

POLITECNICO DI TORINO

Department of Mechanical and Aerospace Engineering

Master degree course in Aerospace Engineering



Master Degree Thesis

*Facial Expression Analysis for Cognitive State
Estimation in Aerospace Human-Machine Systems*

Advisor:

Roberto Sabatini

Manuela Battipede

Co-Advisor:

Alessandro Gardi

Candidate:

Federico Rivalta

ACADEMIC YEAR 2019-2020

April 2020

Abstract

This thesis is part of a collaboration between Politecnico di Torino and Royal Melbourne Institute of Technology RMIT. The developed project was carried out at the Bundoora East campus in collaboration with *THALES Australia* and *Northrop Grumman Corporation*. This project aims to analyse the potential integration of Facial Expression monitoring within a sensor network to evaluate in real time the cognitive state of ATM and one-to-many UAS operators. ATM operators perform a safety-critical work in which it is essential that the situation awareness is not lacking and the workload is not excessive. Most accidents in the aviation field are due to human error so monitoring the cognitive state of the operators would lead to increase the efficiency of air traffic and the safety of operations.

Currently, the Federal Aviation Authority (FAA) has mandated that remote pilots or visual observers are only allowed to operate or command one Unmanned Aerial Vehicle (UAV) at any time (14 CFR 107.35), so current UAS operations require multiple human operators to manage one UAV, known as 'many-to-one' operations. This research therefore aims to analyze the possibility of managing multiple UAVs with a single operator through a system in which the trusted autonomy is based on a bio-sensing network.

These technologies are also fundamental for the development of Single Pilot Operations (SPO), which in recent years have been studied in order to have passenger aircrafts with only one pilot.

The sensor network allows to monitor in real time the operator in order to evaluate the cognitive state and adapt the level of automation of the software according to the latter. It is composed of several sensors that monitor different features in order to give greater reliability. The to date monitored parameters are: Breathing Rate, Blink Rate, Visual Entropy, Heart Rate, EEG which are used to define a parameter: the workload. The objective of this research is to evaluate a potential relationship between workload variation and facial contractions and to form the basis for a potential inclusion of Facial Expression monitoring in the sensor network.

FE have been studied for many years to assess emotional state but little research has been done so far to assess cognitive state and workload. Through ATM and OTM experiments the psycho-physiological response of the operators has been evaluated and various types of analysis have been carried out to find correlation between the variation of the workload and the physiological response.

These studies have been carried out in aerospace field for ATM, OTM and SPO applications but they can be applied in many other fields such as automotive.

Ringraziamenti

Desidero ringraziare la Prof.ssa Manuela Battipede, il Prof. Roberto Sabatini ed il Senior Research Fellow Alessandro Gardi per l'opportunità formativa che mi hanno offerto.

Un grazie speciale alla mia famiglia che in questi anni ha sempre creduto in me.

A mia Madre e mio Padre che sono sempre stati pronti a dedicare del tempo per ascoltarmi e consigliarmi, lasciando però che facessi le mie scelte e credendo sempre in me. A mio fratello che è stato un esempio ed una guida.

Ai mie compagni, in particolare Filippo, Erika, Simone ed Elena. Abbiamo creato un Team fondato sulla condivisione ed il sostegno reciproco che mi ha insegnato molto ed ha fatto la differenza nel raggiungimento di questo traguardo.

Agli amici che mi sono stati vicino anche quando ero lontano, che hanno saputo comprendere le mie assenze ed il mio distacco nei momenti di difficoltà, un grazie speciale a Marco e Davide.

Contents

Introduction	1
1 Facial Expressions	5
2 State of the Art: The relation between Human and Machine	13
2.1 Human Machine Interface Interaction.....	15
2.2 CHMI2 architecture.....	16
2.2.1 Tasks Analysis.....	17
2.2.2 Cognitive Task Analysis.....	18
3 The Sensing layer	23
3.1 Cardiorespiratory sensor	24
3.2 Brain waves sensor	27
3.3 Eye sensor	29
4 Cognitive state and Workload	32
4.1.1 NASA-TLX.....	34
4.1.2 Wickens model.....	35
4.1.3 Sperandio Model.....	35
4.1.4 Adopted model and considerations	37
5 The Software: Open Face	38
5.1 Constrained Local Neural Field	39
5.2 Action Unit detection.....	42
5.3 Software Interface.....	43
6 Performance evaluation	45
6.1 Backwards and forwards movement	47
6.2 Rotation.....	51
6.2.1 X axis	51
6.2.2 Z axis.....	51
6.3 Blink Rate	52
6.3.1 Outlier rejection	56
7 Experiments	57
7.1 OTM experiments: Bushfire-fighting	57
7.1.1 GCS interface	59
7.1.2 Test Scenario	60
7.1.3 Assumptions	63
7.2 Experiment activities.....	64

8	Online and offline analysis	65
8.1	Filtering	66
8.2	Prominence.....	67
8.3	Variance and Covariance.....	69
8.4	Correlation coefficient.....	69
9	Data analysis	71
9.1	Objective parameters: Secondary tasks	71
9.2	Sensor data smoothing.....	72
9.3	Correlation coefficient.....	73
9.3.1	Cross correlation.....	75
9.3.2	Dynamic time warping.....	82
10	Psychophysiological parameters	88
10.1	Eye features	88
10.1.1	Visual Entropy	90
10.2	Bioharness.....	91
10.3	Electroencephalography	96
11	ATM experiment	99
11.1	Objective parameters: number of aircrafts and control inputs.....	101
11.2	Eye features	103
11.2.1	Visual Entropy	105
11.3	Cardiorespiratory features	106
11.4	Electroencephalography	108
12	Protocol	110
13	Conclusions	120
	Bibliography	122

List of Figures

Figure 1: Sensor Network Structure.....	3
Figure 2: Research Methodology.....	4
Figure 3: FACS Action Units	6
Figure 4: Action Units and Emotions 1.1[26]	8
Figure 5: Action Units and Emotions 1.2[26]	9
Figure 6: Definition of affective states[33]	11
Figure 7: Frequency of appearance of emotions[33]	11
Figure 8: Content requirements for a Critical Tasks Analysis Report	18
Figure 9: Yerkes-Dodson law	21
Figure 10: CHMI2 structure	22
Figure 11: Zephyr Bioharness 3.....	24
Figure 12: Poincare plot analysis.....	26
Figure 13: Electrode cap actiCAP Xpress	28
Figure 14: GP3 Eye Tracker installation.....	29
Figure 15: Pupil radius detection.....	30
Figure 16: Saccades pattern.....	30
Figure 17: Sperandio's model.....	36
Figure 18: Regulation of Workload by the use of different strategies [62]	36
Figure 19: WL objective assessment model	37
Figure 20: Open Face operational block diagram.....	39
Figure 21: LNF structure	40
Figure 22: Landmark Detection.....	42
Figure 23: Software interface.....	43
Figure 24: Open Face outputs.....	44
Figure 25: Webcam Logitech C270	45
Figure 26: Geometric method for determining camera's FoV	47
Figure 27: Adopted Reference System	48
Figure 28: Forward, geometrical parameter.....	49
Figure 29: Forward, Action Unit.....	49
Figure 30: Backward, geometrical parameter	50
Figure 31: Backward, Action Unit	50
Figure 32: Dx to Sx complete rotation	51
Figure 33: Exp1. Low to Normal BR comparison	53
Figure 34: Exp2. Low to Normal BR comparison	53

Figure 35: Exp3. Low to Normal BR comparison	53
Figure 36: Exp1. High to Normal BR comparison	55
Figure 37: Exp2. High to Normal BR comparison	55
Figure 38: Exp3. High to Normal BR comparison	55
Figure 39: Task flowchart for the tactical coordination	58
Figure 40: GCS Interface	59
Figure 41: AOR partitioning	60
Figure 42: Secondary Tasks boundary classification	61
Figure 43: Kappa coefficient for single Action Unit[29]	66
Figure 44: Peaks evaluation.....	67
Figure 45: Prominence determination	68
Figure 46: Task trend smoothing process.....	72
Figure 47: AU9 Raw	72
Figure 48: AU9 trend	74
Figure 49: AU6 FE-RL	76
Figure 50: AU9 FE-RL	76
Figure 51: AU12 FE-RL	76
Figure 52: AU15 FE-RL	77
Figure 53: AU25 FE-RL	77
Figure 54: Phase correction AU6	79
Figure 55: Phase correction AU9	79
Figure 56: Phase correction AU12	79
Figure 57: Phase correction AU15	80
Figure 58: Phase correction AU25	80
Figure 59: Phase correction AU6 AU9.....	81
Figure 60: Phase correction AU12 AU15.....	81
Figure 61: Phase correction AU25	81
Figure 62: Participant 4 MFE AU17 Phase 3.....	82
Figure 63: AU6 Warped	83
Figure 64: AU15 Warped	83
Figure 65: Dynamic Time Warping path	84
Figure 66: DTW without constrains	84
Figure 67: AU6 derivate warping	85
Figure 68: AU12 derivate warping	86
Figure 69: AU15 derivate warping	86
Figure 70: Blink Rate Ph2.....	88

Figure 71: Blink Rate Ph3.....	89
Figure 72: VE Phase 2.....	90
Figure 73: VE Phase 3.....	90
Figure 74: Partecipant 1 Phase 2 Bioharness AU6	92
Figure 75: Partecipant 1 Phase 2 Bioharness AU12	92
Figure 76: Partecipant 1 Phase 2 Bioharness AU15	92
Figure 77: Partecipant 1 Phase 2 Bioharness AU25	93
Figure 78: Partecipant 1 Phase 3 Bioharness AU6.....	93
Figure 79: Partecipant 1 Phase 3 Bioharness AU12	94
Figure 80: Partecipant 1 Phase 3 Bioharness AU15	94
Figure 81: Partecipant 1 Phase 3 Bioharness AU25	94
Figure 82: Partecipant 2 Phase 3 Bioharness	95
Figure 83: Partecipant 4 Phase 3 Bioharness	96
Figure 84: Partecipant 1 Phase 3 AU9	98
Figure 85: Partecipant 1 Phase 3 AU25	98
Figure 86: ATM simulation environment	99
Figure 87: Hardware architecture	100
Figure 88: n aircraft and control input trends.....	102
Figure 89: WL and AU25 trends.....	103
Figure 90: AU25- Blink Rate comparison.....	104
Figure 91: AU6- Blink Rate comparison.....	105
Figure 92: AUs-VE comparison.....	105
Figure 93: AU15 trend	106
Figure 94: AU17 trend	107
Figure 95: AU25 trend	107
Figure 96: AU9-AU15 trends	109
Figure 97: AU17-AU25 trends	109
Figure 98: Calibration experiment	111
Figure 99: Neural Fuzzy Network architecture	115
Figure 100: ATM experiment.....	118
Figure 101: OTM Partecipant1	118

Introduction

This Thesis project has been carried out in collaboration with *THALES Australia* and *Northrop Grumman Corporation* for the development of Psychophysiological-Based Integrity Augmentation (PBIA) systems in Air traffic Management (ATM) and One-To-Many (OTM) operations for Unmanned Aircraft System (UAS) to fulfil the evolving aircraft certification requirements and future goals in aeronautic and defence field.

The interaction between human and machine has always been a field of strong research to support humans, reduce the probability of accidents and optimize effectiveness and efficiency of operations. The objective is to have the machine perform repetitive operations and thanks to the machine learning and artificial intelligence, it is increasingly possible to entrust the machines with the task of reasoning. Obviously, however, it is necessary to have the presence of human as a supervisor. In particular fields such as Air Traffic Management (ATM) and One-To-Many (OTM) operations, automation support is essential to increase traffic capacity and support human operators in their duties. In fact, the constant growth of air traffic has generated an increasing need to support the ATM operators(ATMo) with adaptable human-machine interfaces which can reduce their workload so as to increase not only the efficiency but also the safety of air traffic. Furthermore in recent years the trend towards Single Pilot Operations(SPO) has become more and more evident leading to the need to have systems that monitor the pilot during operations because it is no longer present the second pilot, this replacement of the second pilot by a machine obviously implies the need to create monitoring systems that allow a high level of integration and reasoning by the machine itself.

However, the rapid advancement of the technology has also led to an increase in the stress of operators who find themselves operating in a complex work systems in which they must adapt their decision making and performance in the face of more and more dynamic and ever-changing environments.

The controllers are responsible for managing different zones like control and approach areas in a complex mixture of air traffic from commercial, general, corporate, and military aviation. There's no need to point out that their role is vital to maintaining air traffic under its jurisdiction in a safe and timely manner. In spite of technical improvements in aircraft performance, and ATM facilities and their operational betterment, aircraft accidents continue to occur. If Air Traffic Management organizations are to meet future demand safely, better models of controller workload are needed to manage concurrent task demands, time pressure, and tactical constraints that operators are supposed to fulfil. It is known that the major cause of air accidents and incidents is the human error, within this category it can be said that a large part of these accidents and incidents are attributable to negative factors associated with cognitive load, such as fatigue and stress. This factors can severely impact pilots, ATMo and OTM operators performance, potentially compromising the ability to accomplish their duties thereby representing risks to flight safety. It is reported that approximately 80% of all aircraft accidents are results of human error. Human error, rather than technical failure, now represents the greatest threat to system reliability and safety in socio-technical systems like aviation. So the goal is to increase the reliability of the 'human' who is the main decision-maker and in the future will be more and more helped by

machines. For efficient aircraft operations, including flight safety enhancement, delay reduction, and fuel usage improvements, the leading agencies in aviation such as the Federal Aviation Administration (FAA), International Civil Aviation Organization (ICAO), and EUROCONTROL are actively establishing policy and criteria governing all aviation activities. Their goals are to improve navigation aids, airspace management, ATM and airport operations, emergency aircraft handling and human factor aspects.[1]

The new frontier of human-machine interaction is a human monitoring systems which capture and analyze psycho-physiological parameter in real-time like gaze, blink-rate, heart-rate, prefrontal lobe oxygenation, brain waves, neuro-physiological signals and others. The outcomes of all these sensors are gather to evaluate the cognitive states of the operators. Numerous studies carried out on the monitoring of these psycho-physiological parameters have demonstrated the potential applicability of these systems and the need of further research to better understand the physiological responses of operators performing safety-critical operations. [1-12]

Hence had been developed a Sensor Network composed by many devices that communicate with each other to extrapolate the cognitive features of the operator instant by instant like mental workload (MWL), attention, situation awareness(SA) and fatigue. The workload is the most related to human-machine system adaptation, therefore it's one of the most studied especially in the aerospace field. Workload varies as a function of task demands placed on the human operator and the capacity of the operator to meet those demands. For example, research to date has shown that the blink rate is one of the most effective measures of mental workload and it has been proven to decrease with the growth of MWL. Electroencephalography (EEG), as a physiological measure of the momentary functional state of cerebral structures, provides information on inattention and high cognitive workload. The real-time monitoring will be carried out adopting a combination of wearable and remote devices such as eye-tracking systems and face micro-expression detection cameras[13, 14]

All airlines are pressing for more and more aircraft to be operated at the same time, of course, without jeopardising the safety of operations. It is unthinkable to create a static model to help controllers because ATM operations are difficult to predict because of unexpected traffic increases, severe weather, malfunctioning of equipment, and so on. This leads to the need to have high levels of efficiency and automation of the ATM facilities.[15]

As aforementioned, the Air Traffic Management operator (ATMo) can have significant variations in workload during the day and it has to be avoided that the overall workload raise beyond his capabilities. Additionally, more the oscillation in WL is and more the operator could experiment fatigue and stress. New automation support features have been studied to create an environment in which human and machine work in synergy and the latter adapts to the workload levels perceived by the operator so that the human can have a quite constant workload during his working day. It is necessary to specify that the level of automation must not be unreasonably high because it would lead to a loss of ATMo's situational awareness. Generally, ATMo's performance in decision making degrades when operational complexity increases due to environmental difficulties or unfamiliarity with the situation.

In the field of Human Factor engineering, therefore, great efforts are being made to have more and more efficient Human-Machine Interfaces and Interactions (HMI2) and in recent years the innovative concept of Cognitive HMI2 (CHMI2) has become more and more studied. In the latter the interaction between human and machine is

focused on , and managed in function of, the cognitive state of the operator. Numerous sensor monitor the operator’s psycho-physiological parameters such as heart activity, gaze patterns, brain evoked potential, EEG/fNIRS, Blink Rate and Visual Entropy with the aim of determining in real time the cognitive state of the operator and adapt the level of automation, the graphical interface and whatever is deemed necessary to optimize the efficiency of the latter and reduce potentially disastrous distractions. [12, 16-19]

All the sensors mentioned above are affected by background noise, measurement uncertainties propagation, disturbances and the main issue is that each person is different so sensor outcomes are hardly interpretable. Consequently the adoption of different sensors which operate in a network is both a natural and necessary evolution to effectively exchange, synchronise and process measurement data within a customisable operational network architecture composed by multiple monitoring devices that can communicate with each other. From this need comes therefore the object of this research that integrate in the sensor network a further element: Micro-Facial Expression Detection and Analysis. [18]

According to research conducted so far and available in the literature, no studies have yet been done on the cognitive states analysis through facial micro-expressions (FE). FE have been studied in the psychological field for many years, a close relationship between emotions and FE has already been found and demonstrated but no research has been done on the relationship between Facial Expressions and cognitive load, especially for aerospace purposes in HMI2, CHMI2 in ATM and OTM operations. The cognitive state is composed by various variables and in particular this work is focused on Workload evaluation. Needless to say that this study is conducted in the aeronautic field but could be applied in a wide range of environment like space[18, 20] and automotive, to improve safety and security. [21]

Figure 1 shows the top level structure of the developed sensor network.

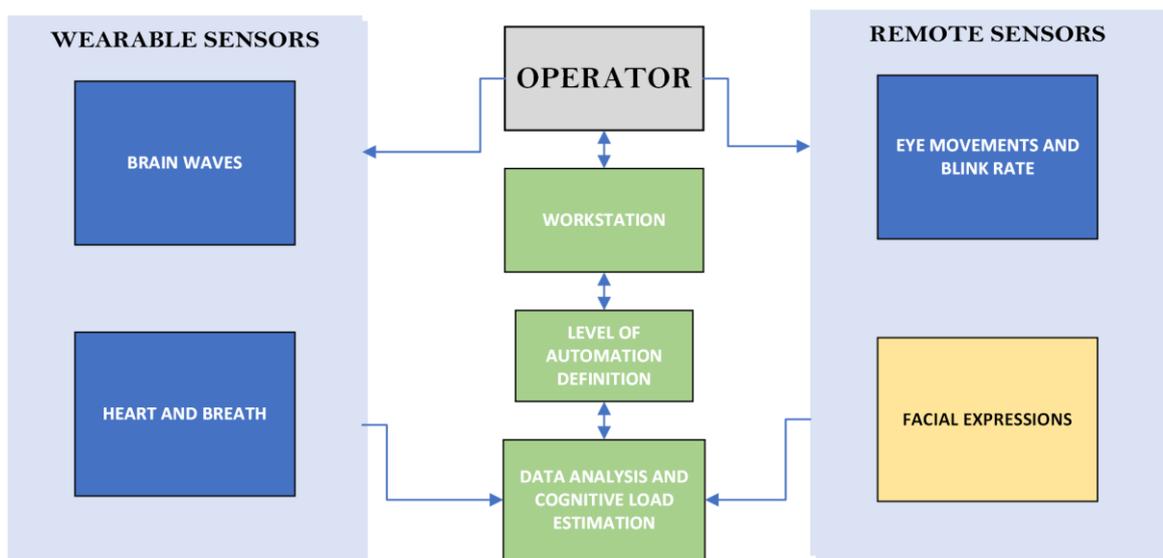


Figure 1: Sensor Network Structure

Following an extensive literature review and in line with the objectives of the research project of which this thesis was part, the following research questions have been defined:

- How do the ATM/OTM operator FE vary under different operational conditions?
- Are FE monitored data in accordance with Taskload?
- Are FE monitored data in accordance with other Sensors recordings?
- Can the FE monitoring be integrated in the Sensor Network?

