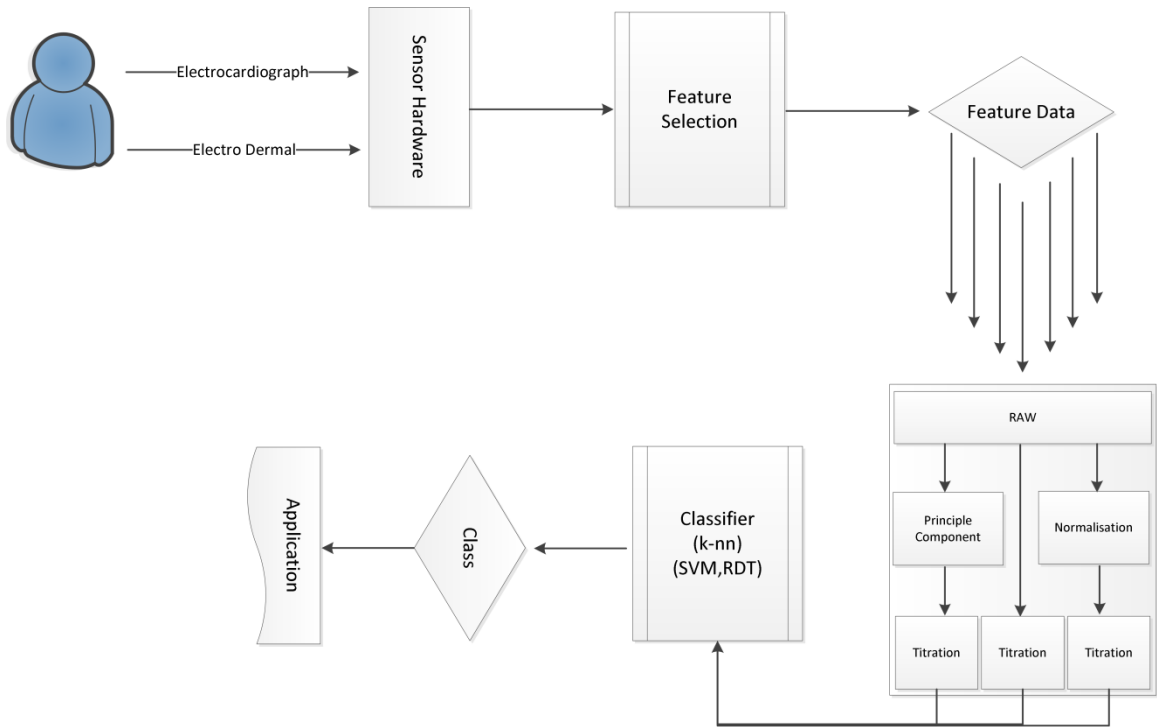


Figure 7.2 Experimental Analysis Framework

Analysis groups were created to allow for a comparison of classification accuracies between the two modes of labelling images, to form training sets for use with the classification algorithms. Mode one used images labelled using generic IAPS ratings. Mode two used those that were “personalised” using titration. Group A consists of 10 images in the High, Medium and Low categories labelled within categories using standard IAPS arousal ratings. Group B consists of 10 images in the High Medium and Low categories titrated using participant subjective arousal ratings. Groups C and D were created as reduced datasets of three images from each category to focus on data that were completely representative of each of the categories.

Table 7-5 presents the structure of the analysis approach used for comparing datasets.

	Category	High	Medium	Low
Analysis		#Images	#Images	#Images
A	IAPS	10	10	10
B	Titration	10	10	10
C	IAPS	3	3	3
D	Titration	3	3	3

Table 7-5 KNN analysis table

The analysis table allowed the following hypotheses to be tested:

- That on first exposure to the experimental visual stimuli, participant physiological responses will be of sufficient magnitude (in terms of variance per stimulus epoch) to allow the KNN classifier to successfully discriminate between three classes of response (high, medium and low) based on generic IAPS ratings.
- The accuracy of the KNN classifier will increase when the images are categorised according to subjective rating responses (via the titration process) as opposed to classification via generic IAPS ratings, i.e. accuracy for B and D should be higher than A and C respectively
- The accuracy of the KNN classifier will increase when the images are representative examples of each category of classification, i.e. accuracy for C and D should be higher than A and B respectively
- When applying the KNN classifier to two classes of data (high and low) as opposed to three categories, predicted accuracies will increase
- Applying Principle Component Analysis and Z-score normalisation techniques to reduce the variance of individual physiological responses within the data will increase KNN classification accuracies in all cases.

7.5. Results

For the measures of HR and EDA separately and then combined, discrete classifier analysis from A-D were completed on the resulting datasets. Both measures were then analysed post hoc using principle component analysis and normalised change score data, the classifier was applied to these datasets using the same method. Figure 7.3 displays the KNN classifier percentage chance accuracies over the full range of data. The results from analysis D (Table 7-5) for KNN were used as a basis for comparison against results gained from the SVM and RDT classifiers for the same data.

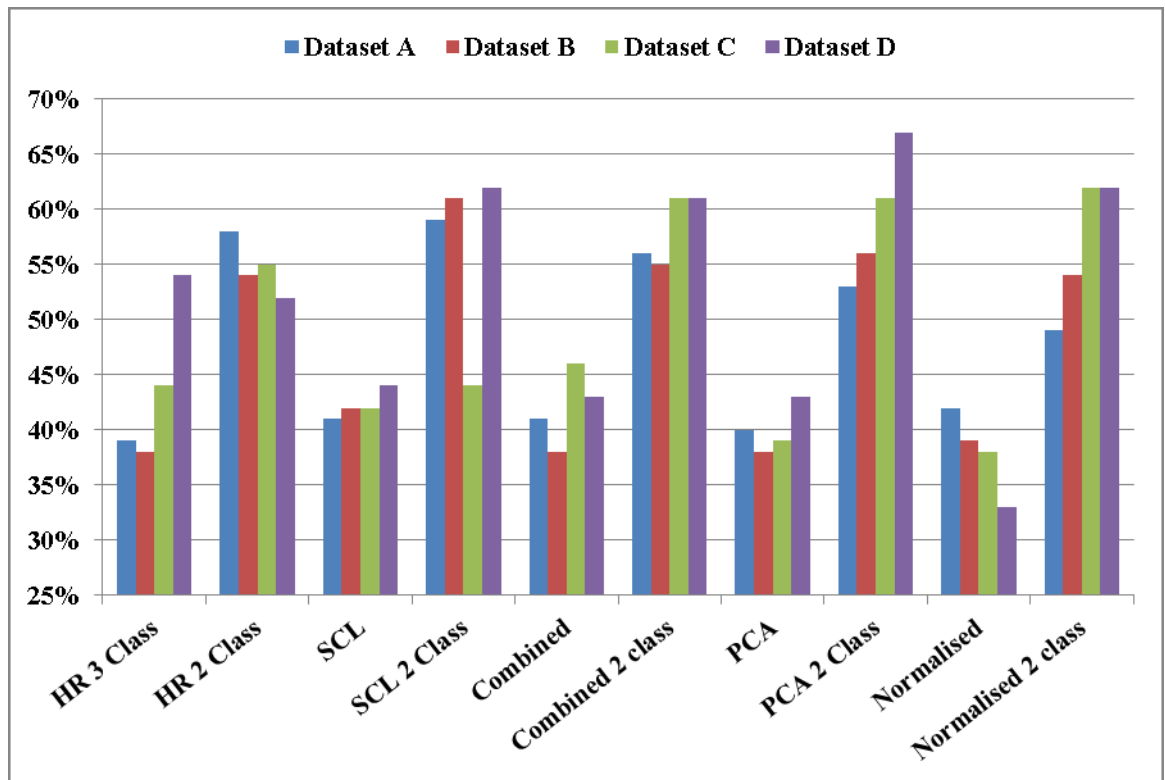


Figure 7.3 KNN Classifier accuracies 3 classes vs. 2 classes all analyses

For categorisation problems involving the three classes (high, medium and low) the KNN classifier reported poor accuracies regardless of whether the data is subjectively titrated or pre-categorised using IAPS scoring. Using the features of electrodermal activity titrated using subjective scoring for the three optimal images per category as input; the maximum accuracy of the classifier in a three class problem was 44%. In contrast, the maximum accuracy classifying three classes for features of heart rate for the same titration was 54%. This appears to show that heart rate is the most sensitive measure for three class problems when data is individuated using titration. These results indicate that the titration process is more effective at accurately depicting participant interest levels.

Classification accuracy improved when the data was presented as a two class problem (High vs. Low activation). The titration process further improved accuracy. Using the EDA feature data as input, a classifier accuracy of 62% was reached, compared with a 58% accuracy using the heart rate feature data. Removing the Medium category of data presents the classifier with physiological data with a higher degree of differentiation between the classes. The lack of physiological, and therefore class, differentiation is best represented in Figure 7.4, which displays that participant physiological responses in the Medium and High categories of image become indistinguishable leading to classification errors.

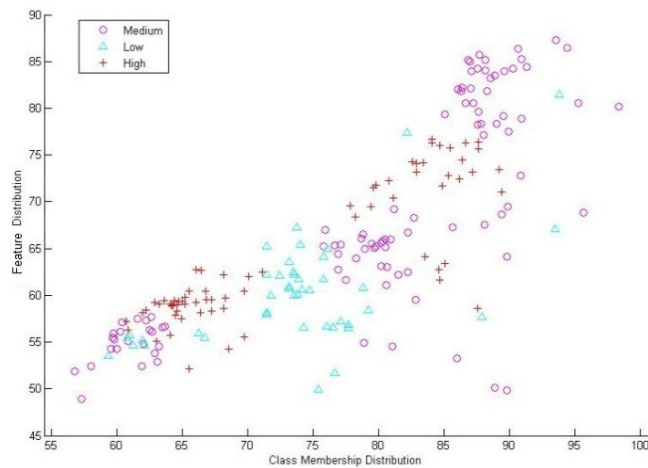


Figure 7.4 KNN classification showing no clear differentiation between classes

When HR and EDA data was combined into one dataset representing the three optimal images per category (Datasets C and D), the classifier reported the same low accuracies for the three class problem as reported previously, with the best achievable accuracy (46%). However, when KNN was used to classify C and D using the two category model (High vs. Low) accuracy increases to 61%.

On the premise that a reduction in feature dimensionality would improve classifier accuracies by providing factors that represented the most significance within the data, the combined HR and EDA data was pre-analysed with principal component analysis (PCA) (Cacioppo *et al*, 2007). However, for the three class problem, using the same titrated dataset, the classifier reported a maximum accuracy of 43% compared to a maximum of 39% for IAPS scoring. The results from the two class problem are more positive, classifier accuracy reaches 61% and 67% for IAPS scoring and Subjective titration respectively. The subjective titration accuracy represents the highest classifier accuracy overall for all analyses completed using this physiological data, as shown in Figure 7.3.

The observed rise in accuracy can be explained by the reduced feature dimensionality and inherent normalisation and significance tests that PCA performs during the analysis. The purpose of PCA is to reduce feature dimensions based on those features that are highly correlated and output n number of factors (in this case four factors combining eight features) that represent the significant features. Consequently, the data provided to the classifier was highly correlated and normalised to reduce the spread of individual psychophysiological responses, allowing for better spatial separation between the physiological responses to the categories of stimulus. The increased separation between the two classes of physiological response is correlated with the observed rise in classifier

accuracy. These results were promising for two classes. However, for a problem of three classes, the classifier was unable to differentiate with any degree of accuracy between the high, medium and low activation state responses.

7.5.1. *Comparing KNN and other classification algorithms*

The aim was to compare the classification performance of KNN with the RDT and SVM classification algorithms on dataset D, which consists of the 3 optimal subjectively rated images for the high and low categories of arousal.

RDTs work by creating a tree structure that maps observations about a value to assumptions about that value’s target class. Within the tree structure, a leaf represents a class label and a branch represents a binary decision that denotes the class label.

Table 7-6 compares results from the PCA two class KNN analyses to those from the SVM and RDT classifiers. KNN accuracies compare favourably with RDT for the subjective titration and poorly with the IAPS scoring.

Subjective Titration KNN	67%
Subjective Titration SVM	83%
Subjective Titration RDT	67%
IAPS Scoring KNN	53%
IAPS Scoring SVM	81%
IAPS Scoring RDT	67%

Table 7-6 Classifier Accuracy Comparison

The similarity of reported accuracies for KNN and RDT reflects the poor differentiation in physiological responses, in that both classifiers require data with a high degree of separation in order to function optimally. The complexity of a decision “tree” is directly correlated with the amount of spatial separation within the data; less separation is associated with more complexity and more classification errors.

Neither KNN nor RDT compare favourably with the SVM classifier which reported a very positive 83% and 81% accuracy for the two class three optimal images per category data. The type of SVM used to analyse this data was a radial basis kernel function (RBF) as this type has been shown to handle physiological data more effectively (Rani et al, 2006, Frantzidis *et al*, 2010).

It can be concluded from the results, that the KNN classifier is sensitive to noise within the physiological data. This conclusion parallels the literature on this issue (Petrantonakis &

Hadjileontiadis, 2010). Noise, in this context, was implicit in the medium category of images and expressed in the participants' physiological responses (shown in Figure 7.4) and further conveyed in the classification output. Using heart rate features alone, the KNN classifier reported low accuracies for both three and two class data. Accuracies scaled to a maximum of 54% using subjective titration, barely above that expected by chance. When the classifier is applied to electrodermal features, the three class accuracies are again poor. However, a positive increase in accuracies was observed, with a maximum accuracy of 62% for two classes using subjective titration. This result shows that EDA contributed highest to any significance within the physiological data and is consistent with the literature (Rani et al. 2006, Villon & Lisetti 2006, Kreibig 2010) and suggests that EDA is a measure sensitive enough for inclusion in future experiments.

The results for KNN compare favourably with RDT, in that output accuracy is similar. The RDT is better able to differentiate between data categorised using IAPS scoring and subjective titration, negating the effect of titration for a 1% difference in classifier accuracy between the two types of data categorisation. However, similar to KNN, the effects of noise within the data for RDT manifest in low accuracy output and an increase in the complexity of the decision tree. The increase in complexity is an important factor when considering the application of classifiers in real-time environments, in that complexity has a concomitant computational cost in terms of processing and storage requirements.

The maximum reported accuracy of 83% for the SVM shows its superiority for classifying physiological signals. The SVM overcomes noise within the data by adding support vectors and implicitly allowing for some misclassification of data points. The SVM uses, the distance from the hyperplane to insert support vectors, which in essence add "weight" to new values (based on the distance from the respective support vector) pushing them into the correct class. However, there is an implicit trade-off when using SVMs on noisy data, as there is a direct correlation between the number of support vectors required for classification and the computational cost required for new classifications. The increased cost is acceptable in a laboratory environment, but may present problems when applied in a real-time context using embedded computing or mobile devices.

7.6. Conclusion

From the results of the analysis, it is clear that in the context of this study (classifying psychophysiological responses to still imagery), the K nearest neighbour classifier is unable to differentiate between the three conditions representing a high, medium or low psychophysiological activation state with any degree of accuracy. Reporting a maximum 42% accuracy for three classes,

when subjective score data titration was used. The high degree of response overlap between the psychophysiological responses to the medium category of image and the psychophysiological responses for the low and high categories accounted for this lack of classification accuracy. When psychophysiological data is presented as a two class problem representing low and high activation states, in which the medium category of data is removed, the classifier reports higher accuracies of 67% for data processed using titration and principal component analysis. Despite its simplicity, transparency and low computational requirements, the poor accuracy reported here indicates that the KNN classifier is unsuited to real-time applications, where due to environmental factors psychophysiological data may contain noise or other artefacts. Similarly, the results from the RDT classifier were closely aligned with those of KNN when applied to 2 and 3 class problems, indicating that the accuracy of this classifier also suffers when data is less linearly separable, resulting in complex decision trees and a high misclassification rate.

The high classification result from the SVM classifier indicates agreement with those results reported in the literature (Novaks et al 2012), showing that the SVM is an algorithm capable of dealing effectively with data that has less than perfect differentiation. This capability may prove useful in future studies and making the SVM worthy of further investigation in both laboratory and field contexts. However, the high classification result obtained in this study was obtained through lengthy data processing techniques (e.g. z-scoring and subjective titration), and this additional data processing indicates that subject-independent classification techniques may be not suitable for real-time application. Furthermore, the increase in classification accuracy achieved using the subjective titration process indicates that subject-independent classification techniques should be investigated further. This outcome is in-line with the results reported in the literature review, which highlights that subject-dependent classification techniques are to date the most utilised and accurate way to classify psychophysiological responses.

8. Study Two: Test Retest classification of autonomic activation using Support Vector Machine (SVM)

8.1. Abstract

The aim of this study was to investigate cross-session classification of autonomic activation wherein a support vector machine classifier is trained on session one and applied to data from session two before re-testing with data from a third session. The classification algorithm is applied to autonomic responses (heart rate, skin conductance and respiration) to art images (paintings), recorded in a laboratory setting using data from 10 subjects. Two classification schemas were investigated by training classifiers with either survey or subjective assessment class labels. The effects of normalisation and dimension reduction of the physiological data on classification performance was examined and compared to physiological data with no data processing using both subject independent and subject dependent approaches. It was shown in this instance that autonomic reactivity was greatest during initial exposure to a set of stimuli and that reactivity will decline with subsequent exposures. In addition, it was shown that normalising the physiological data providing negligible benefit in terms of classification accuracies and that PCA is useful tool for identifying uncorrelated features within psychophysiological feature data which can lead to greater classification accuracies. Furthermore, the results showed a marked difference when comparing classification accuracies between a classifier trained using survey labels versus those provided by subjective judgment and that in this instance subject dependent classification provided the highest classification accuracies when compared to subject independent classification.

8.2. Introduction

The previous study demonstrated that the support vector machine algorithm produced the best performance for classification of autonomic activation when working on a subject-dependent basis. One of the goals of this programme of research is to posit a protocol for real-time classification, to achieve how classifiers perform in a variety of use contexts needs to be explored, such as performance within single sessions or across sessions / days. Consideration must be given to the effect of repeated testing within a subject-dependent context. To this end, determining whether a classifier built during sessional use on day one can be applied to consecutive sessions (day two, day three etc.) is an issue crucial to the real-time operation of a classification engine. The design implication is that if the classifier cannot generalise across sessions of use, it will need to be trained with each episode of usage. There are two main issues to be considered: (1) the test-retest validity of psychophysiological measures i.e. are the measures of physiological variance stable over

repeated testing, and (2) the ability of the SVM classification algorithm trained during one session to generalise to similar stimuli at a later point of time.

There is a rich body of research literature in the field of psychophysiology using the picture perception methodology to induce psychophysiological responses to imagery in a single experimental session (Lang et al, 1999). However, there is a paucity of literature to address the issue of test-retest validity of psychophysiological classification using machine learning algorithms. Test-retest reliability (stability) is an important trait for both physiological features and classifiers i.e. when some element of physiology is measured repeatedly under the same conditions, the resulting features should not vary “too much” as this would jeopardise a classifier that was trained on an earlier data-set. The test-retest reliability of many physiological parameters can be found on the literature e.g. quantitative EEG features (Tomarken et al 1992, Gudmundsson et al 2007); heart rate variability and respiration rate (Guijit et al 2007); autonomic nervous system measures (e.g. heart rate electrodermal activity) and EMG (electromyography) (Arena et al 1983, et al 1989). This second experimental study aims to investigate the test-retest reliability of machine learning algorithms by utilising the SVM algorithm as a determinant statistical tool. The working hypothesis of this study is that classification accuracy will remain high if a feature or subsets of features of psychophysiological reactivity are stable across sessions. If however, the classifier is unable to generalise across different stimuli or different experimental sessions, this could be indicative of poor classifier reliability due to the inherent variability in psychophysiological data or specificity of the protocol used to generate training data for the machine learning algorithm. The consequence of poor test-retest reliability is that a subject-dependent classifier will not generalise to different stimuli or across different experimental sessions and must be trained every time that the system is used.

The current experiment represents an advance on the previous study by applying a SVM classification algorithm to indices of autonomic activation in a series of two-class categorisation problems, i.e. a discrimination of high from low; baseline from high and baseline from low activation. These two class problems are based on a model which maps the IAPS (Lang et al., 1999), image model of arousal / valence space, into one of high or low autonomic physiological activation. This experiment uses images selected from the Digital Image Visual Aesthetics Survey (DIVAS) (Kreplin, 2014), a survey of paintings that have been classified using subjective data into high and low activation. In addition, we wished to explore issues surrounding the creation of a real-time classification engine for use in adaptive systems. For instance, would it be possible to expose participants to standard material in order to train a system for subsequent classification - and if so, how often would the system require calibration?

The purpose of this experiment is:

- To assess the stability of psychophysiological measures of autonomic activation across experimental sessions using the SVM as a discriminant tool
- Assess classification accuracy by applying a Support Vector Machine (SVM) approach to autonomic activation
- To test machine learning algorithms with cultural heritage material (paintings)
- To test whether the SVM trained on one dataset can generalise to a second dataset
- To assess test-retest validity of the SVM by training the algorithm to an image set and repeating exposure to the same image set at a later point in time

8.3. **Methods**

8.3.1. *Participants*

Ten participants 6 female (aged 20-40) took part in the experiment, a mixture of undergraduate and postgraduate students at Liverpool John Moores University. All participants provided signed consent form and the protocol was approved by the University Research Ethics Committee.

8.3.2. *Experimental design*

The experiment was designed as a repeated measures, laboratory investigation i.e. participants were exposed to high/low activation images from a standard database on three separate occasions, investigating the independent variable(s), level of activation with two conditions: High and Low activation with the following psychophysiological variables as dependent variable(s): heart Rate (HR), respiration (RSP) and skin conductance (SC).

8.3.3. *Experimental Measures*

The Nexus X Mk II (MindMedia Inc.) sensor hardware was used to collect psychophysiological data. A three-lead (lead 2 configuration) electrode connected to the torso was used to capture an ECG signal, which was filtered between 0.5 and 35Hz and sampled at 512Hz. The skin conductance signal was collected from two fingers on the non-dominant hand via the SCL channel of the Nexus hardware, providing an unfiltered signal sampled at 512Hz. An elasticated respiration band was worn around the chest by participants to collect respiratory rate and recorded using the RSP channel of the Nexus hardware providing an unfiltered signal sampled at 256Hz. Recorded data was saved to digital file and exported to AcqKnowledge 4.2 (BIOPAC systems Inc.) for processing.

8.3.4. *Experimental Material*

The 24 image stimuli used for this experiment were chosen from the Digital Image Visual Aesthetics Survey (DIVAS) (Kreplin, et al., *um*) in which 1023 participants gathered from the internet, rated relatively unknown contemporary artworks. Participants were asked to sit comfortably (approx. distance 1 meter) in front of a large 32" television screen and view images chosen from, after each image participants were asked to interact with a keyboard to record a subjective measure of "high" or "low" activation using a simple scale.

To present the stimulus images and record subjective responses an application was developed using the E-Prime experiment design environment (PSTinc 2011). Twelve images were selected from the database to represent high activation and a second group of twelve used to represent low activation these are shown in Tables (8-(1-4)).


Image Name	Activation Score	
afremove	6.08	
magicrealism	5.26	
blood	5.49	
realmofsense	5.24	
watertrap	5.34	
forbiddenspark	5.75	

Table 8-1 Image Set 1, High activation, mean activation score: 5.53, SD 0.301.


Image Name	Activation Score	
tincan	2.71	
armchair	3.19	
lavoisier	3.85	
earthsky	3.48	
candle	3.89	
sidestreet	3.85	

Table 8-2 Image Set 1, Low Activation, mean activation score: 3.50, SD 0.451.


Image Name	Activation Score	
artistinlove	5.30	
pursuithappiness	5.42	
elytron	5.16	
summerwine	5.55	
repetition	5.26	
prevenit	5.96	

Table 8-3 Image Set 2, High activation, mean activation score: 5.44, SD 0.261.

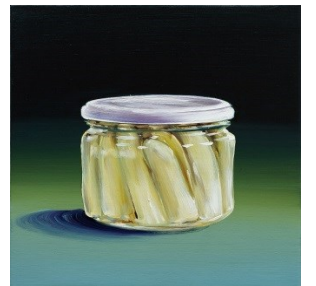
Image Name	Activation Score	
lightafterstorm	3.29	
masquelavoisier	3.88	
swimmer	3.69	
spargel	2.84	
rootsarmchair	3.85	
planetsmos	3.95	

Table 8-4 Image Set 2, Low Activation, mean activation score: 3.58, SD 0.396.

Both high/low activation groups were split into two sets of six and combined to yield: image set 1 that was presented to participants on the first (training) and third (retest) session and image set 2

that was presented on the second test session. Image set 1 was used to initially train the SVM algorithm during the first session, this algorithm was applied to classify a new image set (2) at session two and then re-applied to image set 1 during the third session (see Figure 8.1). Thus, a classifier trained on image set one and tested using image set 2 creates a test of subject-dependent generalisation performance. Whereas, a classifier trained using image set 1 and tested using image set 1 separated temporally creates a test of classifier reliability. When combined as an experimental protocol these data create a test-retest classification assessment scenario.



Figure 8.1 Sequence of testing for SVM experiment (a) and (b) examples of low and high activation images (left to right)

8.3.5. *Procedures*

Instruction about the experimental procedure was given and participants were asked to complete a consent form in accordance with the Liverpool John Moores Ethical Committee, and then fitted with the sensor hardware. Participants were asked to sit in a relaxed position approximately half a meter in front of a 32 inch high definition screen.

During each test session, a 10 second baseline (focus on fixation image) was collected prior to a 10 second exposure to each stimulus image. The images were presented in two blocks (High vs. Low Activation) and the order of presentation was counterbalanced across participants. Participants attended the second test session between 24-36 hours after the first session and the same time gap was used between the second test and the retest session.

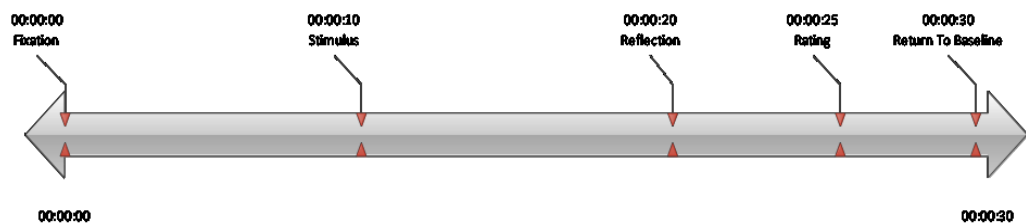


Figure 8.2 The stimulus presentation timeline

Figure 8.2 details the stimulus presentation timeline; after a 5 minute period in which participants were allowed to become comfortable wearing the sensor hardware the experimental procedure began with the display of two test images after which a fixation image lasting 10 seconds was displayed; this was followed immediately by the stimulus image which lasted 10 seconds. After 15 second reflection period, participants were then asked to provide a rating of high or low activation; this was followed by a 15 second relaxation period, to allow participant physiology to return to a pre-stimulus baseline state.

8.3.6. *Support vector machine parameterisation & accuracy estimation*

In this study a subject dependent approach was taken to analysing psychophysiological data to determine the recall accuracy of the SVM classifier. The SVM classifier implemented for this study was part of the bioinformatics module within Matlab. To evaluate classifier performance, over the training data, the *sequential minimal optimisation* (SMO) (Platt, 1998) and hold-out cross-validation methods were chosen. To provide the optimal settings for the box constraint and sigma values of the SVM radial basis function (RBF) kernel a loose grid search algorithm was developed and applied outside of the hold-out cross-validation procedure see Algorithm 1. The hold-out cross-validation method partitions the data into two parts, by randomly assigning data to either training or testing sets, ensuring that the classifier is trained and tested with novel data and is analogous to a real world task. This method of cross-validation has been shown to provide a more accurate assessment of potential classifier performance in comparison to k -fold cross-validation when applied to small datasets, such as those gained from real-time applications (Isaksson, et al, 2008).

ALGORITHM 1. Holdout Cross-validation using n by n grid search (*loose*)

Input: *Physiological data, Class labels, max Box-constraint, max Sigma*

Output: *Optimal Box-constraint; Sigma; accuracy*

$\sigma = 0.1;$

$\text{box-constraint} = 0.1;$

$\text{Counter} = 1;$

Create array for box-constraint; sigma and accuracy values

[*optimalValues*];

for n to max *box-constraint* **do**

for n to max *sigma* **do**

 Create two class problem

 Create a 60/40 split of *Physiological Data* as training and test data with associated *Class labels*: [*train, test*]

 Initialise a performance tracker

 Get instances of training data: $\text{trainIdx} = [\text{train}];$

 Get instances of test data: $\text{testIdx} = [\text{test}];$

 Train SVM using training data, current value of *box-constraint* and *sigma*

 Test the SVM model using test instances of training data

 Gather performance statistics

$\text{optimalValues} = [\text{box-constraint}, \sigma, \text{accuracy}]$

$\text{Counter} = \text{Counter} + 1;$

$\sigma = \sigma + 0.1;$

end

$\sigma = 0.1$

$\text{box-constraint} = \text{box-constraint} + 0.1$

 Store performance statistics

$\text{Optimal} = [\text{optimalValues}]$

end

Find optimal settings

$\text{Criteria} = \max[\text{Optimal}(\text{accuracy})]$

Output optimal settings

$\text{Parameters} = [\text{box-constraint}, \sigma, \text{accuracy}]$

Accuracy is determined by the number of true classifications plus the number of true negative classifications divided by the number of true plus false negative classifications plus the number of true negative classifications in the form of:

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

8.4. Analysis

8.4.1. Feature Extraction

Prior to commencing classification analysis of the physiological data, features were derived from measures of heart rate, skin conductance and respiration for a 10 second stimulus epoch. In total 11 descriptive features were derived from the psychophysiological measures for each of the stimulus events and used as classification vectors for the SVM (see Table 8-5). These features were then aggregated into a single observation, such that each observation set creates a unique classifier feature vector for each stimulus event, giving 24 observations per session. This training set is then used as the basis for classifying new instances of data into its respective class.

Measure	Derivative			
Heart Rate	<i>Mean</i>	<i>Stdev</i>	<i>iBi-Mean</i>	<i>iBi-Stdev</i>
Skin Conductance	<i>Mean</i>	<i>Area</i>	<i>Stdev</i>	
Respiration	<i>Rate Mean</i>	<i>Rate Stdev</i>	<i>Amplitude Mean</i>	<i>Amplitude Stdev</i>

Table 8-5 Features derived from physiological recordings

Data from each participant were analysed separately to determine the recall accuracy of the SVM classifier when compared against DIVAS survey labels and for individual subjective labels. Labels derived from a survey of population rankings benefits from statistical standardisation making them efficient when applied in a classification context. However, these labels while statistically balanced may not always be representative of an individual's subjective judgment. Whereas, individualised labels are fully representative of the subjective judgment. However, these labels are labour intensive to collect and may be implicitly biased towards an individual's preference. Recall accuracy in the context of classification is determined by first validating the SVM model over the training data to derive the original SVM classification, this SVM model is subsequently tested over novel instances of test data. In a laboratory context, the labels associated with the test observations are known to the experimenter but unknown to the SVM model, thus recall accuracy is calculated by comparing SVM model classification output (in terms of class) and comparing those to the known labels, the result is how well the SVM model recalled the class of the observation.