Various models of mental Workload's definition have been evaluated in order to choose the model that best reflects the research objectives. The approach to experimental verification involves the decomposition of the research questions into smaller, individually testable hypotheses each with its own assumptions, conditions and theoretical predictions.

Figure 2 shows the adopted research methodology and the four main objectives of this research. Six OTM and one ATM experiments were conducted to assess the relationship between FE and Workload. This evaluation can be divided into two main approaches:

The FE-Workload correlation through objective parameters such as the Taskload and the FE-Workload correlation through biometric analysis conducted by the other sensors of the sensor network. These evaluation have the goal to evaluate the potential ability of the FE to provide useful parameters for workload definition and thus evaluate the potential integration of the FE sensing system in the sensor network. This integration must consider that each individual has a different physiological response to stimuli, which is why the ultimate objective of this research is to define an empirical protocol of FE sensing adaptation to the operator.

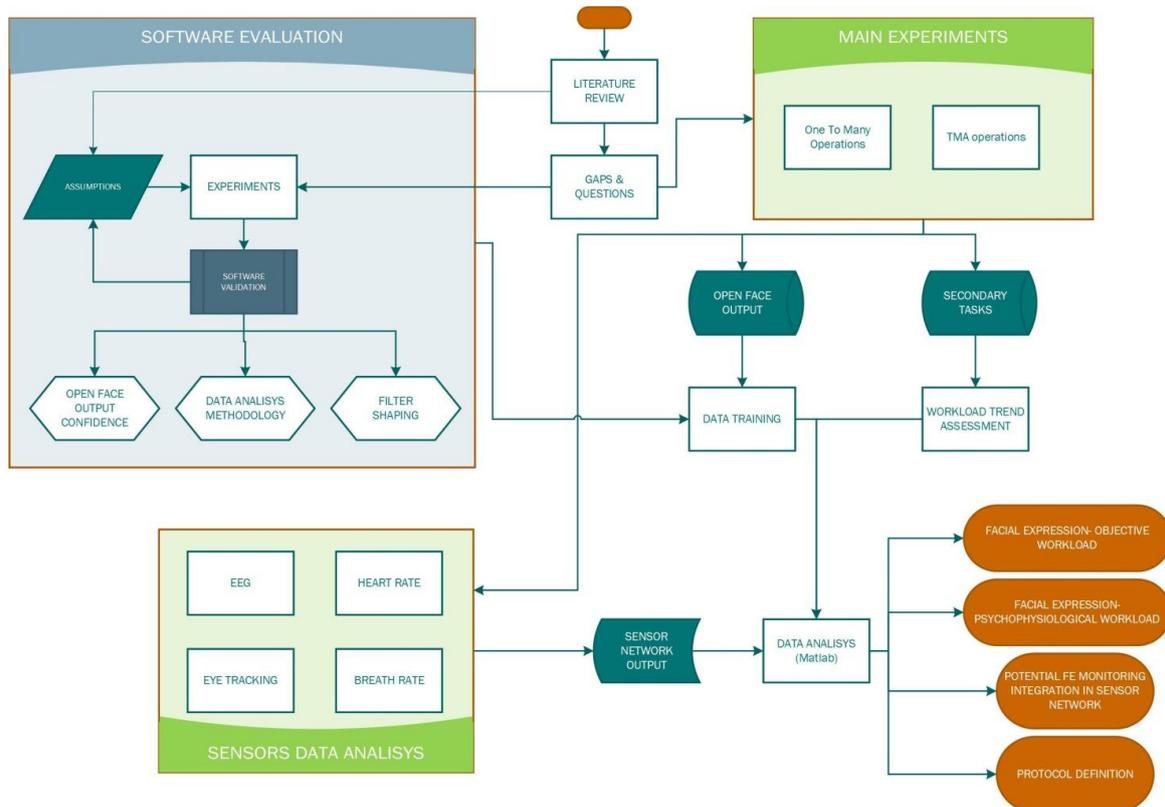


Figure 2: Research Methodology

1 Facial Expressions

Studies on Facial Expression(FE) and their link to emotional states started over a hundred years ago (Darwin, 1872/1998) and the first in-depth study was carried out by E. A. Haggard e K. S. Isaacs in 1966 for the analysis of non-verbal behaviour in psychotherapy. Based on the work of the Swedish anatomist Carl-Herman Hjortsjö, Paul Ekman and Wallace V. Friesen developed one of the most influential methods to objectively code facial behaviour in 1978, which was further fine-tuned in 2002. These studies have been carried out in the psychological field to understand the bodily reactions induced by emotions but in recent years other research had been carried out to study the relationships between FE and other non-emotional aspects such as frustration and cognitive states. It is essential to point out that what are called Facial Expressions tend to be voluntary and easily visible because they are part of non-verbal communication. When human beings relate to themselves they use them consciously such as smiling and winking. The Facial Micro-Expressions (FME) are different, they are involuntary and generally are unaware contractions of the facial muscles. The magnitude of muscle contraction is very small so only a careful and trained eye can perceive them or, as in our case, a specific software which uses image processing algorithms and a high-resolution camera. MFEs reveal unconscious aspects of which the subject is sometimes not even aware or does not want to show through the so-called lapses. For this reason, MFEs are also used to discover lies. [22].

It is also important to note that mood and emotion are different states. The mood is an emotional state that can last up to hours or days whereas the emotion can last even a few milliseconds such as a shock or disgust for an image that we have seen.

Ekman's neurocultural theory defines basic emotions, genetically determined, universally diffused and distinct from each other on a physiological and psychological level: anger, fear, disgust, contempt, joy, sadness, surprise. Each would have a particular pattern of facial behaviour, conscious experience, physiological basis and cognitive functions. For this reason, the meaning of facial expression would not change depending on the context in which it is perceived. It has also been demonstrated that in different groups of subjects belonging to very different cultural environments there are the same mimic patterns for the expression of emotions and the process of decoding presents a wide margin of homogeneity in different cultures. [23]

Our face is one of the most complex signal systems available to us. It includes over 40 structurally and functionally autonomous muscles which can be triggered independently of each other. The facial muscular system is the only place in our body where muscles are not only attached to the bone and facial tissue (other muscles in the human body connect to two bones) but also to facial tissue only such as the muscle surrounding eyes or lips. This makes the facial movements more unpredictable and 'noisy' because they don't generate a physical displacement of another bone (for example, contraction of the biceps that acting on humerus and radius/ulna generates the flexion of the forearm in a controlled, geometrically determinable and easily repeatable way).

Humans are able to produce thousands of slightly varying sets of facial expressions – however, there is only a small set of distinctive facial configurations that almost everyone associates with certain emotions, irrespective of gender, age, cultural background and socialization history.

The Facial Action Coding System (FACS; Ekman & Friesen, 1978; Ekman, Friesen, & Hager, 2002) represents a fully standardized classification system of facial expressions based on anatomic features. FACS itself is descriptive and includes no emotion-specified descriptors. Hypotheses and inferences about the emotional meaning of facial actions are extrinsic to FACS. Furthermore, FACS is the most comprehensive, psychometrically rigorous, and widely used method of Facial expressions in which they are a combination of elementary components called Action Units (AUs). AUs are anatomically related to the contractions of specific facial muscles and they are identified by a number (AU1, AU2, etc.). All facial expressions can be broken down into their constituent AUs. Assumed that facial expressions are “words”, AUs are the “letters” that make up those words. Figure 3 shows the main AUs that are used in facial expression detection software.

Upper Face Action Units					
AU1	AU2	AU4	AU5	AU6	AU7
					
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU41	*AU42	*AU43	AU44	AU45	AU46
					
Lip Droop	Slit	Eyes Closed	Squint	Blink	Wink
Lower Face Action Units					
AU9	AU10	AU11	AU12	AU13	AU14
					
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU15	AU16	AU17	AU18	AU20	AU22
					
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU23	AU24	*AU25	*AU26	*AU27	AU28
					
Lip Tightener	Lip Pressor	Lips Parts	Jaw Drop	Mouth Stretch	Lip Suck

Figure 3: FACS Action Units

AUs can occur either singly or in combination. When AUs occur in combination they may be additive, in which the appearance of each action unit is independent or non-additive, in which they modify each other’s appearance. An example of an additive combination in FACS is AU1+AU2, which often occurs in surprise. An example of a non-additive combination is AU1+AU4.

An applicable approach is the frequency of co-occurrence i.e. assessing the AUs that usually appear together as a single event. Analyses made from past research on

reliability for occurrence/non-occurrence ([24]) are reported in the *Experiments Chapter* and they have been used to define the adopted data filter. It also addresses the problem that some action units may linger and merge into the background(ex. slight but persistent presence of AU12) or can be permanent or transient.

Examples of permanent features are the lips, eyes, and any furrows that have become permanent with age. Transient features include facial lines and furrows that are not present in a *neutral face* but appear with facial expressions. Thanks to the Action Unit coding three categories can be defined:

- **Macro-expressions or Facial Expressions:** typically last between 0.5 - 4 seconds, they can be easily seen and occur in daily interactions.
- **Micro-expressions** last less than half a second, occur when trying to consciously or unconsciously conceal or repress the current emotional state.
- **Subtle expressions:** associated with the intensity and depth of the underlying emotion. Unlike micro-expressions, they are associated with the intensity of the emotion that is occurring and not with the length of time that they occur, the intensity of these facial actions constantly varies. Subtle expressions denote any onset of a facial expression where the intensity of the associated emotion is still considered low.

Because of its importance to the study of emotion, and recently cognitive states, several observer-based systems of facial expression measurement have been developed.

Usually, many facial expression recognition software, don't capture all the AUs because of the non-additive property which can lead to measurement uncertainties or errors. For example, especially in our case, it is better to have the AUs related to lid and brow (AU1-AU7) clearly defined and don't capture AU41-46 (squint, blink, wink) which are very similar and capture them with a specific eye-tracking sensor. In fact, as it will be detailed below, the adopted software does not give the possibility to monitor all the AU but the main and universally recognized in other coding systems and that still allow to make a complete analysis. It can therefore be summarized as follows:

- Facial behaviour occurs not continuously but rather as episodes (events) that typically manifest themselves as discrete events.
- Action units that occur together are related in some way and form an event. AU 1+2+5 is an example of action units that frequently co-occur.
- The consequence of dimensional accounts is that linear transitions from one expression to another will be accompanied by characteristic changes in identification. Any transition between expressions lying at opposite points in the emotion space will need to pass through a neutral expression. For example, a transition from a *happy* face to an *angry* face will need to pass through a neutral face. Whereas transitions between expressions which do not involve entering the region of another emotion can be relatively abrupt.[25]

It's therefore evident that many studies have already been done on the relationship between MFE and emotions and very few on cognitive states. Some efforts have been made but they are limited to empirical and conceptual analysis and they are not satisfactory and useful in the engineering field. For this reason, in order to analyse the relationship between MFE and cognitive state, the first approach adopted is to

study the AUs that come together during an emotional state and then relate these AUs events to a particular cognitive state. In Figure 4 and Figure 5 are and reported the AUs related to emotional states ([26]).

Emotion	Example photo	Action units	Physical description
Fear		1+2+4+5+7+20+25	Eyebrows raised and pulled together, upper eyelid raised, lower eyelid tense, lips parted and stretched
Happiness		6+7+12+25+26	Duchenne display
Interest		1+2+12	Eyebrows raised, slight smile
Pain		4+6+7+9+17+18+23+24	Eyes tightly closed, nose wrinkled, brows furrowed, lips tight, pressed together, and slightly puckered
Emotion	Example photo	Action units	Physical description
Amusement		6+7+12+25+26+53	Head back, Duchenne smile, lips separated, jaw dropped
Anger		4+5+17+23+24	Brows furrowed, eyes wide, lips tightened and pressed together
Boredom		43+55	Eyelids drooping, head tilted, (not scored with FACS: slouched posture, head resting on hand)
Confusion		4+7+56	Brows furrowed, eyelids narrowed, head tilted

Figure 4: Action Units and Emotions 1.1[26]

Contentment		12+43	Smile, eyelids drooping
Coyness		6+7+12+25+26+52+54+61	Duchenne smile, lips separated, head turned and down, eyes turned opposite to head turn
Desire		19+25+26+43	Tongue show, lips parted, jaw dropped, eyelids drooping
Disgust		7+9+19+25+26	Eyes narrowed, nose wrinkled, lips parted, jaw dropped, tongue show
Embarrassment		7+12+15+52+54+64	Eyelids narrowed, controlled smile, head turned and down, (not scored with FACS: hand touches face)
Pride		53+64	Head up, eyes down
Sadness		1+4+6+15+17	Brows knitted, eyes slightly tightened, lip corners depressed, lower lip raised
Shame		54+64	Head down, eyes down
Surprise		1+2+5+25+26	Eyebrows raised, upper eyelid raised, lips parted, jaw dropped
Sympathy		1+17+24+57	Inner eyebrow raised, lower lip raised, lips pressed together, head slightly forward

Figure 5: Action Units and Emotions 1.2[26]

It can be seen that there are not only reported the 7 primary emotions but also the secondary ones which can be defined as primary ones strained by personal experience. In fact, personal experience has a key role in the cognitive state. Gross proposed five emotional regulation strategies and one of them is ([5]):

‘Cognitive change means change the understanding of emotional events and the cognition of the personal significance of the emotional time’

From the literature review carried out, the key contents that have been selected as a starting point to analyse the data and achieve the purpose of the research are reported [27-32]:

- Numerous studies of the correlations between facial expressions and learner emotions have identified a link between confusion and AU4, which is the “Brow Lowerer” movement. Thus, confusion may function as an essential intermediary state on the path of deep learning.
- Flow e Confusion are strictly related to problem-solving.
- Positive emotional expressions: AU 12 had to receive a lower intensity rating (2 to 5) if it co-occurred with AU 6, whilst AU 12 had to receive a bit higher rating (3 to 5) if it doesn’t appear with AU 6.
- AU 23 and AU 24 are both controlled by the same muscle and co-occur frequently.
- Negative emotional expressions were defined by the absence of AU 12 and the presence of at least one of the following AUs: 1+4(pulling the medial portion of the eyebrows upwards and together), 9, 10, unilateral 14, 15 and 20.

But not all the negative emotions generate a reduction in cognitive state. For example, certain negative emotions, such as confusion, can have a beneficial effect because when the subject experiences a state of confusion he/she becomes aware of the extent of the problem and start to think in order to resolve troublesome impasses. The effectiveness of problem-solving in promoting learning at deeper levels of comprehension can also be attributed to the deployment of key cognitive and meta-cognitive processing:

- happiness and delight when tasks are completed
- eureka moments when challenges are unveiled and major discoveries are made, and flow-like states when they are so engaged in problem solving that time and fatigue disappear.

Figure 6 briefly explain what these emotions mean.

Affective State	Definition
Anger	negative affect toward material or person to an extreme degree
Anxiety	nervousness, anxiety, negative self-efficacy, embarrassment
Boredom	uninterested in the current problem
Confusion	poor comprehension of material, attempts to resolve erroneous belief
Contempt	annoyance and/or irritation with another person
Curiosity	desire to acquire more knowledge or learn the material more deeply
Disgust	annoyance and/or irritation with the material and/or their abilities
Eureka	sudden realization about the material, a ha! moment
Fear	feelings of panic and/or extreme feelings of worry
Frustration	difficulty with the material and an inability to fully grasp the material
Happiness	satisfaction with performance, feelings of pleasure about the material
Neutral	displays no visible affect, at a state of homeostasis
Sadness	feelings of melancholy, beyond negative self-efficacy
Surprise	genuinely does not expect an outcome or feedback

Figure 6: Definition of affective states[33]

According to the experiments available in the literature only some emotions can be related to cognitive load.[34]

These emotions don't occur in the same frequency during cognitive load tests. Figure 7 shows that boredom, confusion, curiosity and frustration appear more frequently instead sporadic emotions are anxiety and happiness. Finally, the other emotions like anger, contempt, disgust, eureka, fear, sadness and surprise appear in exceptional cases and they can't be strictly related to a specific cognitive load.

Affective States	Frequencies		Proportions		One-sample t-test		
	<i>N</i>	<i>P</i>	<i>M</i>	<i>SD</i>	<i>t</i> (40)	<i>p</i>	<i>d</i>
Routine							
Boredom	39	.951	.106	.108	3.14	< .010	.49
Confusion	36	.878	.092	.062	4.01	< .001	.63
Curiosity	33	.805	.138	.142	3.85	< .001	.60
Frustration	39	.951	.105	.071	4.68	< .001	.73
Sporadic							
Anxious	25	.610	.042	.045	-.162	.112	-.24
Happiness	35	.854	.055	.049	.300	.763	.04
Exceptional							
Contempt	17	.415	.027	.047	-.351	< .010	-.55
Eureka	21	.512	.025	.039	-4.58	< .001	-.72
Anger	17	.415	.022	.039	-5.12	< .001	-.79
Disgust	26	.634	.030	.037	-3.97	< .001	-.62
Fear	6	.146	.008	.028	-10.3	< .001	-1.6
Sadness	13	.317	.012	.024	-11.1	< .001	-1.7
Suprise	25	.610	.027	.031	-5.52	< .001	-.84
Neutral	41	1.00	.311	.208			

Figure 7: Frequency of appearance of emotions[33]

N is the number of students that experienced the state at least once. P is the proportion of students that experienced the state at least once. So the 90% of the subjects has experienced the first 4 emotions.

Furthermore anger, anxiety, boredom, confusion, curiosity, disgust and frustration are the more persistent emotions. Persistence means how much an emotion that occurs at an instant t persists even at an instant $t+1$ during a cognitive effort.

Boredom is a state that is alleviated when a new problem is presented. On the other hand, confusion and curiosity are most frequently observed in the midst of problem-solving, followed by the presentation of a new problem. Frustration and happiness are another pair of affective states with similar occurrence patterns and rarely occurred during the process of deriving a solution to the problem. These affective states seem to be complementary to confusion and curiosity in that they occur *after* the resolution has been reached. Consequently, confusion and curiosity appear to be related to the problem-solving *process*, while frustration and happiness are related to the problem-solving *outcome* (or *product*). Furthermore, boredom induces shorter response time unlike confusion.

Some research on frustration has proved useful data [35]. In fact, this emotional state can be strictly related to the alteration of a cognitive process, triggering negative effects, especially with regard to workload or attention. Reducing frustration during the performance of ATC tasks is an important step towards improving Ait Traffic safety.

Studies from Human Computer Interaction (HCI) linked frustration to increased facial muscle movement in the eye brow and mouth area([30, 35, 36]).

In addition, a recent study investigating facial activity of frustrated drivers found that muscles in the mouth region (e.g., tightening and pressing of lips) were more activated when participants were frustrated compared to a neutral affective state.

The 5 most frequent AU during cognitive efforts in which appear frustration and learning gains are AUs 1, 2, 4, 7, and 14. Upper face movements are predictive of learning, engagement and frustration while mouth dimpling is a positive predictor of learning and self-reported performance.

In general, it can be said that AU1 (inner brow raising) and AU2 (outer brow raising) are related to frustration, whilst AU4 (brow lowering) and AU7 (eyelid tightening) are related to confusion. It's further necessary to say that AU4 introduces a conflicting movement of the brow that may impact the detection of the expression.

The correlations found between AUs and frustration are now listed in summary form:

- Brow lowering (average) intensity (AU4) positive predict frustration. AU4 has been correlated with confusion in prior research [37] and interpreted as a thoughtful state in other research [27, 38].
- A lesser sense of being hurried or rushed is predicted by a reduction of the AU2 frequency.
- Action Unit 14 (mouth dimpling) was positively correlated with both frustration and learning gain. [39] [35] and performance is positively predicted by AU14 frequency.
- AU2 frequency and intensity is a positive predictor of frustration

It can be deduced that AU4 and/or AU14 better represent a thoughtful and contemplative state. [38]

2 State of the Art: The relation between Human and Machine

In this chapter will be explained the starting point for understanding the interaction between human and machine to create a global view of the problem that will be analysed and provide the necessary awareness to understand all aspects to be taken into account in a human-machine interface.

The interaction between man and machine has been studied for many years and while on one hand, it is easy to predict the response of a machine and change it in function of needs, on the other hand, it is very difficult to model the response of man that is very complex and unpredictable.

Many efforts have been made so far and researches continue to be carried out on human cognitive processes because the dozens of factors involved make it very difficult to predict the responses of human behaviour and, above all, it is often not even easy to monitor them because they are affected by disturbances especially if monitored by wearable sensors. For this reason, this research work has been focused on facial micro-expressions, a rapidly developing field in which many types of research have been carried out in the psychological field since the seventies, but in the aerospace engineering field, no FE cognitive analysis has been carried out so far. One of the great advantages of facial micro-expressions is that it has been shown that they can be catalogued and assumed to be constant with age, ethnicity, cultural and social spheres. Therefore FE monitoring can be performed using non-wearable sensors, which are therefore less affected by disturbances and, above all, they do not bother the operator both physically and cognitively.

An important aspect of the interaction between man and machine is that as much as possible the machine must adapt to man and not vice versa. Asking an ATMo with years of experience to change its habits or to wear monitoring equipment could lead to an increase in stress (distress not eustress) of the latter and thus go against what is the objective of monitoring. So another potential winning aspect emerges, the hardware required for monitoring is very simple and economical because a simply camera is required.

Now the real evaluated scenarios will be presented, such as ATM, SPO, OTM to figure out the research study approach.

Controlled airspace is divided into sectors. An en-route sector is a region of airspace that is typically situated at least 48 kilometres far from an airport for which an associated ATCo has responsibility. ATCos have to accept aircraft into their sector; check aircraft; issue instructions, clearances, and advice to pilots; and hand aircraft off to adjacent sectors or airports. The radar screen displays characteristics of the sector (e.g., boundaries and airways), the spatial position of aircraft, and vital flight information (identifiers, altitude, speed, flight destination). When the aircraft leaves the airspace assigned to the ATCo, control of the aircraft passes on to ATCo controlling the next sector (or to the tower ATCo). As is typical in many real-world complex systems, this environment imposes multiple concurrent demands on the operator.[13] One goal of workload modelling is to allow ATC providers to predict workload levels ahead of time in order to allow them to adopt workload management strategies. For example, this may include splitting a sector or introducing flow restrictions. However, to date, dynamic density metrics have been unable to

accurately predict ATCo workload ahead of time (Kopardekar & Magyarits, 2003; Majumdar & Ochieng, 2002; Masalonis, Callahan, & Wanke, 2003).

It can therefore be deduced that the task demand is very much linked to the workload, thereafter will be explained how workload could be assumed as a number of task function.

Currently in commercial transport the flight crew is composed of a Pilot Flying (PF) and a Pilot Non-Flying (PNF). The presence of both has always been considered fundamental because in case of distraction, fatigue or illness of one of the two, the second can intervene both in flight and on the ground. In this way, what can be called 'shared between human responsibility' has always been guaranteed the safety of air traffic. The new frontier is to eliminate one of these two crucial elements and this obviously requires a great deal of effort to be able to guarantee the same level of safety supporting enhanced synergies between the human and the avionics systems. These synergies yield significant improvements in the overall performance and safety levels. In this way the primary duty of the single on board pilot (PF) is still controlling the aircraft. It is necessary to point out that this would involve not only assisting the human more while is flying but also on the ground because also the Airline Operations Centre (AOC) should manage more tasks.

In case of emergency, AOC operators upgrade their roles to ground-based first officers, who assist the on board pilot by real-time voice coordination with the FD and control of the aircraft through the HMI2 in the ground workstation. A novel Cognitive Pilot-Aircraft Interface (CPAI) concept, which introduces adaptive knowledge-based system functionalities to assist single pilots in the accomplishment of mission-essential and safety-critical tasks in modern commercial transport aircraft. The CPAI working process is subdivided in three separate stages: sensing, estimation and reconfiguration. The level of automation can be adapted based on the operator's workload monitored on real-time detection of the pilot's physiological and cognitive states allowing the avoidance of pilot errors and supporting enhanced synergies between the human and the avionics systems. Moreover, it is not only important the system adaptation but also the alerting system because an alone pilot in an alerting situation could react in an unsafe way if alerting system saturates his/her cognitive and stress management skills.

Consequently, the transition to SPO require substantial increases in automation support both in the flight deck and on the ground as well as significant changes in the roles and responsibilities of pilots and Air Traffic Management (ATM) operators. Suitable mathematical models are introduced to estimate the mental demand associated to each piloting task and to assess the pilot cognitive states.

In particular, the primary duties of pilots are progressively shifting towards supervisory roles, intervening only when necessary.

"The consensus among research and operational communities is that it is important to understand the factors that drive mental workload if they are to improve airspace capacity ."

(Christien, Benkouar, Chaboud, & Loubieres, 2003; Majumdar, Ochieng, McAuley, Lenzi, & Lepadatu, 2004).

Subsequently, the so far carried out studies on cognitive states analysis with wearable and remote sensors will be presented.

2.1 Human Machine Interface Interaction

Adaptive Human-Machine Interfaces and Interactions (HMI2) are closed-loop cyber-physical systems comprising a network of sensors measuring human, environmental and mission parameters, in conjunction with suitable software for adapting the HMI2 (command, control and display functions) in response to these real-time measurements. [40, 41]

According to what has been said so far, it is clearly deductible that a real-time monitoring is necessary to take into consideration the dynamic variations in cognitive task loads. Cognitive HMI2 are a particular subclass of HMI2 systems, which support dynamic HMI2 adaptations based on the user’s cognitive state. Therefore, it is necessary that the System can independently decipher the cognitive signals and adapt its level of automation. At this point, it is necessary to make a brief incision on two concepts: Automation and Autonomy.

Automation is the ability of a system to perform well-defined tasks without human intervention using a fixed set of “hard-coded” rules/algorithms to produce predictable, deterministic results. The automated tasks may be sub-tasks of a larger activity that involves human intervention—in which case, the overall activity is only partially automated to a lesser or greater degree. [3, 42]

Autonomy is the ability of a system to perform tasks without human intervention using methods, usually emergent, that arise from its interaction with the external environment. Such behaviours include reasoning, problem solving, goal-setting, self-adaptation/organisation, and machine learning, and may not be deterministic. In our context, the degree of autonomy exhibited by the system has to be dynamically variable.

Autonomous systems are distinguished from highly automated systems by their ability to respond to the environment and adapt their behaviour without being explicitly programmed to do it. Being non-algorithmic, they are often implemented using heuristics and non-deterministic AI techniques such as machine learning , deep neural networks, fuzzy logic, and genetic algorithms.

Characteristic	Automation	Autonomy
Augments human decision-makers	Usually	Usually
Proxy for human actions or decisions	Usually	Usually
Reacts at cyber speed	Usually	Usually
Reacts to the environment	Usually	Usually
Reduces tedious tasks	Usually	Usually
Robust to incomplete or missing data	Usually	Usually
Adapts behaviour to feedback (learns)	Sometimes	Usually
Exhibits emergent behaviour	Sometimes	Usually
Reduces cognitive workload for humans	Sometimes	Usually
Responds differently to identical inputs (non-deterministic)	Sometimes	Usually
Addresses situations beyond the routine	Rarely	Usually
Replaces human decision-makers	Rarely	Potentially
Robust to unanticipated situations	Limited	Usually
Adapts behaviour to unforeseen environmental changes	Rarely	Potentially
Behaviour is determined by experience rather than by design	Never	Usually
Makes value judgments (weighted decisions)	Never	Usually
Makes mistakes in perception and judgment	N/A	Potentially

Table 1: Automation and Autonomy feature ([43])

During the development of a closed loop monitoring system it is important to consider some key aspects for a correct evaluation of the parameters:

- Distinction and characterization between normal and atypical operations.
- Qualitative and quantitative model for the definition of biometric parameters such as attention, mental workload, fatigue for:
 - Support real-time decision-making in high-stress dynamic conditions
 - Dynamic reallocation of functions between humans and machines
- Definition of a graphic interface that makes the operator aware of the level of automation of the software so the human knows what the software is doing.

Therefore, an important aspect is not only to help the operator adapting the level of automation but also to make him aware of the levels of workload, fatigue and situation awareness because often humans are not aware of what they are experiencing so, for example, if they are affected by a lack in situation awareness, making them aware about it can already go to solve the partially situation without necessarily having to increase or adapt too much the autonomy of the software.[44, 45]

Current HMI2 are static, obviously represent an important help for the operator but given the future forecasts of the evolution of the aeronautical and aerospace field these new roles require a corresponding evolution in the HMI2. Additionally, as part of the Technology Horizons project, the United States (US) Air Force identified that natural human capacities and advanced technologies become increasingly mismatched and humans will be the weakest component in the generalised processes and systems by 2030. [12]

Developments in wearable and remote sensing technologies led to the development of sensors with a high enough reliability to be implemented in a sensor network and they are now a strong object of study. [2]

The switch from static HMI2 to CHMI2 is therefore a necessary consequence. Furthermore the cognitive state is the set of all those factors that can be considered also closely related to the emotional state. For this reason the research conducted so far on detecting the emotional state through MFE had been used as a starting point.

2.2 CHMI2 architecture

For the determination of the cognitive state, many factors must be taken into account, both personal factors related to the emotional state and objective ones such as the task load. The final objective is to determine the cognitive load of the operator that in the simplest and most intuitive form can be defined as:

$$CL = f(WL, A, SA, Fa, Fr, E, Env) \quad (2.1)$$

It is a function of: Mental Workload or Workload (WL) as explained in the following chapter, Attention (A), Situation Awareness (SA), Fatigue (Fa), Frustration (Fr), Emotional state (E) and *Env* which is the environment that obviously plays also a key role. This research is focused on the evaluation of Workload, we have taken into account the environment as a source of possible measurement errors by the sensors without considering its influence on the cognitive load.

Two ways of analysis have been pursued, the first is the evaluation of cognitive state through the Workload using objective parameters such as Taskload or number of aircraft, the second is the evaluation of cognitive state through biometric

evaluations of the sensors that make up the sensor network referring to studies carried out by us and available in the literature.

It is therefore essential to perform a Task Analysis to highlight the number and difficulty of tasks during the execution of operations. Kirwan and Ainsworth's *A Guide to Task Analysis* (Kirwan and Ainsworth 1992) provides a good reference on performing a task analysis. The types of Task Analysis are now briefly presented, which will be taken up later in *Experiments* Chapter

2.2.1 Tasks Analysis

Tasks Analysis (TA) is a necessary and fundamental procedure to define the workload of the operator and therefore to understand how to develop the software interface, the Level of Automation (LOA) and then how to implement the monitoring carried out by the sensor network and its influence on the system. The first step to write a TA is to allocate tasks in a list according to their level of criticality through Hierarchical Task Analysis (HTA).

The HTA is a means of systematically defining tasks and functions from the user's perspective to organize them in a hierarchal manner.

Task procedures can be summarized in three levels:

- User level: it's the top level task referred to what the operator has to do without defining yet how and when to do it.
- Platform level: are the procedures imposed by the interface. Obviously the interface must be created in such a way to make intuitive and ergonomic the execution of tasks (click, pull-down menus, typing). This level is also generic in that many different high-level goals can be accomplished using various combinations of low-level procedures.
- Application level: this is the level that have the greatest impact. The development of the system determines what information are shown on the screen and which are more or less visible so how the operator performs the low-level interface procedures to accomplish top-level goals.

The Hierarchal task analysis can therefore summarized as follow:

1. Identify goals
2. List the step that the operator have to perform to accomplish goals
3. Improve the procedures(sequence of steps)

Block diagrams are commonly used to represent the top-down relationships between objectives and tasks.

Top-level objectives are decomposed into lower-level sub-objectives, which can themselves be decomposed into lower-level tasks and subtasks. A task is usually broken down into the sequence of actions that need to be performed to accomplish it in a systematic and sequential manner. The sequence of actions that is created to perform a task can be evaluated in a more or less coarse way depending on key factors such as complexity or criticality of the task, the impact of a particular design on the human-machine interaction or the risk of human error.

The drafting of the task hierarchy allows to have a first workload estimation. [46]

Subsequently, the tasks must also be evaluated according to their criticality so their impact on the safety of operations, this analysis is carried out through what is called Critical Task Analysis (CTA).

A CTA is specified by the Federal Aviation Authority (FAA) as an approach for evaluating the human factors involved in mission-critical tasks when designing/implementing a system. CTA involves the application of task analysis techniques to tasks critical to safety, integrity, and environment, to facilitate the identification of uncontrolled or poorly controlled error risk. Identifying critical tasks is the key to deciding when the operator has to be more monitored and setting the thresholds of software autonomy. The process of critical task analysis is similar to cognitive task analysis, tasks are represented through tables or diagrams to show which are most affected by performance-influencing factors and to improve the awareness of failure modes. This analysis is done a priori before developing a system and then verified with the aim of improve performance and minimize risk. Figure 1 shows what are the key parameters to consider when performing a CTA.

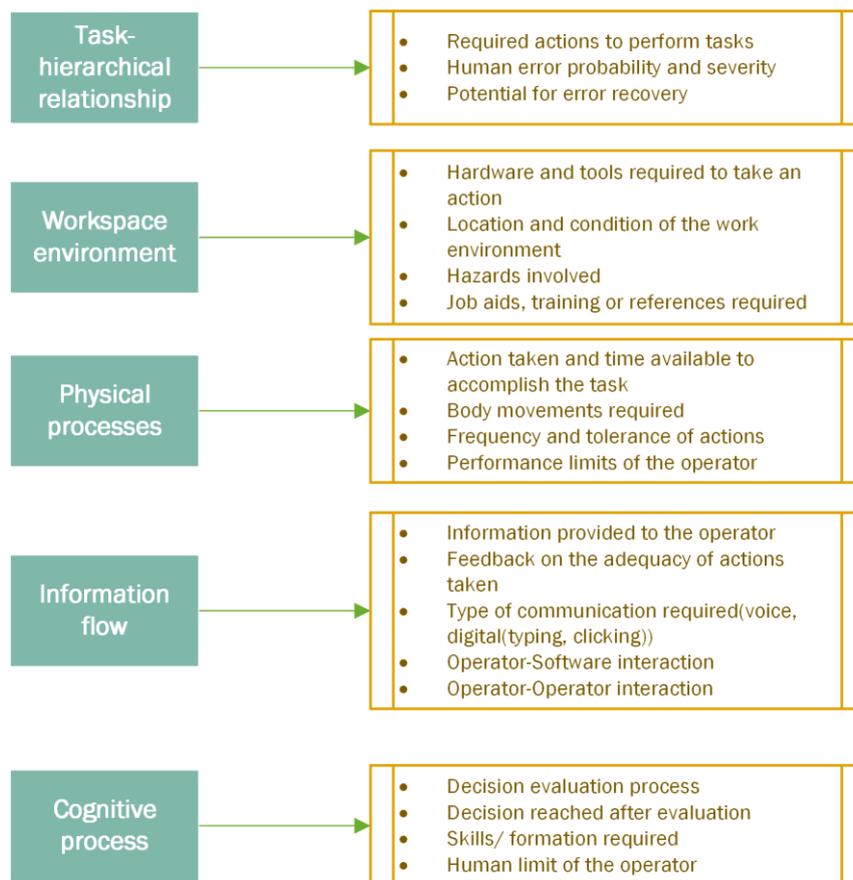


Figure 8: Content requirements for a Critical Tasks Analysis Report

2.2.2 Cognitive Task Analysis

The greater the number and complexity of tasks, the greater the importance of understanding what is the mental process that must be implemented to perform a task. As previously mentioned the experience plays a key role and also the training because for example in air traffic control the operator knows specific procedures to be implemented to perform tasks but when the tasks repetitiveness is excessive a reduction of the situation awareness occurs or on the contrary when a new problem happen this leads to a sudden increase of workload because the process of understanding the problem and determining the solution requires more effort. In

fact, in the following chapters it will be explained in detail what is meant by workload as a variable that also takes into account the cognitive effort required to perform a task. It's necessary to point out that a key factor is not only the number of tasks but also how much memory is required from the operator (information that the operator must keep in mind) and the process of understanding and processing a solution.

This analysis can start with a list of activities, such as one obtained from a gross task analysis (Miller 1953), which describes the top-level tasks, relevant sub-tasks and user control actions. This list is developed with the aim of defining the operator's mental process by understanding the succession of information managed to determine a decision-making strategy. This approach is also useful to determine the requirements specifications that can be used to define or improve the software interface at platform level and application level.