A number of analysis trials were prepared in order to determine the effect of various data treatment methods on SVM classification accuracy:

- Trials 1&2 compare raw untreated and normalised physiological data with labels derived from survey (effects of normalisation on classification)
- Trial 3 shows the effect of feature dimension reduction using principle component analysis (effects of dimension reduction on classification) using survey derived labels
- Trial 4 shows the effect of using raw feature data with labels derived from subjective ratings from each participant, comparing all features with a feature set with dimensionality reduced by PCA (effects of using subjective labels from each individual to train the classifier with and without dimension reduction)
- Trial 5 – presents a generalised model to assess the future viability of the subject dependent classification approach compared to subject independent classification accuracy using survey derived labels

8.4.2. *Trial 1 Raw Physiological Feature Analysis using Survey Labels*

This analysis trial consisted of untreated raw feature data, in which the features listed in Table 8-5 were fused into three data sets; each dataset is representative of one of three experimental sessions, these were then split into the three experimental conditions, giving a total of 9 datasets containing 12 observations, such that each condition (high and low) was compared against “baseline” data (using feature data from the fixation image). This partitioning resulted in two datasets that consisted of 6 condition images and 6 fixation images for a two condition discrimination analysis, high from baseline and low from baseline. To test high from low, the baseline images were removed leaving feature data for 6 High and 6 Low condition images. Reducing the data in this way ensured no bias of class labels was introduced during the training of the classification algorithm (SVM). For the raw feature data analysis all class labels used within the training phase of the algorithm were provided by the survey data of (Kreplin 2014).

8.4.3. *Trial 2 Normalised feature analysis using Survey Labels*

This analysis trial was completed to observe the impact of normalisation of the feature data on classification accuracy, the results from experiment one showed that a marginal increase in accuracy can be achieved by rescaling feature data to be within a standard range (0 to 1). The normalisation calculation was applied to columnar values (Novak et al. 2012), such that each feature was normalised to a standard score in the form of:

$$z = \frac{X_i - \mu}{\sigma}$$

Where X_i is the value to be scored μ is the columnar mean of the dataset and σ the columnar standard deviation. As with the raw feature analysis a total of 9 datasets containing 12 observations were created to test from condition discrimination.

8.4.4. *Trial 3 Dimension reduction using principal component analysis*

Based on findings from study one, which showed improvement in classification accuracies when the dimensionality of feature data was reduced. Raw feature data from all participants for day 1 were combined into one dataset and analysed using principal component analysis. Before completing this analysis HR-rate was removed from the dataset due to its direct correlation with HR-iBi. This analysis resulted in a total of 3 principal components HR-iBi – mean and standard deviation, and SC-level mean and area; SC-level area was added to the data set based on its high Eigenvalue (.976), a value of 1 is the generally accepted norm for acceptance within a PCA model. All respiration features were removed on the basis of the results from the PCA. Table 8-6 displays the amount of variance explained by the principal components. For each experimental session (train, test, retest) datasets for classification were constructed for each participant, from their individual psychophysiological data to correspond with the features identified by the PCA.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative	Total	% of Variance	Cumulative
1	2.824	31.382	31.382	2.824	31.382	31.382
2	2.651	29.452	60.834	2.651	29.452	60.834
3	1.227	13.634	74.468	1.227	13.634	74.468
4	.967	10.743	85.211			
5	.475	5.278	90.488			
6	.330	3.666	94.155			
7	.242	2.690	96.845			
8	.211	2.348	99.193			
9	.073	.807	100.000			

Table 8-6 Total Variance Explained, Principle component analysis

8.4.5. *Trial 4 Raw data analysis with associated subjective labels*

The purpose of this trial is to compare the effect of subjective response labels to those provided by the image survey on the performance of the classifier. The working hypothesis for this analysis is an expectation of greater classification accuracy across experimental sessions, given that the subjective label represents a more accurate indicator of the activation state for that particular participant in response to each image stimuli. This trial was completed using raw feature data and each observation was associated with the subjective response for that observation. Due to the nature of subjective ratings, the class labels provided by each participant contain a level of implicit class bias, for this reason, only psychophysiological data from the high and low classes of image were used to construct two datasets. Using these datasets two analysis runs were completed, firstly using raw untreated data and all derived features and secondly using raw untreated data and PCA derived features. For each analysis dataset, day 1 visit 1 data was aggregated and used to build the SVM classifier. The remaining data from day 2 and 3 were then used as novel data to test the generalisation (across session) performance.

8.5. Results

In the following set of tests (trials 1-3), the SVM attempted to classify psychophysiological responses to three different conditions: baseline (i.e. viewing a fixation point for 10 sec), low activation images and high activation images. The SVM classified a series of two-class problems (a discrimination of baseline from high, baseline from low, high from low activation), which were performed on each test session training, test and retest. The final test (trial 4) uses subjective labels associated with physiological response data, in one test all data is aggregated (inclusive of baseline

condition data); in the second test data from the high and low conditions is used. The accuracies obtained from the classification from each of the trials, is recall accuracy, in that the class labels for each observation (classification vector) are known in the experimental context but remain unknown to the classifier. Thus, after a correlation test, the performance of the classifier in real-terms can be obtained by comparing the how accurately it recalls the label associated with each observation. The mean classification recall accuracies and variation of classifier performance for each analysis type and all participants are illustrated in the following text.

8.5.1. *Trial 1 Classification of raw physiological features using survey labels*

Overall, it was anticipated that discriminating high activation from baseline would present highest classification accuracies. In the first analysis trial this is shown to be true (Table 8-7(1)), with the training session showing a respectable classification recall accuracy of 88.33% (range 83.33-100%) for the high from baseline condition. However, the mean accuracy results from the remaining two discrimination conditions, low from baseline and high from low show only a marginal decrease in accuracy of 85% (range 66.67-100%) and 83.33% (range 66.67-100%) respectively. Moving to the second visit analysis for the same conditions, which tests the ability of the classifier trained on day one data to generalise across days and to novel stimuli (using image set 2), we can see a dramatic decrease in recall accuracies. The magnitude of the decrease in performance was striking, in this instance the high from baseline condition performs best with 53.33% (range 50-66.67%) barely above that of chance levels. In the retest analysis accuracies rise slightly over all conditions, again the high from baseline condition performs best reporting 53.33% (range 33.33-75%).

Classification	Raw Data (1)			Normalised Data (2)		
	Train	Test	Retest	Train	Test	Retest
Baseline vs. High	88.33	53.33	53.33	80.00	54.16	54.17
Baseline vs. Low	85.00	50.83	52.50	86.67	55.00	60.83
High vs. Low	83.33	49.17	52.50	88.33	54.17	59.00

Table 8-7 Trial(s) (1), (2) Average SVM classification recall accuracy over the three test sessions, (raw and normalised feature data) n =10.

8.5.2. *Trial 2 Classification of normalised features using survey labels*

The results presented in Table 8-7(2) show the normalised feature data classification results, similar to the raw data analysis, validating the training data shows favourable recall accuracies for all test conditions. However, unlike the previous analysis the standout classification performance is in the high from low discrimination test with a mean recall accuracy of 88.33% (range 66.67-100%), highlighting a possible advantageous effect from data normalisation. This effect proves modest in the test and retest classifications however, normalisation providing results of just 54.16% (range 33.33-66.67%) and 59% (range 33.33-100%) for both test and retest conditions respectively.

8.5.3. *Trial 3 Effects of dimension reduction using principal component analysis*

As stated previously, a principal component analysis was completed to reduce the dimensionality of the data, consistent with this analysis, hr-rate mean, stdev; scl stdev and all features of respiration were removed from the physiological response data. The resulting dataset was then analysed in the same fashion as before (raw data vs. normalised), the effect of PCA on classifier performance can be seen in Table 8-8(a), (b).

Classification	Raw Data (a)			Normalised Data (b)		
	Train	Test	Retest	Train	Test	Retest
Baseline vs. High	76.67	50.83	49.17	80.00	57.50	55.83
Baseline vs. Low	75.00	57.17	54.17	83.33	52.50	57.50
High vs. Low	86.67	59.17	45.83	91.67	63.00	51.94

Table 8-8 Trial 3(a), (b). Average SVM classification recall accuracy over the three test sessions, for PCA derived features (raw and normalised) n=10.

Comparing the results from the PCA derived raw feature analysis (Table 8-8(a)) to the previous raw feature classification trial (Table 8-7(1)) there can be seen a decrease in the average training cross-validation accuracy across the three discrimination tests, and an increase in classifier accuracy variance across participants. However, even with the increased variance the results from the high from baseline and high from low test conditions remain above chance at 76.67% (range 66.67-100%), 86.67% (range 66.67-100%) respectively. Moving to the PCA derived normalised dataset (Table 8-8 (b)), the cross-validation of the training data presents with high accuracies and moderate inter-participant variance, with the high from low condition scoring 91.67% (range 66.67-100 %) average recall accuracy. The remaining conditions also score well at 80% (range 66.67-100%) and 83.33% (range 66.67-100%) for the high and low from baseline conditions respectively. However, results from the test and retest classification conditions present a far less positive result, similar to those reported in trials 1 and 2 above.

8.5.4. *Trial 4 Classification of features using subjective ratings*

In this trial the effect of training the SVM classifier using labels provided by the participants during the experiment is explored, the first test (Table 8-9(i)) uses all physiological feature data to create the classification vector and the second (Table 8-9(ii)) features are drawn from those selected by PCA as the classification vector, both tests utilise the subjective ratings taken from participants after every image was presented. In both tests, the baseline period measurement data has been removed to give a discrimination test of high from low; the rationale for removing the baseline comparison data is concerned with the bias implicit to subjective ratings, i.e. there is no guarantee of balance within the class labels provided by the participant. This would reduce the amount of data

available to the classifier as an equal number of high or low class labels are required to compare against the baseline data. Recall accuracies for these tests are then calculated by taking the average classification accuracy for all participants with inter-participant variation in classifier performance included. However, even with the baseline data removed class bias is still integral to the dataset as it is truly representative of the subjective judgments given to the stimulus imagery.

Classification	Raw Data (i)			Principal Components (ii)		
	Train	Test	Retest	Train	Test	Retest
High vs. Low	80.43	70.38	70.83	86.27	67.80	72.92

Table 8-9 Average SVM classification recall accuracy over the three test sessions, for PCA derived features (raw and PCA) n=10.

The results obtained from the first test utilising all raw feature data as a classification vector are presented in Table 8-9, from this Table 8-9(i) there can be seen a marked improvement in classifier recall accuracies over the test and retest conditions in comparison to the standardised material results above. In this case the cross-validation of the training data presents with favourable recall accuracies of 80.43% (range 60-100%). Surprisingly, a favourable result of 70.38% (range 50-83.33 %) and 70.83% (50-91.67%) is also observed in both the test and retest conditions respectively. However, inter-participant variation is higher in these tests, dropping to chance levels for both conditions.

When comparing the principal component analysis derived features and associated subjective labels Table 8-9(ii), with those of the previous analysis, classifier mean accuracies remain largely unaffected. However, a greater variation between participants can be seen, shown as a decrease in classifier stability dropping below chance levels in some cases. In the training condition mean accuracies remain well above chance at 86.27% (range 72.73 -100 %). However, as with previous analyses, classification accuracies diminish in the test and retest conditions at 67.80% (range 41.67-91.67%) and 72.92% (54.17-91.67 %) respectively, showing a greater degree of inter-participant classification variance.

8.5.5. Trial 5 Generalised models

To determine whether the SVM classifier is able to generalise across individuals and across experimental sessions, two tests were completed. These tests aggregate all participant data for each visit (a total of 120 observations per condition) and use the full feature set. In the first test the classifier is trained using the standardised survey labels and associated psychophysiological responses, the second test utilises participant subjective labels as a basis for training the classifier. In both cases the classifier is tested using the novel data from visits 2 and 3. The results displayed

in Table 8-10 show the classifier recall accuracies for the discrimination tests that use the standardised survey labels, as can be seen from this table and indeed from all previous results reported here, in the cross validation test of first visit data the classifier presents with high accuracy for all three discrimination tests, with the high from low test reporting the highest recall accuracy of 81.25%. However, the same sharp decline in classifier recall accuracies reported in previous analyses that train the classifier using the standardised survey labels; for the test and retest conditions are again present here, with the classifier reporting accuracies barely above chance levels for any discrimination test.

Classification	Train	Test	Retest
High vs. Baseline	75	52.5	55
Low vs. Baseline	75	55	54.17
High vs. Low	81.25	54.17	55

Table 8-10 SVM classification recall accuracy over the three test sessions, datasets derived by combining all data from all participants and trained using survey labels

The results reported here indicate, that a SVM classifier trained using subjectively labelled psychophysiological data outperforms one trained using standardised survey labeled data, both in terms of consistent mean recall accuracy reports across visits and classification variance between participants.

8.6. Discussion

One goal of this experimental study was to assess the stability of psychophysiological measures of autonomic activation across experimental sessions using the SVM as a determinant tool. The results from the 5 classification trials show that in this instance, the features of heart rate (iBi mean & standard deviation) and skin conductance (mean and area), show moderate stability as indicators of psychophysiological activation across experimental sessions. However, this stability is only apparent in tests where subjective labels derived directly from individual participants in situ were used to train the SVM classifier. In all trials that involved training the classifier using labels from the survey data, the classifier performed poorly in the test and re-test conditions. Evidence to support this position is apparent from high training data crossvalidation accuracies and a sharp decline (below chance levels in some instances) in the test and retest conditions e.g. trial 1 baseline vs. high; train: 88.33%; test: 53.33%; retest 53.33% (Table 8-7). One could argue for a possible methodological issue involving order effects within the data; however image presentation was counterbalanced across participants, leaving only the effects of training the SVM classifier using different label types, classification model over-fitting, or possible habituation effects on psychophysiological reactivity to explain the decrease in classification performance. Overfitting

occurs when a classification model begins to fit only to the training data rather than to generalising to new data from the presented training examples. However, every effort was taken to prevent classifier over-fitting through appropriate parameter tuning and the use of balanced classes. Furthermore, the decrease in classification performance post training is most apparent in the classification tests involving the training of the classifier using survey labels when compared to the test and retest results classifiers trained using subjective labels, possibly highlighting an interaction between training label type and psychophysiological habituation as the possible cause.

The results show a marked difference when comparing classification accuracies between a classifier trained using survey labels and those provided by subjective judgment. For example, trials 1 (survey) and 4 (subjective) which both utilise raw physiological data and all features to discriminate high activation from low, we see similar high training data crossvalidation accuracies of 83.33% and 80.43%, however it is in the test and retest conditions where the effects become more pronounced; the classifier trained using survey labels reporting 49.17% and 52.50% accuracy for test and retest conditions respectively; however, the classifier trained using subjective judgments shows significant stability reporting 70.38% and 70.83% respectively for the same conditions (Table 8-9(i)). This finding highlights the importance of aggregating a good training data set for supervised classifiers as noted by Novak (2012). Furthermore, these results indicate that training supervised classifiers using psychophysiological data should take into account the subjective judgments of the individual user, that is, good training data is derived from *valid* psychological manipulations and one way to ensure that validity is to calibrate the manipulation to the individual rather than rely on group statistic as proof of validity. These results make a strong argument that a subject-dependent approach should be employed when calibrating a classifier in the context of real-time physiological computing.

The habituation effect represents the decrease in physiological response to stimulus after repeated presentations. Recall, that in this repeated measures study, experimental sessions one and two (separated by a number of days) used the same image stimuli for presentation. In an experiment presenting pleasant, unpleasant and neutral picture stimuli repeatedly Lang, Bradley & Cuthbert (1993), compared the habituation response patterns of heart rate, electrodermal activity and facial corrugator muscle responses. They found that for the first (novel) instance of image presentation, psychophysiological responses showed a high degree of differentiation, all subsequent trials displayed a marked decreased (habituated) response in the indices of psychophysiological variance. This finding holds true in the current experimental context and is supported by high classification accuracies during the training phase followed by a steep drop in accuracy across test and retest conditions for all of the analysis trials completed. However, this effect is only clearly delineated from other possible confounds (such as internal SVM mechanics), in the generalised model trials

(trial 5) in the instance of classifying high from low activation using subjective labels; which shows a moderate habituation response over the full cohort of participants in the accuracy of the classifier for these data reporting 81.44% and 75.42% for the training and retest conditions respectively.

There is a supposition in the literature that applying principal component analysis and normalisation to feature data, to reduce dimensionality and the scale of the data, should potentially provide improved classification performance (Novak 2012; Gudmundsson et al., 2010). This improved performance is only partially demonstrated in the results reported here. The effect of normalising the data resulted in higher crossvalidation accuracies, but poor test and retest accuracies coupled with higher inter participant accuracy variance, mean accuracy 54.16% (range 33.33-66.67%) and 59% (range 33.33-100%) respectively. When considering the results from the classification trials that involve PCA (trial 3), the effect of dimension reduction on the classifier is similar; training condition cross-validation accuracies improve in some test conditions, specifically in the high from low activation discrimination; the same improvement is not observed in the test or retest conditions, however.

Comparing the performance of a classifier trained using the subjective labels, and the full feature set with one trained using the PCA derived feature set (trial 4), mean accuracies over all three conditions compare favourably. However, there can be seen an increased level of inter-participant accuracy variation over the test and retest conditions of 67.80% (range 41.67-91.67%) and 72.92% (54.17-91.67 %) respectively for the PCA test compared to 70.38% (range 50-83.33 %) and 70.83% (50-91.67%) for the raw feature data test. These findings indicate that applying PCA to reduce data dimensionality and increase classifier performance shows promise. Care must be taken however, to ensure that too much information about the physiological response is not removed. For example, a set of features may prove statistically significant in determining proof of “effect”, in terms of variance between conditions. However, when those features are used to train a classifier performance may not be optimal, as the reduced feature set does not represent the actual pattern of physiological response to stimuli (classifiers are essentially pattern recognition algorithms); thus, the application of PCA to feature data may remove elements of the “pattern” of psychophysiological responses and while these elements may be redundant in a statistical sense they may prove essential within the classifier as truer representations or patterns of response to stimuli. For example, the features of respiratory activity were added back into the feature data for comparison with the PCA derived features (trial 4), which acted to stabilise the classifier over repeated sessions with the effect of reduced inter-participant variance in classification accuracies.

Mean accuracy reports can be seen as indicators of classifier performance when comparing a classification methodology across individuals, the results from this study indicate that a truer representation of performance would be a measure of accuracy variance; the lower the variance, the better the classifier will perform when trained to each individual; this is an issue of methodology and parameterisation not computational approach (in terms of type of classifier). The results reported here demonstrate, that higher classification accuracies can be achieved by adopting a subject-dependent methodology where the vagaries of an individual's psychophysiological response are integral to the training and use of the classifier, when compared to classifiers built with generalised deployment purposes in mind. Classifiers that generalise across populations are the "Gold Standard" in the fields of affective and physiological computing; however the purpose of a physiological computing system generally, is applied to individuals and specific contexts. Therefore, any classification methodology that must adapt to and classify an individual's psychophysiological responses, would require a calibration period were the classifier is trained at runtime, using both the psychophysiological responses and associated subjective judgments before deployment in a real-time task context in order to perform optimally.

Overall, the results from this study indicate that training classifiers using a subject-dependent approach to classify user interest using measures of autonomic activation can prove successful. However, this success is contingent on the aggregation of good training data that is wholly reflective of the individual's response to stimuli. The modest level of classification accuracy reported here may reflect the moderate level of psychological stimulation provided by passive viewing of a still image.

8.7. Conclusion

This second experiment focused on classification within the context of psychophysiological responses to image material standardised by survey, using the support vector machine algorithm. In this study, the image presentation experimental paradigm was extended to include cultural heritage material (paintings), and a comparison of classifier accuracy when trained using labels provided by internet survey and subjective responses to the same stimuli. The results from this study and those of study one clearly indicate that training the SVM classifier using standardised labels meet with less than optimum results. However, when the SVM classifier is trained using subjective judgments, both the accuracy and the stability of the classifier is improved, and in the current study this improvement in classification performance can be observed across the training, test and retest conditions.

The application of the SVM to data obtained during the retest session was included to assess the stability of the algorithm over time. It may be argued that autonomic reactivity is greatest during initial exposure to a set of stimuli and reactivity will decline with subsequent exposure due to familiarity, habituation etc. In the case of these data, classification accuracy fell by approximately 10% for all comparisons between the training and retest sessions. This finding suggests that habituation towards repeated stimuli may indeed be a major factor, leading to classification instability over repeated long interval test sessions with the same stimuli and raises the question of when to train the classifier.

It is apparent from these results that PCA is useful tool for identifying uncorrelated features within psychophysiological feature data i.e. those features that will potentially lead to more accurate classifications, such as heart rate inter-beat interval mean and standard deviation, skin conductance mean and area in this case. However, it can be seen from the results that utilising PCA derived features *as* feature sets for use in training and testing SVM classifiers results in an unstable classifier, from which classification accuracies can vary considerably between participants.

It seems clear from the experimental results that classification accuracy tended to decline sharply when the SVM was applied to material that was similar but different to the training set. The exception to this finding was the classification of the subjectively labelled data, which showed a less dramatic decrease of 10% between the test and retest conditions relative to the training condition. With the exception of the training validation tests, in the case of the survey labelled datasets, no test performed significantly above chance levels for the test and retest conditions. Two conclusions can be drawn from a comparison between the test and retest sessions: (a) train the SVM using material that is representative of test material if possible, and (b) train and apply the SVM on the same day as testing to control for intra-individual differences.

9. Study 3: A Virtual Heritage installation²

9.1. Abstract

The goal of this experimental study was to examine the cultural heritage experience then posit a three dimensional psychological model of interest as a potential driver of this experience and operationalise the model as multiple measures of psychophysiological activation. A three dimensional model of interest consisting of activation, cognition and valence was developed based upon a distillation of the four factors of cultural heritage experience described by Pine and Gilmore (1998). The interest model was then operationalised using psychophysiological measures to derive features of autonomic, cognitive and emotional activation, these data were then used to train and test a SVM classifier using both a subject-dependent and subject-independent classification methodology. Ten participants a mixture of students and cultural heritage patrons took part in study which used genuine cultural heritage material in the form of audio narratives presented in a simulated cultural heritage environment. The results show that in this instance the combination of psychophysiological interest with the SVM algorithm provided accurate and reliable classification using a subject dependent approach.

² This work was published Karran, A.J., Fairclough, S.H., K Gilleade "Towards an adaptive cultural heritage experience using physiological computing" In proceedings of: CHI 2013 "Changing Perspectives", Volume: CHI 2013 Extended Abstracts, April 27, May 2, 2013, Paris, France

9.2. Introduction

In the previous studies the results showed that the support vector machine classifier performance was good, providing the classifier was trained using data that was generated by the individual. However, these studies used material generated for laboratory based studies and focused on recording measures of activation as captured by a range of autonomic measures, and therefore this chapter will move towards a test case for cultural heritage in terms of setting and stimulus material.

Psychophysiology, physiological computing and machine learning can provide a unique way to operationalise the covert psychological experience of media by measuring, analysing and classifying psychophysiological responses. Physiological computing systems monitor the physiology of the user and use these data as input to a computing system (Fairclough, 2009). The passive monitoring of spontaneous changes in physiology indicative of cognition, emotion or motivation is used to adapt software in real time. These systems are constructed around a biocybernetic loop (Fairclough & Gilleade, 2012) that handles the translation of raw physiological data into control input at the interface. Passive monitoring of user psychophysiology can be used to inform intelligent adaptation, thus permitting software to respond to the context of the user state to deliver personalised media. For example, an application designed to deliver media to users within a cultural heritage (CH) environment. In this environment the physiological computing system could monitor the CH experience in real-time by quantifying and classifying the psychological state of the visitor and using these data to personalise the experience by autonomously adapting information to provoke a state of interest or detect states reflecting low interest. To perform this act of personalisation, the physiological computing system must be sensitive to the psychological dimensions that underpin a CH experience.

9.2.1. *The cultural heritage experience*

De Rojas and Camarero (2008) described the cultural heritage experience in terms of satisfaction, which in turn is determined by positive expectations of the visitor being fulfilled. The optimal CH experience has been defined in conceptual terms as a “total experience” that incorporates aspects of leisure, culture and social interaction (Pine & Gilmore, 1998, De Rojas & Camarero, 2008). There are several routes to the creation of memorable experiences. The visitor may supply a cognitive and emotional resonance by actively encoding the visit with their own personal meanings.

The analysis of cultural heritage experience described by Pine and Gilmore (1998) provided a deeper level of analysis by describing four crucial drivers of visitor experience:

- entertainment (leisure, narrative)
- educational (knowledge transfer)
- aesthetics (pleasure)
- escapist (immersion)

The first factor refers to capacity of cultural heritage artefacts to engage the visitor in a cognitive and affective manner. The educational component of the CH experience represents the process of knowledge transfer by which the visitor is informed about artefacts. The aesthetic aspect of cultural heritage is perhaps the most difficult to understand because cultural artefacts are capable of evoking a range of aesthetic responses. Previous definitions of aesthetic experience have emphasised both information processing and emotional responses (Leder et al, 2004), i.e. a cognitive perceptual process accompanied by a dynamic affective state.

The final factor (escapist) is associated with the degree to which the visitor is immersed within a mixed reality (i.e. past – present, new technology – ancient artefact). The concept of immersion is often associated with a sense of presence in a three-dimensional virtual reality (VR) (Russell, 2003); however, the same concept may be applied to mixed reality systems such as augmented reality. The degree of immersion may be characterised within three levels: (1) engagement (lack of awareness of time), (2) engrossment (lack of awareness of the real world), and (3) total immersion (sense of being within a computerised environment) (Jennett et al, 2008). Immersion has clear implications for creating stimulating experiences in CH contexts, particularly using technology to engage and engross the visitor in a particular artefact.

The experience of a cultural heritage environment, regardless of whether it is a museum or gallery, is shaped by exploratory behaviour driven by the interest and curiosity of the visitor. The next section details a conceptual model of interest distilled from the study of the CH experience.

9.2.2. Conceptual model of interest

The concept of interest as a psychological entity was described by Berlyne (1960) in terms of increased arousal and sensation-seeking, i.e. objects inspire curiosity via novelty and emotional conflict. This concept was expanded by Silvia (2008, 2010) to incorporate a cognitive dimension, i.e. interest driven by stimulus complexity. Both cognitive and emotional facets of interest were explored by Hidi and Renninger (2006) who referred to the former as perceptual/representational processes, which was accompanied by a sense of positive emotion derived from intellectual engagement; they argued that positive emotion occurred even during engagement with negative material.

The proposed model of interest distils the four elements of the Pine and Gilmore into two important elements, cognitive factors (education and knowledge transfer) and affective influences (aesthetics). Cognitive factors are defined here as stimulus features that drive the curiosity of the viewer, such as novelty and complexity, whereas affective influences are defined in a two dimensional space, similar to the circumplex model of Russell (1980) as activation and valence. It is proposed that activation, cognition and valence serve an interactive role in the CH experience with cognitive stimulation playing a primary role in the educational aspect and activation and valence capturing the emotional and aesthetic aspect of visitor experience.

The model consists of three dimensions of perceptual representational processes, which are mapped onto a unidimensional scale ranging from high to low interest:

- Cognition, which captures the novelty and complexity of the stimuli i.e. familiarity vs. unexpectedness and intricacy vs. simplicity
- Activation, which captures how stimulating the stimuli is
- Valence, to capture the level of positivity or negativity towards the stimuli

9.2.3. Operationalising the model

The cognitive component of interest is identified with activation of the rostral prefrontal cortex i.e. Brodmanns area (BA) 10, which has been linked to working memory and attentional control (Ramnani & Owen, 2004). BA 10 has also been associated with a wide range of cognitive process, ranging from the selection and judgement of stimuli held in short term memory (Petrides 1994) to

reversal learning and stimulus selection (Dobbins et al 2002); of specific import to the interest model is the association with the ‘elaboration encoding’ of information into episodic memory (Henson, et al 1999, Wagner, et al 1998), another area of import that overlaps this region of the prefrontal cortex is BA 8, a part of the medial pre-frontal cortex, that has been associated with processes that involve the motivational or emotional value of incoming information (Tataranni 1999, Rolls 2000) and a link proposed between asymmetry of frontal alpha activation and emotional states (Davidson 1990).

Operationalising the cognitive and valence components of the interest model using this research as a template, gives four cortical regions, that can be mapped simply using the international 10-20 system (Jasper, 1958), for cognitive activation FP1 and FP2 corresponding with BA 10, for valence F3 and F4 corresponding with BA 8. These can be seen as potentially encapsulating the capture of responses to CH material, equating to the aesthetic and educational elements of the Pine and Gilmore (1998) model of CH experience. The measurement of cognition is captured using spontaneous measures of electrocortical activation (EEG), it has been shown that there is an inverse relationship between the level of alpha activity and brain activation (Goldman et al, 2002), i.e. higher alpha activity is associated with reduced brain activation, thus cognition becomes a ratio derived from activity in the beta band (12-30Hz), divided by activity in the alpha band (7-11Hz) at each site. Capturing valence will be through the level of frontal hemispheric asymmetry expressed as a ratio, subtracting right from left hemispheric alpha band activity. It has been hypothesized that greater left activation of the prefrontal cortex is associated with positive affect whereas greater right side activation is linked to negative affect (Davidson et al, 1990, Henriques & Davidson, 1990, Lang, 1995, Silbermann & Weingartner 1996, Davidson, 2004).

The activation component is captured via the level of skin conductance (SCL) and supplemented by measuring heart rate (HR); SCL is highly sensitive to sympathetic activity (Boucsein, 1992) and HR captures both sympathetic and parasympathetic components of the autonomic nervous system. This array of physiological measures is designed to deliver a multidimensional representation of the psychological state of interest, to quantify the interest level of an individual in a dynamic fashion.

9.2.4. *Study Goals*

The aim for this study was to build and assess the performance of a subject-dependent classifier trained using psychophysiological responses to audio commentary in a cultural heritage scenario. The goal of the classifier was to discriminate between those audio segments where interest was high. To achieve this goal, subject-dependent classification was used to examine the impact of psychophysiological response data from each dimension of the interest model singly or in combination upon classification accuracy.

The study was designed with a threefold purpose:

1. To measure and classify psychophysiological reactivity in response to cultural heritage content presented as image and audio stimuli
2. To classify the psychophysiological variance as a two condition level of interest (high or low)
3. To evaluate the accuracy of the SVM classifier output when compared to subjective response data

9.3. **Methodology**

9.3.1. *Participants*

Ten participants 8 female (aged 19-75) took part in the experiment, a mixture of undergraduate/graduate students at Liverpool John Moores University and patrons of a heritage institution. In accordance with the universities lease of ethical approval participants signed a consent form and were in good health.