One of the outputs of the cognitive task analysis is the *decision ladder* (Rasmussen 1976), describing the information processing activities involved at each stage of the decision process.

The cognitive task analysis is ultimately used to drive the systems-level definition of the related human factors requirements.

Once the tasks list has been drawn up and the cognitive tasks analysis has been carried out, the interaction between Human and machine must be specifically evaluated in order to define the decision authority shared between both the human operator and automated system. Obviously the objective is that the tasks are carried out safely and efficiently at any level of complexity required.

Endsley (1987) developed a LOA hierarchy in the context of the use of expert systems to supplement human decision making. This list has been developed precisely those operating environments in which a system provides information to an operator who has to perform certain tasks at the established automation level. There is therefore an interaction between Human and machine on four domains that can be performed to different degrees by human or software: Monitoring, Generating, Selecting, Implementing. The first stipulated list was composed of five levels and was later expanded to 10 to better consider psychomotor tasks required during real-time control, aircraft piloting, advanced manufacturing and tele-operations. ([47]). The 10 levels of automation are therefore presented:

Monitoring (M), Generating(G), Selecting(S), Implementing(I)

- I. Manual Control (MC):** The operator performs all tasks M,G,S and I so the level of automation is null.
- II. Action Support (AS):** The operator is assisted by the system only for certain operations such as M and I while human control actions are required for G and S.
- III. Batch Processing (BP):** the human generates and selects the options to be performed and they are autonomously carried out by the system(automated implementation).
- IV. Shared Control (SHC):** is the first level where even the system generates possible decision options and the operator can select one of them or generate his or her own options. The selecting role is still fully carried out by the Human.
- V. Decision Support (DS):** From this level the implementation is carried out by the system. This is the typical level that characterizes the decision support

systems that provide option guidance which can be selected or not by the operator.

- VI. **Blended Decision Making (BDM):** Is the first level where also the selecting is partly done by the computer and the operator can approve the generated selected option or select others from a list generated by the system.
- VII. **Rigid System (RS):** This level is representative of a system that presents only a limited set of actions to the operator. The operator's role is to select from among this set.
- VIII. **Automated Decision Making (ADM):** From this level also the selecting role is carried out by the system which selects the best option to implement. Represents the autonomous decision making.
- IX. **Supervisory Control (SC):** At this level the system generates options, selects the option to implement and carries out that action. As the name of the level says, the Human is only supervisor. From this level on, in particular, problems may arise in reducing situation awareness and therefore the need for monitoring the cognitive state.
- X. **Full Automation (FA):** this level, the system carries out all actions. Also the monitoring is done by the computer and the human can not intervene so it is completely out of the control loop. This level is used in systems where errors do not compromise safety and therefore human processing is not deemed to be necessary.

As previously mentioned, the level of automation must be chosen with care and adapted not only to the number and difficulty of tasks but also to the type of person because a too high level of automation could lead to boredom phenomena and therefore a reduction in arousal that greatly affects performance.

A dynamic tasks allocation based on a continuous assessment of the human cognitive states and the estimated task mental demand associated with environmental and operational conditions is therefore needed.

The tasks allocation has also to take into account the environmental and operational external conditions such as, in ATM and airlines constraints aircraft velocity, position and attitude. All of them affect the human performance significantly so the CHMI² shall consider these parameters in real time with a very short delay time between appearance and system response.

The relationship between performance and arousal levels was first described by Yerkes and Dodson (1908), in what is now called the *Yerkes-Dodson law*, showed in Figure 9. It states that a high level of arousal can enhance performance on an easy task, but on a difficult task performance is an inverted U-shaped function of arousal the more difficult the task is, the lower the arousal level at which performance peaks. ([48]) At the same time an excessive arousal leads to a reduction in situation awareness so a decrease in the number of monitored cues.

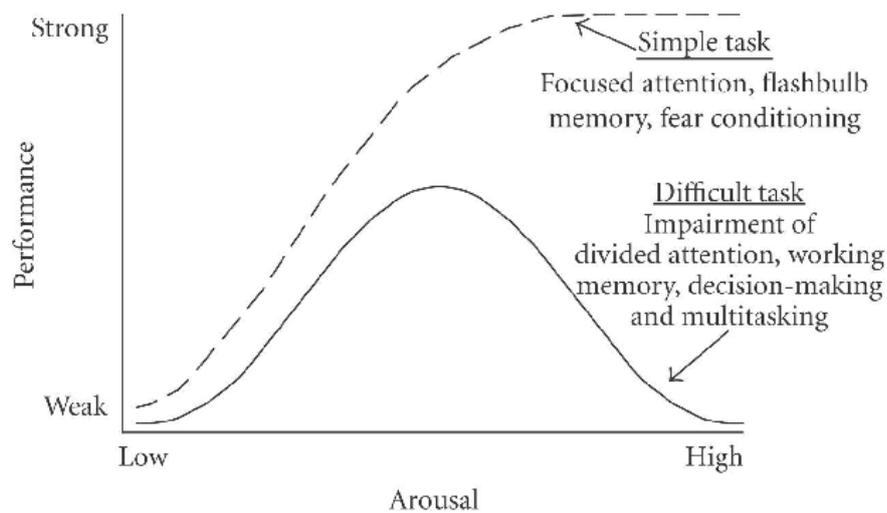


Figure 9: Yerkes-Dodson law

A CHMI2 system must therefore monitor in real time the human physiological and cognitive states in order to adapt the automation level to reach three main goals: minimise the cognitive load of the human operator, enhance the situational awareness of the human operator, minimise reliance on human memory for any task and system operating procedure. In case of unsatisfactory performance, the system must warn the operator providing caution or warnings. Therefore, a CHMI2 system can monitor the Human-environment situation awareness between the operator and the tasks he/she has to perform and at the same time provide Human-Human situation awareness and make the operator aware on his/her cognitive state through warnings.

Traditional alerting involves visual and auditory cues and usually the forms of alerting are fixed. However, studies in this regard have shown that stressed individual tends to focus more on solving a task that he or she considers a priority and neglects to attend to other important information or tasks. This phenomenon is called Attention Tunnelling and usually even if people are very experienced and trained can not realize that they have entered into this condition and may not even notice the warnings or alerts that are presented because they are used to those modes. So adaptive alerting can be designed to prevent such occurrences. Typically, people cannot avoid or solve this and similar problems by themselves so adaptive alerting can be designed to prevent such occurrences; for example integrating haptic cues.

Once defined the mode of evaluation of the cognitive state and the levels of automation of a system, the CMHI2 design requirements are now analyzed.

During the monitoring of physiological and cognitive states in real time the main goal is to minimize the cognitive load of the human operator and enhance the situational awareness. The key aspects to accomplish this goals are:

- minimize reliance on human memory for any task and system operating procedure;
- clear features and unambiguous display formats and functions of system modes and sub-modes

Moreover, being aware that these technologies will continue to evolve and must be adapted to the environment in which they are applied, it is essential to structure

the system with a modular architecture allowing the accommodation of additional input parameters, cognitive state variables, new formats and functions.

It is therefore clear that the process of using a sensor network for monitoring physiological features is developed starting from the detection of biometric parameters to estimate the cognitive state and then take measures to correct the level of automation or warning according to a defined model. The purpose of this analysis is not to define the mode of adaptation of the software so we focused on the previous steps, from sensing to the definition of cognitive states and in particular the workload.

The typical structure of a CHMI2 system is now presented, which can be divided into three layers: sensing, cognitive states estimation and adaptation. [17] [16]

- Sensing: retrieves environmental/operational observables and biometric measurable. This level can be divided into actual sensing in which numerous sensors detect parameters such as blink ratio, brain frequencies or Facial Expression and extraction in which the raw detected data are processed to identify and extract meaningful variables.
- Cognitive states estimation: In this layer the sensor data are interpreted to define the cognitive states in real time. This layer is the one on which this research has focused, in which the relationship between the data collected by the sensors has been analyzed to verify the possible integration of FE in the network.
- Adaptation: the integration of the sensor network in the system to close the loop and modify the system interface on the estimation of the cognitive state made in the classification layer to optimize the human-machine teaming effectiveness. It can be made by simple haptic alerts, triggered by excess or undesirable cognitive states, or can involve in more complex interactions where adaptive levels of decision support and assistance is rendered to the operator through well-defined decision logics.

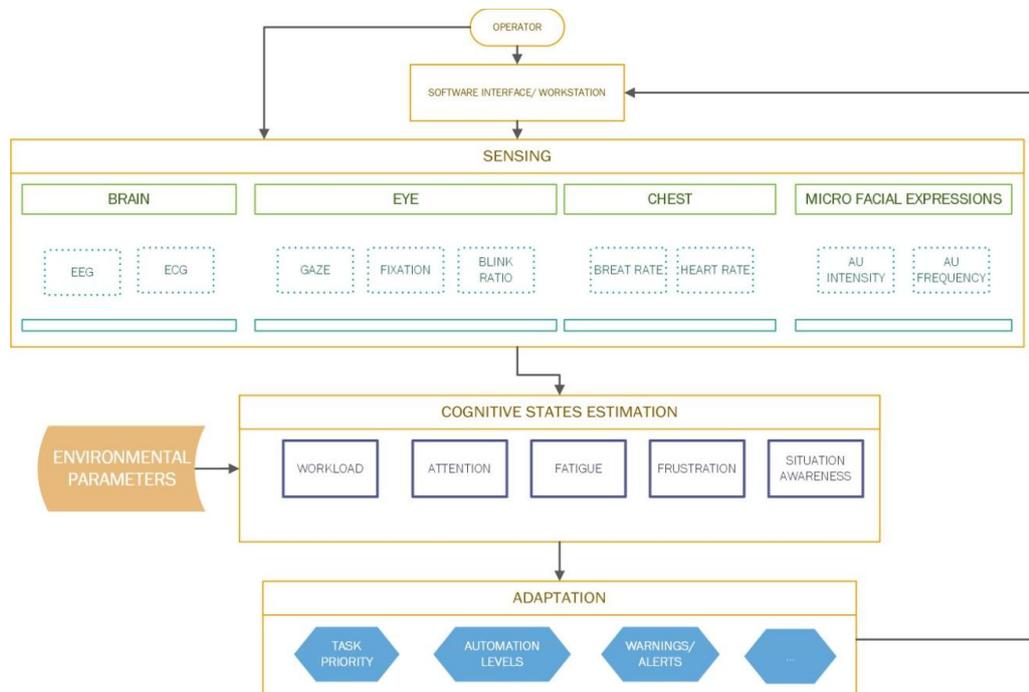


Figure 10: CHMI2 structure

3 The Sensing layer

The first task of the system is performed by the sensing layer that monitors in real time the operator's various parameters. The sensing layer comprises the hardware (sensors and systems) for acquiring the data as well as the software to pre-process the data.

Rubio's list of criteria is commonly used to evaluate the suitability of the measured parameters:

- *Sensitivity*: the ability of the parameter to detect changes in the operator's cognitive states.
- *Diagnosticity*: the ability of the parameter to discern the reason for the changes in these states.
- *Selectivity*: the sensitivity of the parameter in relation to a particular state (i.e. a parameter which is sensitive to changes in multiple cognitive states is not selective).
- *Intrusiveness*: the required amount of interference imposed on the operator to obtain the parameter.
- *Reliability*: the consistency of the parameter in reflecting the cognitive state of the operator.
- *Implementation requirements*: the time, hardware and software requirements for identifying the parameter.

Sensors can be divided into two main categories, wearable and remote sensors. The first ones require the operator to wear the sensor and this generates positive and negative aspects. Wearable sensors allow the biometric parameters of the operator to be directly observed. However, they can also induce a sense of discomfort or unease, thereby resulting in potentially corrupted measurements.

This aspect is essential because monitoring has not to lead to a counterproductive aspect such as increased frustration or reduced comfort that could severely affect cognitive states. Remote sensors such as the Gazepoint GP3 and Open Face have the great advantage that they are transparent to the operator that is not aware of being monitored, the negative side is that during the execution of tasks the operator can rotate the head or generally move the body over the sensors field of view. Eye tracking data and observations on the operator behavior and facial expression are particularly important to develop the specific CHMI² knowledge base.

These aspects, in relation to Open Face, will be discussed in *Performance evaluation* chapter. Wearable sensor are now presented.

3.1 Cardiorespiratory sensor

The monitoring of cardiorespiratory parameters has been conducted using the Zephyr Bioharness 3 showed in Figure 11.



Figure 11: Zephyr Bioharness 3

Two devices are installed on the side strap: the Heart Rate and Breathing Rate sensor. The raw monitored data are stored in a internal memory or can be transmitted to an external receiver to allow the real time evaluation without using uncomfortable cables. Further processing is performed to extract the cardiorespiratory features. This device allows to monitor a large number of data such as: inter-beat interval RR, Heart Rate HR, Standard deviation of NN intervals, Root Mean Square of Successive NN Differences, Percentage of successive NN pairs that differ by more than x milliseconds, Low Frequency component of HRV, High Frequency component of HRV, Minor axis of Poincaré plot(SD1), Major axis of Poincaré plot(SD2). It is necessary to note that this sensor has been designed for sports use but it has an acceptable accuracy for our purposes. The most related parameters to the cognitive state are presented in Table 2:

Parameter	Frequency (Hz)	Range (BPM/ms)	Accuracy	Unit
Heart Rate	250	0-240	(±2 BPM)	Beats per minute
Heart Rate Variability	250	0-280	(±1 ms)	millisecond
Breathing Rate	18	0-120	(±3 BPM)	Breath per minute

Table 2: Bioharness monitored features

The accuracy is referred to a static activity like in the analyzed scenario.

The instantaneous Heart Rate is described in units of beats per second as follow:

$$HR = \frac{60}{RR} \quad (3.1)$$

RR or inter-beat interval is the output of the sensor and means the time between two consecutive heart beats. The heart rate is usually described in beats per minute and can be derived by counting the number of RR peaks within a 60-second window which occur at discrete intervals between 800ms-1500ms. The obtained signal is very variable and assumes a step trend so usually to analyze and compare it with other sensors it is made smoother by applying a moving mean window with an amplitude varying between 5 and 10 seconds depending on the type of analysis and the sensor with which the Heart Rate is related.

Many other parameters can be defined by evaluating the Heart Rate Variability. These parameters can be classified into geometric, time and frequency domains. The main geometrical features by converting RR intervals into geometric plots are now presented:

$$SD1 = \sqrt{0.5 \cdot Var_n(RR_i - RR_{i+1})} \quad \text{short term HRV characteristics} \quad (3.2)$$

$$SD2 = \sqrt{0.5 \cdot Var_n(RR_i + RR_{i+1})} \quad \text{long term HRV characteristics} \quad (3.3)$$

n: sample window set at 30 seconds.

Minor axis of Poincaré plot SD1 and Major axis of Poincaré plot SD2 are obtained displaying the correlation between consecutive RR intervals where RR(i) is plotted on the x axis whereas RR(i+1) on the y axis. A Poincaré plot is a type of recurrence plot used to quantify self-similarity in processes into a higher-dimensional state space. The obtained points assume an elliptical distribution as shown in Figure 12.

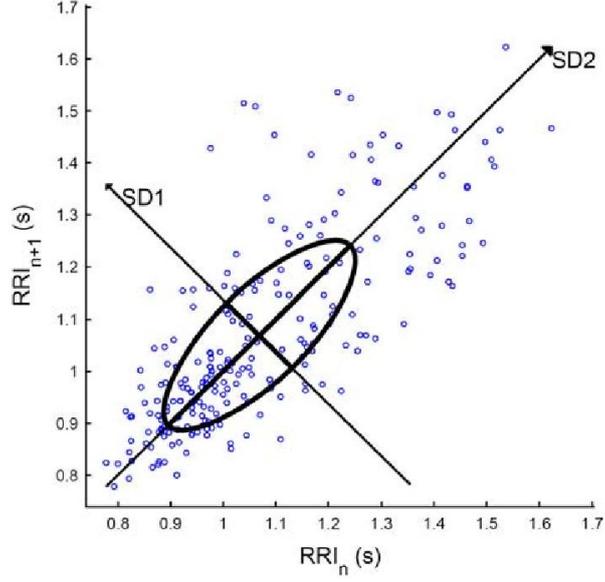


Figure 12: Poincare plot analysis

The x and y coordinates of the ellipse are given by the parametric equation:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \sqrt{2} \cdot \begin{bmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{bmatrix} \cdot \begin{bmatrix} SD2 \cdot \cos(\theta) \\ SD1 \cdot \sin(\theta) \end{bmatrix} + \begin{bmatrix} \overline{RR}_i \\ \overline{RR}_{i+1} \end{bmatrix}, 0 < \theta < 2\pi \quad (3.4)$$

The main three feature in the time domain are the percentage of successive RR pairs that differ by more than x milliseconds (pNN_x):

$$pNN_x = \frac{\text{count}_{n-1}(|RR_{i+1} - RR_i| > x \text{ ms})}{n - 1} \quad (3.5)$$

The root mean squared difference of successive RR intervals (RMSSD):

$$RMSSD = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} (RR_{i+1} - RR_i)^2} \quad (3.6)$$

The standard deviation of RR intervals (SDNN):

$$SDNN = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (RR_i - \overline{RR})^2} \quad (3.7)$$

The Spectrum Power Density (PSD) $f(\lambda)$ of the RR interval in a given RR time series, divided in 4 bands, is used to analyse data in the frequency domain [49]:

- Ultra-Low Frequency (ULF) ($1e^{-4}$ Hz to $3e^{-3}$ Hz)
- Very-Low Frequency (VLF) ($3e^{-3}$ Hz to 0.04 Hz)
- Low Frequency (LF) (0.04 Hz to 0.15 Hz)
- High Frequency (HF) (0.15 Hz to 0.4 Hz)

The frequencies of greatest interest that have been evaluated are HF and LF because they represent respectively the heart's control of the sympathetic and parasympathetic branches of the autonomic nervous system. Sympathetic system is traditionally described as a component that performs an attack/fugue function and is related to HF. The parasympathetic system is responsible for rest and digestion responses, i.e. all activities that occur when the body is at rest and is related to LF.

$$LF = \int_{0.04\text{hz}}^{0.15\text{hz}} f(\lambda)d\lambda \quad (3.8)$$

$$HF = \int_{0.15\text{hz}}^{0.40\text{hz}} f(\lambda)d\lambda \quad (3.9)$$

A cognitive effort or in general mental activities induce an increase of LF power in proportion to the HF band so the ratio:

$$PR = \frac{LF}{HF} \quad (3.10)$$

Can be an indicator of the mental strain.

3.2 Brain waves sensor

The monitoring of brain waves is done through a portable device called actiCAP Xpress which is designed for real time processing. In fact the collected data are sent to the BrainVision Recording software which can be interfaced with other software tools like C++ o Matlab.

Figure 1 shows how the cap is made. An elastic helmet allows to easily wear the device on which are installed several electrodes (16 active electrodes, one ground electrode and one reference electrode). The electrodes do not require conductive pastes because they contain low noise preamplifiers with a high precision.