9.3.2. *Experimental Design*

The experiment was designed as a repeated measures, laboratory investigation i.e. participants were exposed to a digital image reproduction of a CH exhibit and audio narrative, investigating the independent variable(s), level of interest with two conditions: High and Low interest with dependent variable(s), heart Rate (HR), respiration (RSP) skin conductance (SC) and electrocortical activation (EEG).

9.3.3. *Apparatus and Experimental Measures*

Physiological responses from the autonomic system were measured during experimental sessions, using the Electrocardiogram (ECG, sampled from the torso) and SCL (distal phalanges, second and forth finger, non-dominant hand) channels of the Mind Media Nexus X Mk II (sampled at 512Hz). Four channels of electroencephalographic (EEG) data were recorded, measuring alpha (11-12Hz) and beta (13-30Hz) activity, using the Enobio wireless 4-channel sensor (sampled at 250Hz) with ground contacts on left ear lobe and inner ear (Starlabs Inc). A Biosemi EEG cap was fitted and aligned to ensure sensor placement, electro-conductive gel was added to sites FP1, FP2, F3 and F4 and electrodes attached. Recorded data was saved to digital file and exported to AcqKnowledge 4.2 (BIOPAC systems Inc.) for processing.

9.3.4. *Task definition*

Participants were asked to stand in a relaxed position approximately 2 meters in front of a 3*2 meter projection screen, giving an image size of approximately 103 inches in width and 78 inches in height, giving a 130 inch 4:3 aspect ratio screen. This was followed by the audio-visual presentation of the Valencia kitchen, lighting was dimmed throughout the presentation and audio was reproduced via a Dolby 5.1 surround sound speaker arrangement, at moderate easy listening volume (approx. 30dB). The presentation of the kitchen stimulus was linear and timed to progress through the narrative, giving four stories (average 17s in length) consisting of 3 factual elements. The audio commentary was divided into four ‘stories’ consisting of three discrete ‘facts’. The four stories were composed around elements in the still image *refreshments*, *the Lady of the House*, *the ceramics* and *the dog*. To draw the gaze of the viewer specific fragments of the mosaic were highlighted (see Figure 9.1). When the presentation was completed each participant was asked to rate which two stories were perceived to be the most interesting out of the four that were presented.



Figure 9.1 The still image used in the experiment with highlighted sections that were linked to the audio

The full text for each story is presented below:

Story 1 –Refreshment?

1. According to the sources of the time a “refreshment” was a snack or light meal served in the afternoon with drinks, sweets and chocolate.
2. “Refreshments” were served at social functions in upper-class families by the household servants. To the right of the Lady you can see a male servant is carrying the first dish of the depicted “refreshment”.
3. The dish the servant is holding is called a salver or “footed dish” which is reserved for fruit, sweetmeats, marzipans or sorbets.

Story 2 – The Lady of the House

1. For the occasion the Lady of the house is wearing a French style dress comprising of: a jacket of long tails, a bodice and a petticoat made of a silk cloth.
2. The dress fabric is decorated with a motif of interlaced flowers and leaves, and the cuffs and edges are trimmed with lace. The Lady is also wearing a fichu, or kerchief, around her shoulders to compliment her dress.
3. This fashion style was very popular in Spain at the time and can be seen today in the traditional costume of Valencia.

Story 3 – The Kitchen

1. In Valencia, at the end of the 16th century there was a stable production on painted tiles which were used in the construction of the Kitchen.
2. By the 18th century Valencia had become famous for the production of these tiles. From the ovens of Valencia's factories came tiles decorated in a variety of styles including baroque and rococo aesthetics.
3. The scenes depicted are painted with a trompe l'oeil effect. Today we categorise these decorative walls as "traditionalist"

Story 4 – The Family Pet

1. To the right of the Lady a dog can be observed trying to attract her attention.
2. Dogs were commonly kept as pets in upper-class households unlike cats which were used as a means of pest control.
3. The dog's collar was used to symbolise the social status of the household, and so the more elaborate the collar the more important the family.

9.3.5. Procedure

After receiving instruction about the experimental procedure, participants were asked to complete a consent form in accordance with the Liverpool John Moores Ethical Committees lease of ethical approval. Participants were then fitted with a wearable pouch to hold the nexus sensor hardware at the hip. Electrodes for ECG were placed on the torso. The Biosemi sensor cap was fitted and electrodes attached. Participants were asked to stand in a relaxed position approximately 2 meters in front of a 2*3 meter projection screen and lighting was dimmed. The still image of the kitchen was displayed onto the projection screen using a ceiling-mounted projector. This was followed by the audio-visual presentation of the Valencia kitchen. The presentation of the kitchen stimulus was linear and timed to progress through the narrative; giving four stories consisting of 3 factual elements (see Figure 9.2). On completion of the presentation each participant was asked to rate which two stories were perceived to be the most interesting out of the four that were presented, these ratings were subsequently used as class labels within the SVM classifier.

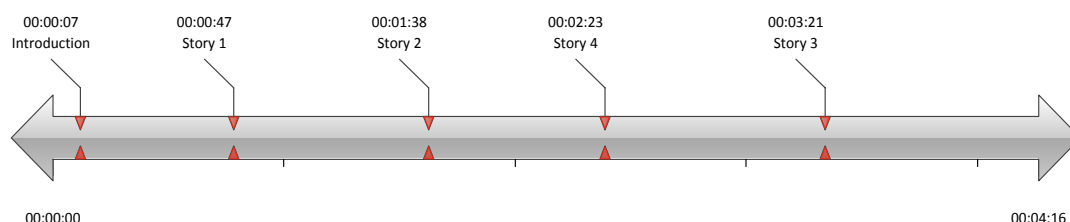


Figure 9.2 Experimental stimulus timeline

9.4. Analysis

EEG (beta/alpha ratio) and autonomic data (mean and standard deviation of IBI (HR) and SCL) were extracted from an epoch that equated to each of the three facts (average 17s) related to each story. Based on the subjective assessment of the participant, six of these facts were classified as ‘high interest’, i.e. they were associated with the two stories assessed to be of most interest for that particular individual. All feature data was derived from raw signal output using AcqKnowledge 4.1 (BioPac 2012).

9.4.1. Feature Derivatives

Autonomic measures (activation) of heart rate and skin conductance level (SCL) were collected. A three-lead electrode connected to the chest was used to capture an ECG signal, which was filtered between 0.5 and 35Hz. The SCL signal was filtered at 35Hz only. Heart rate was captured as the mean and standard deviation of the inter-beat interval (IBI); the same descriptive statistics were used to represent SCL.

EEG data were collected using four channels. Dry electrodes (i.e. no gel) were placed at FP1 and FP2 on the forehead. Electrodes were also placed at F3 and F4 using a small amount of electro-conductive gel and an electrode designed to make contact through hair. The resulting EEG signals from all four channels were filtered at 0.05 and 35Hz. These data were subjected to a power spectral density analysis (Hanning window) to yield power in the *alpha* (8-12Hz) and *beta* (13-30Hz) bands. A ratio measure of cortical activation was obtained (*beta/alpha*) where a higher number is equated with increased activation, i.e. alpha activity = inverse of cortical activation. All features were derived from a single stimulus epoch representing a single fact (approx. 17 seconds). For Cognition: Where the ratio : x is expressed as β (power) divided by α (power) at sites (fp1/fp2) and (f3/f4).

$$: x = \left(\frac{y_{\beta}^i}{y_{\alpha}^i} \right)$$

For Valence: Where the ratio : x is expressed as the natural log of α (power) subtracting right from left hemispheric activity at sites (FP2-FP1) and (F4-F3) (Coan & Allen, 2003).

$$: x = \ln(z_{\alpha}^i) - \ln(y_{\alpha}^i)$$

All features were extracted from a stimulus epoch that equated to the length of each of the three facts related to each story, these feature derivatives represent continuous values and not traditional change score from baseline, this approach was purposeful to be more representative of a real-time environment.

Measure	Derivative			
Heart Rate	<i>iBi-Mean</i>	<i>iBi-Stdev</i>		
Skin Conductance	<i>Mean</i>	<i>Stdev</i>		
EEG	<i>Ratio β/α FP1</i>	<i>Ratio β/α FP2</i>	<i>Ratio β/α F3</i>	<i>Ratio β/α F4</i>
	<i>Ratio α FP1-FP2</i>	<i>Ratio α F3-F4</i>		

Table 9-1 Features derived from physiological recordings

9.5. Results

Prior to commencing classification analysis using the psychophysiological data, features were derived from measures of heart-rate, skin conductance and EEG (see Table 9-1). This resulted in a total of 10 features for each of the 12 stimulus events. These features were further subdivided into the three components interest model, activation 4 features (*HR, iBi mean and standard deviation; SC, mean and standard deviation*); cognition 4 features (beta/alpha power at sites FP1, FP2; F3, F4); valence 2 features (alpha power FP2-FP1; F4-F3), such that each set of psychophysiological features created a unique classifier feature vector for each of the components.

The analysis was completed in two stages, subject-dependent testing to determine the recall accuracy of the SVM classifier for individual participant responses and subject-independent testing to test the generalizability of the SVM models across the population of participants. The SVM classifier is a supervised pattern recognition algorithm, requiring an n dimensional vector (observation) and an associated label (class) for training. In this instance based on the subjective assessment of the participant, six facts (two stories) were classified as ‘high interest’, i.e. they were associated with the two stories assessed to be of most interest for that particular participant were used as class labels. This training set was then used as the basis for classifying new instances of data into its respective class. For this experimental study the SVM implementation within the *matlab* 2012Rb bioinformatics module was used and optimal parameterisation was achieved using a loose grid search algorithm. Each feature set was tested using the *hold-out* cross-validation method given in algorithm 1 see chapter 8 pp63.

This approach has a number of advantages, each feature vector is identified as a separate element of the model; feature sets can be combined as a fusion of features; and the effect of each feature set or fusion of features on classifier class recall can be evaluated for both subject dependent and independent SVM models. Fusion refers to the combination of feature data (Novak et al. 2012) into a vector that represents either single or multiple dimensions of the interest model. Table 9-2 displays the feature sets, subject-dependent classification accuracy, mean recall accuracy and stability (as standard deviation) of the classifier for each fusion of features, for each participant and

each dimension of the interest model. Feature sets are denoted by: *A* (activation); *C* (cognition); and *V* (valence).

Feature(s)	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Mean Recall	Stdev
A	83	83	83	100	100	83	83	83	100	100	90	8.17
A, C	83	67	83	100	100	100	100	83	100	100	92	11.18
A, V	100	83	83	100	100	83	100	100	100	100	95	7.64
A, C, V	100	83	100	100	100	83	83	83	100	100	93	8.17
C, V	100	83	83	83	100	67	83	83	83	100	87	10.00
C	100	83	83	83	67	83	67	83	83	100	83	10.54
V	100	83	67	83	83	67	67	100	100	100	85	13.84

Table 9-2 Classification recall accuracy (%) for all participants presented across each source of psychophysiological data using holdout crossvalidation

The feature sets (activation, cognition and Valance) were classified alone and in combination, to determine which permutation of features provided the best class recall accuracy over all participants. The data table indicates that the combination of activation and valence features afforded the best mean classification recall accuracy of 95%. Similarly, the combination of activation and cognition or all three components together performed well with 92% and 93% respectively, showing a negligible difference in recall accuracy between these three feature vectors.

As discussed previously in the second experimental study another way of viewing classifier performance and methodological validity is to examine the inter-participant variation within recall accuracies, Table 9-2 shows that classifier performance is most stable when the SVM is tasked with classifying three feature vector combinations; activation features alone resulted in a 90% classification accuracy and low variance (σ 8.17); the combination of activation and valence 95% (σ 7.54) and finally the combination of activation, cognition and valence 93% (σ 8.17). These results indicate that the classifier is most accurate and stable across individuals in situations where multi-dimensional feature data is used when compared with data from single dimensions of the interest model.

Moving to the test of generalisation, which assesses the ability of the classifier to generalise from the population to new individuals, the results in Figure 9.3 show a sharp decrease in classification recall accuracies for all component features and fusions of features that represent the interest model, in this generalised SVM model the combinations of activation and cognition, activation cognition

and valence report the highest accuracies of 66.1%, while single dimensions of the interest model suffer lesser recall accuracy.

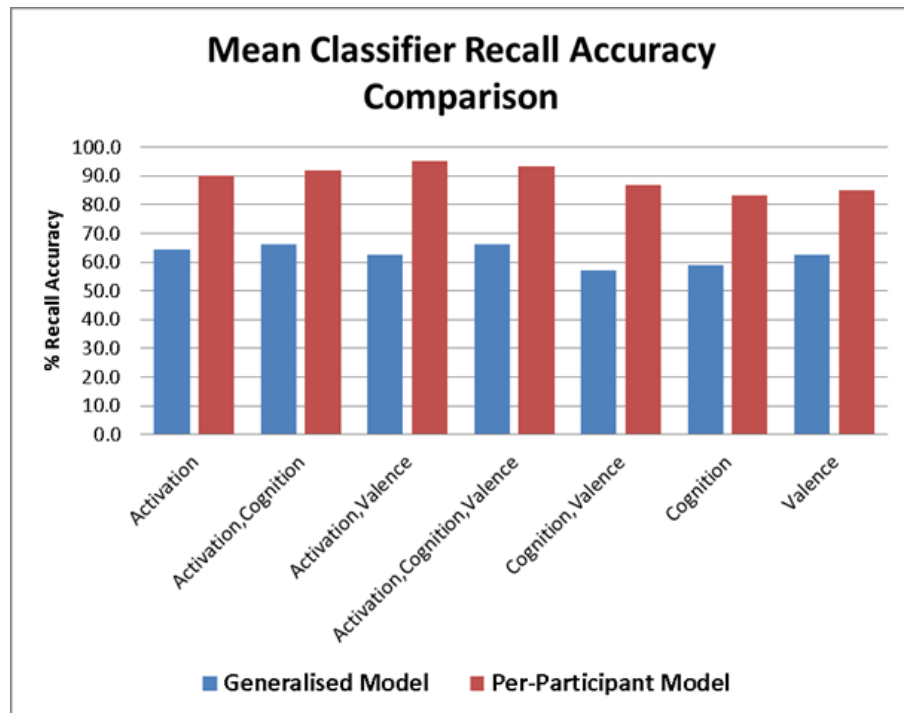


Figure 9.3 Mean Classifier Recall Accuracy comparison generalised model vs. subject dependent model

It is clear from this figure that in this instance, the subject-dependent classification model provides the greatest level of classification accuracy when compared to the generalised model, highlighting the strength of the subject-dependent methodological approach.

9.6. Discussion

In this study a multidimensional psychophysiological model of interest was tested in response to audio material in a cultural heritage context, the results showed that the subject-dependent classification accuracy of high/low interest was high and inter-participant accuracy variance was low for three combinations of feature data. The combination of autonomic (activation) and EEG frontal asymmetry (valence) gave the greatest accuracy (95%), followed closely by the feature set that combined all three (activation, cognition, valence) features of the interest model (92%). However, the classification of autonomic features alone reported a significant 90% accuracy and low inter-participant variation. Furthermore, the results showed that the generalised model produced a significantly lower level of classification accuracy when compared to the subject-dependent model. This study represented an evolution of methodological approach taken in the previous studies, whereby the derived psychophysiological data was used without the use of any processing techniques (such as normalisation). One of the major problems underlying normalisation is that the range of psychophysiological responses may fluctuate significantly due to inter-individual differences (personality) and intra-individual variability (e.g. transfer effects, time of exposure, time of day). The results show that the subject-dependent classification approach taken in this study negates these confounds by embracing these differences as co-factors within the classification context, that is, the variability *is* the pattern of response for that individual and classifiers are built to suit the individual.

With respect to the classification results, it was apparent that accuracy was determined by the feature and source of psychophysiological responses used to train the SVM i.e. features specific to the cognitive, autonomic or valence components of the interest model. The combination(s) of metric (feature) and source (component) yielded positive classification accuracies with low variability across individuals. Closer inspection of the classifications accuracies for each individual revealed that autonomic activation produced the best accuracy for any single psychophysiological dimension of the interest model when compared with cognition and valence. However, when the classification variability across all participants is taken into account, the combination of features derived from cognition and valence subcomponents delivered highest recall accuracy and lowest variance across all permutations.

This study was conducted to investigate the suitability of psychophysiological classification as a means of adapting information in a cultural heritage context such as a museum. The results are positive in this respect, particularly for the subject-dependent model. With regard to the model of interest, activation (Table 9-2) was the strongest single predictor of participant interest levels, and this could have been due to a high level of stimulation provided by the audio narrative. This finding is in line with research from (Bradely et al., 2008, Codispoti et al., 2008, Demaree et al, 2004) who concluded that measures of the autonomic nervous system are sensitive to some stimulus types such as video and audio-visual presentations. Furthermore, these results are very promising from a cultural heritage standpoint, when viewed from the perspective of a museum in which participants are stationary but standing naturally while viewing / listening to material about exhibits.

The results reported here support the viability of subject-dependent multidimensional measurement and classification of interest using a machine learning classifier, e.g. in the case of classifying the independent measures autonomic activation and EEG asymmetry, which when combined provided the highest level of classification accuracy and stability across participants. This finding is supported in the literature where it is found that subject dependent approaches are generally more accurate than subject independent approaches (Novak 2012).

A potential weakness which may have enhanced the classification results was the “forced choice” within the subjective assessments of interest, that is, participants were asked to give a binary decision of interest or no interest. In other words participants were required to make a binary choice, which may not be a realistic portrayal of a subjective assessment given in a cultural heritage environment. Furthermore, the task scenario itself was idiosyncratic in that a participant’s subjective preference was used to train classifiers as opposed to standardised scoring which may have favoured a subject dependent approach.

While the generalised model classification results are poor in comparison to subject-dependent results, it remains a subject of further study as to whether this is an artifact of small training data set size. The developed experimental methodology is geared towards real-time deployment and application within a CH context, which necessitates the use of small datasets, in such datasets individual differences may be more pronounced with a larger proportion of outliers when compared to a training set with a high number of participant data aggregated. In a larger data set the inter-personal variation may possibly be averaged into appropriate ranges allowing for better differentiation between classes of physiological response and thus greater classification accuracies. However, aggregating a large dataset for this purpose would require a large n which may not be practical for real-time classification where individual preference recognition is the primary goal.

These results provide support to a subject-dependent multidimensional modelling approach for interest and the use of psychophysiological measures that are relatively independent of one another i.e. electrocortical vs. autonomic activity. When taken as a whole, the results from the multidimensional classifications are very promising, providing evidence of a possible many to one inference between a psychological state of interest and indices of physiological activation. However, whether this represents the fact that the measures are indicative of a correctly classified state of interest or merely an effect of this experimental study alone requires further research.

9.7. Conclusion

This third experimental study focused on psychophysiological classification within the context of a virtual cultural heritage exhibit. This study provides a second step towards a framework that utilises psychophysiological data and a psychological model of interest for use in a real-time physiological adaptive system that could potentially be deployed within cultural heritage institutions to adaptively curate information according to a visitor's level of interest towards artefacts or exhibits.

A three dimensional model of interest consisting of activation, cognition and valence was posited based upon a distillation of the four factors of cultural heritage experience described by Pine and Gilmore (1998). This interest model was then operationalised using psychophysiological measures to derive features of autonomic, cognitive and emotional activation, these data were then used to train and test a SVM classifier using both a subject-dependent and subject-independent classification methodology.

The results show that in this instance the subject-dependent classification model provided higher classification accuracies when compared with the independent model. These high classification rates provide some evidence in support of a psychophysiological operationalisation of interest within an experimental context. The combination of psychophysiological interest with the SVM algorithm provided accurate and reliable classification using a subject dependent approach. Moreover, the results show that it may be possible to utilise the operationalised model of interest in the field, using ambulatory sensor hardware and the SVM classification algorithm, to provide the basis for a real-time adaptive physiological computing application.

10. Study Four: Liverpool FACT Study – Classification of multimodal cultural heritage material

10.1. Abstract

This fourth experimental study focused on psychophysiological classification within the context of in situ measurement of psychophysiological indices of interest using cultural heritage material presented as mixed media (audio, text, images and video). Responses from 8 participants all patrons of a cultural heritage institution were recorded and a process pipeline was developed to analyse these data and output vectors suitable for classification retrospectively. A framework (Interest as Binary or Interest as State (IBIS)) for a biocybernetic loop was proposed to take in psychophysiological measurement at one end and output classifications of user interest at the other. Two classifier training protocols utilising subjective judgements as classifier training labels were proposed and tested, and the results showed that including subjective judgements in the classifier training process results in high accuracies and stable classifiers. Furthermore, the results show that in this instance combining the features activation and valence of the interest model provided the highest classification rates.

10.2. Introduction

Psychophysiology, physiological computing and machine learning can provide a unique way to operationalise the covert psychological experience of cultural heritage material by measuring, analysing and classifying psychophysiological responses towards cultural heritage artefacts. If classification can be achieved in real-time a physiological computing system could be created where information provision such as the type or depth of information is personalised to the individual in a form of adaptive curation. For example, content for a museum audio guide could be selected based on the listener's psychophysiological responses to topical keywords. Whenever the listener responded with interest to a topic an adaptive guide pushes new content based on this theme. If the listener responded with no interest the system would skip this topic. In this scenario, the adaptive system requires only a binary classification of interest to be effective.

To perform this act of personalisation, the physiological computing system must be sensitive to the psychological dimensions of interest and to the user's subjective judgement of interest. Measuring the dimensions of interest may be complicated by the different media used to convey information to the user, e.g. voice narration, still images, video and combinations thereof. However, in order to create this type of adaptive system, the process of psychophysiological measurement and classification must be conceptualised within the context of a working system. Interest is measured with respect to three psychological dimensions (activation, valence, cognition) and these dimensions may be described as dichotomous (high vs. low, positive vs. negative) consisting of a combination of signals from EEG, SCL and ECG which are used to generate features for each dimension. Each signal has its own frequency range and minimum time window to provide a sensitive response, coupled with distinct stimulus epochs, for example an audio narrative lasting 30 seconds split into six 5 second windows (to correspond with a SCL response).

Operationalising the interest model within a working biocybernetic loop for cultural heritage applications requires a framework that takes these dichotomous physiological inputs and outputs binary classification judgements in a format useable in an adaptive systems context. One issue with providing inputs for system adaptations is how to turn a multimodal representation of a user interest state into binary classification output that is diagnostic of the interest state. Another issue concerns the stimuli itself, in that dynamic media (such as audio narrative or video) may cause interest levels to fluctuate; one strategy for dealing with this could be to increase the frequency of classification to enable the system to respond faster to those dynamic fluctuations. These issues increase in complexity when the subjective judgement of user interest is taken into consideration when training classifiers, previous studies utilised standardised or "forced" choice labels (i.e. which stories were most interesting) to train classifiers. However this method of forcing a binary choice or

relative choice on the user may have been restrictive, in that the system is making frequent binary judgements and in doing so may distort the nuanced response from the user. This issue could be addressed by capturing a more reasoned subjective response from the user before or while the system is in use and applying these responses during classifier training.

Classifications can then be output based on how the classifier is trained, such as a singular binary output representing an overall interest rating (high or low) or multiple classifications representing each component of the interest model which are then combined to project interest upon a scale. However, the type and number of classifications would have a significant impact on the adaptation model needed to provide the cultural heritage experience which is informed by the rationale of the system, such as to provide infotainment or a “memorable” experience. Capturing this more reasoned response from the user could be performed by using Likert scales during the classifier training process in order to increase a user’s freedom of expression and capture a more nuanced subjective assessment of interest. However, while this captures the user experience more effectively, it presents issues for the system which must then decide how to partition these data into high and low categories.

10.2.1. *The Interest as Binary - Interest as State (IBIS) Framework*

The framework developed to integrate the operationalised interest model in a working system is shown in Figure 10.1. At its core the framework utilises two classification methods, which receive psychophysiological input from three component processors, each component processor represents one dimension of the interest model e.g. activation, cognition and valence. When the psychophysiological data is classified, the outputs can represent either Interest as Binary (low or high) or Interest as a discrete State (low to high). In a real-time context, inputs from the physiological sensors are forwarded to the component processors; these processors derive features from the physiological data to create feature vectors used within the classifiers; the feature vectors are then associated with a training label derived from a user’s subjective judgement and output to the classifier; feature vectors are either truncated into a single “composite” classification vector used to classify interest as a binary state, or expanded using a classification from each component to create three binary classifications (activation, cognition and valence) which would apply propositional logic to create a series of discrete interest states. In this framework the classification process can be seen as an interpretive layer in which classifications can be a composite (as a binary high or low state) or a component (as in three discrete classifications of high or low states), it is upon the adaptation model to utilise these outputs effectively.

The IBIS framework enables two classifications of the interest state to be completed concurrently as a *composite* model (single classification vector) and a *component* model (multiple classifications,

single discrete state). Using a cultural heritage exhibit as an example, psychophysiological signal data is captured from a user and processed to produce feature vectors, the exhibit experience consists of 20 second narrations and associated video. In this scenario the classifier is built using a single composite classification vector which comprises of the psychophysiological measurements (activation, cognition and valence) and a training label derived from a subjective rating of “interest”, thus the composite model delivers one classification for each 20 second stimulus segment, representing an “overall” interest classification; the component model classifiers are built using the psychophysiological measurement of activation, cognition and valence individually with a training label derived from a subjective judgement given for each component of the model to deliver 3 classifications each 20 second segment to represent interest upon a scale. That is, logic is applied to classifier output in the form of IF activation = high AND cognition = high AND valence = high THEN interest = high, and so forth.

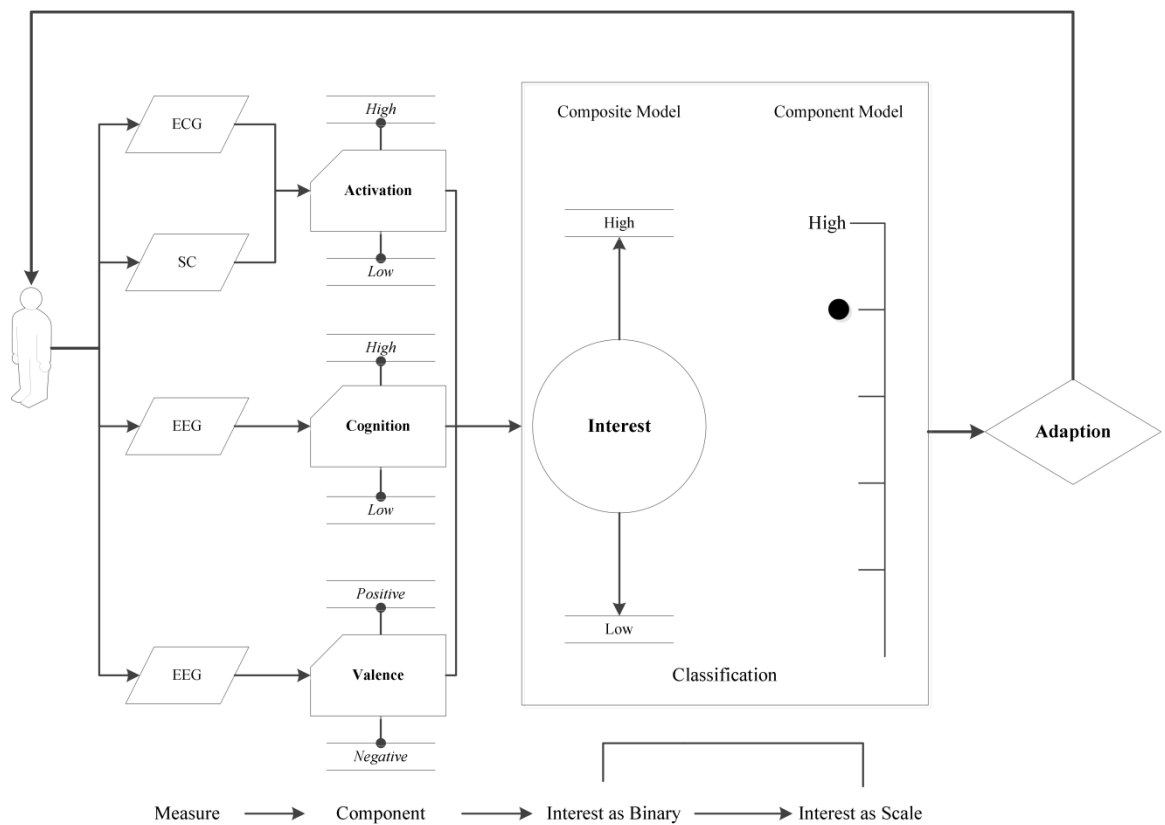


Figure 10.1 The Interest as Binary, Interest as Scale (IBIS) classification framework

In operation within a cultural heritage context psychophysiological measurements are taken from sensors attached to the user and then processed into classification vectors, these vectors are used to train the classifier which outputs classifications as either a single binary state or multiple classifications with a majority vote leading to a discrete interest state ranging from very low to very high interest.

10.2.2. *Psychophysiological Signal Processing Pipeline:*

For this experimental study, a feature extraction processing pipeline was developed to duplicate the methods required in the IBIS framework to import raw signal information from digital storage and output feature vectors for use in training and testing the classifiers used to classify the interest state. The goal was to remove the reliance on external signal analysis software; previously base methods for signal analysis, measure derivatives and feature vector creation were completed using the Acqknowledge (Biopac Inc.) signal analysis software and a spread sheet.

Figure 10.2 shows the data flow of pipeline processes used to output classification feature vectors. The pipeline starts with a module to import the physiological data; this module draws from two pools of digitally stored data which originates from the Nexus © and Enobio © ambulatory sensor hardware. These data are then stored internally and the process pipeline forks into two top level processes; process autonomic data and process EEG data. The autonomic data processor includes filters for both electrocardiogram (ECG) and skin conductance level (SCL) of 0.5 to 35Hz and 35Hz respectively. The ECG data is then forwarded to a beat detection process to determine the inter-beat-interval (iBi) of heart rate and an epoch analysis process to produce the two derivatives, mean and standard deviation of iBi. The filtered SCL data is forwarded to the epoch analysis module to produce the two derivatives, mean and standard deviation of SCL. The resulting derivatives from ECG and SCL are then forwarded to a feature store for eventual output.

The EEG data processor performs filtering (Bandpass 0.05-35Hz) and epoch analysis before forwarding the signal data to a Fast Fourier Transform (FFT) which transforms EEG data from 3 sites of electrocortical activity FP1, FP2, FPz to determine the total amplitude spectra of the signal in the alpha (8-12Hz) and beta (13-30Hz) bands. The data from the FFT are forwarded to two top level processes; calculate cognitive activity and calculate valence response and then subject to temporal analysis (2s Hanning window over 30s stimulus epoch). From this analysis cognitive activity is calculated as beta divided by alpha at sites FP1, FP2, FPz to give the ratio of beta to alpha and valence (hemispheric asymmetry) is calculated as natural log alpha (power) subtracting FP2 from FP1. The resulting 5 derivatives from the EEG data are then forwarded to the feature store and combined with the autonomic derivatives and exported to digital file as feature vectors.