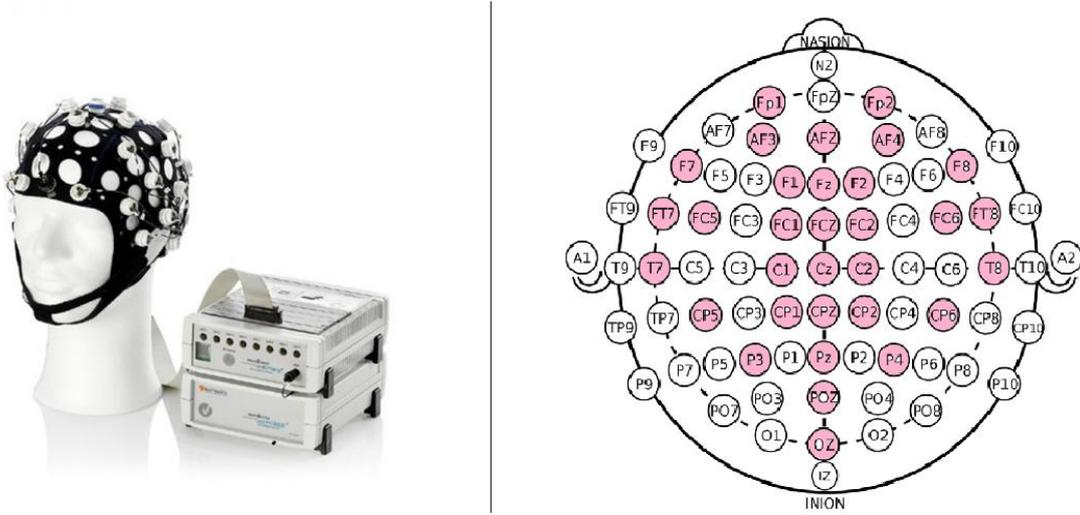


Figure 13: Electrode cap actiCAP Xpress

The sample rate is 2000 Hz, the measurement range is from -410 mV to 410mV with a resolution of $\pm 0.05 \mu\text{V}$. The brain waves can be divided into 5 main groups: Delta, Theta, Alpha, Beta, Gamma. Research conducted on brain waves has shown a close relationship between some brain waves and cognitive states such as workload, attention, engagement and fatigue so it was chosen to monitor the following 3 features[50] [51] [43]:

Alpha ($\alpha_F, \alpha_C, \alpha_P$): characterized by a frequency ranging from 8 to 13.9 hertz, they are typical of moments before falling asleep and can be divided in lower and upper alpha.

Theta ($\theta_F, \theta_C, \theta_P$): range from 4 to 7.9 hertz, characterize stages 1 and 2 of NREM sleep and REM sleep.

Beta ($\beta_F, \beta_C, \beta_P$): ranging from 14 to 30 hertz, are recorded in a waking subject, during intense mental activity. They can be divided in low, middle and high wave.

All of them are captured in the frontal, central and parietal sections of the brain. A useful parameter that allows to relate these waves is the EEG index, defined as:

$$EEG\ index = \frac{\theta_{F4+C4}}{\alpha_{O1+O2}} = \frac{\int_{4Hz}^{8Hz} f(\lambda)d\lambda}{\int_{8Hz}^{12Hz} f(\lambda)d\lambda} \quad (3.11)$$

The *EEG index* is calculated at each 5 second interval and the a linear detrending is applied. These pre-processed data are then filtered by a 50Hz notch filter, a band-pass filter and finally the Power Spectrum Density(PSD) is obtained. The obtained filtered sample is then integrated over different frequency intervals to determine the band power. Once all channels have been processed, the band powers of specific channels are summed and then divided to derive the *EEG index*.

3.3 Eye sensor

The sensor used to monitor eye parameters is the GP3 Eye Tracker. This sensor is a desk-mounted remote eye-tracker that provides numerous data such as raw eye tracking, blink rate and fixation. Being a remote sensor has the advantage of not being invasive to the operator, it is easy to install and can be mounted on the screen without affecting the operator's visibility on the monitor as shown in Figure 14. The sensor is a single hardware device composed by an infrared camera and an illuminator that send the raw monitored data to a computer via a simple USB port. Data are then processed by a dedicated software, the Gazepoint control software.



Figure 14: GP3 Eye Tracker installation

The stated tracking accuracy of the GP3 is between 0.5 to 1 degrees.

The two main features in the eye tracing studies are pupillometry and gaze features. Pupillometry is the science that studies the variation of pupil size and recent studies ([52]) have shown a correlation between the difficulty of the tasks performed by the operator and the pupil radius, in fact it is known as the pupil size varies as a function of illumination but this is a phenomenon that occurs at a low frequency (0.1Hz-2Hz). If instead the variation of the pupil radius is analysed at high frequency (2Hz-6Hz) it can be seen that it is related to the variation of cognitive states. Studies have been carried out on this subject that have highlighted this relationship but mathematical models that can generalize typical behaviour to all individuals are not yet available in the literature. Other parameters that have been analyzed in the field of pupillometry are eye closure and blink rate.

The parameter who describe the pupil dilatation is the dilation spectral power P_{dil} which is a function of the power spectral density estimate of the pupil radius time series $r(\lambda)$:

$$P_{dil} = \int r(\lambda)d\lambda \quad (3.12)$$

P_{dil} is the area under the PSD. Figure 15 shows the pupil radius detection accomplished by the sensor.

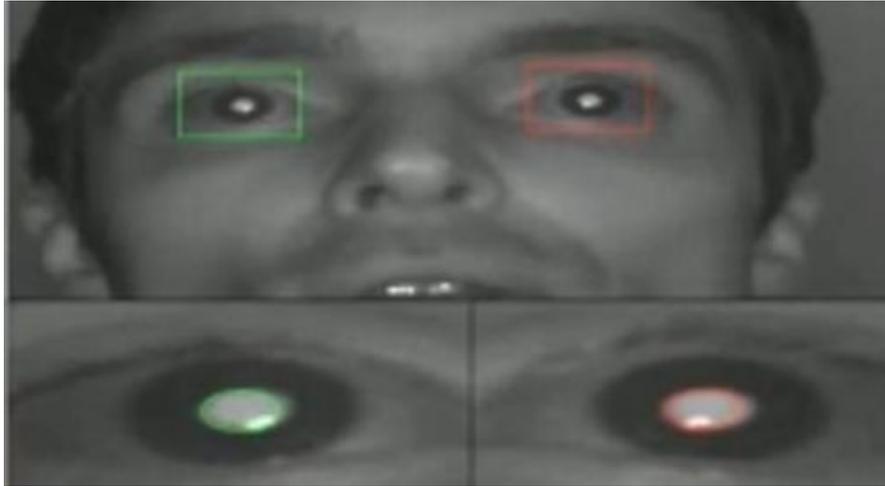


Figure 15: Pupil radius detection

When the human eye has to go to look at a monitor or read a text, it makes sudden movements of the eyeball by dwelling in places called saccades. In fact, even if you think that while scrolling through a text our eye reads every single letter in the text, it is not so. Figure 16 shows an example to explain the phenomenon.

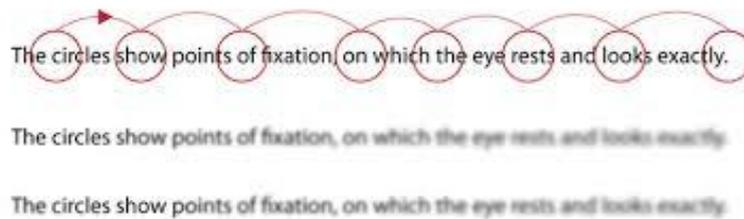


Figure 16: Saccades pattern

Eyes fix (fixation) a row and skim it in certain points (saccades) along it that vary according to the individual. A trained reader needs fewer saccades to read a row than an occasional reader. So the process works as follows: once the row is fixed certain points called saccades are selected and thanks to peripheral vision the brain uses its self-correction ability dictated by experience to interpret/estimate what is written around the saccade. The difference between saccades and fixations is that the former are higher frequency eye movements from one fixation to the next, usually lasting 20 ms to 200 ms.

Gazepoint GP3 identifies fixations based on the separation of groups of gaze points, if the dispersion (D) of the detected gaze points of the group overcome a given threshold it means that a new fixation appear. [53]

The sensor span consecutive data points along x and y axing into sliding window of typically 100ms to check for fixations.

$$D = \sqrt{(x_{max} - x_{min})^2 + (y_{max} - y_{min})^2} \quad (3.13)$$

Where x and y are the coordinates for the gaze points within the window. If D lies below a given threshold D_{max} , a fixation is registered and the window expands until the dispersion is calculated to exceed D_{max} . D_{max} is usually derived from gaze angles between 2° to 3° . In the conducted experiments the operator distance from the desktop is about 0.6m, this determines a range of D_{max} between 21mm and 31mm.

The gaze path randomness can be measured by three parameters: Nearest Neighbour Index (NNI), the explore/exploit ratio and the Visual Entropy(H). The Visual Entropy computes the randomness based on gaze transitions between predefined Regions of Interest (ROI). [54]

H is determined from gaze transitions between different ROI, the number (or probability) of these transitions are typically represented in a n-by-m matrix.

$$H = - \sum_{i=1}^n p(X_i) \sum_{j=1}^m p(Y_{ij}|X_i) \log_2 p(Y_{ij}|X_i) \quad (3.14)$$

H is a function of the probability of fixation p where X_i and Y_{ij} are respectively the previous and present fixation region. The more random is the fixation path the higher is H, on the contrary, during a cognitive effort the operator focus on a particular task so the H value decrease.

Among all these parameters Visual Entropy and Blink Rate have been thoroughly analyzed because is firmly proved their relationship with the cognitive load. The relationship between VE,BR and MFE has been therefore analyzed.

Blink Rate(BR) is known as one of the most effective measures of mental workload and it has been shown that increasing mental workload leads to decreasing blink rates ([11]). As will be seen in the following chapters, the relationship between BR and MFE has been studied because Open Face and GP3 monitor completely different parameters but the physiognomy of facial muscles causes eyelid contraction to induce disturbances on the rhinolabial muscle and the sides of the mouth that can be interpreted by Open Face software as disturbances. In the same way, however, these two software can work in synergy to increase the reliability of the interpretation of psychophysiological parameters. For example, if a cognitive state detected by GP3 is also proven by Open Face it means that the detection is verified and the warning and automation management system can react accordingly.

4 Cognitive state and Workload

Studies in workload started as early as 1930, but evaluated tasks mostly had a physical component that required the manipulation of machines (Sheridan, Simpson 1979). The advent of computers and the increase of automation in workplaces, had led to a more and more difficult definition of the workload, especially in HMI2. Even though there is no universally accepted definition of mental workload (Cain 2007), if it is considered at a top-level, it refers to the measurement of the mental processing demands placed on a person during the performance of a task (Gopher, Donchin 1986) but if analyzed at a deeper level the most difficult component to consider and model is the chain of relationships that determines how the performance of a task according to an individual cognitive process affects the occurrence of other tasks and then the mental processes to solve them. For this reason the workload has been more and more evaluated as a close loop.[55] The general psychological model identifies three main stages(Johnson, Proctor 2004):

1. Perception
2. decision making/response selection
3. response programming/execution

When the human being has to perform demanding physical and/or mental tasks his cognitive state varies and many parameters such as fatigue, situational awareness, attention, memory and trust come into play.[56] As already mentioned, the objective of the sensor network is to monitor the cognitive state in order to define an objective parameter that can be implemented in an adaptation model of the automated system, the Workload (WL). ATCOs' workload is one of the most important factors determining the maximum airspace occupancy in today's operations (Majumdar & Ochieng, 2007).

Because the controller's job is primarily related to cognition and information processing, this term generally refers to the mental workload needed to accomplish tasks. [1] The most used and validated theories so far assume that tasks require the allocation of the operator's attentional resources for efficient execution, and that the operator workload reflects the overall level of demand for those resources (Wickens, Mavor, and McGee 1997) .

So, What Workload means? There are many possible definitions but at a top-level it can be said:

'Workload is the difference between the required capacity and the available capacity of a human operator for executing tasks demands'

(Gopher & Donchin, 1986; Moray, 1979)

In this chapter will therefore analyze some of the methodologies to evaluate the workload. Many studies have been carried out in order to first of all give a definition of Workload by evaluating what are the factors that influence it. Casali and Wierwille (1983) noted, mental workload must be inferred as it cannot be directly observed [48]. But the technologies available to date have paved the way for a new approach to WL evaluation. Most of the developed methods and models adopted in the past are subjective, they consist on giving a test to the analyzed subject at the end of the experiment or during the session.. However, this approach is not applicable to an adaptive system whose

purpose is to make an online evaluation in real time but it is fundamental to understand the methodology of analysis and which parameters have to be considered.

The task demand can be considered as a fundamental workload evaluation parameter, but an evaluation time window for mental demand has to be defined. This means don't has to be considered only tasks that are performed or are pending at a given time because it may happen that some tasks are not seen or the operator is aware that they have to be performed but are not priority tasks so he/she will perform them later. This last case is the one in which the mental demand is greater because the operator is asked to keep in mind a lot of information and can run into the saturation of his mnemonic capacity leading to the worst scenario, the operator forgets to perform them (excessive WL). In ATM the factors that greatly influence the mental demand are Dynamic complexity factors such as aircraft count, traffic density, and proximity measures between pairs of aircraft. Elements to consider but which have a lower weight instead are for example *aircraft transition factors*, reflecting an aircraft's change in vertical and/or lateral position, speed, and heading.

But not only high levels of workload can lead to situations that reduce the safety, but this is not the case. Low workload is almost as crucial as high workload levels. This is because an excessive reduction of WL can lead to boredom, distractions, reduced vigilance and attention that result in slower response times, poor decision making, loss of situation awareness, change in decision criteria and a failure to detect relevant signals. In the case of Air Traffic Control the operator must perform two main operations which are:

1. Ensure that aircraft under jurisdiction adhere to International Civil Aviation Organization (ICAO) mandated separation standards
2. Ensure that aircraft reach their destinations in an orderly and expeditious manner.

These goals require the ATCO to perform a variety of tasks, including monitoring air traffic, anticipating loss of separation (i.e., conflicts) between aircraft, and intervening to resolve conflicts and minimize disruption to flow. [57, 58]

To evaluate Workload in an interactive and adaptive environment it is necessary to consider some fundamental effects that may occur:

- *difficulty insensitivity*: performance degradation is not directly proportional to the difficulty level of the task. (Kantowitz, Knight 1976).
- *perfect-time sharing*: it occurs when two tasks do not affect each other when performed together (e.g., Schumacher *et al.* 2001).
- *structural alteration*: happens when the performance of one task depends on the response of another.

These effects show that the same task can demand different resources depending on the operating conditions both at a given time t and how previous tasks were performed. These and other findings have led to the development of the concept of multiple resources (Wickens model).

All WL evaluation methods can be grouped into three categories: physiological, performance-based and subjective. [48]

Subjective ratings have been used to obtaining feedback from the operator on his/her self-assessed performance, to evaluate the flight handling qualities (Cooper-Harper scale (1969)) or interface usability. They are the most commonly used because they are cheap, direct and easy to use but the base their evaluation un a quite strong assumption. They assume that operators can reliably rate several aspects of the

tasks. Two of the most popular ones are the *NASA Task Load Index* or NASATLX and the *Subjective Workload Assessment Technique* or SWAT.

Performance-based measures assess mental workload through task performance. The base assumption in these models is that as workload increases, time to complete tasks and errors increase as well while accuracy decreases (Huey, Wickens 1993). Therefore, it is possible to assess workload by tracking performance in a task with different difficulty levels.

However performance can not be affected even if the workload is high thanks to strategy adjustments, the major performance drop occurs usually under too high and too low workload. For this reason, a secondary-task methodology is often used, the emphasis is on the secondary task and the degradation in performance measured in the primary task (loading task technique). The goal of the secondary task is to use up the resources left over by the primary task and they are carried out at the same time as the primaries and the time-sharing effect is measured.

Physiological and psychological measures have also been used to empirically determine operator performance (Kramer 1991; NATO 2004), and include measures of heart activity, brain evoked potential, and gaze patterns. [46]

Some of the most common and used models are now briefly reported to evaluate the workload analysis benchmarks.

4.1.1 NASA-TLX

The NASA requires people to rate the task from low to high on each of six scales: mental demand (MD), physical demand (PD), temporal demand (TD), performance (OP), effort (EF) and frustration (FR) level. A weighting process takes the individual differences between the scales into account to compute an overall workload score. Il punteggio del WL viene determinato attraverso weighting scales and is used extensively in aviation research. The following formula is used to calculate mean workload scores.

$$[MDrating(s) * weight(s) + PDrating(s) * weight(s) + TDrating(s) * weight(s) + OPrating(s) * weight(s) + EFrating(s) * weight(s) + FRrating(s) * weight(s)]$$

These weights are asked questions to the operator as a result of a test so there are many disadvantages: reliance on memory, the task variability effect (people tend to use the whole rating scale, independently of the stimulus range), susceptibility to operators' bias (Johnson, Proctor 2004). Finally, the method has been criticized because it considers emotional aspects (frustration and anxiety) that are very hard to weight. In fact often when the NASA-TLX method is used the weighting steps are skipped and the average or sum of all ratings are taken into account. This method can therefore be a starting point, but since the rating is given after the task is done, it doesn't allow the WL evaluation in real time. Other subjective methods are: the Subjective Workload, the Visual, Auditory, Cognitive, Psychomotor method, the Workload Index method, the Multiple Resource Questionnaire, the Defence Research Agency Workload Scale. [1, 20, 59]

4.1.2 Wickens model

The *Multiple Resources Theory* or commonly known as Wickens model is one of the most influential theories to evaluate performance analyzing the relationship between WL and tasks. It has been validated neuro-physiologically and often used as a guideline by human factors designers to define the tasks of a system. It is developed in three dimensions and the first one is made up of two information processing stages: resources involved with perceptual-cognitive activity are functionally different from those related to response processes (Wickens 1991). For example perceptual-cognitive tasks like reading and voice comprehension can be easily time-shared with simple physical movements like pressing a button. The second dimension is related to perception, cognition and response, which are not shared and are in fact associated with different cerebral hemispheres. It refers to the resources used in processing spatial (manual responses like using a stick or a mouse usually involve spatial codes) and verbal *codes* (ATC radio).[48]

The third dimension consists of the different perceptual (visual or auditory) *modalities*. This level refers to the already verified concept that attending both visual and auditory sources is easier than manage two simultaneous auditory or visual messages. This issue was demonstrated by Latorella (1999) who showed in a simulated flight deck that, during visual tasks, auditory interruptions are more disruptive than visual interruptions. This is because it is easier for the human to store visual than auditory information so when an auditory information is presented the operator tend to give more attention to it, at the expense of other concurrently presented visual stimuli. Modality dimension has been criticised in recent years but such an in-depth analysis is not the goal of this research.