To test the validity of the processing pipeline, three participant data was subject to a correlation test with those derived from Acqknowledge (Biopac inc.), the results of this test were positive with a 0.97 correlation between the data.

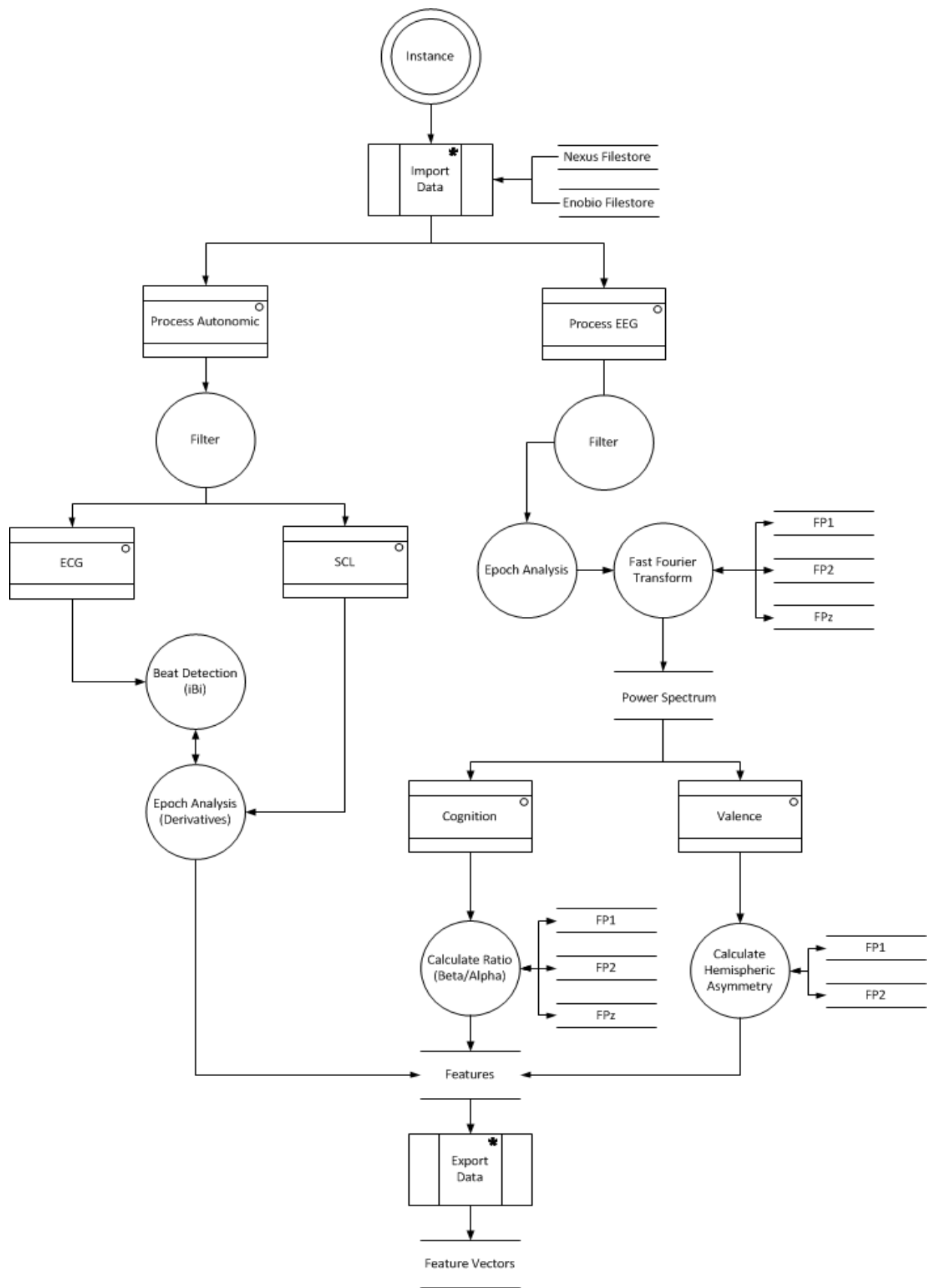


Figure 10.2 Meta Process Pipeline: feature extraction processing

10.2.3. *Study Goals*

This study was conducted to further iterate the subject-dependent classification methodology as a means of adapting information in a cultural heritage context such as a museum. The emphasis is placed on working with the composite model of interest as defined by the framework and data processing pipeline, using multimedia cultural heritage material produced by a partner institution presentation and measured with ambulatory physiological sensor hardware to provide a facsimile of a real-time sensor environment. The SVM classification algorithm will be applied to the classification of psychophysiological data from each dimension of the interest model separately and in combination and trained using participants subjective assessment of interest derived from Likert scales.

The study was designed with the following goals:

- To replicate the classification of cultural heritage material using the three component model of interest from study three
- To test the interest model functions with an extended range of media (beyond audio narrative)
- To explore classification performance using two approaches to labelling cases for classification (composite vs. component)

10.3. **Methods**

10.3.1. *Participants*

16 participants 8 female (aged 19-75) took part in the experiment; all participants were patrons of the host heritage institution (FACT). However, only 8 participant data were used for analysis, 4 participants data was rendered unusable due to sensor failure, and the remainder due to a loss of signal fidelity due to ecological factors (such as cosmetics), of the remaining participant population 5 were male 3 female (aged 20-40). In accordance with the universities lease of ethical approval participants signed a consent form and were of good health.

10.3.2. *Experimental Design*

The experiment was designed as a repeated measures, laboratory study completed at a cultural heritage institution i.e. participants were exposed to a series of multimedia presentations from a proposed cultural heritage exhibit; the level of interest of participants for the exhibit was generated retrospectively from subjective questionnaire data.

10.3.3. *Experimental Measures*

Physiological responses from the autonomic system were measured during experimental sessions, using the Electrocardiogram (ECG, sampled from the torso) and SCL (distal phalanges, second and forth finger, non-dominant hand) channels of the Mind Media Nexus X Mk II (sampled at 512Hz). Three channels of electroencephalographic (EEG) data were recorded, measuring alpha (11-12Hz) and beta (13-30Hz) activity, using the Enobio wireless 4-channel sensor (sampled at 250Hz) with ground contacts on left ear lobe and inner ear (Starlab Inc). A mobile sensor forehead band was fitted and nasion aligned to ensure sensor placement at FP1, FP2, FPz and electrodes attached. Once extracted, the feature data was imported into Matlab 2012Rb for manipulation and classification using the SVM algorithm native to the Matlab environment.

10.3.4. *Materials*

Stimulus material took the form of multimedia presentations of the work of three living film directors (see Table 10-1), the presentation of each directors work lasted 2 minutes and 30 seconds; director one, 4 segments; director two, 6 segments; director three, 5 segments, for a total of 15 segments (7 min 30 sec). The presentations were displayed on a 22” computer LCD screen and audio was reproduced through stereo speakers at an easy listening volume of 70 dB placed on the floor approximately 45” in front of the participant. The presentation took the form of a documentary narrative, detailing the context, work and style of each director. Each narrative lasted 30 seconds (see Figure 10.3). After each director presentation was complete, participants were asked to provide subjective judgements using a provided questionnaire consisting of three Likert scales ranked 1 – 10. These scales aligned to the 3 dimensions of the interest model; *Activation*: “how did this content make your feel” *tired passive 0 to activated alert 10*; *Cognition*: “How would you rate your level of mental activity (thinking, understanding, effort)” *low 0 to high 10*; and *Valence*: “how did this content make your feel” *sad angry 0 to happy cheerful 10*.

The presentation order of the director narratives was counterbalanced within director; the first narrative presented was used to prime participant physiology and not included in the classification analysis.

Director	Context			Work			Style	
Krzysztof Wodiczko	Video Audio	Video Audio		Video Audio	Audio	Video Audio	Video Audio	Video Audio
Content	Artist work			Artist work	Interview	Interview	Interview	Artist work
Length(sec)	30			30	30	30	30	
Ken Loach	Video Audio	Video Audio	Video Audio	Image Audio	Image Audio	Image Audio	Audio	Audio
Content	Interview	Interview	Interview	Artist work	Artist work	Other films	On realism	On politics
Length(sec)	30			30	30	30	30	30
Apichatpong Weerasethakul	Video Audio	Video Audio		Audio	Video Audio	Audio	Video Audio	Video Audio
Content	Artist work			Artist work	Artist work	Artist work	Interview	
Length(sec)	30			30	30	30	30	

Table 10-1 FACT study stimulus material

10.3.5. Procedures

After receiving instruction about the experimental procedure, participants were asked to complete a consent form in accordance with the Liverpool John Moores Ethical Committee. Electrodes were placed on the torso for ECG and on the distal phalanges of second and forth finger of the non-dominant hand for SC. Participants were asked to sit comfortably but remain as still as possible, approximately half a meter in front of a 22", 16:9 aspect computer LCD screen. This was followed by the multimedia presentation of the CH material (Figure 10.3), which was counterbalanced and timed to progress linearly through one directors' material until exhausted. After each director presentation, participants were asked to complete a questionnaire comprising self-ratings of physiological activation, cognitive engagement and emotional valence. A second screen was provided to allow the participant to review the material during the subjective judgement segment to allow for recall and assessment.

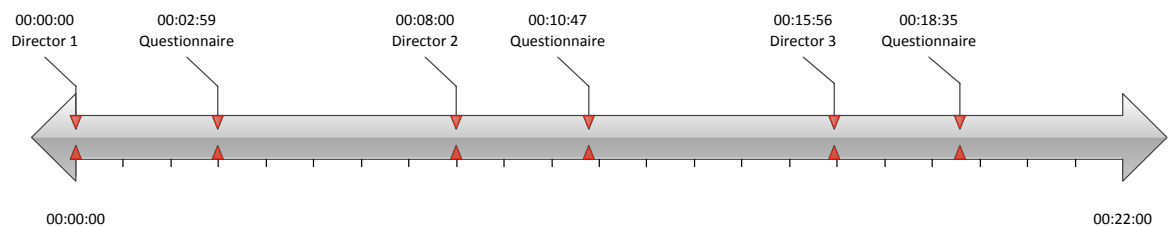


Figure 10.3 The procedure and stimulus timeline

10.4. Analysis

EEG (beta/alpha ratio) and autonomic data (mean and standard deviation of IBI (HR) and SCL) were extracted from an epoch that equated to 30s for a total of 15 stimulus events. All feature data were derived from raw signal output using the data process pipeline³.

10.4.1. Feature Extraction

For this study, 8 features were derived from physiological signals (see Table 10-2), for HR, mean, and standard deviation of IBI; for skin conductance level, mean and standard deviation; for EEG- electrocortical activity was derived from a fast *Fourier* transform (FFT) of total amplitude spectra using a 2 second *Hanning* window for each stimulus epoch (approx. 30 seconds), where the ratio $:x$ is expressed as *beta* (power 13-30Hz) divided by *alpha* (power 8-12Hz) at sites FP1, FP2, FPz. Hemispheric asymmetry was measured as the ratio $:x$ expressed as *alpha* (power) subtracting right from left hemispheric activity at sites (FP1-FP2), this results in n 2 second psychophysiological responses (observation) stimulus segment.

For Cognition: Where the ratio $:x$ is expressed as β (power) divided by α (power) at sites (FP1, FP2, FPz) for each 2 second window.

$$:x = \left(\frac{y_{\beta}^i}{y_{\alpha}^i} \right)$$

For Valence: Where the ratio $:x$ is expressed as the natural log of α (power) subtracting right from left hemispheric activity at sites (fp1, fp2) for each 2 second window.

$$:x = \ln(z_{\alpha}^i) - \ln(y_{\alpha}^i)$$

<i>Component</i>	<i>Measure</i>	<i>Derivative</i>		
Activation	Heart Rate	iBi-Mean	iBi-Stdev	
	Skin Conductance	Mean	Stdev	
Cognition	EEG	Ratio β / α FP1	Ratio β / α FP2	Ratio β / α FPz
Valence		Ratio α FP1-FP2		

Table 10-2 Features derived from physiological recordings and the relationship with the interest model

³ Software code provided under contract by Research Assistant K Gilleade

10.4.2. *Classification Trials*

The derived feature data were grouped according to the three dimensions of the interest model, such that each feature set created unique feature vectors for training the SVM classifier. Each dimension (referred to as a component) of the interest model has corresponding psychophysiological measures (Table 10-2) and labels (either as composite “interest” or individual component responses). This approach has a number of advantages, each feature vector is identified as a separate component of the model; feature sets can be combined as a fusion of features; thus the effect of each feature set or fusion of features on classifier class recall can be evaluated.

Accuracy in the context of classification in this study is determined using two methods. The first method uses the holdout method (Isaksson, et al, 2008) described in chapter 8 (pp63) for SVM parameterisation and cross-validation. This method of cross-validation uses the entire dataset as both training and testing data by splitting the data arbitrarily according to criteria; that is, data is randomly assigned to either training or testing according to the “set size” determined before classification (in this case 60% training, 40% testing). The dataset contains both the classification vectors (observations) and its associated label (subjective judgements), testing the SVM model involves classifying the remaining (40%) novel instances of test data, to determine accuracy. In a laboratory context, the labels (subjective judgements) associated with the test vectors (observations) are known to the experimenter but unknown to the SVM model, thus accuracy is calculated by comparing SVM model classification output (in terms of class) and with the known class labels.

As the current study is based exclusively upon the use of subjective feedback to provide class labels for training the classifier, careful consideration must be given to the balance of classes within the training data. In this instance these data are implicitly unbalanced due to the nature of subjective ratings, thus the commonly used accuracy metric which denotes the “hit rate” performance of the classifier; i.e. the percentage of correct classifications, which may be higher if classes are biased towards one class or another (in this case high or low) and thus hit rate accuracy may not reflect the whole of classifier performance in this instance. To counter any possible class bias effects, the second method uses the same holdout crossvalidation method as previously described. However, the accuracy output in this case uses the f_1 score, the f_1 -score (Powers 2011) measures classification accuracy using the statistics precision and recall. Precision is the ratio of true positives (TP) to all predicted positives (TP + FP). Recall is the ratio of true positives to all actual positives (TP + FN). The F_1 score is achieved by calculating:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Or more formally:

$$F_1 = \frac{2 * TP}{2 * TP + FP + FN}$$

Where TP = the number of true positive classifications, FP the number of false positive classifications and FN the number of false negative classifications. The classification of the psychophysiological data was completed within three trials; each trial uses a unique set of labels derived from the questionnaire data to train the SVM classifier.

10.4.3. *Deriving Binary Labels from Likert Scales*

For this study, participants were asked to complete a questionnaire consisting of three Likert scales ranked 1 – 10. These scales aligned with the 3 dimensions (referred to as components) of the interest model; Activation: tired passive 0 to activated alert 10; Cognition: low 0 to high 10; Valence: sad angry 0 to happy cheerful 10. Two forms of classification labels were derived from the questionnaire data to train the classifiers with, one which represents the overall level of “interest” towards the stimulus material and one that represents the individual component responses to the stimulus material, both ranked high or low. To derive the binary class labels for “interest” used within the classification analysis, the Likert scores for each participant and each stimulus segment were normalised in the form of:

$$y^i = \sum_{i=3} x_i \left(\frac{x_i - minC}{maxC - minC} \right)$$

Where y^i is the sum of subjective scores for each dimension of the model (activation, cognition and valence) combined, $minC$ and $maxC$ are the minima and maxima of the population of scores for each stimulus segment. The result y^i is a population of normalised scores. To set the threshold for class assignment, the median of this population was calculated. Above the median was labelled as high interest and below as low interest. Class labels for the individual components, were derived by modifying the above method to remove the sum component, thus x_i becomes the population of scores for each of the components.

The result, in the first instance is a class label (either high or low), that represents a single subjective judgement as a composite “interest” score for each stimulus segment within each content block (director). In the second instance, the class label (high or low) represents the level of response for each component of the interest model individually. These labels are then associated

with the psychophysiological data for that stimulus segment and once combined these data become the feature vectors used to train and test the classification algorithm.

10.5. Results

10.5.1. *Trial 1 Composite classification using “interest” labels*

In this trial, feature data from each participant is classified individually and classification was performed in such a way as to iterate through each combination of components of the interest model. The labels used for training the classifiers were calculated from the questionnaire data as the total subjective level of “interest”. These labels represent the composite of the scores provided for each component (activation, cognition and valence) of the interest model which were then combined to yield a unidimensional scale of interest, which was divided into binary states of high or low. The results displayed in **Error! Reference source not found.**, show that in this instance the activation component presents with the highest mean accuracy and stability of accuracy across participants (0.87, σ 0.07). The combination of activation with valence and activation and cognition features both present with high mean accuracy (0.81 σ 0.10, 0.82 σ 0.09) respectively, however the higher accuracy variance for these two feature classifiers shows these classifiers to be more unstable across some individuals. The classifier created to combine activation, cognition and valence to reflect the full interest model, reports a mean accuracy of 0.81 77.02% (σ 0.11) the higher variation in accuracy for this case is indicative of a classifier that is unstable across individuals. It is worth noting however, that the lowest reported accuracies still remain above chance levels.

Participant (Subjective Judgement “interest”) F ₁ -score										
Dimension	P1	P2	P3	P4	P5	P6	P7	P8	Avg. F ₁ -Score	StDev
A	0.93	0.83	0.97	0.84	0.91	0.73	0.87	0.89	0.87	0.07
C	0.74	0.79	0.74	0.69	0.74	0.71	0.83	0.83	0.76	0.05
V	0.68	0.58	0.59	0.71	0.75	0.70	0.77	0.69	0.68	0.07
A,C	0.73	0.79	0.97	0.76	0.82	0.68	0.87	0.89	0.81	0.09
A,V	0.65	0.80	0.99	0.81	0.93	0.73	0.78	0.89	0.82	0.10
C,V	0.81	0.78	0.64	0.71	0.79	0.63	0.77	0.81	0.74	0.07
A,C,V	0.67	0.81	0.97	0.69	0.90	0.69	0.85	0.89	0.81	0.11

Table 10-3 Classification Accuracy (F₁-score) across Individuals and Feature Sets (Activation (A), Cognition (C), Valence (V), Participant (P) – composite model

10.5.2. Trial 2 Component classification with individual response labels

In this trial, feature data from each participant is classified individually and a classifier for each component of the interest model is trained using labels provided by the individual for each component of the interest model individually. These labels were derived from the questionnaire data by normalising (range 0 to 1) to the min and max of all the subjective scores given for each component (activation, cognition and valence) of the interest model by each individual, for each stimulus segment from a particular director. The resulting binary label of high or low for each stimulus segment is derived from a unidimensional scale, where a value of 0.5 or above would be a labelled high. The results displayed in Table 10-4 show that in this instance the classification of the features of activation present with favourable accuracy (mean F_1 0.90 σ 0.05), additionally the classifier displays a high degree of stability across participants in this instance. The classifiers for the features of cognition and valence both report low accuracy when compared to activation of F_1 0.66 (σ 0.12) and F_1 0.69 (σ 0.16) respectively, both classifiers show low accuracy variance across participants however, creating a stable, if inaccurate classifier.

Dimension	Participant (Subjective Judgement “component”) F ₁ -Score								Avg. F ₁ -Score	StDev
	P1	P2	P3	P4	P5	P6	P7	P8		
A	0.89	0.84	0.98	0.91	0.89	0.86	0.97	0.86	0.90	0.05
C	0.75	0.79	0.56	0.40	0.63	0.66	0.73	0.74	0.66	0.12
V	0.69	0.49	0.89	0.56	0.85	0.79	0.45	0.78	0.69	0.16

Table 10-4 Classification Accuracy across Individuals and Feature Sets (Activation (A), Cognition (C), Valence (V); Participant (P)) - component model

10.5.3. Trial 3 Generalised model using both “composite” and “component” labels

In this trial all participants’ data is aggregated to create one large data set, and the classifier is trained using either the composite “interest” class label or the “component” specific label. This trial tests the classifiers ability to generalise across individuals and represents an estimate of the predictive potential of the classifier to generalise to new individuals using this data set. Here again the results displayed in **Error! Reference source not found.** show that the classifier for the ctivation component presents with the most favourable mean recall accuracies F_1 0.73 and F_1 0.85 for classifiers trained using “interest” or “component” class labels respectively. Moving to the classification of the cognition and valence components, there can be seen a significant drop in accuracy reported for both components, for either labelling schema.

Generalised Models		
Dimension	Composite	Component
A	0.73	0.85
C	0.66	0.54
V	0.59	0.67
A,C	0.71	
A,V	0.65	
C,V	0.68	
A,V,C	0.70	

Table 10-5 Generalised Classification Performance: Feature Sets (Activation (A), Cognition (C), Valence (V))

Turning to the results from the classification of the components combined, it can be seen that in this instance the activation and cognition components combined provided the highest accuracy F_1 0.71, and the classification of full combination of components that make up the interest model presented a modest accuracy of F_1 0.70 .

10.6. Discussion

The goals of this study set out to replicate the subject-dependent approach taken in study three with respect to feature selection and a further test of the validity of the interest model posited in study three. Additionally this study set out to explore how a wider range of media types influenced the ability of the psychophysiological features to differentiate between states of high or low interest. Furthermore, novel methods of generating classifier training labels were explored with respect to subjective self-report data from questionnaire. A data analysis process pipeline was developed to perform the operations necessary to filter physiological signals and create feature vectors for classification from digital storage.

The results indicate that the combination of physiological measures, psychological model of interest, derived psychophysiological features and classifier training protocol explored within this study provided classification rates of user interest in excess of 80%. This combination of psychological model and physiological features proved robust, despite the addition of a wider range of media types used as stimulus material. Additionally, these high classification rates support the use of the SVM algorithm and hold out training-validation methodology, which is now integral as an addition to the developed processing pipeline. Furthermore, classification rates gained from the activation component, which is based upon features of the autonomic nervous system, were shown to be the highest when classified alone and in combination with other components.

The results from the first classification trial which utilised the “composite” interest labels for classifier training, showed that the subject-dependant classification of high/low interest was highest in four classifiers, activation 0.87 (σ 0.07); activation and valence 0.82 (σ 0.10); activation and cognition 0.81 (σ 0.09); activation, cognition and valence 0.81 (σ 0.11). The classifiers built using the features of activation alone or activation and valence together proved to be the most stable and best performing classifiers, reporting high accuracy and low accuracy variance. A noteworthy result within this classification trial is the stability reported by the other classifier variants, all report low but above chance classification accuracy, however the stability of these classifiers is superlative when compared to the best performing variants; leading to classifiers that exhibit lower accuracy but remain stable across participants.

The classification results from trial one using the composite model were broadly equivalent to those results from trial two in which the labels used to train the classifier were derived from each component of the interest model. In this second trial the activation component reports the highest accuracy 0.90 (σ 0.05). The cognitive 0.66 (σ 0) and valence 0.69 (σ 0.16) component classifiers perform at a much lower accuracy yet display excellent stability in terms of accuracy variance across participants. An interesting finding is the lack of appreciable difference in classifier accuracy between trials one and two for component level classifications. This may be due to a high correlation between the two forms of label derivation from the subjective survey data, leading to almost identical classifier training outcomes for the classifiers built to classify each dimension of the interest model using composite and component labels.

The final classification trial tested the ability of the classifier to generalise across individuals, that is to utilise all of the participant data and labels (both “composite” and “component”) to build a classifier that would potentially generalise to new individuals after a single training session, and in this respect the classifiers performed poorly compared to subject-dependent classification. However, contrary to the expectation that the classifiers would generalise poorly, the classifiers for activation (0.73) and activation plus cognition (0.71) report respectable accuracies (i.e. well above chance), this result resonates with those gained from the subject-dependant classifications, in that the effect of the stimulus content on the psychophysiological state of the participants resulted in a strong delineation between the two classes of interest under observation. It remains an issue for speculation and investigation, as to whether the user of a system would perceive a difference in the quality of interactions with systems that were either 73% or 87% accurate. The standout result from the generalised classification trial comes from an activation classifier trained using the “component” labels which reported an 85% classification accuracy. However, this result may be suspect given that the classification results from the classifiers for cognition and valence both report marginally above chance levels of classification accuracy. These divergent classification

results between the three components could be due to the nature of the questions asked in questionnaire used to gather subjective judgements. Participants may have not understood fully what was meant by mental activity or what they felt (in terms of valence), it would appear from the results that participants grasped more readily what was meant by activation however.

The results for this study show that in terms of the interest model, the classification of activation and the fusion of activation and valence features, classified with highest accuracies, and the effect of either labelling schema used for training the classifiers produced only marginal differences in mean recall accuracy. For this study, classifiers trained using the “composite” labels derived from subjective questionnaire data present with the lowest variation in accuracy across individuals, showing that classifiers trained using this methodology are more stable across individuals when tasked with classifying indices of autonomic and hemispheric asymmetry activity specific to a state of interest. These results remain consistent within the context of the interest model, in that the advancement of the measurement protocol, shows a strong association between the measures of activation and a user judged level of interest. Furthermore, the results show that it is possible to achieve good classification accuracy using a range of ambulatory sensors to measure the interest level of the individual in a cultural heritage context.

However, given the small sample size of data collected for this study, these findings must be interpreted with caution; in that, the ecological field viability of the sensor technology is not a proven factor, and this can be seen as data loss due to sensor failure. This failure may have arisen from factors that cannot be controlled for in an ecologically valid real-time environment (such as a museum), such as cosmetics, heat and humidity, low battery power (which results in corrupt or inaccurate data). These issues can only be addressed by better sensor design and more robust error signalling, which would allow power to be topped up or display signal quality degradation. Furthermore, data from this study was analysed retrospectively using a process pipeline and the effect of analysing and classifying psychophysiological responses using this pipeline in real-time remains an issue still to be investigated.

Interpreting these results within the domain of the IBIS classification framework and the wider context of a biocybernetic control loop for cultural heritage applications, it can be seen that the composite model of classification can reliably output high classification rates for a binary classification of high or low interest, and this output has the potential to be used readily within an adaptation model. However, the component model classification approach displayed poor results in comparison to the composite model, which despite its potential to output a more nuanced classification of interest means that this form of classification requires more development in terms of classification accuracies and second stage processing before it can be seen as a candidate for

inclusion in to a system of this type, an example of this development can be seen in Table 12-4 (pp 158). An issue of import to the overarching goal of the research in this thesis is one of user perception; we have seen from the results of this study that including subjective judgements in the training of classifiers can have a large effect on classifier accuracy; how users perceive classifier accuracy and system utility or performance represents a research question still to be explored.

10.7. Conclusion

This fourth experimental study focused on psychophysiological classification within the context of in situ measurement of psychophysiological indices of interest using cultural heritage multimedia material. A process pipeline was developed to analyse these data and output vectors suitable for classification retrospectively, and a framework for a biocybernetic loop was proposed to take in psychophysiological measurement at one end and output classifications of user interest at the other. Two classification training protocols that included subjective judgements as part of the classifier training process as to provide training labels were proposed and tested, and the results show that including subjective judgements in the classifier training process results in high accuracies and stable classifiers. Furthermore, the results show that in this instance combining features of activation and valence provides the highest classification rates.

11. Study 5: classifying the interest state in real-time

11.1. Abstract

This fifth experimental study is focused on the subject dependent classification psychophysiological of indices of interest in real-time within the context of a laboratory. A real-time application framework was developed that integrates the process pipeline developed in study four (Chapter 10 pp. 99) and the classification output proposed within the IBIS model. A proof of concept application based on this framework was used to capture, measure and classify user interest responses to multimedia stimuli in the form of movie trailers; 16 participants, all students took part in the study. The aim was to ascertain the nature of the relationship between mathematical accuracy as reported by the SVM classifier and the users' perception of that accuracy; this was achieved using a "wizard of Oz" interaction paradigm. Classifiers were trained subject dependently over a series of four builds using subjective feedback to provide classifier class labels. ROC analysis revealed that while machine accuracy remains stable across four classifier training builds, user perception of that accuracy fluctuates across all four training builds, culminating in a perceived accuracy that on average exceeded that of the users' initial perception; and that the classifier developed high to excellent discriminatory power over the series of training builds, in terms of recognition of the users interest preference towards movie trailers. Furthermore, the results showed that it is viable to classify and output indices of user interest as preference in real-time and that these outputs would be transferable to other elements of a biocybernetic loop.

11.2. Introduction

Two of the key components of the biocybernetic loop are psychophysiological measurement and classification of the psychophysiological response, most work in the field is devoted to the identification of psychophysiological measures and offline classification. The implicit assumption is that good measures and accurate classifications will deliver biocybernetic adaptations that will improve the user experience, and by applying “knowledge” (i.e. psychophysiological and subjective responses) from previous interactions biocybernetic systems have the potential to “learn” user preferences, styles of work and levels of activity during interaction and task completion. Current research concerning the use or potential use of biocybernetic control is mostly concerned with verifying that physiological data can be used for controlling or informing the system to make changes (see Novak 2012, van de Laar et al. 2013). As a consequence of this focus, assessing the interaction between the mathematical accuracy of the system and the perception of that accuracy by users’ remains largely unknown and under researched.

There are a number of issues which can affect the adoption of biocybernetic control in systems. such as the use of psychophysiological data which is seen in some cases to be unreliable input when used as the driver of the system (van de Laar et al 2013); the intentionality of the user and the measured psychophysiological response; and critically the issue of how to assess the accuracy of the system. Assessing system accuracy and performance for biocybernetic control systems, is an issue that can be split into two dimensions, one “hard” which assesses the mathematical accuracy of the system (in terms of classification accuracy) and the second “soft” which assesses the accuracy of the system in terms of user perception of system accuracy, and in this case soft accuracy can vary as a function of hard accuracy. However, soft accuracy can also be affected by perceptual bias i.e. a lack of trust in the system (Lee and See, 2007), or be linked to the cost of errors to the user or in system judgements when operating the system, for example in the case of classification within medical applications where misclassifications can affect diagnosis or treatment or in the case of entertainment (such as gameplay) where the cost of errors is a small annoyance to the user with low impact on the overall operation of the system.

The issue of the unreliability of psychophysiological input as the driver of biocybernetic systems is concerned with the detection of artefacts within the measured signal such as, movement, electrical noise and biological factors such as inter and intra personal difference. These factors can potentially lead to “interaction noise”, in which biocybernetically controlled systems offer interactions or judgements to users that are not based on context, system rationale or accurate

classifications of user states, but rather upon the detection of artefacts as valid inputs (a false positive as it were). Interaction noise can be defined as those instances where psychophysiological classification is perceived as inaccurate by the user. It may originate from several sources: (1) artifacts or poor sensor connections that distort the psychophysiological signal with consequences for classification, (2) non-stationarity and other factors that compromise the integrity of the data used to train the classifier, (3) clarity of feedback from the system at the interface, and (4) user perceptions of system accuracy. The effects of Inter and intra personal differences have been investigated in studies presented in this thesis (see chapter(s) 7, 8 and 9), and the results of these studies has shown that the impact of these factors can be decreased significantly through the application of a subject-dependent measurement and classification approach within systems, and possibly controlled fully if the subject-dependent approach is expanded to include training and calibrating a system to the user each time it is used.

Solving issues such as artefact correction is a relatively mechanistic procedure involving various forms of signal filtering, artefact separation and removal (see Sweeny et al. 2012). However, artefact detection and removal from physiological signals is far from trivial and is currently an active area of research. Good system design can alleviate some of the issues that create interaction noise, specifically in dealing with the intentionality of the user and the measured psychophysiological response and reducing the effects of signal artefacts; in that a well-planned system rationale and specific task definition coupled with a psychophysiological model that is both sensitive and diagnostic of the user state, within the rationale and task, reduces the scope in which the system makes judgements, thus reducing the area of effect in which errors and thus interaction noise applies.

Good design can also have a positive effect in how trust is engendered in and towards the system, trust has been shown to reduce the “noise” in interactions between the user and the system (Lee and See 2004), whereby implicit input is accepted as being synergistic with the user, the system and the task. However, in this instance interaction noise is seen a function of computational efficiency, in which the time taken to calibrate a system to the user and the timeliness and frequency of classifications are key variables. Thus, a system that requires a long time to calibrate to the user, takes a long time to generate classifications and classifies responses infrequently can create interaction noise by reducing the synergy that is created between user and system when information and interactions are timely. Manipulation of these key variables may prove to be an important factor that affects whether a user will accept and trust in a system and its purpose.

Determining the accuracy of biocybernetic control and the users perception of system accuracy, has clear implications for systems in both the design stage and while in operation. Assessing the performance of a system that has been built to calibrate itself to a user while it is in operation is an area of investigation in the field of (BCI) brain computer interfaces (van de Larr et al. 2012, Bos et al. 2012) where system performance (as classification accuracy) is defined as the amount of control perceived by the user during interaction with the system. However, this metric is difficult to evaluate as perceived control can vary as a function of classification accuracy and understanding this relationship is further complicated by the sheer range of classification accuracy measures available, such as, standard mathematical accuracy output (which is a percentage of correct interpretations); error rate; precision (the fraction of retrieved instances that are relevant); sensitivity (the fraction of relevant instances that are retrieved). However, of the metrics discussed here none contain a means of dealing with the bias implicit in systems that use the human operator as a means of calibration, in that these methods assume a balanced dataset from which performance is assessed (i.e. an equal number of high to low class data). The issue of class bias presents a dilemma when attempting to assess the performance of a classifier, in that creating balanced training data makes for a mathematically sound classifier. However, in the case of user-dependent biocybernetic control, bias represents a human factor. Therefore, even if a system is initially built using perfectly balanced data, with time (assuming retraining), when it is re-trained using data gathered in the field, bias will eventually creep into the training data. Removing the possibility of bias in the case of user-dependent systems to create the perfect training data set, would amount to an artificial calibration phase which would no longer represent real human responses but rather a stylised version of the ideal human response which could possibly lead to an increase in interaction noise.

A different approach to assessing system performance in terms of “hard” and “soft” accuracy and one that could provide a more descriptive view that targets the relationship between hard mathematical and soft perceived accuracy for systems, is the receiver operator curve (ROC) and associated (AUC) area under the curve. The ROC displays the ratio between the true positive rate (fraction of true positive classifications) and the false positive rate (fraction of false positive classifications). The AUC value of the ROC gives the probability that the target state (e.g. a state of high interest) has a higher confidence than a non-target state (Fawcett 2006). In this the mathematical accuracy of the system can be mapped against the perceived accuracy of the system, highlighting the breakpoints of system accuracy and the corresponding user perceived accuracy rate, which could transfer into how likely a user is to accept or trust the system.

11.3. Classifying the Interest Response in Real-time

11.3.1. *The Interest as Binary - Interest as State (IBIS) Model*

The model developed in experiment 4 has been further iterated to overcome the issues identified with the component level classifications; this revised framework is shown in Figure 11.1. At its core the model utilises two classification methods, which receive psychophysiological input from three component processors, each component processor represents one dimension of the interest model e.g. activation, cognition and valence. When the psychophysiological data is classified, the outputs can represent either Interest as Binary (low or high) or Interest as a State (low to high). In comparison to the previous iteration of the IBIS output from the component processors has been truncated into the component model, which now outputs two forms of classification based upon the same input in parallel. In an online context, inputs from the physiological sensors are forwarded to the component processors; these processors derive features from the physiological data to create feature vectors used within the classifiers; the feature vectors are then associated with a training label and output to the classifier; feature vectors are either truncated into a single classification vector used to classify interest as a binary state, or expanded into multiple classifications (of binary states) utilising a majority vote to create a series of discrete interest states. In this model the classification process can be viewed as an interpretive layer applied using two time frames, one in which the classifier outputs a binary high or low state infrequently (e.g. at the end of a stimulus event) or two as a ratio of classifications of high to low which can be frequent (e.g. a continuous updating ratio) or infrequent (e.g. a singular ratio interpretation), it is upon the adaptation model to utilise these outputs effectively.

Thus, the IBIS model enables two classifications of the interest state to be completed concurrently as a composite model (single classification vector) and a voting model (multiple classifications, single discrete state or continuous “interpretation” of the interest state). Using a cultural heritage exhibit as an example, psychophysiological signal data is captured from a user and processed to produce feature vectors, the exhibit experience consists of 20 second narrations and associated video. In this scenario the composite model delivers one classification for each 20 second stimulus segment, representing an “overall” interest classification; the voting model delivers n classifications each 20 second segment to represent a nuanced state of interest for the stimulus segment. That is, a majority vote applied to the classification output forms a ratio of binary high to low classifications e.g. 8 high 2 low = highly interested or 5 high 5 low = neutral interest and so forth and this can be continuous or singular depending on the needs of the system.

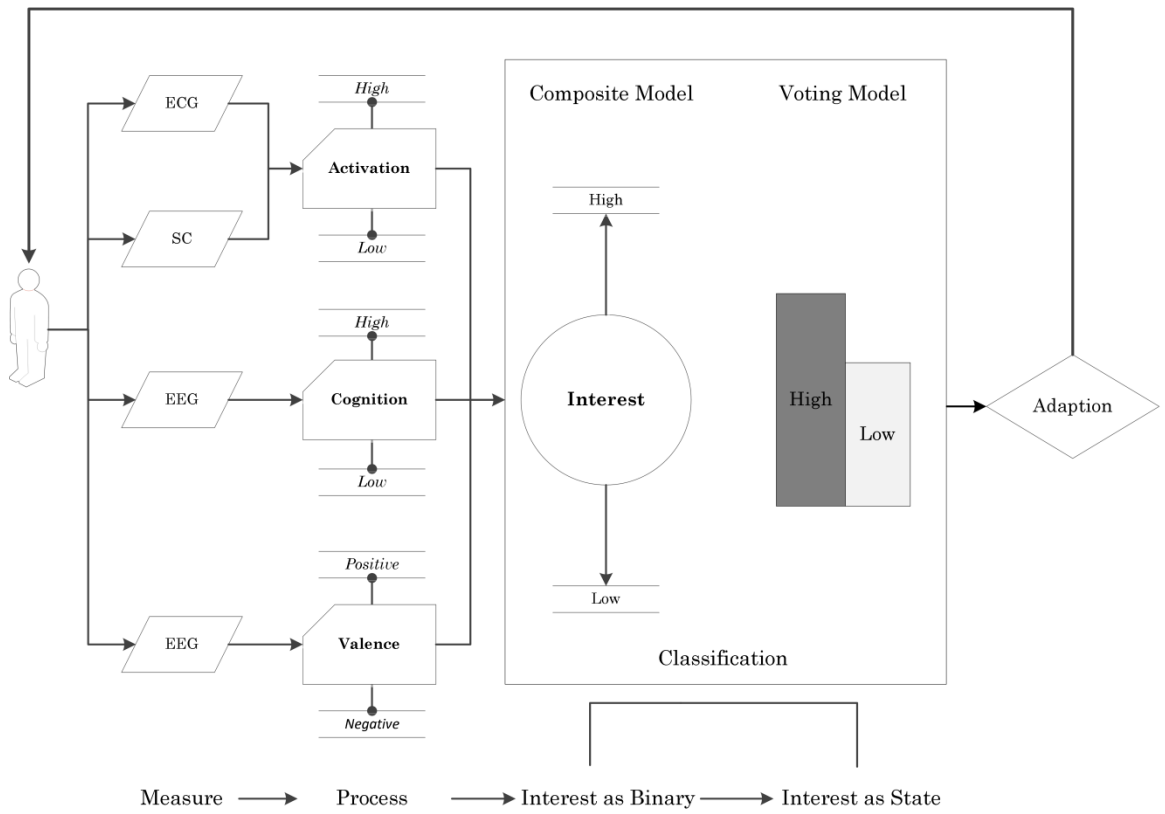


Figure 11.1 The revised IBIS Framework

11.3.2. *Applied Real-Time Interest Classification Framework (ARTIC)*

For this experimental study, the feature extraction processing pipeline developed for study four 4 (see chapter 10 of this thesis) is integrated into a software application capable of duplicating the methods required in the IBIS model to import raw signal information from digital storage and output feature vectors for use in training and testing the classifiers used to classify the interest state. The goal of the application is to integrate the methods and techniques for signal processing identified previously into a single application capable of calibrating a classifier to a user and then classifying that users interest state in real-time.

Due to the complexity of the application framework some elements require extracting from the overall structure to highlight how these processes function within the application. Figure 11.2 shows the function which displays the video content via display unit to the user, and takes the form of a video player sub-window, which also acts as the means for gathering and processing the subjective responses to each video after it has been viewed. To begin with the video player is executed and draws from the pool of video material, after a video is displayed a new window is displayed which asks the user of the system for a number of subjective responses. These responses are then processed and forwarded to the export module for aggregation and association with the psychophysiological responses for that video. The function then iterates until all video content is exhausted or the exit function is enabled.

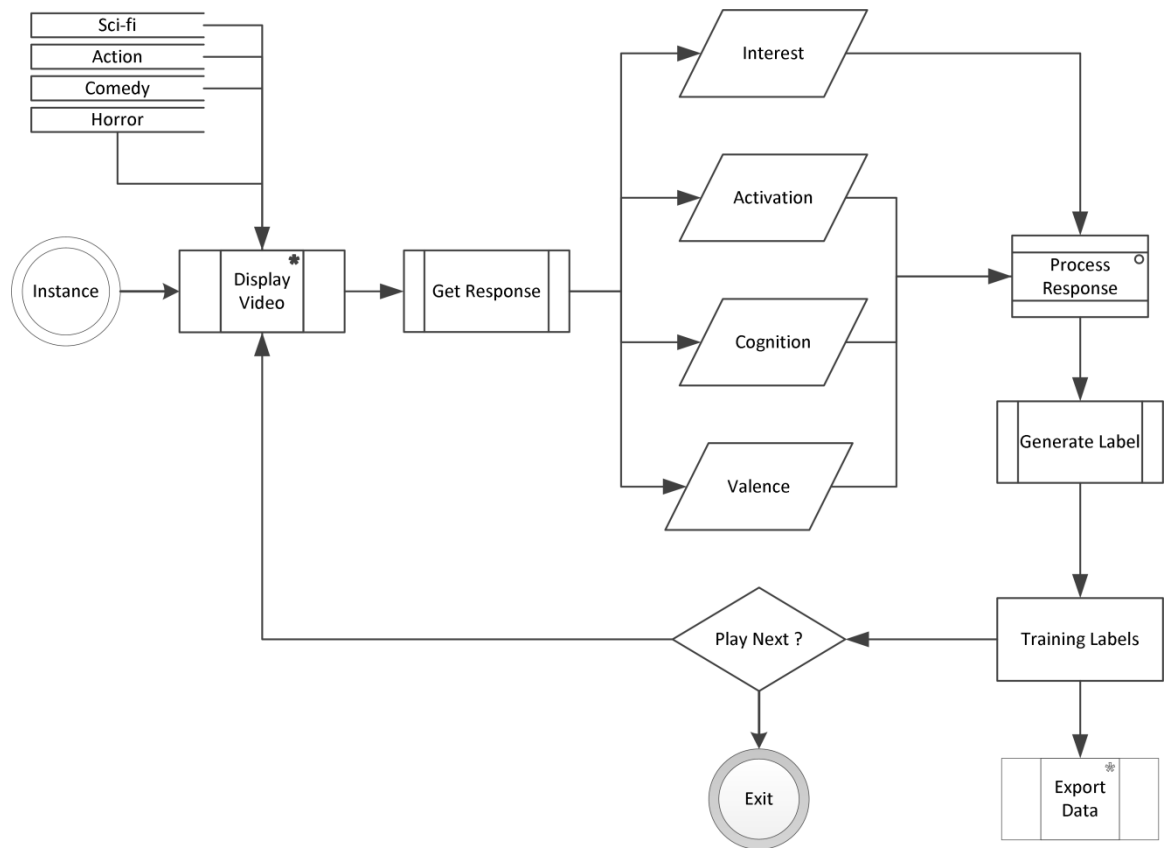


Figure 11.2 The display video and subjective feedback process

The data flow for the train classifier process is shown in Figure 11.3. This module takes the feature vector output from the data export process and checks if a classifier needs constructing; if true, a check is performed to determine if the current request is for the first build of the classifier. If this condition is satisfied, a further check is performed to determine if two full instances of two classes (i.e. two examples of high and low class data) exist within the data. If two examples of high and low class data are found, a classifier is built using the “composite” model of interest data (i.e. all features from activation, cognition and valence plus the class label). However, if a classifier already exists and a new classifier build is required, then data collected for the current stimulus period is added to the existing training set and a new classifier is constructed based upon this new training set and applied to classify new instances of data. If no new classifier build is required, the train classifier process is bypassed and new vectors are classified and output. The requirement to check if a classifier build is required functions on the premise that there are 40 videos in the stimulus pool, the check works by subtracting the number of videos used to train the classifier initially and dividing the remainder by 4; If the remainder cannot be divided by four then the application exits; else each time the counter reaches the new build query value, the system performs a classifier build, integrating all new physiological feature vectors into the new training data set and then classifying fresh instances of data until the stimulus pool is exhausted.

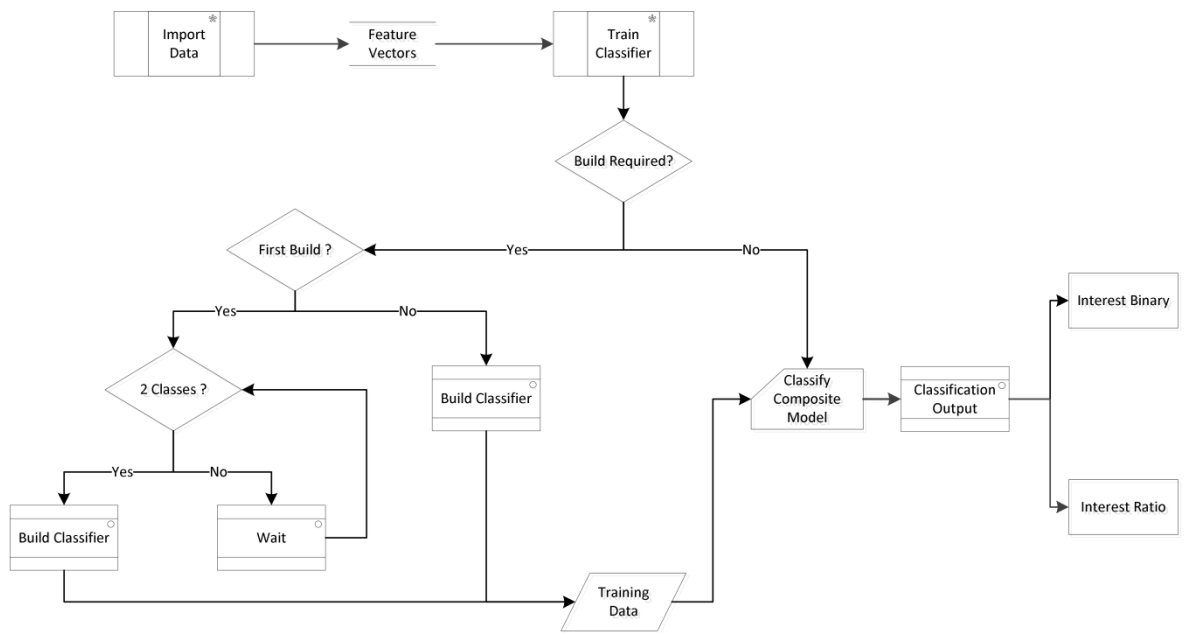


Figure 11.3 The train classifier process

The applied real-time interest classification (ARTIC) application framework is shown in Figure 11.4, this details the data flow of pipeline processes used to output interest classifications. The pipeline starts with a module to import the physiological data; this module draws from two ambulatory physiological sensor technologies in real-time, the Nexus © (used to capture autonomic ECG and SCL responses) and the Enobio © (used to capture EEG responses). These data are then buffered internally and the process pipeline forks into two top level processes; process autonomic data and process EEG data. The autonomic data processor includes filters for both electrocardiogram (ECG) and skin conductance level (SCL) of 0.5 to 35Hz and 35Hz respectively. The ECG data is then forwarded to a beat detection process to determine the inter-beat-interval (iBi) of heart rate and an epoch analysis process to produce the two derivatives, mean and standard deviation of iBi. The filtered SCL data is forwarded to the epoch analysis module to produce the two derivatives, mean and standard deviation of SCL. The resulting derivatives from ECG and SCL are then forwarded to a feature store for eventual output.

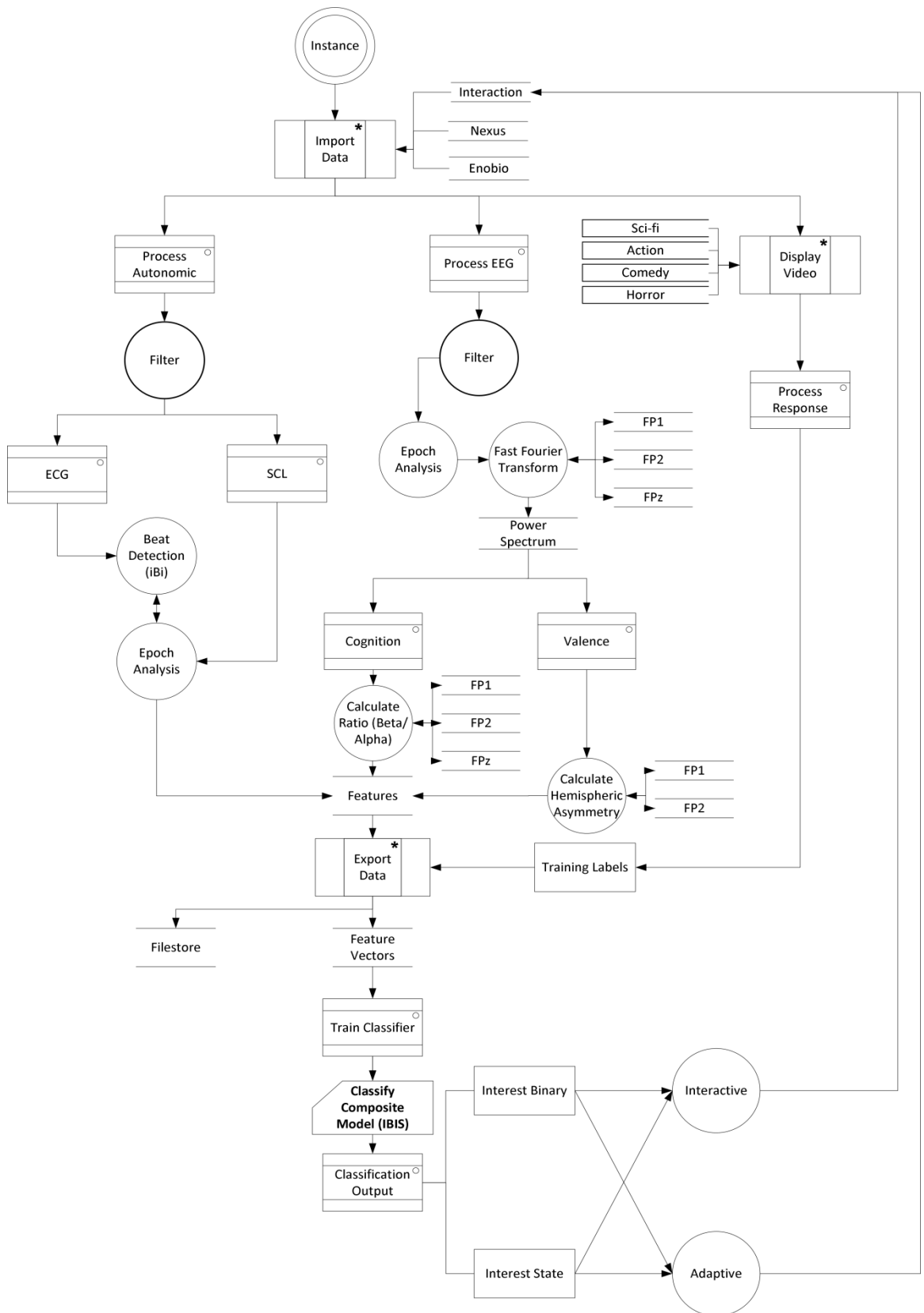


Figure 11.4 The Applied Real-Time Interest Classification Framework (ARTIC)

The EEG data processor performs filtering (Bandpass 0.05-35Hz) and epoch analysis before forwarding the signal data to a Fast Fourier Transform (FFT) which transforms EEG data from 3 sites of electrocortical activity FP1, FP2, FPz to determine the total amplitude spectra of the signal in the alpha (8-12Hz) and beta (13-30Hz) bands. The data from the FFT are forwarded to two top level processes; calculate cognitive activity and calculate valence response and then subject to temporal analysis. From this analysis cognitive activity is calculated as beta divided by alpha at sites FP1, FP2, FPz to give the ratio of beta to alpha and valence (hemispheric asymmetry) is calculated as natural log alpha (power) subtracting FP2 from FP1. The resulting five derivatives from the EEG data are then forwarded to the feature store and combined with the autonomic derivatives and exported to the train classifier process (detailed in Figure 11.3). The training of the classifier takes place within Matlab using the deployment command line processor for real-time data interaction; once trained the same command line processor is used to classify the feature data and export the classification output back into ARTIC. To train and ascertain estimated performance of the classifier in real-time, the sequential minimal optimisation (Platt, 1998) and hold-out cross-validation methods are used on the aggregated training data. When coupled with a loose grid search algorithm (Algorithm 1 see chapter 8 pp63) these methods form the basis for the training and parameterisation of the SVM in real-time, providing the optimal settings for the box constraint and sigma values of the SVM radial basis function (RBF) kernel for each new instance of training data as the system is used. That is, for each new build of the system, training data is aggregated and cross-validated to create a new classifier in real-time, in this instance to prevent over fitting of the classifier to the training data and reduce computation time the box constraint and sigma values are set to a maximum of 2.

11.4. Study Goals

This study was conducted to iterate and utilise the subject-dependent classification methodology from previous studies in a real-time system. The emphasis is placed on the capture and measurement of psychophysiological responses to movie trailers within the composite model of interest, as defined by the IBIS framework (see Figure 11.1). The system is designed to apply the IBIS framework in a real-time application, and calibrate by creating a classifier and training data sets that are tailored to the user, by the user, across a series of four builds during runtime, and subsequently provide feedback to the user about the results of classifications, to assess the level of agreement with subjective assessments. Thus, to assess the efficacy of the system the research goals are:

- Verify the psychological construct “interest” consisting of three components activation, cognition and valence in real-time
- Determine if classification accuracy improves with additional training data (i.e. does “machine learning” occur)
- Determine the effect of additional training data on user perception of the systems accuracy
- Determine the relationship between system accuracy and user perception of system accuracy

11.5. Methods

11.5.1. *Participants*

16 participants 9 female (aged 19-25) took part in the experiment; all participants were from the student body at Liverpool John Moores University. However, only 14 participant data were used for analysis, 2 participants data was excluded based on an application exit due to not meeting the classifier training criteria, that is these participants did not produce the required instances of high and low classes over the maximum 12 videos needed to train the system. In accordance with the universities lease of ethical approval participants signed a consent form and were of good health.

11.5.2. *Experimental Design*

The experiment was designed as a repeated measures design (i.e. the same participants took part in all build sessions). A “Wizard of Oz” (Kelley 1983) real-time interaction prototyping approach was derived in order to provide feedback to the user. An real-time interactive application was used to gather and classify participant responses, the application required four build phases to complete the experiment, build 1: initial classifier training, requiring responses (both psychophysiological and subjective) from at least two of each of the target classes (high and low), once built the classifier begins to classify responses; build 2, which aggregates the responses from build 1 and all responses

gathered up until that point into a training data set then begins to classify responses based on these new training data; build 3, which aggregates the responses from builds 1 and 2 and all responses up until that point into a new training data set then begins to classify responses based upon these new training data; build 4 which aggregates the responses from the previous 3 builds and all responses up until that point to create a final training dataset then begins to classify responses based upon these new training data.

11.5.3. *Experimental measures*

Physiological responses from the autonomic system were measured during experimental sessions, using the Electrocardiogram (ECG, sampled from the torso) and SCL (distal phalanges, second and forth finger, non-dominant hand) channels of the Mind Media Nexus X Mk II (sampled at 512Hz). Three channels of electroencephalographic (EEG) data were recorded, measuring alpha (11-12Hz) and beta (13-30Hz) activity, using the Enobio wireless 4-channel sensor (sampled at 250Hz) with ground contacts on left ear lobe and inner ear (Starlabs Inc). A mobile sensor forehead band was fitted and nasion aligned to ensure sensor placement at FP1, FP2, FPz and electrodes attached. All data was collected and analysed in real-time using an application developed using ARTIC framework as its basis⁴.

11.5.4. *Materials*

The stimulus material used for this study took the form of movie video trailers from four genres of film: science fiction, comedy, action and horror (see Table 11-1). The presentation of each movie trailer lasted 60 seconds; each genre contained 10 trailers. Videos were displayed on a 42" LCD TV screen at 720p resolution and audio was reproduced through television stereo speakers at an easy listening volume of 70 dB. Participants sat at an approximate 1 meter distance directly in front of the television and within easy reach of a computer connected mouse. Video display and user interactions were captured using a computer with two display outputs; one screen output the video and subjective response collection application interface and the other displayed the classifier interface. The presentation order of the movie trailers was randomised for each participant, with each video presentation drawing from the pool of 40 until all material was exhausted.

⁴ Application software and code based on the ARTIC framework was provided under EU FP7 project No.270318 (ARtSENSE)

Genre	Length	Number
Science Fiction	60s	10
Action	60s	10
Comedy	60s	10
Horror	60s	10

Table 11-1 Genre, length and number of stimulus videos

11.5.5. *Procedures*

After receiving instruction about the experimental procedure, participants were asked to complete a consent form in accordance with the Liverpool John Moores Ethical Committee. Electrodes were placed on the torso for ECG and on the distal phalanges of second and forth finger of the non-dominant hand for SC. Participants were asked to sit comfortably but remain as still as possible, approximately 1 meter in front of a 42” 16:9 aspect television screen. The experimental procedure (see Figure 11.5) was completed in two parts, mode one (build) and mode two (classify).

During the build mode a video trailer of 60 seconds duration randomly chosen from a pool of 40 and displayed upon the television. After each video trailer presentation, participants were shown a simple interface on screen to interact with, which asked 4 questions and provided buttons for answers to each question in the form of:

- Was this content interesting? Yes or No
- Did you find this content activating? Yes or No
- Did you find this content mentally engaging? Yes or No
- Did you feel positive or negative about this content? Positive or Negative

Once feedback was given another interface screen appeared to allow the next video in the sequence to be played

- Play next video? Yes or No

This procedure was repeated until (upon the second opaque screen) a message was received by the experimenter, that the system was building a classifier. The number of videos used to build the classifier was noted by the experimenter, and using a simple division ratio $(40 - n) / 4$ determined the amount of videos needed to complete 4 builds of the classifier within the bounds of the experimental material. If too many videos (i.e. >12) were used during the initial build process the

application was terminated and the experiment stopped, given that the remaining stimulus material would produce unequal classifier build phases (i.e. classifier training data set sizes would be unequal between builds making accuracy comparison between builds impossible). If however, the ratio of videos left allowed for 4 equal classifier builds the experiment proceeded.

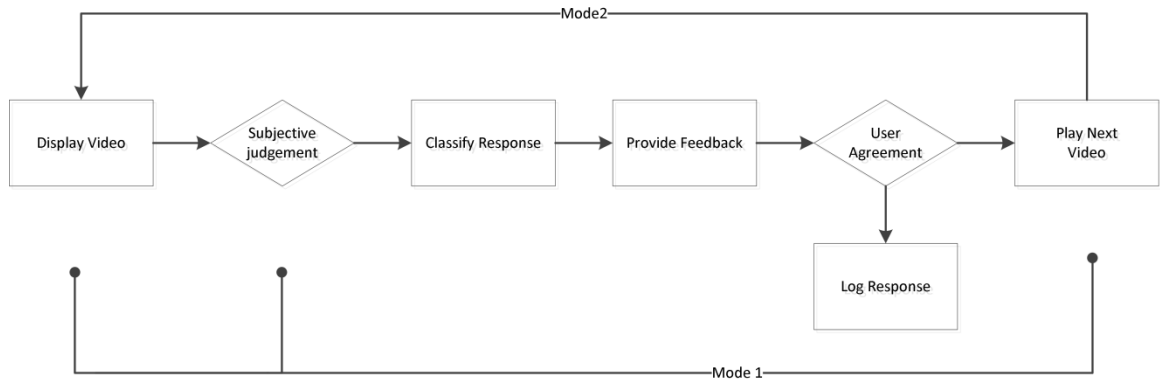


Figure 11.5 Experimental “Wizard of Oz” interaction procedure

Mode 2 (classify) involved all of the procedures required in mode 1, however extra stages are included to qualify the “wizard of Oz” experimental protocol. In this mode after each video trailer is displayed and subjective feedback is given and before the next video is displayed, the psychophysiological interest response is classified by the system, this classification is displayed to the experimenter on a separate screen opaque to the participant. The classification analysis is then told to the participant, who is asked verbally if they agree with the systems classification of their interest response, this response is then noted by the experimenter and associated with the systems response. This sequence of events continues until the new build query point is reached, whereupon after the previous classification has been given to the participant and noted and before the next video trailer is played a system build order is completed to aggregate the classification vectors into a new training data set, and the procedure begins again.