The fourth dimension corresponds to the focal and ambient visual *channels*. Focal vision is required for pattern/object recognition and high acuity perception, whereas ambient vision is involved in orientation and movement perception of oneself. [60, 61]

For this reason, numerous researches have been carried out on the relationship between gaze and workload and an eye tracking device has been inserted in the sensor network.

4.1.3 Sperandio Model

A well known model was developed by Sperandio (1971) who conducted several studies on air traffic controllers with the aim of determining the cognitive processes adopted by operators and the fundamental parameters to define the Workload. For this reason many researches made so far refer to Sperandio's model which is particularly suitable to evaluate the relationship between task demand and workload. In fact the adopted strategies by Air Traffic Control Operators (ATCOs) can vary widely from individual to individual because they are in function of skills, training, experience, fatigue.[62] In function of that it is very useful to have a method to quantify the skills level of the ATCOs, this can be done through simulator tests with the purpose not only to train them but also to determine/monitor their skills level.

Figure 17 shows the Sperandio's model structure(1971). It proposed that ATCo strategy is an intervening variable between task demand and the work achieved. This means that the process cannot be considered as a single feedback loop, but two feedback loops are necessary, one related to the variation of the mental workload according to the applied solving strategies and the second related to the tasks sequence.

The first takes into account the variation of the mental workload according to the adopted strategy to accomplish tasks (primarily resource management), the adopted strategy in turn affects the needed/applied work methods in the forthcoming cycle. The second cycle, on the other hand, takes into account the fact that the tasks cycle are a function of how the the previous task was performed. A time delay has to be considered for those tasks that can be 'paused' for a while as explained above. Sperandio emphasized that it is the change in workload, not the change in task demand, that explains the change in strategy. [13] The latter depends on three factors: individual characteristics(training, motivation, age, health, etc.), task characteristics (i.e., its requirements, including work conditions) and workload levels.

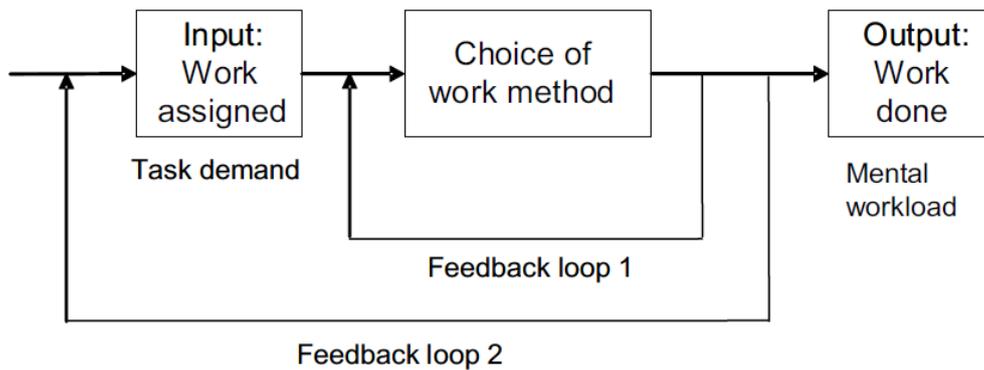


Figure 17: Sperandio's model

Studies conducted by Sperandio show that taskload is the fundamental parameter to define the workload but it is conditioned by the adopted strategy which in turn depends on training and experience. Figure 18 shows how the task demand is directly proportional to the workload but the use of successively more economical strategies S_1 , S_2 and S_3 determines the reduction of the slope of the curve. From this model it can be assumed that the Task demand determines about the $\frac{3}{4}$ of the Workload.

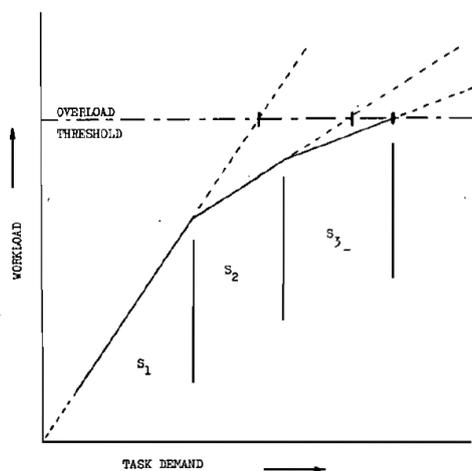


Figure 18: Regulation of Workload by the use of different strategies [62]

4.1.4 Adopted model and considerations

As could be understood there are various methods and approaches to determine the Workload and there are even different definitions of the latter. The hardest part in workload evaluation and assessment is that human factors play a significant role as experience, fatigue, concentration and others purely personal factors related to the specific operator. In general, however, could be assessed that the relationship between task demand and workload depends on the capacity of the controllers to select priorities, manage their cognitive resources, and regulate their own performance.

Analyzing the above mentioned models and other studies ([1, 13, 15, 63-65]) it can therefore be concluded that they consider Taskload and strategy development as fundamental parameters for the WL evaluation. So the adopted method to evaluate the workload in our experiments assess the cognitive state with a task-base measure. A simple definition like Taskload = Workload is incorrect because neglect the component related to the strategy, so it can say that the 70%/80% of the Workload is defined by Taskload. It is expected that the greater the skills and experience of the operator, the more stable and repeatable is the physiological response monitored by the sensor network. The experiments conducted in our laboratory have demonstrated this assumption because, as will be explained in the following chapters, an individual with training and experience in air traffic shows a trend of AU much more similar to the Taskload trend than a neophyte one.

Our experiments were conducted on six individuals and they show that an operator who has air traffic training and therefore adopts well-defined traffic management strategies also has a greater awareness of how to monitor the situation and manage it. This obviously reflects on the physiological response, between the six subjects analysed 'Alex', the one with more experience, is taken as a reference. His facial expressions response follow a given Taskload trend with a much greater repetitiveness than for example 'Nicha' who has less experience and shows a more random trend of her biological parameters.

In conclusion, consistent with the models and studies available in literature, in this research, we adopted a simplified one-loop model in which the benchmark is the Taskload, in particular the secondary tasks. The adopted model takes as reference the Sperandio one and is shown in Figure 19.

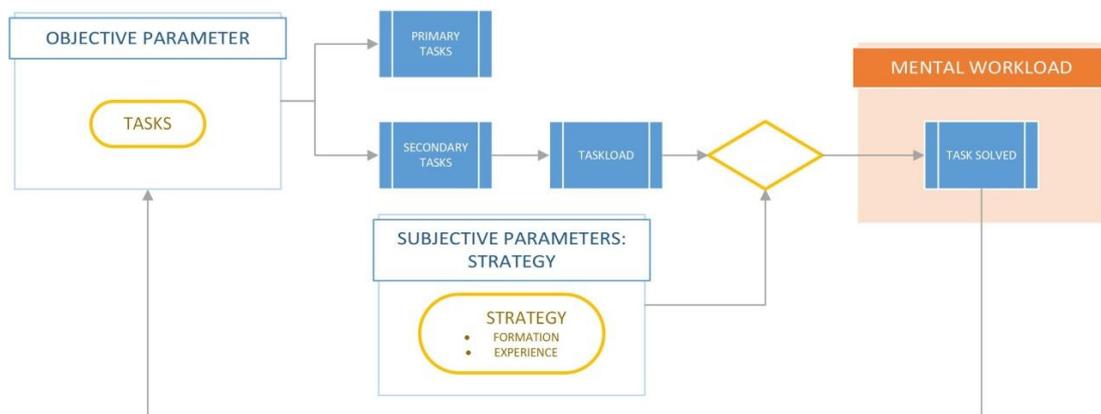


Figure 19: WL objective assessment model

In *Experiments* chapter will be explained how these secondary tasks are assessed.

5 The Software: Open Face

Face representations need to be resilient to intrapersonal image variations such as age, make-up, styling but capture interpersonal image variations between different people. For this reason, the first feature that a software for the FME recognition and analysis must have is a good ability to distinguish the characteristic features of the subject from those due to age.

Facial Expression detection software usually work developing three main steps:

- **Face detection:** The position of a face is found in a video frame or image.
- **Facial landmark detection and registration:** Within the detected face, facial landmarks such as eyes and eye corners, brows, mouth corners, the nose tip etc. are detected. It's like an invisible virtual mesh that is put onto the face of the respondent to model the main feature through geometrical shapes defined by points. Whenever the respondent's face moves or changes expressions, the face model adapts and follows instantaneously.
- **Facial expression and emotion classification:** Once the simplified face model is available, position and orientation information of all the key features is fed as input into classification algorithms which translate the features into Action Unit codes.

The adopted software is Open Face, an open-source tool that allows analyzing the MFE by dividing them into Action Units (AU). Many tools for AU recognition are available but most of them have prohibitive cost, unknown algorithms, and often unknown data training. [66, 67] Furthermore, considering that the ultimate goal of this research is to be able to apply the study of MFE in various aerospace environments some tools have been avoided because they can be used only on a single machine. The software is cross-platform and has been tested on Windows, Ubuntu and Mac OS X. Finally, and most importantly, commercial products may be discontinued resulting in poor data repeatability due to lack of product transparency (this is illustrated by the recent unavailability of FACET). Today's top-performing face recognition techniques are based on convolutional feed-forward neural networks and Facebook's DeepFace and Google's FaceNet([68],[69]) are the best available on the market but belong to the category just mentioned because they have a prohibitive cost. It was therefore chosen Open Face because it offers levels of accuracy very similar this two private state-of-the-art tools but is open source and the code (C++) is easily accessible and editable.

Given an input image with multiple faces, face recognition systems typically first run face detection to isolate the faces. Each face is pre-processed and then a low-dimensional representation (or embedding) is obtained. A low-dimensional representation is important for efficient classification to avoid too long processing times that would not allow real time analysis.

A simple Conditional Local Neural Fields (CLNF) is used for landmark detection and tracking. CLNF is mainly composed by two elements: Point Distribution Model (PDM) which captures 68 landmark shape variations; patch experts which capture local appearance variations of each landmark.[70]

A Three layer Convolutional Neural Network(CNN) is used as a validation step to predict the expected landmark detection error.

Figure 20 shows the operational block diagram of Open Face.

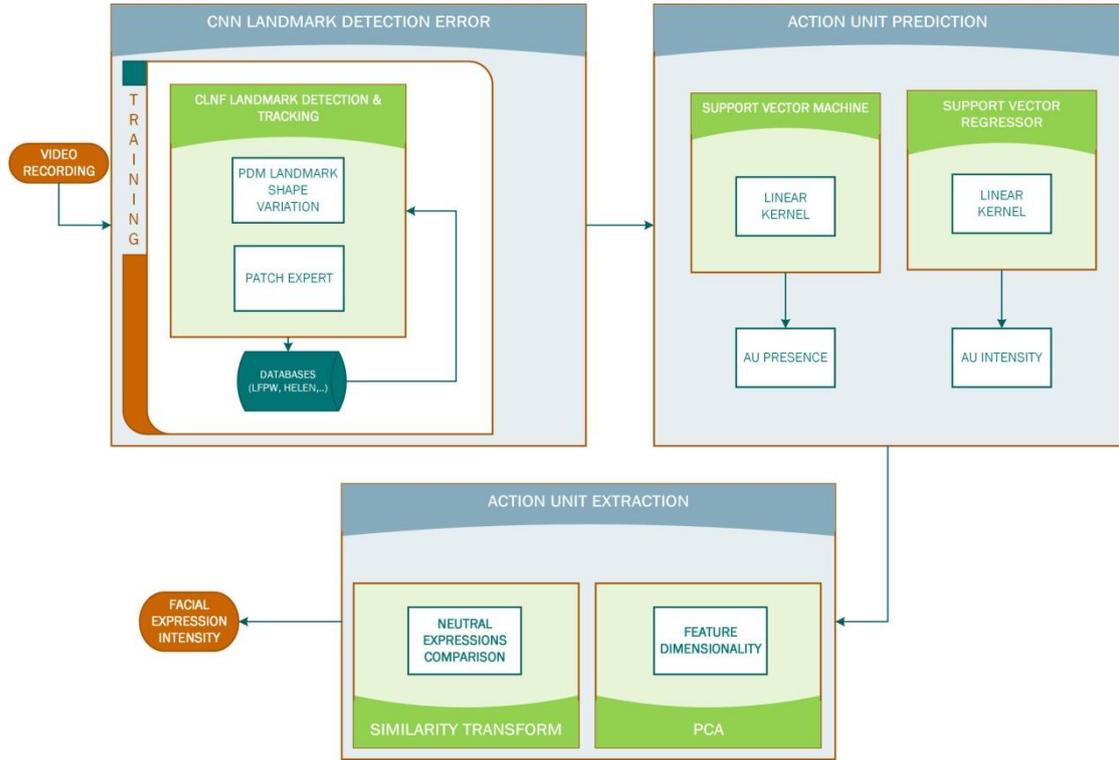


Figure 20: Open Face operational block diagram

The CNN is needed especially to track a face over a long period of time and to follow it if the monitored person move in front of the camera or goes out of sight. The PDM is trained on two data set called *Labeled Face Parts in the Wild* (LFPW,[71]) and Helen and it models the location of facial feature points in the image using 34 non-rigid and 6 rigid shape parameters. 28 sets of patch experts are obtained training them for different views. This allows to identify the main points despite variations in pose, lighting, expression, hairstyle, subject age, subject ethnicity, partial occlusion of the face, camera type, image compression, resolution, and focus.

The Constrained Local Neural Field model is now briefly presented.

5.1 Constrained Local Neural Field

The model is based on Local Neural Field patch expert which learns the nonlinearities and spatial relationships between pixel values and the probability of landmark alignment. The LNF can capture the relationship between pixels whether they're near or far from each other and using a neural network layer can capture complex non-linear relationships between pixel values and the output. The patch expert (also called local detectors), has to capture the *spatial similarity* gk (pixels nearby should have similar alignment probabilities) and the sparsity lk obtaining the peaks number in the evaluated area. The sparsity is forced to have only one peak in the area in which the patch expert is evaluated. In Figure 1 the structure of the LNF is represented, consisting of observed input variables $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, $\mathbf{x}_i \in R^m$ is a vectorised pixel intensity in the patch expert support region (e.g. $m = 121$ for an 11×11 support region). $\mathbf{Y} = \{y_1, y_2, \dots, y_n\}$ is a set of output variables that has to be predicted $y_i \in R$ expressing the probability that a patch is aligned, and n is the area in which the patch expert is evaluated.

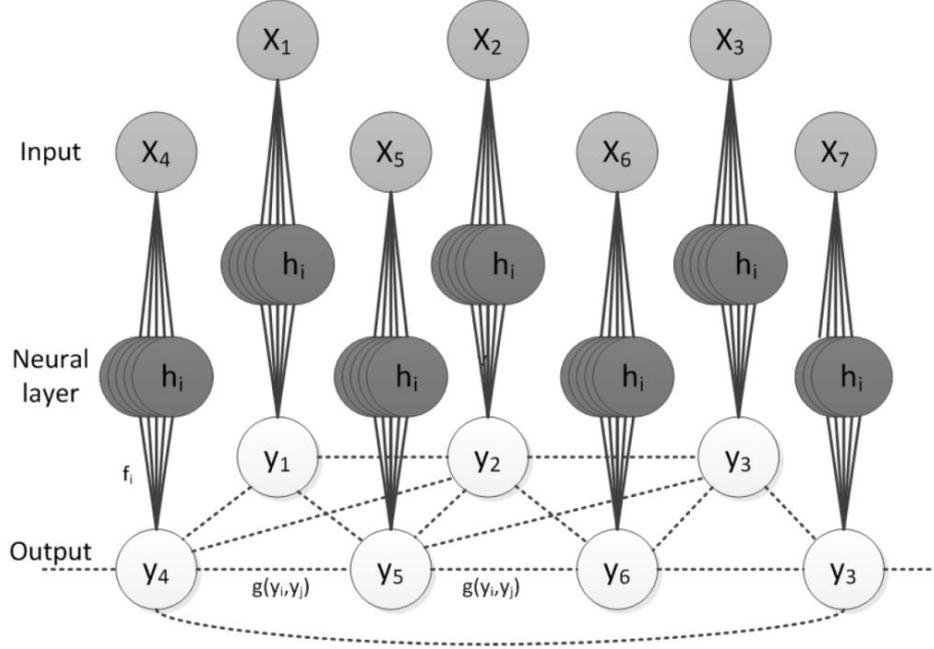


Figure 21: LNF structure

The probability density:

$$P(y|X) = \frac{\exp(\Psi)}{\int_{-\infty}^{+\infty} \exp \Psi \, dy} \quad (5.1)$$

Affect the model that is a conditional probability distribution. The potential function is:

$$\Psi = \sum_i \sum_{k=1}^{K1} \alpha_k f_k(y_i, X, \theta_k) + \sum_{i,j} \sum_{k=1}^{K2} \beta_k g_k(y_i, y_j) + \sum_{i,j} \sum_{k=1}^{K3} \gamma_k l_k(y_i, y_j) \quad (5.2)$$

Where α , β , θ e γ are the vertex weight, they are learned and used for inference during testing. Three potentials types are defined: vertex features f_k and edge features g_k, l_k :

$$f_k(y_i, X, \theta_k) = -(y_i - h(\theta_k, x_i))^2 \quad (5.3)$$

$$h(\theta, x) = \frac{1}{1 + e^{-\theta^T x}} \quad (5.4)$$

$$g_k(y_i, y_j) = -\frac{1}{2} S_{i,j}^{g_k} (y_i - y_j)^2 \quad (5.5)$$

$$l_k(y_i, y_j) = -\frac{1}{2} S_{i,j}^{l_k} (y_i + y_j)^2 \quad (5.6)$$

The neural network is therefore composed of one layer, the input \mathbf{x}_i are connected to the scalar output \mathbf{y}_i through a neural layer (θ) and the vertex weights.

The terms of the equations are now described:

- θ_k is the weight vector for a particular neuron k , it can be thought of as a set of convolution kernels that are applied to an area of interest.
- The vertex weight α_k for vertex feature \mathbf{f}_k represents the reliability of the k th neuron (convolution kernel).
- Similarities between observations \mathbf{y}_i and \mathbf{y}_j are represented by the edge feature g_k .
- Sparsity constraint between observations \mathbf{y}_i and \mathbf{y}_j are represented by the edge features \mathbf{l}_k .
- The neighbourhood measure $\mathbf{S}(\mathbf{g}_k)$ points out where the smoothness is to be enforced ($\mathbf{S}(\mathbf{g}_k) = 1$ if i and j are direct (horizontal/vertical), otherwise 0)
- The neighbourhood measure $\mathbf{S}(\mathbf{l}_k)$ points out the regions where sparsity should be enforced (1 only when two nodes i and j are between 4 and 6 edges apart).