11.6. Analysis

EEG (beta/alpha ratio) and autonomic data (mean and standard deviation of IBI (HR) and SCL) were extracted from a 60 second stimulus epoch for a total of 40 stimulus events. All feature data were derived from raw signal captured from sensor hardware in real-time and subject to filtering and epoch analysis (see Figure 11.4). For autonomic measures features were captured using a 12 second data window with a moving window of 6 seconds (from which averages are taken). For EEG, features are captured using a 12 second data with moving 6 second window (from which power transformations are taken). This approach constructs a feature vector every 6 seconds resulting in 10 -1 (due to the overlapping data 12 second data windows) per sixty seconds stimulus epoch. This gives a potential of $360 - (n * 9)$ classification vectors, where n equals the total number of vectors used to train the classifier initially.

11.6.1. Feature Extraction

For this study, 8 features were derived from physiological signals, for heart rate - mean, and standard deviation of IBI; for skin conductance level, mean and standard deviation; EEG features were derived from a fast *Fourier* transform (FFT) of total amplitude spectra using a 12 second feature window with an overlapping *Hanning* window of 1 second to construct a moving average every 6 seconds, where the ratio $:x$ is expressed as *beta* (power 13-30Hz) divided by *alpha* (power 8-12Hz) at sites FP1, FP2, FPz. Hemispheric asymmetry was measured as the ratio $:x$ expressed as *alpha* (power) subtracting right from left hemispheric activity at sites (FP1-FP2), this results in ten, six second psychophysiological responses (observation) per stimulus segment.

For Cognition: Where the ratio $:x$ is expressed as β (power) divided by α (power) at sites (FP1, FP2, FPz).

$$:x = \left(\frac{y_{\beta}^i}{y_{\alpha}^i} \right)$$

For Valence: Where the ratio $:x$ is expressed as the natural log of α (power) subtracting right from left hemispheric activity at sites (fp1, fp2).

$$:x = \ln(z_{\alpha}^i) - \ln(y_{\alpha}^i)$$

The table of features used for this study is shown in Table 11-2

Component	Measure	Derivative		
		Activation	Heart Rate	iBi-Mean
Skin Conductance	Mean		Stdev	
Cognition	EEG	Ratio β / α	Ratio β / α	Ratio β / α
Valence		Ratio α FP1-FP2		

Table 11-2 Features derived from physiological recordings and the relationship with the interest model

11.7. Results

Across all participants it took on average 7 videos (min 4, max 12) to calibrate the classifier to the participant (build 1), classifier builds took on average 3.75 seconds and were completed in parallel with playing a video trailer, recording and analysing physiological signals. Each build used an average of 8 videos (min 7, max 9) worth of data per build (see Table 11-3). A classification was output continuously every 6 seconds and took on average 0.175s to complete depending on system load, these classifications were visualised within the application interface as a voting ratio (interest as state) but final judgement was output as a hard binary classification (interest as binary).

	Avg	Max	Min
Build 1	7	12	4
Build 2	8	9	7
Build 3	8	9	7
Build 4	8	9	7

Table 11-3 Build statistics average number of videos used per build

The first part of the results analysis consists of receiver operating characteristic curves and the associated area under the curve measure, the ROC displays the ratio between the true positive rate (fraction of true positive classifications) and the false positive rate (fraction of false positive classifications) while the AUC value of the ROC gives the probability that the target state (e.g. a state of high interest) has a higher confidence than a non-target state given the null hypotheses (a value of 0.5). If the curve follows the diagonal, the null hypothesis is met then the tested classifier is little better than chance. In sum, ROC analysis provides information about the diagnosticity of a classifier: the closer the apex of the curve toward the upper left corner, the greater the discriminatory ability of the classifier (i.e., the true-positive rate is high and the false-positive [1 - Specificity] rate is low) and this is value is captured quantitatively by the AUC measure.

The confidence of AUC values for a classifiers ability to discriminate are:

- .5 to .6 : no usefulness
- .6 to .7 : poor to moderate
- .7 to .8 : moderate to good
- .8 to .9 : good to excellent
- .9 to 1 : excellent

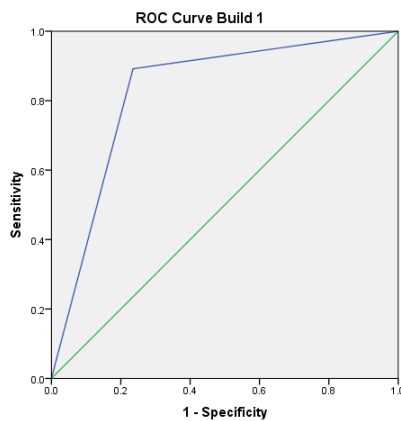


Figure 11.6 ROC Build 1

Area Under the Curve				
	Std. Error ^a	Asymptotic Sig. ^b	Lower Bound	Upper Bound
Area	.048	.000	.735	.922

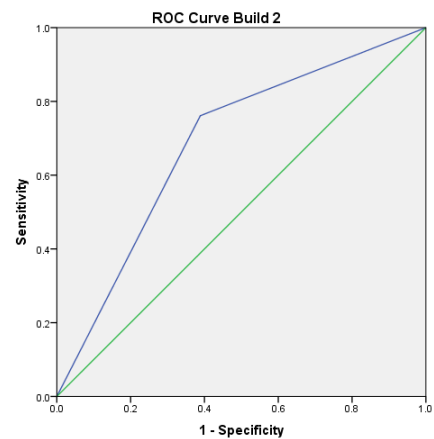


Figure 11.7 ROC Build 2

Area Under the Curve				
	Std. Error ^a	Asymptotic Sig. ^b	Lower Bound	Upper Bound
Area	.073	.013	.544	.828

The ROC curves displayed in Figure(s) Figure 11.6 – Figure 11.9 represent the accuracy of the system as perceived by the users of the system as it is trained over four training sessions (builds). As can be seen from Figure 11.6 & Figure 11.7, during the calibration phase (build 1) the AUC value of .828 shows a classifier with good to high discriminatory power for most participants, the values for (LB) lower bound .735 and (UB) upper bound .922 (which can be interpreted as classifier stability) show some variation, however this variation is not significant (AS =.000). The ROC plot for build two shows a sharp decline in perceived classifier discriminatory performance, and this is reflected in a low AUC value of .686, the results also show a significant (AS=.013) loss of classifier stability (LB .544, UB .828), placing this classifier in the category of poor to moderate usefulness with classification in some cases barely above chance.

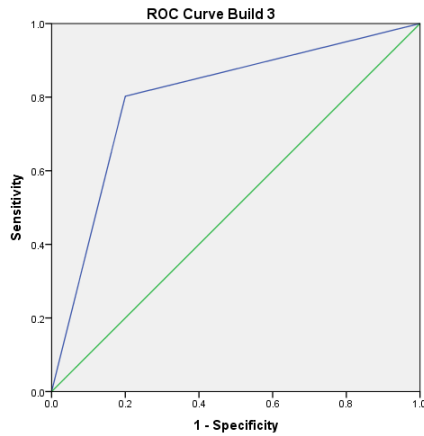


Figure 11.8 ROC Build 3

Area Under the Curve				
	Std. Error ^a	Asymptotic Sig. ^b	Lower Bound	Upper Bound
Area	.058	.000	.688	.915

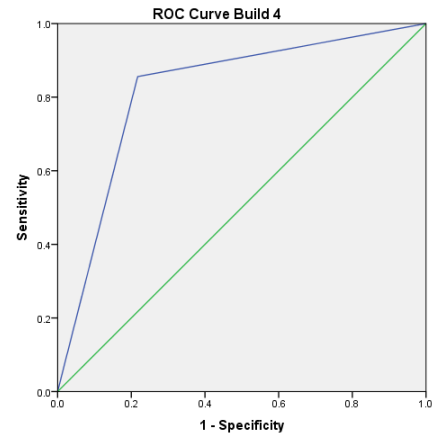


Figure 11.9 ROC Build 4

Area Under the Curve				
	Std. Error ^a	Asymptotic Sig. ^b	Lower Bound	Upper Bound
Area	.054	.000	.714	.924

For build three (Figure 11.8), the ROC plot shows (AUC .801) that the perceived accuracy of the system increases compared to build two, and variance decreases to non-significant levels (LB .688, UB .915. AS =.000). Similarly, perceived system accuracy for build four (Figure 11.9) also increase with an AUC of .819 (LB .714, UB .924 AS=.000) showing a significant decrease in variability.

The plots shown in Figure(s) 11.10 and 11.11 highlight these findings more clearly and clearly portray the user perception of system accuracy decreasing sharply after build one before building to a peak at build four, this peak in perceived accuracy surpasses that observed for build 1. Comparing these plots with the mean system accuracy output (Figure 11.10) it can be seen that while user perception of system accuracy varies across builds, reported mean system accuracy remains stable within a 3% variation across the four classifier builds.

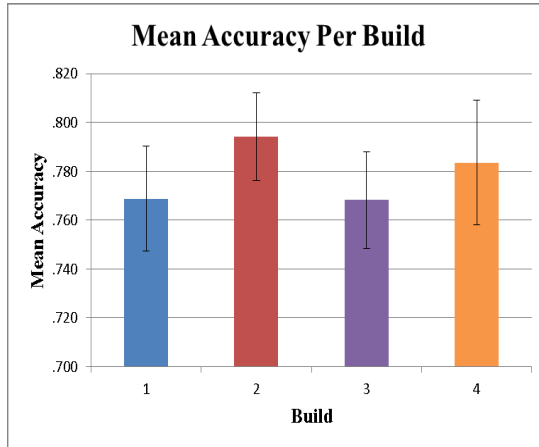


Figure 11.10 Mean system accuracy per build

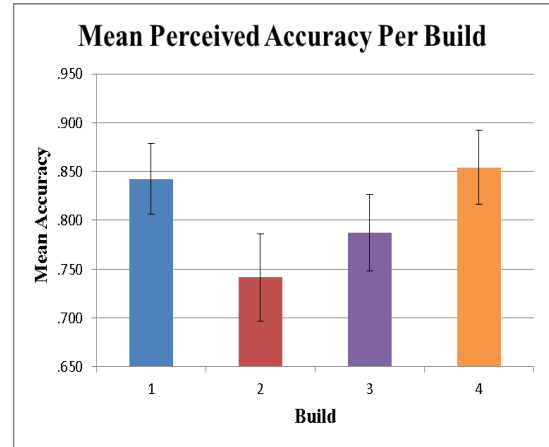


Figure 11.11 Mean perceived accuracy per build

In order to determine if the variance between classifier build was of statistical significance a repeated measures ANOVA was completed. The results from this analysis showed no significant difference between classifier builds [$F(3,11)=0.86, p=0.49$] between participants for system reported accuracy (classifier trained accuracy). For user perceived accuracy the trend of accuracy over each build fell just outside significance [$F(3,11)=3.15, p=0.069$]. Post-hoc testing revealed a decline between builds 1 and 2 (Figure 11.11).

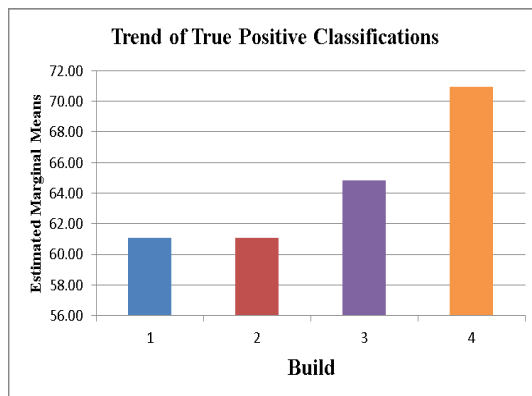


Figure 11.12 Trend of true positive classifications

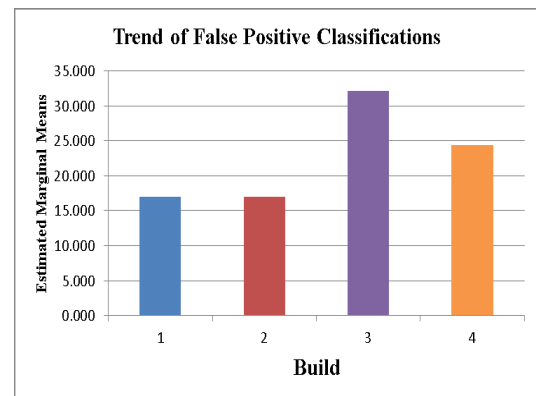


Figure 11.13 Trend of false positive classifications

Looking closer at system versus perceived classification output accuracies, which represent the system in operation; the ANOVA results show no statistically significant variance in true positive [$F(2,12)=0.565, p=0.583$] and false positive [$F(2,12)=1.84, p=0.200$] classifications between builds. However, when the trend for true positive classifications (Figure 11.12) is examined, it can be seen that the number of true positive classifications increased with each classifier build. The trend for the number of false positive classifications (Figure 11.13) shows that the classifier is

stable for builds one and two, with a brief but not statistically significant increase in the number of false positive classifications during operation in build three before reducing in number in build four.

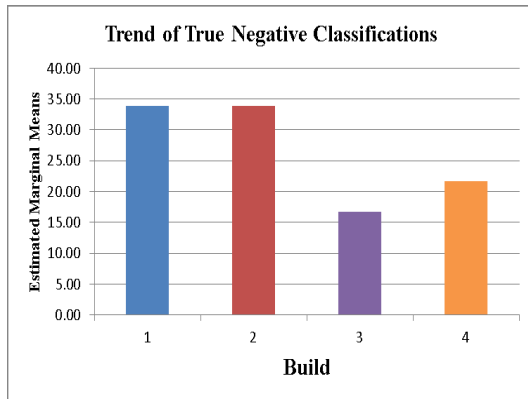


Figure 11.14 Trend of true negative classifications

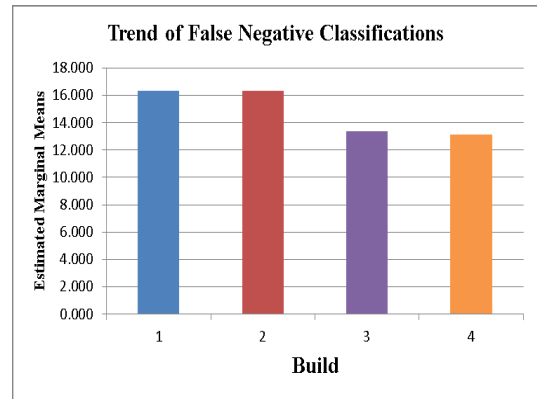


Figure 11.15 Trend of false negative classifications

The ANOVA models for true negative [$F(2,12)=1.86, p=0.20$] and false negative [$F(2,12)=0.21, p=0.81$] classifications revealed no significant variation between the four builds. The trends (Figure 11.14 and Figure 11.15) however, show that in the case of true negative classification these tended to remain stable for builds one and two then decline sharply for build three before increasing slightly for build four. Similarly, in the case of false negatives, erroneous classifications remain stable during builds one and two, before decreasing sharply for build three and decreasing further for build four.

11.8. Discussion

Overall, the results from the ROC and ANOVA analysis demonstrate that the ARTIC system displayed good to excellent discrimination power (i.e. classifying interest as high or low) with no statistically significant difference in system level outputs (in terms of raw accuracy output) between the four build phases. The classification trend plots also support this conclusion, showing that true positive classifications increased continuously after build two and false positive and false negative classifications decreased; this pattern increased the discriminatory potential of the classifier with each successive build. The decline of true negatives corresponds to the increase of true positives due to user preference, which is neither stable nor unbiased. From the standpoint of users, the system was perceived to be accurate in determining their interest state during build one, and then during build two users perceived accuracy to fall; however, perceived accuracy recovered and slightly exceeded build one during the fourth and final build.

With respect to mathematical accuracy (Figure 11.10) values peaked at build two, which is in contrast to the decrease in perceived system accuracy by users. This is possibly an effect of the increase in size of the training dataset, with an average of 8 videos giving 72 new feature vectors (plus initial training data) and not implicitly due to a more accurate representation of the user within the training data, indeed this can be seen in the results of the ANOVA test for perceived system accuracy (Figure 11.11) which decreased during build two. The increase in reported system accuracy and decrease in perceived system accuracy may be associated with the way in which the classifier is trained. The classifier crossvalidates internally over the training data to determine optimal parameters for classification, such that the bias of subjective responses during build one is inherently smaller than at build two. This factor has the effect of rendering the system accurate for build one (both objectively and subjectively) and less accurate for build 2 (subjectively) while objectively becoming more accurate (due to a larger training dataset); successive builds however become more “balanced” to the individual as the pattern of the bias becomes more explicit even though classifier accuracies remain stable on average. It should be noted that, even when perceived accuracies dropped significantly during operation at build two, classification accuracies remain above chance, a factor illustrated in the ROC analysis results (AUC 68.6) perceived accuracy mean (74.15).

The inclination towards greater classification stability and accuracy through the aggregation of more balanced training data across each successive build is highlighted in the following table (Table 11-4) , which shows the comparison between the average perceived accuracy of the system by the user and the average systems (F₁ score) accuracy output.

	Perceived Accuracy	F ₁ score
Build 1	0.85	0.89
Build 2	0.75	0.82
Build 3	0.79	0.86
Build 4	0.85	0.90

Table 11-4 Effects of classifier retraining on average perceived accuracy compared to system accuracy output

The trend in user perceptions of system accuracy; that system accuracy changed from build to build could be explained as a purely human factor, in that subjective judgements and agreement changed negatively even in the face of no statistically significant difference in reported classifier accuracy. Comparing the mean classifier accuracies between builds one and two (76.9% and 79.4%) there can be seen a moderate increase in classifier accuracy, which one would expect from a classifier receiving more data from which to train and make judgements from. However, it was noted in the experiment that a number of participants exclaimed negatively upon hearing system judgements

during system operation after build two, for example, system judgement: “system indicates you were interested in this content”, user: “I was interested, but I do not like sci-fi movies”, thus the feedback provided to train the system in this case and other cases where user preference is challenged is negative. This negative feedback to system judgements transforms into positive training for the classifier however, which may help to explain how user perception of system accuracy steadily increased after build two as system judgements thereafter are based on user preferences, making classifications of user interest transpose into judgements of preference (using interest as an index).

Another possible explanation for the trend in perceived accuracy of the system is the engendering of trust in the system and its judgements i.e. users over the course of the experiment time may have come to recognise that the system was responding to the feedback they provided, as system judgements began to mirror user preference. However, this could also be an effect of the experimental fatigue as the novelty of the system wore off over time, leading to users agreeing with system judgements instead of stating their true preferences. As with all experiments involving subjective feedback this effect is hard to control for. However, this experimental study could be repeated with the addition of rest breaks to control for bias due to fatigue.

There were some methodological issues which may have had an impact on the experimental results, the system was initially planned as a fully automated classification engine, which would output system judgements and receive feedback from users without the assistance of the experimenter. Thus, the “wizard of Oz” protocol may have affected user feedback to system judgements given that human to machine to human interaction presents a different dynamic from pure human to machine interactions. The effects from this form of interaction are undetermined at this time. Another issue concerns the use of binary feedback input, as this represents a “forced” choice and not a nuanced subjective response, reducing the fidelity of the response. This could be alleviated somewhat through the expression of the majority vote as an interest scale or number of discrete interest states.

A factor unreported here and a possibility for further research is the classification outputs, classifications were output as infrequent (i.e. once per stimulus segment) binary classifications (high or low interest). However, the developed application (in accordance with the IBIS model) had the capacity to display classification output using both forms of classification judgement i.e. interest as binary (as used) and interest as state (unused). That is, the majority vote system used to output the binary classification was underutilised and the addition of some simple propositional logic to interpret the votes as a ratio, has the potential to output a number of discrete states which may be more reflective of a user state of interest.

Another factor that may have had an effect on system function, training and output is one of bias. Bias in classification problems presents serious implications for training classifiers, in that classifiers function optimally with an equal number of vectors for each class type. However, for this study, bias is not only uncontrolled for, it is embraced as part of the classifier training methodology. Here classifiers are trained for the individual by individual from the ground up, in that classification vectors are constructed from data captured from the individual, during the task context and then associated with a class label by the individual. Therefore, class bias was an implicit and necessary element in training a classifier for the purpose of determining a level of interest in the stimulus material. This methodology however, makes the developed system context dependant and not generalisable to other tasks; it should be noted that the stimulus material may also have played a part in increasing class bias, in that movie trailers are designed to engender interest in the viewer regardless of preference and hence bias responses towards the high interest category.

This study demonstrated that additional data used to train classifiers has the effect of stabilising classification accuracies rather than significantly increasing classification accuracies. However, the effect of additional training data (as provided by users of the system) upon the perceived accuracy of the system, increased over time exceeding initial perceptions. The relationship between machine level system accuracy and perceived accuracy of the system is a complex one that suggests that; at least in the context of this study; that if systems are built with good to excellent mathematical accuracy, the perception of a systems accuracy by the user and thus acceptance of the system will increase, providing that accuracy reflects a user's preference.

11.9. Conclusion

This fifth experimental study focused on the subject dependent classification psychophysiological of indices of interest in real-time within the context of a laboratory. A real-time application framework was developed to integrate both the process pipeline developed in study four (chapter 10 pp. 98) and the classification output proposed within the revised IBIS model. A proof of concept application based on this framework was used to capture, measure and classify user interest responses to multimedia stimuli in the form of movie trailers; 16 participants took part in the study, which further aimed to ascertain the nature of the relationship between mathematical accuracy as reported by the classifier and the users perception of that accuracy. The results indicated that while machine accuracy remained stable across four classifier training builds, user perception of that accuracy fluctuated across all four training builds, culminating in a perceived accuracy that on average exceeded that of the users' initial perception. Furthermore, the results showed that it is viable to classify and output indices of user interest as preference in real-time and that these outputs would be transferable to other elements of a biocybernetic loop.

12. Discussion

The goal of this thesis was to develop a biocybernetic loop to adapt and personalise information for the individual in a cultural heritage setting. This involved the design and development of a real-time data processing pipeline that translated raw psychophysiological data into control input for adaptive information provision or media tagging. A psychological construct was posited and operationalised as physiological measures of the autonomic and central nervous system to create an inference model for a state of interest. Machine learning algorithms were investigated to determine the efficacy of psychophysiological classification in both offline and online contexts. A series of experiments was conducted to explore the design and implementation issues within two components (inference model and classification) of the biocybernetic loop culminating in a framework that integrated each of the components into a real-time proof-of-concept application.

- Study one explored a psychophysiological inference (as autonomic activation) using a range of autonomic measures and classification algorithms under laboratory conditions.
- Study two investigated cross-session classification of autonomic activation, wherein a support vector machine classifier was trained on session one and applied to data from session two.
- Study three was concerned with the classification of multiple psychophysiological measures recorded using ambulatory sensor apparatus in response to audio material, in a cultural heritage setting.
- Study four represented a replication of study three using multiple sources of media (audio, video, still image and combinations thereof) in a cultural heritage setting.
- Study five was a culmination of the previous studies integrating all findings to create a real-time classification protocol to capture high or low interest in response to video material in real-time.

Thus the work in this thesis has been aimed at understanding biocybernetic control through the use of the biocybernetic loop applied in a cultural heritage material and mixed media context. Five studies were performed to explore this aim and develop methodologies to inform future research into biocybernetic control for uses other than workload management and efficiency.

The components of the biocybernetic loop are: inference, classification, adaptation and interaction, the results of the studies reported here will be discussed within this context. The main findings of the thesis will be discussed with reference to these components in the following sections.

12.1. The inference model

The psychophysiological inference is concerned with the quality of the operationalisation of the target psychological state using a physiological measure or range of physiological measures (Cacioppo, et al 2007). The selection of the psychophysiological features that form the inference model is central to the effectiveness of a biocybernetic control loop. If the physiological measures fail to capture the psychological construct with sufficient sensitivity and reliability then the biocybernetic loop does not encompass the clear link between the user state and system operation that is required to drive accurate system adaptation. The rationale that provided the systems context is one of a cultural heritage experience, specifically aimed at tasks such as viewing artefacts or paintings. The goal of the inference model was to provide an accurate and practical index of viewer interest, and to use this index to drive a process of adaptation to deliver content based upon the interest level of a visitor.

The theoretical basis of the inference model of interest proposed in this thesis was informed by three perspectives; embodiment as proposed by Cannon-Bard (1927), which posited that responses to emotional stimuli or events occur simultaneously in the brain and body; affective neuroscience which offered methods from which to ascertain affective responses using measures of EEG; and core affect posited by Russell (1980, 2003) which offered the circumplex model as a two dimensional space upon which to “map” autonomic and affective responses. These perspectives were combined to create the foundation upon which to base a psychophysiological inference model of interest. Empirical evidence (Kreibig 2010) for embodiment indicates that psychophysiological reactivity can be recorded from the autonomic nervous system by placing sensors on the body. The experimental evidence from neuroscience indicates that cognitive and affective responses can be taken from the brain and measured using EEG (Coan and Allen 2004), and that psychological phenomenon can be placed upon a uni-dimensional or two-dimensional scale, such that magnitude changes in physiology can be measured and classified into scalar or binary states.

Initially in studies one and two (see chapters 7 & 8), the development of the inference model was one of autonomic reactivity as the model was expressed in terms of physiological activation in response to still imagery (a passive task). Study one employed images taken from the international affective picture system (IAPS) (Lang et al. 2008) database and the second study used images from a database of art paintings (Kreplin 2014.), both image sets were used because their psychological properties had been rated subjectively by a large group of individuals. The results from study one indicated that autonomic measures appeared stable within the context of the IAPS image viewing protocol (i.e. stimulus - response - rate, return to baseline, repeat) specifically, heart rate (HR-mean, max) heart rate inter-beat interval (ibi) (mean, max), and skin conductance level (SCL-mean and

area) were identified by principal component analysis (PCA) as providing the most variance in response to two conditions of image (high or low activation). Similarly, the results from study two also indicated the importance of HR, HR-ibi and SCL when measuring autonomic responses (as activation), in this instance HR (ibi) and SCL mean and standard deviation (stdev) were identified as providing the most variance in psychophysiological responses towards the stimulus material, due to the correlation with HR (ibi) the heart rate (as beats per minute and variants) measure was removed at this stage as part of the process of feature dimension reduction and to aid in classifications of activation responses. The measures identified in these studies combined to form an index of physiological activation were represented as HR (ibi) mean and stdev, and SCL mean, area and stdev.

The next phase of development for the inference model was to extend the psychophysiological operationalisation beyond autonomic activation. The concept of interest as a psychological entity as described by Berlyne (1960) and Silvia (2008, 2010) was investigated and key elements of the cultural heritage experience as described by Pine and Gilmore (1998) were distilled into a model with three dimensions designed to fully encompass an inclusive concept of interest. The model consists of three dimensions of perceptual representational processes (Hidi and Renninger 2006), which are mapped onto a unidimensional scale ranging from high to low interest:

- Cognition, which captures the novelty and complexity of the stimuli i.e. familiarity vs. unexpectedness and intricacy vs. simplicity
- Activation, which captures how stimulating the stimuli is
- Valence, to capture the level of positivity or negativity towards the stimuli

To expand the range of the inference model to include measures from the cognitive domain (as cognition and valence) a third study was completed (see chapter 9), in this study the cognitive component of interest was identified with activation of the rostral prefrontal cortex (rostralPFC) i.e. Brodmann's area (BA) 10, an area of the brain dedicated to working memory, attentional control and novel problem solving (see Ramnani & Owen, 2004). The rostralPFC has also been associated with a wide range of cognitive process, ranging from the selection and judgement of stimuli held in short term memory (Petrides 1994) to reversal learning and stimulus selection (Dobbins et al 2002). The valence component of the interest model was identified with activation of the medial prefrontal cortex or BA 8, this area has been associated with processes that involve the motivational or emotional value of incoming information (Tataranni 1999, Rolls 2000) and a link has been proposed between asymmetry of frontal alpha activation and emotional states (Davidson 1993). It has been hypothesized that greater left activation of the prefrontal cortex is associated with positive

affect whereas greater right side activation is linked to negative affect (Davidson & Chapman, 1990, Chapman & Henriques, 1990, Lang, 1995, Silbermann & Wiengartner 1998, Davidson, 2004). This study utilised genuine cultural heritage material in the form of audio narratives to stimulate psychophysiological responses and the measures included in the model were expanded to include those indicated in the literature from electroencephalography (EEG) i.e. cognitive activation and frontal EEG asymmetry. From these measures six features were derived, four were of cognitive activation derived as a ratio of beta power divided by alpha power at sites FP1, FP2, F3, F4 and two were of valence captured through frontal asymmetry expressed as the natural log of alpha power, subtracting right from left hemispheric activity at sites FP2-FP1 and F4-F3.

Within the context of the research presented in this thesis the three dimensions of the interest model formed a many-to-one inference between the measures of autonomic and central nervous system activation and the psychological construct of interest. However, there were a number of issues raised concerning the stability of the psychophysiological inference in a general sense, which were highlighted by the results reported from study two. This study tested the stability and reliability of psychophysiological responses over a number of experimental sessions using the same or similar stimuli. This material was presented in a test-retest format, in which images were shown on session one, then similar but novel examples of those images were displayed during the second session with the original images being re-tested during the third and final session. During each session autonomic psychophysiological responses were recorded and a classifier was used as a determinant tool; the results from the classifications showed that overall the two measures HR-ibi and skin conductance level (and associated features) were only moderately stable over repeated sessions. That is, classification accuracy was high on session one but subsequently declined during sessions two and three. This result is consistent with earlier work (Arena et al 1983, et al 1989, Waters et al 1987) that psychophysiological responses are reliable for baselines and procedures but these responses are subject to significant variation both within session and across sessions. In this case classification tests were performed on both a subject dependent and subject independent basis and in the case of the subject independent test the decline in classification accuracy was more pronounced. This result could be explained by the presence of inter- and intra- individual differences in psychophysiological responses that may account for the sharp decline in classification accuracy from session to session.

The lack of day-to-day reliability in both psychophysiological response and classification accuracy essentially narrowed the scope in which the interest inference model was perceived to be diagnostic of a viewer's level of interest i.e. the inference of interest is valid only within the context of the current task and not generalisable across sessions but may generalise across stimulus types. This finding had far reaching effects for the design of the proposed system and how to classify the interest response; these effects will be discussed in subsequent sections.

The validity of the inference model to generalise across different stimulus media types was tested in two studies (see chapter(s) 9, 10). Study three was conducted to test the three component interest model with audio narrative stimuli and participant in a standing position. The three components of the inference model activation, cognition and valence were considered alone and in combination in this study. In this instance high classification rates were achieved for all three components of the model when classified separately. However, the highest classification accuracies were achieved when components of the model were combined, specifically when measures associated with activation and valence were combined. In Study four the interest inference model was further refined to reduce the dimensionality of features and stabilise the sensitivity and diagnosticity of the model in real-time environments when recorded using ambulatory sensors. Activation was reduced to four features, two for heart rate and two for skin conductance, cognition was reduced to three features measured from FP1, FP2, FPZ, and valence was reduced to a single feature measured from FP1, FP2. For this study mixed media stimulus material was used and a classifier was applied to the psychophysiological data as a determinant tool. Classification rates for this study were deviated from the results of study three, in that the single component classification accuracies for cognition and valence alone and in combination yielded poorer overall classification. The result for the activation component was in line with those reported in study three, as was the combination of activation and valence, showing in this instance that responses from autonomic activation and valence from hemispheric asymmetry best captured the viewer interest response.