Once the neural network architecture is defined and implemented it has to be trained to estimate the parameters $\{\alpha, \beta, \gamma, \Theta\}$ using a sequence of input \mathbf{x}_i and known outputs variables \mathbf{y}_i from a database. The training goal is to find the $\{\alpha, \beta, \gamma, \Theta\}$ values that maximise the conditional log-likelihood of LNF on the training sequences:

$$L(\alpha, \beta, \gamma, \theta) = \sum_{q=1}^M \log P(\mathbf{y}^q | \mathbf{x}^q) \quad (5.7)$$

$$(\bar{\alpha}, \bar{\beta}, \bar{\gamma}, \bar{\theta}) = \operatorname{argmax}_{\alpha, \beta, \gamma, \theta} (L(\alpha, \beta, \gamma, \theta)) \quad (5.8)$$

After a few mathematical passages, converting equation 8 into a multivariate Gaussian form and using the partial derivatives of the log $P(\mathbf{y} | \mathbf{X})$ the training samples can be obtained sampling an image at various locations.

It is useful to train the model not only with a dataset of different faces but also with different views in fact the training has been done in separate set of patch experts for seven views and four scales obtaining 28 sets. This allow the software to track faces with out of plane motion and to model self-occlusion caused by head rotation. It is therefore important to note that every facial expression analysis tool uses a different database. This is the reason why giving the same inputs to two different software slightly different outputs can be obtained and it is essential to know the reference databases of the adopted software.

In addition, another key parameter that greatly affects the landmark detection and tracking is the resolution and the multi-scale patch experts allow to enhance the precision and the software applications. *Dlib library* face detector is used to initialize the CLNF model for the 68 facial landmarks. Figure 22 present the Open Face landmark detection.

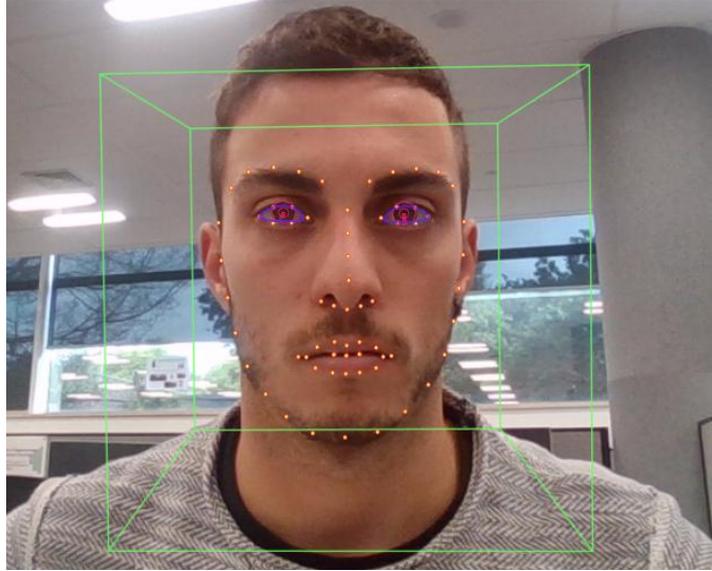


Figure 22: Landmark Detection

5.2 Action Unit detection

Open Face AU detection is based on a recent state-of-the-art AU recognition framework([72],[73]). The software also allows *head pose* and *eye gaze estimation* but for our applications we have not used this software to detect these parameters but dedicated sensors integrated in the network which have a greater accuracy; they will be presented in the following chapters.

People has their own facial expressiveness and of course their own facial geometry so it is essential that the software is calibrated to the monitored person. The two fundamental parameters are the intensity of the AU and how much the software is able to detect its presence. This fundamental step is done through a training on combined datasets, one for the presence and another for the intensity of the AU, using the distance to the hyperplane of the trained Support Vector Machine (SVM) model as a feature for an SVR regressor. This results in a single, with better performance, predictor based on two datasets which uses a linear kernel SVM for the AU presence prediction and a linear kernel Support Vector Regressor(SVR) for AU intensity. Kernel methods map the data into higher dimensional spaces in the hope that in this higher-dimensional space the data could become more easily separated or better structured. The Linear Kernel is the simplest kernel function. It is given by the inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ plus an optional constant c .

$$k(x, y) = x^T y + c \quad (5.9)$$

For further information on the adopted mathematical models, please refer to [43, 72-74].

Finally, a 112 x 112 pixel image of the face with 45 pixel interpupillary distance is obtained from the extracted facial features. Once the landmarks are detected they are compared to the frontal landmarks of a neutral expression using a similarity transform and the feature dimensionality is reduced using a Principal Component Analysis(PCA) model trained on a FE dataset obtaining the final software output, the AU intensity. The entire process is completely automatic and it doesn't need calibration. The automatic coding is accomplished much more objectively than

manual coding where humans, particularly novice coders, tend to interpret the activation of an Action Unit in concert with other Action Units, which significantly alters the results.

Now it is concretely presented how the software has been used and what are the characteristic parameters.

5.3 Software Interface

Once the software has been opened, the measurement can be made by pressing the start/stop logging button. The face is identified through a Facial Landmark Detection that identifies the parts of the face through the differences between one pixel and the other so the higher is the camera resolution the better will be the landmark detection. This system centring method uses both eye gaze and landmarks so the more the gaze is turned towards the camera the higher is the accuracy of the facial feature detection. As already mentioned The Facial Action Unit System (FACS) consists of 64 Action but not all of them can be detected by the software.

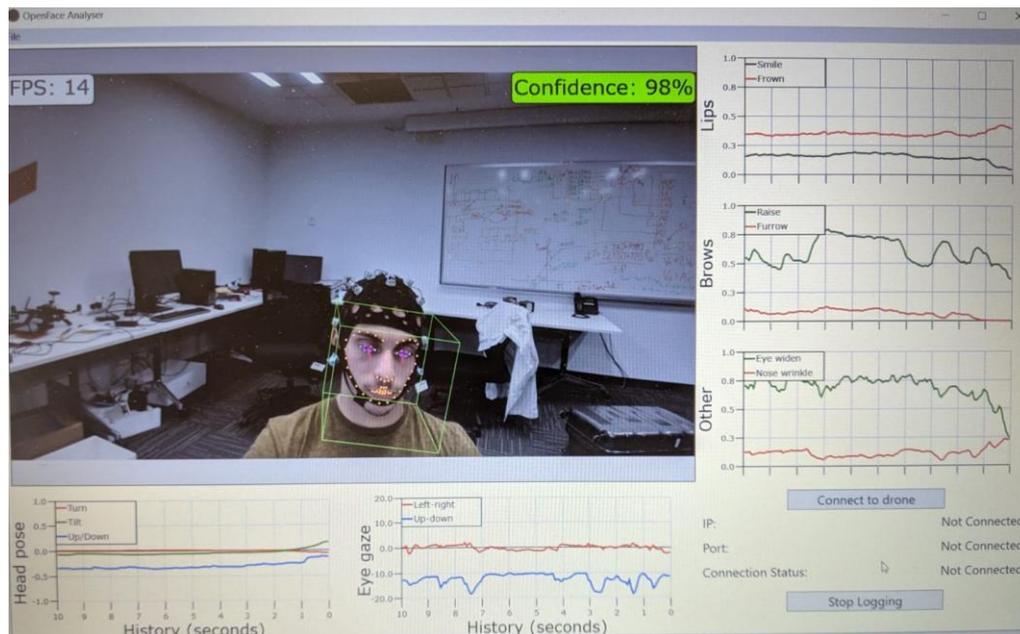


Figure 23: Software interface

Figure 24 shows the AU intensity, it can vary from 0 to about 5 if the AU is captured. Instead for those not captured the software generates by default the value -999. The column 'SYSTIM' represents the time of the computer clock, useful for the synchronization with other sensors especially because the used webcam records with non constant FPS so the time correlation with other sensors could be harder without a clock counting.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
LOG STARTED AT 2019-12-06_04:35:21																
SYSTEM	CONFIDENCE	FPS	SMILE	FROWN	BROWUP	BROWDOWN	EYEWIDEN	NOSEWIDEN	AU1	AU2	AU3	AU4	AU5	AU6	AU7	AU8
35:21.1	0.817	16.3	0.0952	0.2604	0.1337	0.0357	0	0.3215	0.9684	0.2874	-999	0.2408	0	0.2372	-999	-999
35:21.1	0.983	18.4	0.0954	0.2545	0.1345	0.0356	0	0.3173	1.1482	0.2146	-999	0.1763	0	0	-999	-999

R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
AU9	AU10	AU11	AU12	AU13	AU14	AU15	AU16	AU17	AU18	AU19	AU20	AU21	AU22	AU23	AU24	AU25
1.3803	-999	-999	0.7018	-999	-999	0.9142	-999	2.1605	-999	-999	-999	-999	-999	-999	-999	0.4694
1.2295	-999	-999	0.7891	-999	-999	0.8112	-999	2.0796	-999	-999	-999	-999	-999	-999	-999	0.4575

Figure 24: Open Face outputs

The ‘*CONFIDENCE*’ can vary from 0 to 98.3 and usually is higher of 70%, if the monitored subject makes abrupt movements of the head or excessive rotations, the confidence can drop rapidly so that the data in that interval of time are no longer acceptable. Data of the main facial features are also provided from column 4 to column 9 if a raw analysis is needed. Experience has shown that usually the *confidence* is 98.3% during the 93/95% of the recording time during testing but if the subject wears glasses the medium *confidence* can drop to 88%.

According to what has been said so far the two main parameters that must be taken into account when using the software are:

- **Framerate:** The camera should have a stable framerate of 10 fps or higher
- **Resolution:** The resolution of the video should be at least 640 x 480 pixels.

Some cameras might dynamically adjust the frame number dependent on lighting conditions (they reduce the frame rate to compensate for poor lighting, for example).

The original software does not allow to start and end the recording at will so a button had been added to better manage the recording sessions during the experiments. When the recording is finished, the output of the processed data is saved as CSV so that they can be analyzed in post-processing. The software outputs have been evaluated with *Matlab 2019a* remembering that the ultimate goal is the evaluate FE in real time so the developed script, as will be presented in the following chapters, have always had as focus the implementation in real time in the sensor network.

6 Performance evaluation

The performance analysis is a fundamental aspect in order to be able to correctly evaluate the experimental data. Obviously, in the biometrical sensing, there are many disturbances that could affect sensors. [75] Here are reported those that take into account the most incisive variables: blink rate, head rotation and shifting. Each evaluation is based on three experiments on the same subject.

Figure 25 shows the adopted camera, a off-the-shelf webcam Logitech C270.



Figure 25: Webcam Logitech C270

Open Face does not require special types of cameras so a mid-level off-the-shelf camera that meets the software requirements was used.

Technical specifications are reported in Table 3.

Max Digital Video Resolution	1280 x 720 / 30 fps
Computer Interface	USB 2.0, 4 pin USB Type A
Focus type	Fixed focus
Lens technology	standard
FoV	60°
SYSTEM REQUIREMENTS DETAILS	
OS Required	Microsoft Windows 7, Microsoft Windows Vista, Microsoft Windows XP SP2 or later
Processor Speed	1 Hz
Min RAM Size	512 MB
Min Hard Drive Space	200 MB

Table 3: Technical specifications

Open Face does not provide any specification on how the confidence level is computed so different experiments had been conducted to evaluate the parameters that affect the confidence.

The software uses glance as the main parameter of face centering so three experiments have been conducted to verify if a disturb such as a high blink ratio reduces the confidence of the software. Table 4 shows that the confidence remains equal to the maximum value in all conditions so the CNN used by the software is not affected by this disturb. Moreover the lighting conditions between one experiment and the other have been changed causing a different FPS of the camera but also in this case Open Face maintains the maximum confidence.

Blinkrate	Experiment 1	Experiment 2	Experiment 3
LOW	C: 98.3	C: 98.3	C: 98.3
	FPS: 22.2	FPS: 24.9	FPS: 23.8
NORMAL	C: 98.3	C: 98.3	C: 98.3
	FPS: 22.1	FPS: 25.5	FPS:24.2
HIGH	C: 98.3	C: 98.3	C: 98.3
	FPS: 22.6	FPS: 24.4	FPS:23.8

Table 4: C: confidence, FPS: Frame Per Second

A parameter that affects the confidence are glasses, in fact if the subject wears glasses the confidence drops from 7% to 10% .

Another important aspect is the environment, during a general video recording many parameters may vary such as room lighting and therefore corrective actions on gamma, contrast and dynamic range could be needed. In case of high illumination there is the risk of losing details of the recorded face due to the dynamic range but, as explained in the previous chapter, Open Face adopts a CNN that compares the brightness of the pixels and evaluates the Landmark detection in black and white just to avoid dynamic range problems. Different scenarios have been evaluated to understand the variability of the software confidence as the lighting level and the position of the light source vary. In normal lighting conditions the confidence is 98%, the worst case is when the light source is positioned under the face of the subject and leads to a confidence reduction of 7% (91%). It can therefore be concluded that it is not necessary to edit the video recordings that are analyzed by the Software because the confidence reduction is acceptable and above all it occurs in a non-real condition because the light source in the room is from the top and provides a constant brightness.

6.1 Backwards and forwards movement

In this chapter forward (AU57) and backward(AU58) head movements are evaluated, these movements are made remaining within the Field of View of the camera (FoV) which is geometrically defined by a vertical and a horizontal component:

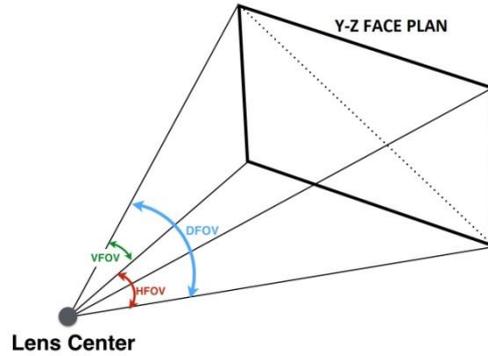


Figure 26: Geometric method for determining camera's FoV

Horizontal FoV (HFOV) and Vertical FoV (VFOV) are calculated using the equations given below.

$$HFOV = 2 \tan^{-1} \left(\frac{l_h}{2d_h} \right) \quad (6.1)$$

$$VFOV = 2 \tan^{-1} \left(\frac{l_v}{d_v} \right) \quad (6.2)$$

l_h and l_v are respectively the horizontal and vertical dimension of the face. These parameters are the dimension of the face along y-axis and z-axis in the adopted reference system as reported in Figure 27. Table 5 shows the typical head anthropometry values for adult men and women, the average value is adopted in the previous formula.

	Men	Women	Mean value
l_h (cm)	15.2	14.4	14.8
l_v (cm)	20.9	19.8	20.35

Table 5: Typical Head Anthropometry

Considering the 60° FoV of the camera the minimum distances $d_h = 12.8 \text{ cm}$ and $d_v = 17.6 \text{ cm}$ are obtained. These values allow to evaluate the uncertainty in the measured FOV through propagation of uncertainty given by the equation:

$$\sigma_{FOV} = \frac{\sqrt{\sigma_l^2 + \left(\frac{l}{d}\right)^2 \sigma_d^2 - \frac{2l}{d} \sigma_{ld}}}{d \left[1 + \left(\frac{l}{2d}\right)^2 \right]} \quad (6.3)$$

Where:

- l : dimension of the face.
- d : distance between the camera and the face.
- σ_l : uncertainty associated with the measurement of the face from anthropometry databases, $\sigma_l = 1.2 \text{ cm}$.
- σ_d : uncertainty associated with the measurement of the distance, $\sigma_d = 0.1 \text{ cm}$.
- σ_{ld} : covariance between the measured distance and length, assumed equal to zero to have a conservative value of σ_{FOV} .

A $\sigma_{FOVh} = 0.07^\circ$ and $\sigma_{FOVv} = 0.05^\circ$ were obtained.

Therefore $d = 20\text{cm}$ is assumed as a conservative value as minimum distance from the camera. This value represents the operating limit of the camera but it has never been reached during the experiments because the closest distance the operator can reach to see the screen in focus even if He/She focuses only one point and loses the global vision of the monitor is 35cm .

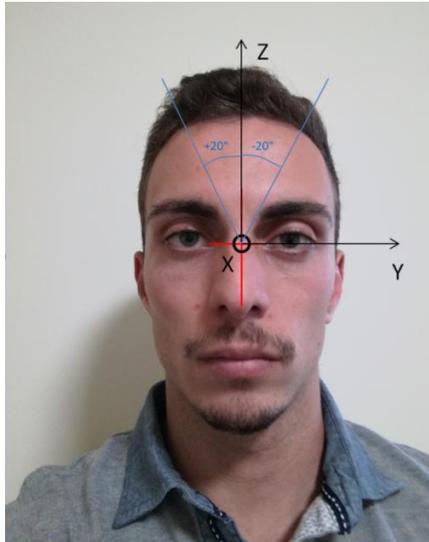


Figure 27: Adopted Reference System

Three experiments have been conducted. During the first 5 seconds of the experiment the subject remained neutral and relaxed to give the software time to calibrate at a distance of 110 cm . Then the operator approached the face to the camera progressively until it reached the closest distance that allowed him to get a good view of the screen (35cm) and remained in that position for 5 seconds.

The AUs trend had been approximated by a second-degree polynomial to clearly show the trend over time. Figure 28 shows the AUs trend for different face areas.