The combined results from studies three and four appear to indicate that interest should be represented not just an autonomic measure of activation but as a multidimensional construct. The multidimensionality enhances sensitivity in the sense that different components can be engaged by specific media i.e. autonomic activation to audio stimuli or different types of material i.e. activation of Fz to material with cognitive challenge or activation of frontal EEG asymmetry to emotional material. Viewed in terms of biocybernetic control and the cultural heritage rationale of the research the inference model developed to capture viewer interest meets the requirements for a stable many to one relationship as defined by Cacioppo, Tassinary and Bernston (2000). Here, interest is captured by features of heart rate skin conductance EEG activation and frontal asymmetry and the link between a viewer's state of interest and physiological reactivity has been

demonstrated with a good to excellent degree of discriminatory power. However, this link gains its validity from basing the selection of measures on solid theoretical evidence and past research in psychophysiology, and classifiers that are trained using subjective judgements. Thus, the validity of the inference must be assessed within that context.

12.2. Classification

Classification concerns the identification of the psychophysiological state in real-time or near real-time. It is important that information passed from this stage be timely if the loop is to function dynamically. The choice of classification algorithm is crucial at this point. The classifier must be capable of processing and categorising information in both an accurate and timely manner. The cost of misclassification of user responses must be considered carefully as ultimately the classifier feeds forward judgements into the adaptation engine and thus shapes the users' perception of the accuracy of the system.

The initial classification of the interest response was completed using the *k*-nearest neighbour algorithm (KNN) (see chapter 7) and the results compared to two other classification algorithms, regression decision trees (RDT) and support vector machine (SVM); KNN was tasked with discriminating between three levels of autonomic activation (high, medium and low), and the classifier was trained using two types of class label; labels provided by IAPS survey (Lang et al 2008); and subjective responses provided by participants of the study. It was concluded from the KNN classification results that the KNN classifier was sensitive to noise within the physiological data which parallels the literature on this issue (Petrantonakis & Hadjileontiadis, 2010). Noise, in this context, was implicit in the medium level of autonomic activation which was barely differentiated from either high or low activation responses. Furthermore, when the classifier was trained using subjective judgements, the number of classes was reduced to a two class problem (involving high and low responses) and the dimensionality of the feature data reduced, the KNN algorithm fared little better outputting classification accuracies of 67% (pp.56 Table 7-6). Thus KNN can be seen to be a poor classifier for physiological data where the magnitude response difference between one class and another is small, or where signal artefacts may be present in the data. Similarly, the results from the RDT classifications demonstrated a lack of robustness when applied to the same data, outputting the same 67% classification accuracy. However, when the SVM was applied to the same data, classification accuracy improved dramatically to 83%; this finding demonstrated that the SVM algorithm was better able to deal with noise within the physiological data. Furthermore, these results indicated that a subject dependent approach to both classifier calibration (i.e. training) and classification may prove to be most accurate as reported in the recent review by Novak et al (2012) (see chapter 3 pp29). The results of the first study both

informed and supported the decision to move to a subject dependent approach to the classification of psychophysiological responses using SVM and this approach was used for all future studies.

The first study also indicated that post processing techniques (such as normalisation) may not be a requirement for high accuracy classifications in the context of a subject dependent approach. The highest classification accuracies in the current work were gained using raw feature data (absolute values) in all classification techniques, and there is some support for this view in the psychophysiological literature (Waters et al 1987), who found that absolute scores were more often stable in comparison to change scores.

The second study was performed as a second test of autonomic activation classification using a subject dependent approach (see chapter 8 pp67). In this case psychophysiological data were classified as raw (untreated) or normalised feature vector variants, and subjected to feature dimension reduction using PCA. As with study 1, classification output remained unaffected by either normalisation or PCA; in the case of PCA the range of autonomic measures used in this study may have been relatively inter-related making the impact of the PCA negligible. The results from the classification trials, which compared training the SVM classifier with labels provided by survey or labels provided by subjective judgement for a two class high from low activation discrimination, are shown in Table 12-1. These data illustrate how a classifier trained using subjective labels provided a clear advantage in terms of mean accuracy when compared to the classification using standardised survey labels. As a contrast a subject independent classification trial was completed and the classifier reported a high training accuracy followed by a sharp decline in accuracies for the test and retest conditions (Table 12-1).

Training Method	Train	Test	Retest
Subjective Labels	80.43	70.38	70.83
Survey Labels	86.67	59.17	45.83
Subject Independent	81.25	54.17	55.00

Table 12-1 Training method and mean classification accuracy for a high from low activation discrimination

The results shown in Table 12-1 which displays mean accuracies for a discrimination of high from low activation suggest that collecting baseline psychophysiological measurements from which to compare for variance may be an unnecessary step in the current context. It was found that classification accuracies were maximised when the classifier was tasked with discriminating between the target high or low activation states as opposed to high or low states from a baseline

state. Thus, removing the baseline state comparisons has the effect of decreasing computational complexity and cost by reducing both the complexity of the training data, and the amount of steps needed before classifications can be performed.

The results from the second study viewed from the standpoint of the day-to-day reliability of a classifier trained using psychophysiological data indicated that both psychophysiological measures (due to intra-personal differences) and subsequent classifications are unreliable over repeated sessions. This finding has clear implications for the training and use of classifiers in the larger context of biocybernetic control as the implication from the second experiment is that a system may require calibration for each user of the system for each session of use. However, this is a subject for further investigation; when taken on face value machine learning classifiers are by definition learning engines and the aggregation of much larger datasets over many sessions may elicit different and more positive results in terms of inter and intra-subject generalisability.

Study three expanded the range of psychophysiological data input for the SVM classifier to include EEG data and utilised naturalistic cultural heritage stimuli involving audio narratives (see chapter 9). This study tested the capacity of the inference model to provide classification vectors that were sufficiently sensitive and diagnostic in a simulated environment when the participant was in a standing position to mimic posture in an actual cultural heritage institute and to test the capability of the interest model to generalise to different stimulus material. The classification of the interest response in the third study moved to a multidimensional model, which allowed a subject dependent classifier to be trained using unprocessed feature data for each aspect of the inference model (activation, cognition and valence) individually or in combination; in addition, classification accuracy was determined by comparing classifier recall accuracy and subjective responses. The results (pp. 90 Table 9-2) demonstrated high classification accuracy for each component of inference model when classified alone 90%, 83% and 85% for activation, cognition and valence respectively. However, by combining the features from each component of the model classification accuracies improved further i.e. Activation and cognition (92%), activation and valence (95%) and activation, cognition and valence together (93%), presented with high classification of recall accuracy and a more stable classifier across all participants. A subject independent classification of the feature data was generated for comparison and in this comparison subject dependent classification achieved higher accuracy, adding further strength to the validity of using a subject dependent approach to classification within the current application. These results indicated that each component of interest model could be classified individually and in combination with a high degree of accuracy in response to audio narratives, and this informed the further development of a classification model which could potentially output a binary judgment as a “composite”

aggregating all three component feature sets together or “component” with each feature set classified separately.

One issue that may have impacted on the high classification accuracies in the third study was the way in which subjective judgements, and therefore the classifier training labels were elicited. In this instance participants were offered a forced choice i.e. participants were asked to rank which audio narratives were interesting and how interesting each narrative was. This choice while mirroring the binary aspect of the classifier may not have captured the participant’s interest response as effectively and with the same level of sensitivity as a more nuanced approach such as a 7-point Likert scale. However, when considering calibrating a classifier for use in a real-time system, such a forced choice may be necessary because this type of subjective choice may deliver more consistency for binary classifications and this approach has some support in the literature (see Levillian et al 2010 and chapter 3 pp 31) who used this form of subjective feedback to ascertain a user’s level of amusement and challenge in a gameplay task.

The fourth study iterated the subject dependent classification and recording methodology utilised in previous experiment and the goal of this study was to classify responses to genuine mixed media cultural heritage material (see chapter 10). The IBIS model (pp. 98 Figure 10.1) of classification was proposed to utilise the composite and component model binary classification output identified in study three. In addition, this study explored classification performance using two types of classifier training schema, a composite schema in which a classifier is trained using labels derived from a “composite” of three likert scales to create a single label, and a component schema which used a training label for each component of the interest model derived from subjective judgments given for each component of the interest model i.e. activation, cognition, and valence. Similar to study three the results for the “composite” trained classifiers showed that the combination of activation and valence (F_1 0.82) and activation alone (F_1 0.87), presented with the highest classification accuracies. The classification output from the full interest model was also above chance (F_1 0.81). The results for the “component” trained classifiers on the other hand, with the exception of activation (F_1 0.90) showed a marked decrease in accuracy for cognition (F_1 0.66) and valence (F_1 0.69).

The results from the fourth study provided further evidence that the composite classification model, which reduced the scores of all three components into a single a binary high or low label, was the best method of classification for these data. The classification rates for component labels with the exception of activation were poor, showing either a potential weakness in the component training schema or the way in which the labels were derived before the classifiers were trained. The results from the generalised model classification of the data which followed the same classification schemas produced accuracies similar to those reported in previous studies i.e. a decrease in classification accuracy. However, in this case the classification of activation alone using either composite (F_1 0.73) or component (F_1 0.85) schemas produced higher levels of classification accuracy, specifically in the case of the component schema. This finding highlights a potential issue in how labels are derived for training the classifiers. The high classification accuracy of activation for both subject dependent and independent methods of classification could be tentatively explained with reference to how well participants understood the process of subjective self-assessment. A better understanding of what activation meant in a conceptual sense and with reference to personal experience would equate to scores more reflective of the activation state, whereas even minimal confusion as to the nature of cognition and valence would result in scores that were less reflective of those states (and those psychophysiological variables associated with those states). Another possible explanation could be inter and intra-personal differences within responses, in that the stimulus materials used for the study may have universally elicited a high physiological activation response in each participant and low responses in terms of cognition and valence.

The fourth study also proposed a process pipeline framework (pp. 100 Figure 10.2) to move away from commercial software and post-hoc data processing and to create a self-contained pipeline for data collection and post-processing. In this respect the experiment was a success, in that the developed process pipeline output physiological features that were of sufficient specificity and diagnosticity to allow for the inference model to be classified with moderate to high accuracy. The IBIS model also proved partially successful with the composite classification schema providing outputs that could be made readily available to other processes in a biocybernetic loop, such as an adaptation process.

Classification for the fifth and final study (see chapter 11) was completed in real-time; using an application developed using a revised IBIS classification model (pp. 118 Figure 11.1). The IBIS model used the composite schema to perform classifications and a majority vote to output the final binary classification and the classifier was trained at runtime over a series of four builds during the course of the experimental procedure. The procedure involved viewing multimedia stimuli (movie trailers) and classifier training was performed by way of subjective judgements given at the end of each segment. After an initial training (calibration) build classifications were completed on average every 6 seconds during a 60 second stimulus epoch. This classifier was rebuilt at key points during the procedure to incorporate more training data into the underlying SVM. The results from this experiment were harder to interpret than previous studies due to the fact that user interactions with the system were iterative and dynamic. The results of the real-time classifications (pp. 132 Figure 11.10) i.e. one of mathematical accuracy as provided by the internal crossvalidation and parameter selection procedure, indicated that the classifier was stable with no significant variation across the four build sessions and achieved accuracy in the range of 76.8 - 79.4%.

The general pattern of the classification results reported in this thesis indicated that classifiers respond best to data with a high degree of separation between the various classes, and that the KNN and RDT algorithms are not suited to real-time classification of psychophysiological data, despite their computational simplicity and transparency as discussed in chapter 3 of this thesis. In terms of classifier calibration the results indicate that classifiers are more accurate and stable when applied to psychophysiological data subject dependently i.e. classifiers trained using psychophysiological data recorded from a subject who then associates the class labels with that data. Furthermore, the results reported in this thesis also indicate that subject dependent classification is best performed on the same day as the classifier is trained, explained by inter and intra-subject response differences in same day and different day responses. Machine learning theory holds that classification accuracies will improve over time given more data to generalise from (Bishop 2006), thus the day to day stability and accuracy of a classifier may be increased if data is aggregated for training after each session of use. However, this lag between the application of the classifier and aggregating training data would mean a greater proportion of misclassifications of the target state until the optimal training set of psychophysiological responses were reached.

A noteworthy issue with subject dependency within classification is one of class bias, in an ideal world classifiers function optimally with data that is highly separable and contains an equal number of classes to prevent over fitting of the classifier to the training data. However, in the case of the research presented in this thesis bias is implicit to subjective responses (interest being wholly subjective phenomena) and these responses were ultimately used to train classifiers tailored specifically to individuals. A methodological issue identified in study four which may have negatively affected the classification results concerned (based upon intra-subject difference) was the number a length of rest and subjective reflection periods, these periods may have had the effect of altering intra-personal physiological responses due to biological changes (such as hunger, fatigue or simply boredom). The results from experiment five (see chapter 11) appear to indicate that this effect can be decreased by calibrating and using a classifier within the same session. However, while the effect of intra-subject difference is lessened this may also indicate that the scope in which classifiers based upon subject-dependency would be limited to same session or same day applications at best.

The following guidelines represent a summary of the findings from the classification analyses performed in this thesis, which may prove useful to designers of similar systems:

- Signal processing, such as filtering and artefact removal are major factors that will affect classification accuracies
- Recording physiological baseline data to compare against an unnecessary step
- Principal component analysis can prove valuable in cases where physiological measures are not highly inter-related
- Normalisation of psychophysiological data is not indicated for use in subject dependent systems and presents increased computational cost with no increased benefit to classification accuracies when compared to absolute values
- Normalisation is potentially useful for use in subject independent systems where generalisation across a population is indicated
- Classifiers perform poorly when tasked with classifying data across repeated sessions and exposures to same stimuli
- In the case of subject dependent applications, classifiers should be trained for each session using a combination of psychophysiological data and subjective assessment for training data captured during that session or just prior
- When designing systems to integrate real-time machine learning classification into biocybernetic control loops, there is a trade-off between the time required for classifier training, accuracy of the resulting classifier and speed of deployment
- Classifiers can more accurately reflect a user's appraisal of psychophysiological state when trained repeatedly during the same session, resulting in more accurate classifications and potentiating an increase in user acceptance or trust towards the system

In sum, a model of classification was posited to output interest as a binary state or interest as a scale (IBIS), this model proved effective in study three when applied to audio stimuli (in the form of audio narratives), but proved less so when applied to mixed media in study four. In this regard only the composite model of classification was proven to be effective in both studies in comparison with the component model. In the final study the IBIS model was revised to integrate the composite model as one with two forms of classification output i.e. interest as binary or interest as state where both these forms of output are interpreted from multiple classification outputs using a majority vote as the final discriminant. The results from the final study showed that the IBIS model output was effective in real-time and that the classifiers built for each user by each user provided a stable inference of a user's interest preference when trained over a number of builds while users were exposed to video stimuli.

12.3. Interaction and Adaptation

In the fifth study, participants were provided with feedback in real-time about the systems classification assessment of their level of interest. Following the component mode of classification from the IBIS model, judgements were given as binary statements e.g. “the system says you were interested/not interested in this content”, “do you agree?” both the systems judgement and the subjective agreement/disagreement with that judgement were recorded. From this recorded data the relationship between machine and perceived classification could be assessed. Table 12-2 shows the mean accuracies from all participants for both machine and perceived accuracy. These results show that from a machine accuracy standpoint the system displays no significant difference in accuracy across the four build sessions.

Accuracy Type	Build 1	Build 2	Build 3	Build 4
Machine	76.9	79.4	76.8	78.4
Perceived	84.3	74.2	78.7	85.4

Table 12-2 Estimated marginal means: machine accuracy versus perceived accuracy

However, users’ perception of system accuracy provided a different perspective on machine-based classification; after initial calibration the system was perceived to be highly accurate but subjective accuracy fell by 10% during build two before increasing to the original level over the next two successive builds (Table 12-2). A ROC-AUC analysis (Table 12-3) was completed on the accuracy data and the area under the curve portrayed a system that initially displays good to excellent discriminatory power falling into the poor to moderate range during use after build two before rising again in builds three and four.

Area Under the Curve					
Build	Area	Std. Error ^a	Asymptotic Sig. ^b	Lower Bound	Upper Bound
1	.828	.048	.000	.735	.922
2	.686	.073	.013	.544	.828
3	.801	.058	.000	.688	.915
4	.819	.054	.000	.714	.924

Table 12-3 ROC-AUC results

The results from the real-time system show that the trend of classifications moves to more accurate determinations of true positive and true negative classifications over the course of the four builds. This trend is mirrored by a decrease in the number of misclassifications (false positives and negatives) as the classifier receives more training data. This increase of accuracy is not only a function of receiving more data however, but rather a complex interaction between the increase of training data and the way in which participants generate subjective class labels associated with that data. Furthermore, the positive results gained from the interaction of classifier and user may indicate that fully embracing subject-dependency within the calibration cycle of a biocybernetic control system leads to more accurate classifications and possibly systems that are more quickly accepted by the user.

Another explanation for why user perceptions trended towards the positive as the experimental session progressed is to consider how users' perceptions may have altered with increased exposure to the system. One factor is the "halo effect" which is the unconscious alteration of judgement in response to information perceived to be authoritative (Nisbett & Wilson 1977), in this case the "machine" as it were is perceived to be authoritative and thus the users assessment of how accurate the system was in comparison to their own may have been biased towards the system early in the experimental procedure. For example:

- Build one, positive bias towards the system in comparison to self-judgements
- Build two, disillusionment as system judgements clash with self-assessments
- Builds three and four, time and exposure have begun to wear down the user making them again positively biased due to ennui.

A secondary explanation could be that by builds three and four time and exposure to the system has "forced" a submission from the user that "the system knows best", increasing the likelihood of agreement and possibly engendering trust. However, the results indicated that the system "learned" a user's interest preference over the course of four builds resulting in positive agreement with system judgements towards the end of the experimental procedure.

Adaptation is concerned with employing the governing rule set or purpose of the loop, that is, what actions should be taken at the interface in response to classification judgements about the user's state. The results from the studies reported in this thesis indicate that the proposed operationalised model of interest can be classified with a reasonable degree of accuracy and could, therefore, be a candidate for use in systems that incorporate a user's interest state into a bio-cybernetic control loop to drive adaptations, such as in the case of cultural heritage

In work completed previously⁵, an example of a possible cultural heritage application was described, that posited a form of biocybernetically controlled “adaptive curation” that adapts content and information depending upon a user’s level of interest. To complete this task the “INTREST framework” was proposed (shown in Figure 12.1). The proposed framework is based loosely around a series of narrative structure. These structures denote the purpose and placement of cultural heritage content (narrative arcs) used to stimulate psychophysiological responses. The INTREST framework is separated into five phases: narrative structure, adaptive story elements, physiological measurement, classification and narrative path. Each phase represents a requirement or process needed to form the system and can be summarised as: INput, sTimulus, REsponse, claSsification, ouTput.

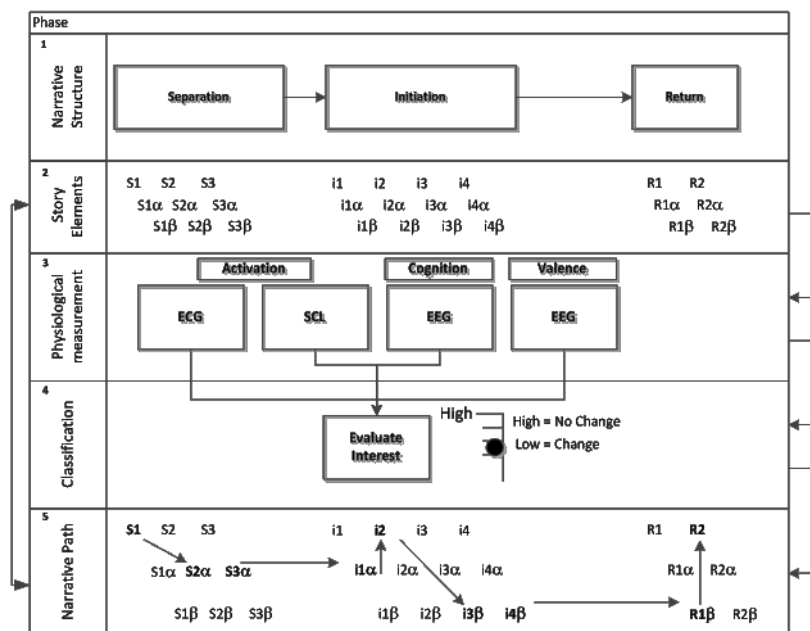


Figure 12.1 A Cultural Heritage “digital curator” Framework

Phase one is concerned with the conceptual narrative structure, each block representing one element of a narrative arc. Phase two consists of adaptive story elements or information blocks created to fit in within the narrative structure, such as semantically linked mixed media content about the exhibit or installation. Phase three is focused on measuring interest in response to story elements. In phase four the classification of the level of interest is completed, outputting either high or low interest in response to story elements. This output is then used in the final phase to pull new content from the adaptive story elements, to create the narrative path of adapted information. When the application is operated a starting point is chosen by the system; the psychophysiological

⁵ Published in Karran, A.J., Kreplin, U., "The Drive to Explore: Physiological Computing in a Cultural Heritage Context", Advances in Physiological Computing, Fairclough, S. F., Gilleade, K (eds) - Springer, (April 2014)

response to this content is then evaluated; if visitor interest is high then the system continues to draw content from the store for that narrative arc; if visitor interest is low, new content is drawn first from the same store, or if responses still indicate low interest content is drawn from the store of content for a different narrative arc in an attempt to elicit the more favourable high response.

Within this scenario the narrative arcs represent elements of the same “story” but with different content databases. For example, consider a Han dynasty vase (Circa 202 BC). The story starts with information about the vase and its provenance, interest is classified as low, new information is provided from further ahead in the story, however the response to this content is still classified as low interest. The system then adapts to these responses by drawing content from a concurrent narrative arc, this time the content is about how the vase is made and the ceramic processes of the period, this elicits a high response, after a number of content blocks are displayed the response is still high and remains that way until the content is exhausted. From this information the system can ascertain, that for this user of the system, content that explains how artefacts are created is of interest. When the user then moves to a different artefact the system utilises this information as the new start point for further content provision, if responses remain high, then the system adapts to give only this type of content. When responses drop to low interest, the entire process begins again.

In this example, the adaptation strategy is one of a fully automated system meant to enhance a museum visitor’s experience of paintings or artefacts and not to track the exposure or educational aspect of cultural heritage, and the results from studies three and four indicate that this classifier training method could prove effective. However, the results from study five also indicate that continuous classifier training could also prove effective; integrating this approach would allow the adaptation strategy to be expanded to create an intermediate stage hybrid semi-autonomous system (Parasuraman, Sheridan and Wickens, 2000) that makes content adaptations using a combination of psychophysiological responses and user decisions that is fully subject-dependent, in that the system is trained using user interest preference.

The choice of adaptation strategy is closely linked with the classification model and accuracy of classifications and these have implications for how the system is perceived when it used. The cost of misclassifications within adaptive systems is both a technical and user perception issue; technical in that misclassifications can cause the system to function incorrectly; and misclassifications can cause perceptual shifts in users causing a lack of trust and an unwillingness to use the system. Within the context of the cultural heritage application, for the fully automated system the cost of misclassification is low possibly resulting in a low opinion of the system but having no real impact upon the visitors experience as the system is interactive only in so much as it adapts to user psychophysiological changes. However, in the case of the semi-autonomous system

in which user interactions are implicit to system function, misclassification costs are higher as users perceive more of the errors while the system is used; this can be alleviated somewhat by the continuous training of the system while it is used so it better reflects the interest preference of the user and by reducing the frequency of classifications.

The timing of classification output can be crucial to the adaptation strategy and to a user's perception of misclassifications, when the cost of misclassification is low, the frequency of classifications can be lowered to reduce the amount of errors received. The results from the fifth study provided an example of an effective strategy as the system was tuned more to the user by training, system level misclassifications were reduced and at the user level the perception of misclassifications was reduced - and as a direct result trust may have been engendered in the biocybernetic loop. However, the engendering of trust highlights a possible issue in the application of adaptive systems. Results showed user agreement with system judgements increased from the starting values, even in the face of reported stable system accuracies across all builds. Although this shows trust has been engendered (i.e. a changed "attitude", Miller 2005), it is not necessarily a positive factor, as users judged system accuracy to be higher than it actually was according to mathematical criteria. This factor would indicate that trust in adaptive systems can be engendered fairly quickly which is a positive in the case of educational applications. However, in the case of safety critical applications, the ease with which a user may trust the system is negative as users may overestimate actual accuracy leading to possible cascades of human-machine errors. This finding while potentially important is preliminary however and requires further investigation with a larger sample size before generalising.

12.4. Limitations

There are a number of limitations to the work presented in this thesis, the first of which involves sample size. Due to the complexity of the experimental procedures and length of data analysis an average of 10 participants per study was used. This makes the indications and conclusions made within this thesis preliminary in nature and less generalisable than if a larger population of participants was used throughout the programme of research.

The amount of features/vectors used for training and classification was an issue for concern in the current thesis. In studies one and two the standard IAPS image viewing protocol was used as the common physiological denominator from which to derive the temporal basis for psychophysiological recording i.e. the stimulus epoch. Thus for a 10 second viewing period 1 feature vector was created (based on a 6 second SC response), this resulted in a low number of feature vectors for these studies. For study three, the impact of audio stimuli was unknown and

additional measures were under investigation. In this study each feature vector represented 17 seconds of audio narrative resulting in a low number of feature vectors for classification, and this may have biased classifier training due to a scarcity of data in the training space, possibly leading to some over fitting. The effect of training set size and classifier performance is an area of much research and debate. However, in each study where classification was performed, the cross-validation procedures used to report classifier accuracies was informed by the literature, as those best suited to small training set sizes indicated in real-time applications (Isaksson, et al, 2008). Another factor concerned with classification, and one that requires extensive investigation, is the effect of class bias. This was discussed in study five (pp. 120-121), in this study class bias was embraced as part of the classifier training protocol on the premise that the labels (which represented subjective feedback from the user) used to train the classifier were representative of a user's interest preference. However, class bias has the effect of forcing classifiers to "over fit" to the data, meaning the likelihood of one class being chosen over another is greatly increased. This factor could be addressed by using a lengthier classifier calibration procedure in which a classifier is only build when equal numbers of the respective classes exists at each stage.

Another limitation of the work involves the "wizard of Oz" protocol used as a proxy for system interaction in study five. In this instance the human-machine-human (as machine) interaction may have biased the human subjective response to system classification judgements in either a negative or positive way. Furthermore, there are a variety of ways in which the system could feedback to the user from the very frequent and explicit to infrequent and implicit to providing no feedback at all, here only one aspect of interaction was explored. This would be an area for further study, with the easiest way to determine effects being to fully automate the experimental procedure to provide various types of feedback and observe the results

Another factor that can be seen as a limitation of the work presented here involves the development of the interest model. Due to limitations on time and resources this model may not have been developed in sufficient depth and the interest model may be too inclusive, i.e. the model is based on an *ad hoc* representation of perceptual representation processes as seen specifically from a perspective of cultural heritage and real-time measurement. As the inference of interest is a heterogeneous pattern, other factors involved in the perception and embodiment of interest such as curiosity, preference, drive etc. could have been expanded upon using other physiological measures. Applying imaging techniques such as functional magnetic resonance imaging (fMRI) or functional near-infrared spectroscopy (fNIRS) and a host of autonomic measures to expand the component level of the model may provide a more detailed picture of what it means to be interested. This could lead to developing a model of interest that reflects both the physical properties of the material (such as volume, pitch of audio etc.) and the type of media at the component level.

In this regard, the interest model as operationalised in this thesis could be seen as measuring more the sensory properties of the stimuli as opposed to a viewer's interest in the stimuli, and this could be a possible explanation of the high classification accuracies reported throughout the work. The material used to stimulate responses may have resulted in something more akin to a "visceral" response and not be as representative of interest as a psychological construct. However, controlling for the sensory properties of stimuli may prove difficult, humans are sensory creatures, and our consciousness is interpretive and heavily biased towards sensory inputs and appraisals of those inputs. Thus, removing the sensory properties may cause unintended effects, such as low intensity responses and inaccurate subjective judgements (Levenson 2003), which could include factors other than interest as the stimulus material may indeed be "less interesting" without those properties.

The subject-dependent model of response recording and classifier training which proved to be successful could itself be seen as a limitation of the work, the small sample sizes used to perform the classification analysis, may have biased the research towards subject-dependency resulting in less focus and development of more generalisable subject-independent classification approaches. However, the decision to move to a fully subject-dependent approach was informed both by the literature (Novak et al, 2012) and the results from studies one and two.

Another limitation not directly of the work itself but more of the technology involved is that of the sensor hardware used to record psychophysiological responses, no research was conducted to ascertain the acceptability and comfort with the sensor hardware, which while ambulatory was still bulky and quite involved when attached. Physiological sensor hardware is still in its infancy; however there is currently a movement in commercial markets to push devices that measure physiology in some way for lifestyle monitoring to consumers. This push may increase the likelihood of user acceptance as devices become smaller and more widespread.

12.5. Future work

There are a number of research threads that could be explored within the context of this work, during an early study a classification method was posited, based on inter-subject response differences. In this method psychophysiological responses were measured based upon dynamic trends (increase/decrease) as features of slope; this slope gave an indication of the “direction” of the response to stimuli e.g. heart rate went up, skin conductance went down, during a stimulus window. The direction of response for each feature was then codified as a string of 0’s or 1’s using second order logic derived after consultation with an expert in psychophysiological recording and the literature (Kreibig 2010). The codification process created a series of physiological “signatures”, which preliminary testing showed that classification accuracies were greatly improved in both subject-dependent and subject-independent models. Furthermore, the preliminary findings showed that out of 360 possible signature combinations only 16 were used in the classification analysis and they were common to most participant data resulting in excellent subject-independent generalisation performance. At the time this analysis was performed, it was determined that the amount of time and resources required to verify the results through further studies was prohibitive in the context of the research project directives, thus the approach was never pursued. Codification has since become a lively sub domain of machine learning research which would indicate there is some merit to investigating further.

Another research thread directly related to the work presented would involve further testing and expansion of the IBIS model of classification, specifically to include the voting ratio of classification model which could provide a more nuanced interest state that may be more reflective of the user interest when applied in real-time. Furthermore the component level classification model which proved unsuccessful in study four could be revisited in the context of a real-time application; this classification model out of the three that were developed has the greatest potential to output and present interest as a scale, requiring only a propositional logic layer on top of the classification output to turn three binary classifications into a scale see Table 12-4 for an example.

Propositional Logic : Interest as a Scale

IF	AND	AND	Inferred Interest
Activation +	Cognition +	Valence +	Very High
Activation +	Cognition +	Valence -	High
Activation +	Cognition -	Valence +	High
Activation +	Cognition -	Valence -	Moderate
Activation -	Cognition +	Valence +	Moderate
Activation -	Cognition +	Valence -	Low
Activation -	Cognition -	Valence +	Low
Activation -	Cognition -	Valence -	Very Low

Table 12-4 Example propositional logic giving interest as a scale

These additions coupled with further exploration of the user-adaptive system trust dynamic could prove a fruitfully area of research useful in a wide variety of contexts such as system where function allocation is dependent on the synergy between adaptive system and user.

The limitations of the interactive protocol identified here could be improved upon with further study; a series of experiments could be designed with separate cohorts of participants. Each study could focus upon one aspect of user-system interaction and contain one or more conditions in which a classifier is trained using a variety of methods, ranging from training with random data to training once with a large balanced training set to multiple times with balanced or unbalanced training data. This would demonstrate the effect of bias on classifications or in the case of random training data, demonstrate if there is indeed a “halo” effect involved with using adaptive systems, as one would assume that given the randomness of the training data user agreement with system judgements would remain low. For the interactive aspect, an example study could be to calibrate the system to the user using a separate procedure utilising example video material. Once calibrated the system displays videos in sequence yet provides no feedback about system judgements until the procedure is completed. The system then provides a report about which videos were the most interesting and requests feedback from the user, this removes any effects the wizard of Oz may have had and ensures the user has no preconceived notions about what the system is doing, possibly leading to more viable subjective feedback.

13. Conclusion

The body of research recounted here explored the biocybernetic control loop in the context of cultural heritage. The psychological construct “interest” was explored and a three dimensional model of interest was posited and operationalised as three components of psychophysiological activation involving:

- Cognition, which captures the novelty and complexity of the stimuli i.e. familiarity vs. unexpectedness and intricacy vs. simplicity
- Activation, which captures how stimulating the stimuli is
- Valence, to capture the level of positivity or negativity towards the stimuli

The psychophysiological inference of interest was classified using the support vector machine classification algorithm using both subject dependent and independent approaches. A classification protocol was posited and developed into a process pipeline and application framework to measure psychophysiological indices of interest and output classification judgements of user interest in real-time. A prototype real-time application was used to successfully verify the process pipeline and classification output.

It was found that subject dependent classification of psychophysiological data and training of the classifier is more accurate than subject independent classification and that classifiers perform poorly when tasked with classifying data across repeated sessions and exposures to same stimuli. In the case of subject dependent applications, it was found that classifiers should be trained for each session using a combination of psychophysiological data and subjective assessment for training data captured during that session or just prior. In addition it was found that classifiers can more accurately reflect a user’s appraisal of psychophysiological state when trained repeatedly during the same session, resulting in more accurate classifications and potentiating an increase in user acceptance or trust towards the system.

14. References

- A. Guijt, J. Sluiter, M. Frings-Dresen, (2007). Test-retest reliability of heart rate variability and respiration rate at rest and during light physical activity in normal subjects, *Archives of medical research* 38 (1) 113-120.
- A. Isaksson, M. Wallman, H. Göransson & M.G. Gustafsson, (2008) *Cross-validation and bootstrapping are unreliable in small sample classification*. *Pattern Recognition Letters*, Volume 29, Issue 14, Pages 1960-1965,
- Abbasi, A. R., Akhtar, H., & Afzulpurkar, N. V. (n.d.). (2010). *Towards context-adaptive affective computing*. *Electrical Engineering/Electronics Computer Telecommunications and Information Technology (ECTI-CON), International Conference on*, 122-126.
- Abdi, H. and Williams, L. J. (2010), *Principal component analysis*. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2: 433–459. doi: 10.1002/wics.101
- Anttonen, J. & Surakka, V. (2005), *Emotions and heart rate while sitting on a chair*, in 'CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems', ACM, New York, NY, USA, pp. 491--499.
- Arena, J. G., Blanchard, E. B. Andrasik, E, Cotch, P. A., & Meyers, P. E. (1983). *Reliability of psychophysiological assessment*. *Behaviour Research and Therapy*, 21, 447-460.
- Arena, J. G., Goldberg, S. J., Saul, D. L., & Hobbs, S. H. (1989). *Temporal stability of psychophysiological response profiles: Analysis of individual response stereotypy and stimulus response specificity*. *Behavior Therapy*, 20, 609-618.
- Arnold, M.B. (1960). *Emotion and Personality* Vol. 1 & 2. Columbia Univ.Press,
- Averill, J.R. (1980). *A Constructivist View of Emotion*. *Emotion: Theory, Research and Experience*, pp. 305-339, Academic Press
- Aysin, B., Aysin, E., (2006). *Effect of Respiration in Heart Rate Variability (HRV) analysis* *Engineering in Medicine and Biology Society*,. EMBS '06. 28th Annual International Conference of the IEEE, vol., no., pp.1776-1779, Aug. 30 2006-Sept. 3 doi: 10.1109/IEMBS.2006.260773
- Bailenson, J.N., Pontikakis, E.D., Mauss, I.B., Gross, J.J., Jabon, M.E., Hutcherson, C.A., (2008). *Real-time classification of evoked emotions using facial feature tracking and physiological responses*. *International Journal of Human-Computer Studies* 66, 303–317.
- Becker, C., Kopp, S., Wachsmuth, I. (2007). *Why Emotions Should be Integrated into Conversational Agents*. In T. Nishida (editor): *Conversational Informatics: An Engineering Approach*, chapter 3, John Wiley & Sons, November, pp 49-68.
- Bishop C., M., *Pattern Recognition and Machine Learning*, (2006) Springer

- Bork, A. (2000). *Tutorial Learning with Computers*. On the Horizon, 8(3), 7-9. doi: 10.1108/10748120010803429.
- Bradley, A.P. (1997) *The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms*. Pattern Recognition, 30. 1145-1159.
- Boucsein W (1992) *Electrodermal Activity*. Plenum Press, New York
- Breazeal. C., (2003). *Emotion and sociable humanoid robots*. E. Hudlika (ed), International Journal of Human-Computer Studies, 59, pp.119-155.
- Brindley, D. N., McCann, B. S., Niaura, R., Stoney, C. M., & Suarez, E. C. (1993). *Stress and lipoprotein metabolism: modulators and mechanisms*. Metabolism: clinical and experimental, 42(9 Suppl 1), 3-15
- Broek V., D., E., L., Joris., H. and Westerink, J., H.D.M. and Schut, M., H., and Tuinenbreijer, K., (2010) *Affective Man-Machine Interface: Unveiling human emotions through biosignals*. In: Biomedical Engineering Systems and Technologies. Communications in Computer and Information Science, 52 (Part 1). Springer Verlag, Berlin, pp. 21-47.
- Broek, E. L. V. D., & Healey, J. A. (2009). *Prerequisites For Affective Signal Processing (ASP)*. Blood, (Box 1).
- Broek, E., L., V., D., Schut, M., H., Westerink J., H. D. M., Tuinenbreijer. Kees., (2009). *Unobtrusive Sensing of Emotions (USE)*. Journal of Ambient Intelligence and Smart Environments. 1, 3 (August 2009), 287-299.
- Broek, E., L., V.,D., Westerink, J., H.D.M. (2009). *Considerations for emotion-aware consumer products* Applied Ergonomics, Volume 40, Issue 6, November, Pages 1055-1064, ISSN 0003-6870, 10.1016/j.apergo.2009.04.012.
- Byrne, E. and Parasuraman, R. (1996). *Psychophysiology and adaptive automation*. Biological Psychology, (42): p. 249-268.
- 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, Sept.
- Cacioppo, J. T., & Gardner, W. L. (1999). *Emotion*. Annual Review of Psychology, 50, 191-214
- Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., & Ito, T. A. (1993). *The psychophysiology of emotion*. In M. Lewis & J. M. Haviland-Jones (Eds.), Handbook of emotions (Vol. 2, pp. 173-191). The Guilford Press.
- Cacioppo, J., T., Tassinary., L., G., and Berntson., G., (2007). eds. Handbook of Psychophysiology. 3rd ed. Cambridge: Cambridge University Press.
- Calvo, R. A. & D’Mello, S. K. (2010). *Affect Detection: An Interdisciplinary Review of Models, Methods, and their Applications*. IEEE Transactions on Affective Computing, 1(1), 18-37.

- Calvo, R.A., Brown, I., Scheduling, 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.
- Cannon, W.B., 1927. The James-Lange theory of emotions: a critical examination and an alternative theory. *American Journal of Psychology* 39, 106–124.
- 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.
- Christie, I.C., Friedman, B.H., 2004. *Autonomic specificity of discrete emotion and dimensions of affective space*. *International Journal of Psychophysiology* 51, 143–153.
- Clark E. (1992), *The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System*. The Institute for the Learning Sciences technical report #32, PhD thesis, Northwestern University
- Coan J. A., Allen J. J. B. (2004). *Frontal EEG asymmetry as a moderator and mediator of emotion*. *Biol. Psychol.* 67, 7–50
- Cover, T., M., and Hart, P. E. (1967). *Nearest Neighbor Pattern Classification*. *IEEE Trans. Inform. Theory*, Vol. IT-13, pp 21-27, Jan
- Berlyne, D. E. (1960) *Conflict, Arousal, and Curiosity*. London: McGraw-Hill Book Company.
- Dalgleish, T., and Power, M., (1999). *Handbook of Cognition and Emotion*. John Wiley & Sons, Ltd.,
- Dalgleish, T., Dunn, B., Mobbs, D. (2009). *Affective Neuroscience: Past, Present, and Future*. *Emotion Rev.*, vol. 1, pp. 355-368,
- Davidson, R., J., Chapman, J., P., Chapman, L., J. & Henriques, J., B. (1990) *Asymmetrical brain electrical activity discriminates between psychometrically-matched verbal and spatial cognitive tasks*. *Psychophysiology*, 27, 528-543. 1990.
- Davidson, R., J. (2004) *What does the prefrontal cortex "do" in affect: perspectives on frontal EEG asymmetry research*. *Biological Psychology* 67, 219-233.
- De Rojas C, Camarero C (2008) *Visitors' experience, mood and satisfaction in a heritage context: Evidence from an interpretation center*. *Tourism Management* Vol 29, pp.525-537
- Damasio, A. (2003). *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*. Harcourt, Inc.,
- Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. John Murray
- Davidson, R., J., Ekman, P., Saron, C., D., Senulis, J., A., Friesen, W., V., (1990). *Withdrawal and cerebral asymmetry: Emotional expression and brain physiology*. *Journal of Personality and Social Psychology*, vol. 58, pp. 330–341,
- Dawson, M. E. (2007). *The Electrodermal System* in Cacioppo, J. T.; Tassinary, L. G. & Berntson, G., ed., *Handbook of Psychophysiology*, Cambridge University Press.
- Ekman, P. (1971). *Universals and Cultural Differences in Facial Expressions of Emotion*. Univ. of Nebraska Press

- Dobbins, I., G., Foley, H., Schacter, D., L., Wagner, A., D., (2002): *Executive control during episodic retrieval: multiple prefrontal processes subserve source memory*. *Neuron* 35:989–996.
- Ekman, P. (1999). *Basic Emotions* (Chapter 3). *HandBook of Cognition and Emotion*. (T. Dalgliesh. & M. Power Eds.) John Wiley & Sons, Ltd. Sussex U.K.
- Ekman, P. and Friesen, W.V. (2003). *Unmasking the Face*. Malor Books.
- Ekman., P. (1984). (Scherer, K. and Ekman, P. (Eds.)) *Approaches to Emotion*. Hillsdale New Jersey: Lawrence Erlbaum, Pp. 319-344.
- Fairclough, S. H. (2009). *Fundamentals of physiological computing*. *Interacting with Computers*, 21(1-2), 133-145. Elsevier B.V. doi: 10.1016/j.intcom.2008.10.011.
- Fairclough, S.H. & Gilleade, K.E. (2012). *Construction of the biocybernetic loop: a case study*. *Proceedings of the 14th ACM International Conference on MultiModal Interaction*. Santa Monica, ACM
- Fasel, B., and Luetttin, J., (2003). *Automatic facial expression analysis: a survey*. *Pattern Recognition*, Volume 36, Issue 1, January, Pages 259-275
- Fawcett, T., (2006) *An introduction to ROC analysis*. *Pattern Recogn. Lett.* 27, 8 (June), 861-874.
- Frantzidis, C. a, Bratsas, C., Papadelis, C. L., Konstantinidis, E., Pappas, C., & Bamidis, P. D. (2010). *Toward emotion aware computing: an integrated approach using multichannel neurophysiological recordings and affective visual stimuli*. *IEEE Transactions on Information Technology in Biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society*, 14(3), 589-97. doi: 10.1109/TITB.2010.2041553.
- Frijda, N., H., (1986). *The emotions*. New York: Cambridge University Press.
- Giuseppe, r., (Ed.) (1997, 1998) *Virtual Reality in Neuro-Psycho-Physiology* © Ios Press: Amsterdam, Netherlands.
- Gudmundsson. S., Runarsson. T., P., Sigurdsson. S., (2012). *Test–retest reliability and feature selection in physiological time series classification* *Computer methods and programs in biomedicine*, 1 January (volume 105 issue 1 Pages 50-60 DOI: 10.1016/j.cmpb.2010.08.005)
- Guyon., I. and Elisseeff., A., (2003). *An Introduction to Variable and Feature Selection* *J. Machine Learning Research*, vol. 3, pp. 1157-1182.
- Henson, R., Shallice, N., T., & Dolan, R., J. *Right prefrontal cortex and episodic memory retrieval: a functional MRI test of the monitoring hypothesis*. *Brain* 122, 1367–1381 1999.
- Hidi, S. and Renninger, K. (2006). *The Four-Phase Model of Interest Development*. *Educational Psychologist* Vol. 42 (2).
- Henriques, J., B., & Davidson, R., J. (1990). *Regional brain electrical asymmetries discriminate between previously depressed and healthy control subjects*. *Journal of Abnormal Psychology*, 99, 22-31.

- Izard, P. C. (1994). *Innate and Universal Facial Expressions: Evidence from Developmental and Cross-Cultural Research*. Psychological Bull., vol. 115, pp. 288-299.
- Jacob, R., J., K., 2001. *Open syntax: improving access for all users*. In Proceedings of the 2001 EC/NSF workshop on Universal accessibility of ubiquitous computing: providing for the elderly (WUAUC'01). ACM, New York, NY, USA, 84-89.
- James D. Foley and Andries Van Dam. (1982). *Fundamentals of Interactive Computer Graphics*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- James, W., (1884). *What is an Emotion?* Mind, Vol. 9, No. 34, pp. 188-205.
- James, W., (1894). *Discussion: The Physical Basis of Emotion*. Psychological Review 1, 516–529.
- Jennett C, Cox AL, Cairns P, Dhoparee S, Epps A, Tijs T, Walton A (2008) *Measuring and defining the experience of immersion in games*. International Journal of Human-Computer Studies, 66, pp.641 – 661
- 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.
- Kelley, J. F., (1983) *Natural Language and computers: Six empirical steps for writing an easy-to-use computer application*. Unpublished doctoral dissertation, the Johns Hopkins University, 1983.
- Kim, K. H., Bang, S. W., & Kim, S. R. (2004). *Emotion recognition system using short-term monitoring of physiological signals*. Medical & biological engineering & computing, 42(3), 419-27.
- Kleinsmith, A., Bianchi-Berthouze, N., & Steed, A. (2011). *Automatic Recognition of Non-Acted Affective Postures*. IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society, (99), 1-12.
- Kothari, R., Dong, M, (2000). *Decision Trees for Classification: A Review and Some New Results*. Singapore: World Scientific
- Krohne, H., W. (2003) *Individual differences in emotional reactions and coping*. Davidson. R., J. (Ed); Scherer, K., R. (Ed); Goldsmith, H. H. (Ed), (2003). Handbook of affective sciences. Series in affective science., (pp. 698-725). New York, NY, US: Oxford University Press, xvii, pp 1199
- Kotsiantis, S., B., Zaharakis, I., D., Pintelas, P., E. (2006). *Machine learning: a review of classification and combining techniques*. Artif. Intell. Rev. 26, 3 (November), 159-190.
- 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(3), 394-421. doi:10.1016/j.biopsycho.2010.03.010
- Kreplin., U.,(2014) Unpublished thesis, Liverpool John Moores University 2014
- Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology* Penguin Group US.
- Laine, T. I., Bauer, K. W., Lanning, J. W., Russell, C. A., & Wilson, G. F. (2002). *Selection of input features across subjects for classifying crewmember workload using artificial neural networks.* *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 32(6), 691-704.
- Leder H, Belke B, Oeberst A, Augustin D (2004) *A model of aesthetic appreciation and aesthetic judgments.* *British journal of psychology*, 95(Pt 4), pp.489–508
- Lang, P., J., (1985). *The Cognitive Psychophysiology of Emotion: Anxiety and the Anxiety Disorders.* Hillsdale, NJ: Lawrence Erlbaum, 1985.
- Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (1999, 2008). “*International affective picture system (IAPS): Affective ratings of pictures and instruction manual.*” Technical Report A-8. University of Florida, Gainesville, FL.
- Larsen. R., & Prizmic-Larsen., Z. (2006). *Measuring emotions: Implications of a multimethod perspective.* In M. Eid & E. Diener (Eds.), *Handbook of Multimethod Measurement in Psychology* (pp. 337–352). Washington, D.C.: American Psychological Association.
- Lee, J. & See, K. (2004). *Trust in Automation: Designing for Appropriate Reliance.* *Human Factors*, 46 (1), 50-80.
- Lee. C., K., Yoo. S., K., Park. Y., J., (2005). *Using Neural Network to Recognize Human Emotions from Heart Rate Variability and Skin Resistance,* *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the,* vol., no., pp.5523-5525, 17-18 Jan. 2006
- Levenson, R. W. (2003). *Autonomic specificity and emotion.* *Handbook of affective sciences*, 2, 212-224.
- 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.
- 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.

- Looney, D., Kidmose, P., Park, C., Ungstrup, M., Rank M., L., Rosenkranz, K., Mandic, D., P. (2012). *The Ear-EEG is born! The In-the-Ear Recording Concept*. IEEE Pulse Magazine, vol. 3, no. 6, pp. 32-42,
- MacWhinney, B., St. James j., Schunn, C., Li, P., Schneider, W., (2001). *STEP—A System for Teaching Experimental Psychology using E-Prime*. Behavior Research Methods, Instruments, & Computers 2001, 33 (2), 287-296
- Mandryk. R.L., and Atkins. M.,S., (2007). *A Fuzzy Physiological Approach for Continuously Modeling Emotion During Interaction with Play Environments*. International Journal of Human-Computer Studies, 6(4), pg. 329-347.
- Matson, S.(2014). *Immersion*. www.sammatson.net/Immersion, accessed Jan 2014.
- Mccarthy, J. (1995). *Making Robots Conscious of their Mental States*. Science, 1-21.
- Mellers, B., Schwartz, A., & Ritov, I. (1999). *Emotion-based choice*. Journal of Experimental Psychology: General, 128(3), 332-345. doi: 10.1037/0096-3445.128.3.332.
- Miller, C. A. (2005). *Trust in adaptive automation: the role of etiquette in tuning trust via analogic and affective methods*. In Proceedings of the 1st international conference on augmented cognition (pp. 22-27).
- 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.
- Nisbett, R., E., Wilson, T., D., (1977). *The halo effect: Evidence for unconscious alteration of judgments*. Journal of Personality and Social Psychology, Vol 35(4), Apr, 250-256.
- Norman, D.A., (2007).*The Design of Future Things*. Basic Books,
- Novak, D., Mihelj, M., Munih, M. (2012). *A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing*. Interacting with Computers, 24, 3, 154-172.
- Panksepp, J. (2004). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford Univ. Press,
- Parasuraman, R., Sheridan T., B., Wickens, C. (2000). *A Model for Types and Levels of Human Interaction with Automation*. Transactions on systems, man, and cybernetics – Part A: Systems and humans, vol. 30, no. 3, May 2000, IEEE
- Parsons, H.M. (1985). *Automation and the individual: Comprehensive and comparative views*. Human Factors, Vol. 27, 99-112.

- 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.
- Petrantonakis, P. C., & Hadjileontiadis, L. J. (2010). *Emotion recognition from EEG using higher order crossings*. IEEE transactions on information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society, 14(2), 186-97. doi: 10.1109/TITB.2009.2034649.
- 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(10), 1175-1191. doi: 10.1109/34.954607.
- Picard, R., W., (1997). *Affective Computing*. MIT Press, Cambridge, MA, USA.
- Picard, R., W., (2003). *Affective computing: challenges*. International Journal of Human-Computer Studies, Volume 59, Issues 1-2.
- Picard, R., W., (2000). *Toward computers that recognize and respond to user emotion IBM systems journal*, vol 39, nos 3&4.
- M. Petrides, *Frontal lobes and behaviour*. Curr. Opin. Neurobiol. 4, 207–211 (1994).
- 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.
- Pine B, Gilmore J (1998) *Welcome to the experience economy*. Harvard business review
- Platt JC (1999) *Fast training of support vector machines using sequential minimal optimization*. In Advances in kernel methods, Bernhard Scholkopf, Christopher J. C. Burges, and Alexander J. Smola (Eds.). MIT Press, Cambridge, MA, USA 185-208
- Pope, A. T., Bogart, E. H. and Bartolome, D. S. (1995). *Biocybernetic system evaluates indices of operator engagement in automated task*. Biological Psychology, 40, 187-195.
- POWERS, D.M.W. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation. Journal of Machine Learning Technologies 2 (1): 37–63.
- Psychology Software Tools INC., <http://www.pstnet.com/>
- Manber, R., Allen, J., J., Burton, B., K., Kaszniak, A., W., (2000). *Valence-dependent modulation of psychophysiological measures: Is there consistency across repeated testing?*. Psychophysiology, Cambridge Journals Online, Volume 37, Issue 05, Pages 683-692, 2000.
- Rainville, P., Bechara, A., Naqvi, N., Damasio A., R., (2006) *Basic emotions are associated with distinct patterns of cardiorespiratory activity*. Int J Psychophysiol 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(1), 58-69. doi: 10.1007/s10044-006-0025-y.

- Rani, P., Sarkar, N., & Liu, C. (2005). *Maintaining optimal challenge in computer games through real-time physiological feedback*. Paper presented at the 11th Human-Computer Interaction International, Las Vegas, USA.
- Rani, P., Sarkar, N., Adams, J., (2007). *Anxiety-based affective communication for implicit human-machine interaction*. *Advanced Engineering Informatics* 21, 323–334.
- Ramnani, N., & Owen, A.M. (2004). *Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging*. *Nature Reviews Neuroscience*, 5, 184–194.
- Redmond, S. J., Basilakis, J., Xie, Y., Celler, B. G., & Lovell, N. H. (2009). *Piecewise-linear trend detection in longitudinal physiological measurements*. Conference proceedings :... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. 3413-6. doi: 10.1109/IEMBS.2009.5332406.
- Regan, L., M., & M., Stella Atkins., (2007). *A fuzzy physiological approach for continuously modelling emotion during interaction with play technologies*. *International Journal of Human-Computer Studies* Volume 65, Issue 4, April, Pages 329-347
- Regan. L., M., & M. Stella Atkins., (2006) *A continuous and objective evaluation of emotional experience with interactive play environments*. *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2006)*, pp. 1027—1036
- 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.
- Rolls, E., T., (2000) *The orbitofrontal cortex and reward*. *Cereb. Cortex* 10, 284–294.
- Russell, J.A. (2003). *Core Affect and the Psychological Construction of Emotion*. *Psychological Rev.*, vol. 110, pp. 145-172.
- Russell. J., A., (1980). *A circumplex model of affect* *Journal of personality and social psychology*, 39, 1161 – 1178
- Gudmundsson, S., Runarsson, T., P., Sigurdsson, S., Eiriksdottir, G., Johnsen, K. (2007). *Reliability of quantitative EEG features*. *Clinical Neuro-physiology* 118 (10) 2162-2171.
- Sakr. G., E., Elhajj. I., H., Huijjer. H., A-S., (2010). *Support Vector Machines to Define and Detect Agitation Transition*. *Affective Computing, IEEE Transactions on* , vol.1, no.2, pp.98-108, July-Dec.
- Scerbo M., W., Freeman F., G., Mikulka P., J., Parasuraman R., Nocero F., Di., and III Lawrence J., P. (2001). *The Efficacy of Psychophysiological Measures for Implementing Adaptive Technology*. Technical Report. NASA Langley Technical Report Server.
- Schreier. 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(2), 93–118.
- Serbedzija, N.B.; Fairclough, S.H., (2009). *Biocybernetic loop: From awareness to evolution*. *Evolutionary Computation, CEC '09. IEEE Congress on*, vol., no., pp.2063, 2069, 18-21.

- 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.
- Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. MIT Man-Machine Systems laboratory, Cambridge, MA.
- Silberman, E.K., & Weingartner, H. (1986). *Hemispheric lateralization of functions related to emotion*. Brain and Cognition, 5, 322-353.
- Silvia, P. J. (2005). *What is interesting? Exploring the appraisal structure of interest*. Emotion, 5, 89-102.
- Silvia, P. J. (2008) *Interest—The Curious Emotion*. Current Directions in Psychological Science, 17(1), 57-60.
- Silvia, P. J. (2010). *Confusion and interest: The role of knowledge emotions in aesthetic experience*. Psychology of Aesthetics, Creativity, and the Arts, 4(2), 75-80.
- 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.
- Sweeney, K.T., Ward, E.T., McLoone, S.F., (2012) *Artifact Removal in Physiological Signals - Practices and Possibilities* IEEE Transactions On Information Technology In Biomedicine, Vol. 00, No. 00, 1
- Tataranni, P., A., et al. *Neuroanatomical correlates of hunger and satiation in humans using positron emission tomography*. Proc. Natl Acad. Sci. USA 96, 4569–4574 1999.
- Takahashi, K., (2003). *Remarks on emotion recognition from bio-potential signals*. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Palmerston North, New Zealand, October 5-8, vol. 2, pp. 1655–1659
- Tomarken, A.J., (1995). *A psychometric perspective on psychophysiological measures*. Psychological Assessment 7, 387-395.
- Tomarken, A.J., Davidson, R.J., Wheeler, R.E., and Kinney, L. (1992). *Psychometric properties of resting anterior EEG asymmetry: temporal stability and internal consistency*. Psychophysiology 29, 576–592.
- van de Laar, B.; Bos, D.P.-O.; Reuderink, B.; Poel, M.; Nijholt, A., (2013) *How Much Control Is Enough? Influence of Unreliable Input on User Experience*. Cybernetics, IEEE Transactions on, vol.43, no.6, pp.1584,1592, Dec. doi: 10.1109/TCYB.2013.2282279
- van Ulzen, N., Semin, Gün., Oudejans, Raoul., Beek, Peter., (2008). *Affective stimulus properties influence size perception and the Ebbinghaus illusion*. Psychological Research Volume 72, Number 3, 304-310 2008, DOI: 10.1007/s00426-007-0114-6
- Vapnik, V., N., Cortes, C. (1995). *Support-Vector Networks*. Machine Learning, September, Volume 20, Issue 3, pp 273-297

- Villon, O., & Lisetti, C. (2007). *Toward Recognizing Individual's Subjective Emotion from Physiological Signals in Practical Application*. Computer-Based Medical Systems, CBMS'07. Twentieth IEEE International Symposium on (pp. 357–362).
- Villon, O., & Lisetti, C. L. (2006). *A user-modelling approach to build user's psycho-physiological maps of emotions using bio-sensors*. Paper presented at the 15th IEEE International Symposium on Robot and Human Interactive communication, Hatfield, UK.
- Vinge, V. (1993). *The Coming Technological Singularity: How to Survive in the Post-Human Era* VISION-21 Symposium NASA Lewis Research Center and the Ohio Aerospace Institute, March 30-31.
- Wagner, J.; Jonghwa Kim; Andre, E., (2005) *From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification*. Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, vol., no., pp.940, 943, 6-6 July.
- Wagner, A., D. et al. (1998) *Building memories: remembering and forgetting of verbal experiences as predicted by brain activity*. Science 281, 1188–1191.
- Waters W, F., Williamson, D, A., Bernard, B, A., Blouin, D, C., Faulstich, M, E., (1987). *Test-retest reliability of psychophysiological assessment* Behaviour Research and Therapy Volume 25, Issue 3, 1987, Pages 213–221
- Wilhelm. F., H., Grossman. P., (2010) *Emotions beyond the laboratory: Theoretical fundamentals, study design, and analytic strategies for advanced ambulatory assessment*. Biological Psychology 2010; 84: 552-569.
- Wilson, G., F., Russell, C., A. (2003). *Operator functional state classification using multiple psychophysiological features in an air traffic control task*. Hum Factors. Fall; 45(3):381-9.
- Wilson, G.F., Russell, C.A., (2007). *Performance enhancement in an uninhabited air vehicle task using psychophysiological determined adaptive aiding*. Human Factors 49, 1005-1018.
- Wu, D., Courtney, C. G., Lance, B. J., Narayanan, S. S., Dawson, M. E., Oie, K. S., et al. (2010). *Optimal Arousal Identification and Classification for Affective Computing Using Physiological Signals: Virtual Reality Stroop Task*. IEEE Transactions on Affective Computing, 1(2), 109-118. doi: 10.1109/T-AFFC.2010.12.
- Yannakakis, G.N., Hallam, J., (2008). *Entertainment modeling through physiology in physical play*. International Journal of Human–Computer Studies 66, 741–755.
- Codispoti, M., Surcinelli, P., Baldaro, B., (2008). *Watching emotional movies: affective reactions and gender differences*. International Journal of Psychophysiology 69, 90–95.
- Demaree, H., Schmeichel, B., Robinson, J., Everhart, D.E., (2004). *Behavioural, affective, and physiological effects of negative and positive emotional exaggeration*. Cognition and Emotion 18, 1079–1097.

- Bradley, M.M., Silakowski, T., Lang, P.J., (2008). *Fear of pain and defensive activation*. Pain 137, 156–163.
- Sweeney, K.T., Ward, E.T., McLoone, S.F., (2012) *Artifact Removal in Physiological Signals - Practices and Possibilities*. IEEE Transactions On Information Technology In Biomedicine, Vol. 00, No. 00, 1
- Bos, D., P-O., Gürkök, H., Reuderink, B., and Poel, M., (2012). *Improving BCI performance after classification*. In Proceedings of the 14th ACM international conference on Multimodal interaction (ICMI '12). ACM, New York, NY, USA, 587-594.
- Chanel, G.; Rebetz, C., Bétrancourt, M., Pun, T. (2011) *Emotion Assessment From Physiological Signals for Adaptation of Game Difficulty*. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol.41, no.6, pp.1052-1063, Nov. 2011