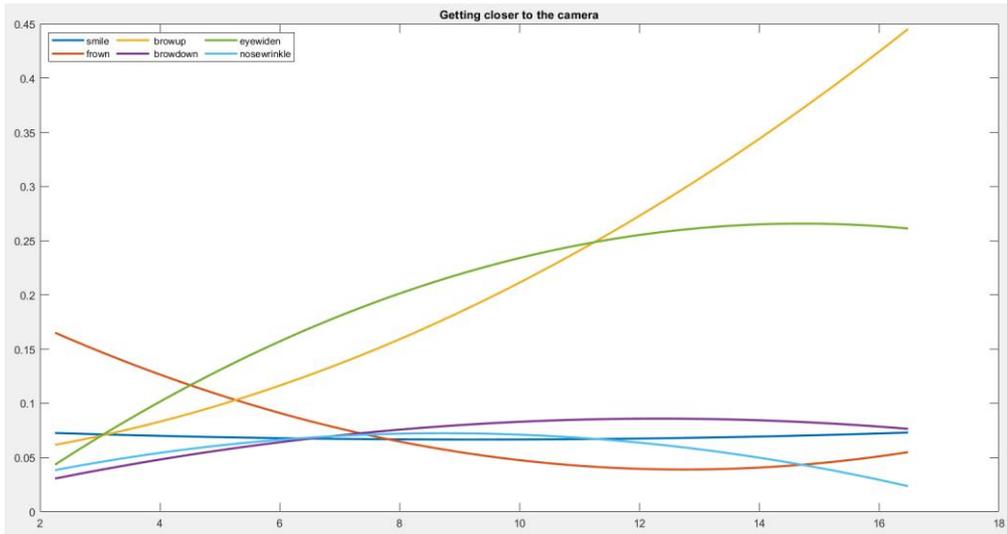


Figure 28: Forward, geometrical parameter

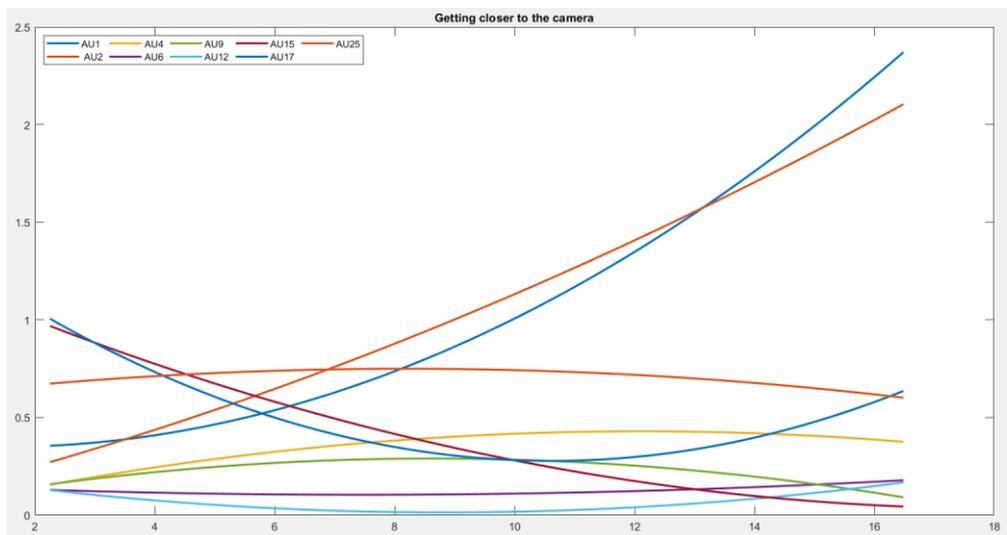


Figure 29: Forward, Action Unit

It can be therefore concluded that the parameters browup, eyewiden AU1,AU2,AU15 are strongly influenced by the distance from the camera. When the operator exceeds about 45 cm of distance from the camera the values undergo a significant increase. This phenomenon, however, is not a problem in operational life because the confidence always remains equal to the maximum value and the operator can make movements of the head but will hardly make a movement along the x-axis of 60 cm, typical movements can reach a maximum of 20cm. The same considerations can be applied on the ‘moving away from the camera’ experiment. In this case, the operator was initially at a distance of about 35 cm from the camera and then moved away up to about 130 cm. Again, he was neutral and relaxed for 5 seconds at the beginning and 5 seconds at the end. Figure 30 and Figure 31 show that moving away from the camera involves a considerable variation in the parameters, but in real-life conditions this occurs only in sporadic cases.

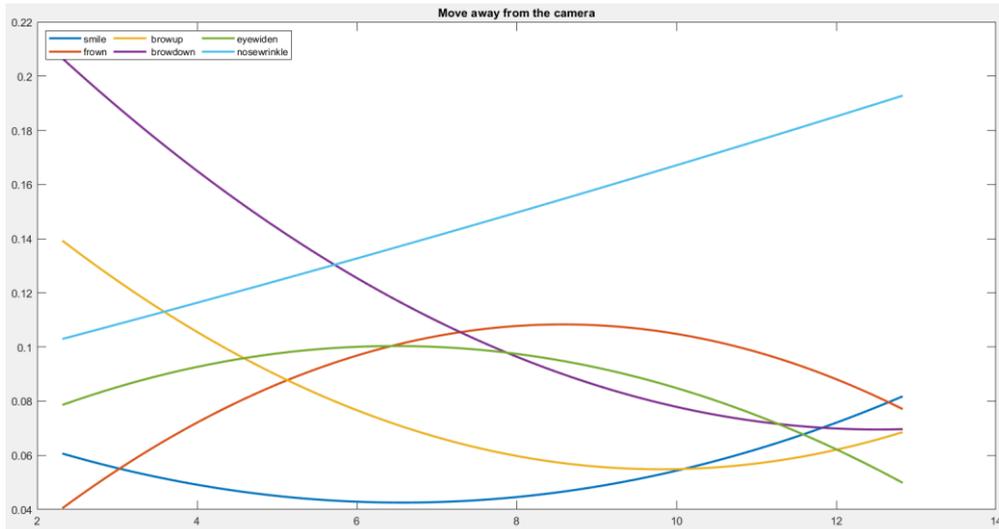


Figure 30: Backward, geometrical parameter

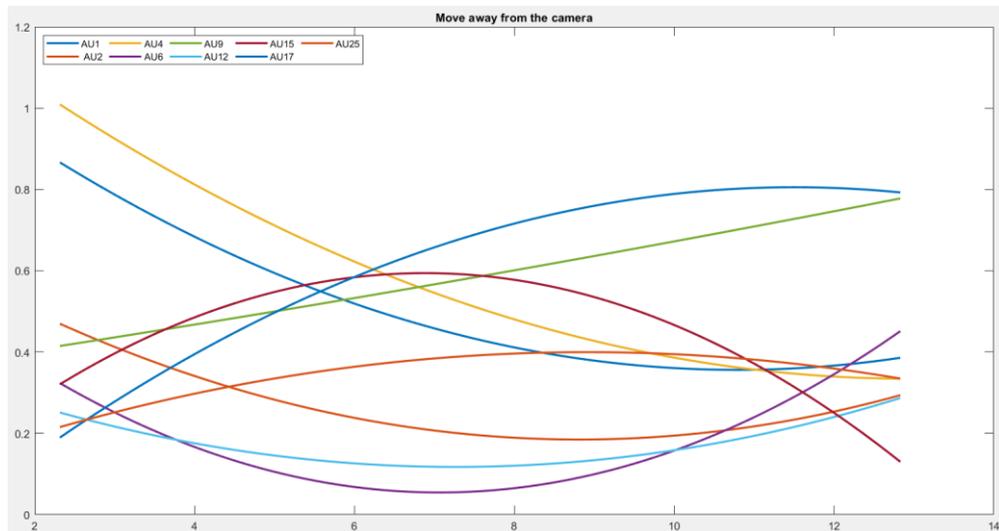


Figure 31: Backward, Action Unit

Nose-wrinkle, AU2, AU6, AU12 and AU15 are characterized by a peak (positive or negative) at about 80/90 cm from the camera. This phenomenon is due to the geometric shapes adopted by the software to model the features of the face. Also in this case the confidence remains constant at the maximum value. It can be therefore deduced that when this system will be integrated in the sensor network, an outlier rejection is needed. If the distance of the operator is between 45cm and 100 cm from the camera the results are reliable, instead if the operators is outside this range the FE parameters has to be rejected. Finally the distance thresholds must be adjusted according to the adopted workstation (distance camera-operator, monitor size, chair height,..)

6.2 Rotation

6.2.1 X axis

The phenomenon of rotation of the head around the X axis (AU55 – AU56) is an unconscious attitude that the humans adopt when they are focused on something and, especially when the auditory apparatus is involved, they tend to rotate the head to the right (positive rotation around the X axis). In this section this phenomenon is analyzed to evaluate potential interferences with the precision of the software. In these experiments as well the operator has been motionless for five seconds before starting to rotate the head gradually from 0 degree (nose aligned with the Z axis) and reached the maximum rotation (about 60 degree referred to Z) and remained in that position for 5 seconds. A second experiment was carried out starting from the head at maximum rotation(+60°) and then moving from 0 degree and bringing it to maximum extension to the opposite side(-60°). In these experiments the confidence is still constant and equal to the maximum value so there are no uncertainties of measurement but as showed in Figure 32, there is a considerable variation of some parameters. Therefore it is possible to define a band of ± 20 degree (marked in the figure) in which FEs detection is affected by a reasonable variation. This assumption is not stringent because it is very rare, even in situations of extreme concentration, that the head is rotated more than 20 degrees, typical values are around 5°/10°.

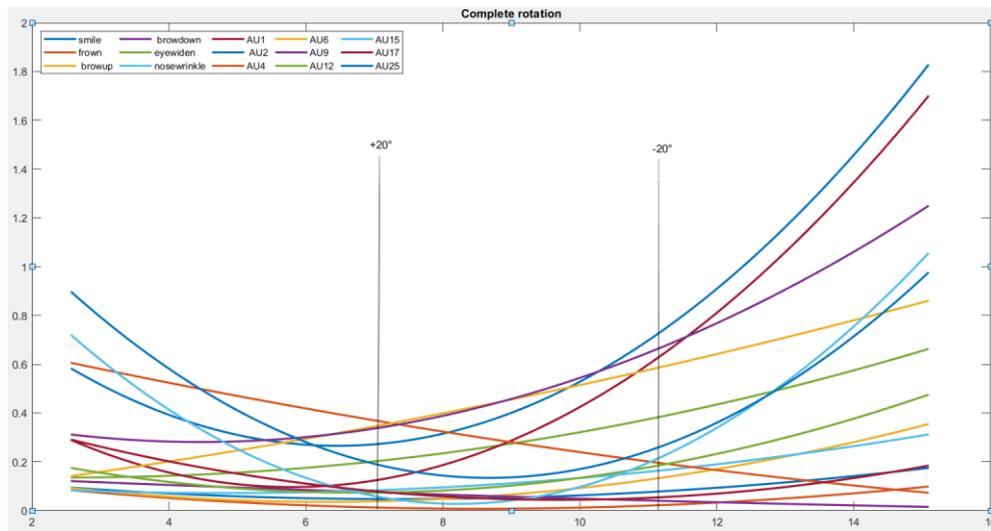


Figure 32: Dx to Sx complete rotation

6.2.2 Z axis

The head rotations around the Z axis (AU51-AU52) involve the greatest error of evaluation because the software determines the geometry of the face using the pupils as a centering system so a rotation of the head around Z leads to a distortion of these geometries and a darkening of a part of the face, which never happens in the cases analyzed so far. In this case the threshold is not referred to value of FE but is related to the confidence of the software. The rotation of the head involves a reduction of the confidence on the data provided by OF for angles higher than about 35°. Two cases may occur, the first is a head rotation keeping the gaze at the monitor, in this case the software captures the pupil, it centers itself and the confidence remains equal to

the maximum value up to 50/60° of rotation. If, on the other hand, the operator looks away from the monitor, a reduction in confidence already appears at 35°. Both cases, however, do not show particular criticality because usually an operator does not rotate more than 15°/20° even if two or three paired screen has to be monitored. If there are more monitor and therefore the operator often has to rotate the head widely around the Z axis, it is possible to install a system with more than one camera selecting the one that shows more confidence.

6.3 Blink Rate

Open Face is developed to don't be affected by the blink ratio, however, the face muscles used to blink are numerous and can trigger disturbances to other areas of the face, so an excessive blink rate can disrupt the AU monitoring.

To analyze the influence of the blink rate (BR) on the output values of the software, three experiments had been conducted considering three different cases: low, normal and high blink rate. In these experiments the monitored subject tried to remain the most impassive to the camera by simply fixing the camera and varying the blink rate. The *NormalBR* condition is the resting condition with typical BR values(10blink/minute), without external disturbances and has been taken as a reference. The low blink ratio is evaluated using the gived equation below:

$$\Delta FE_{LBRi} = mean(FE_{NormalBRi}) - mean(FE_{LowBRi}) \quad (6.4)$$

In Figure 33, Figure 34 and Figure 35 this formula is plotted, each bar represents the average variation over time of each FE(ΔFE_{BRi}) whilst the stem graph represents the maximum value reached by this FE in time. If the subject concentrates on not blinking this would induce unnatural contractions especially of the lower eyelid so the subject was asked just to focus on the camera to experience attention tunnelling. As it can be seen from figures, the comparison between LowBR and NormalBR indicates that the AU 15 and AU17 are always influenced by the blink rate. This phenomenon is not a measurement error but a verification of the correctness and precision of the software because when the subject focuses on a point reducing the BR a lightly contraction the upper and lower eyelids appear and this induce as an involuntary effect, due to facial anatomy, a slight contraction of the muscles of the corners of the lips and an upward pushing of the chin that is captured by the software.

It is demonstrated that a reduced blink rate value indicates an increase in attention that can lead to attention tunnelling and this experiment denote that mouth AUs like AU15, AU17 and AU25 could be evaluated to provide a new parameter to monitor attention tunnelling.

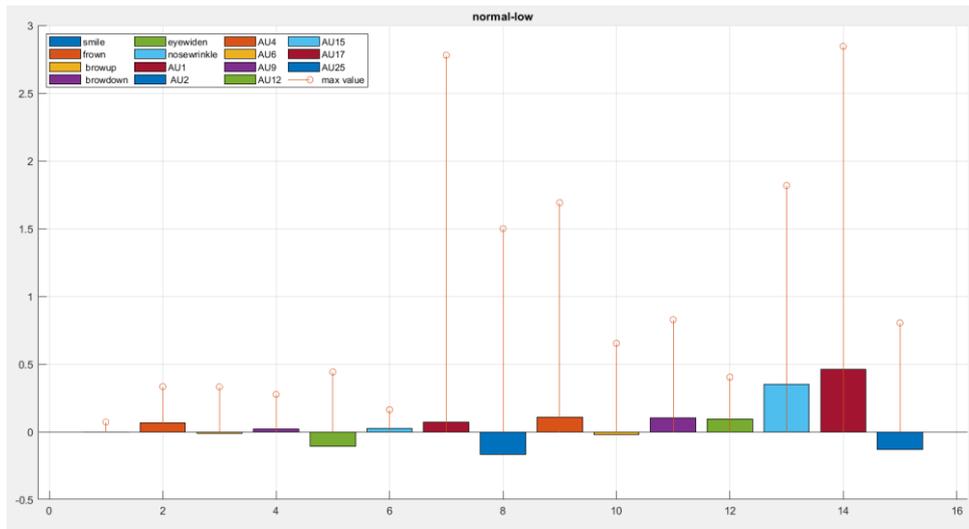


Figure 33: Exp1. Low to Normal BR comparison

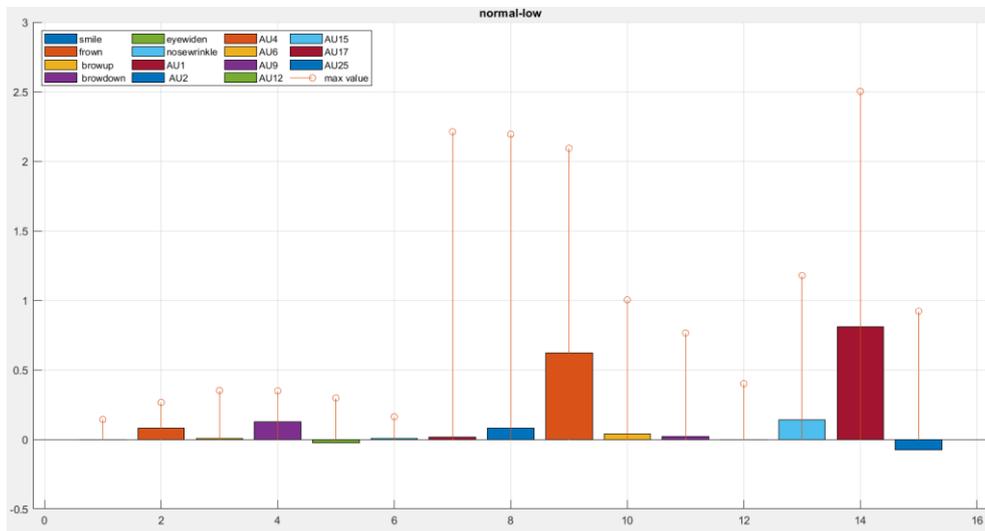


Figure 34: Exp2. Low to Normal BR comparison

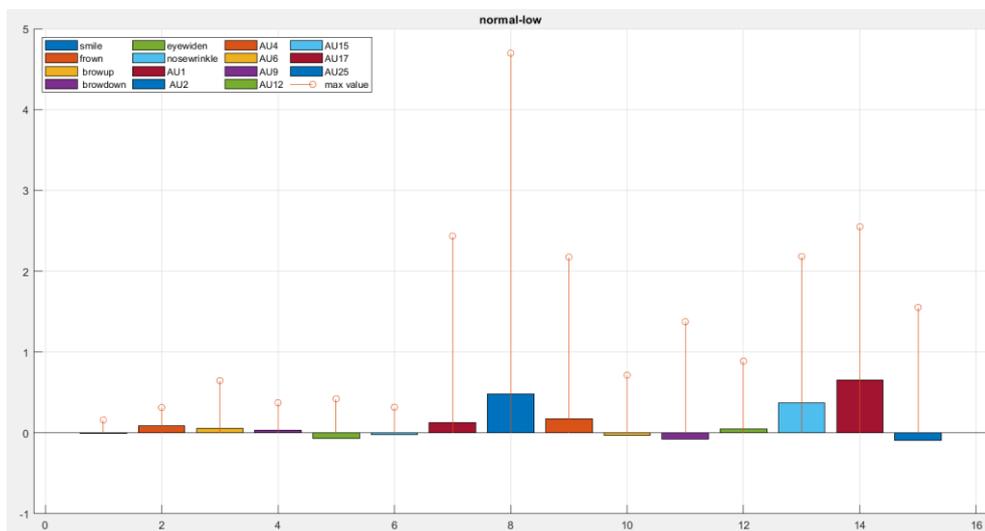


Figure 35: Exp3. Low to Normal BR comparison

AU2 and AU4 show different magnitude in the three experiments because they are related to the eyebrow but it can be deduced that they are not affected by the BR. So the change in the average value of these AUs during the test session can be due to a variation of concentration and tiredness levels. It can be therefore concluded that the only FE that are influenced by BR are AU15 and AU17 whilst the other AU undergo a minimal variation that is not attributable to the variation of BR.

The table below shows $\% \Delta FE_{BRi}$ for AU15 and AU17:

$$\% \Delta FE_{LBR15} = \frac{\max(AU15)}{\Delta FE_{BR15}} * 100 \quad (6.5)$$

That is the percentage change of AU15 and AU17 from their maximum value.

%NormalBR-LowBR	Exp1	Exp2	Exp3
AU15	38.07	31.93	31.85
AU17	48.69	21.65	24.8

Table 6: AU15/17 Normal to Low BR

In the first experiment the BR influences up to 49% the value of AU captured by the software. So to be able to use these AU in the monitoring of the cognitive state is absolutely necessary an integration with the eye-tracker.

The monitoring of these AUs can be embedded in the sensor network as an eye tracking data verification. If the eye sensor software shows an increase in the blink rate and also AU15 and AU17 show an increasing value, an attention tunnelling event could be confirmed.

The *HighBR* is now evaluated, these experiments are conducted in the same way as the Low BR and in all of the the confidence is constantly at its maximum value of 98.3%:

$$\Delta FE_{HBRi} = \text{mean}(FE_{NormalBRI}) - \text{mean}(FE_{HighBRI}) \quad (6.6)$$

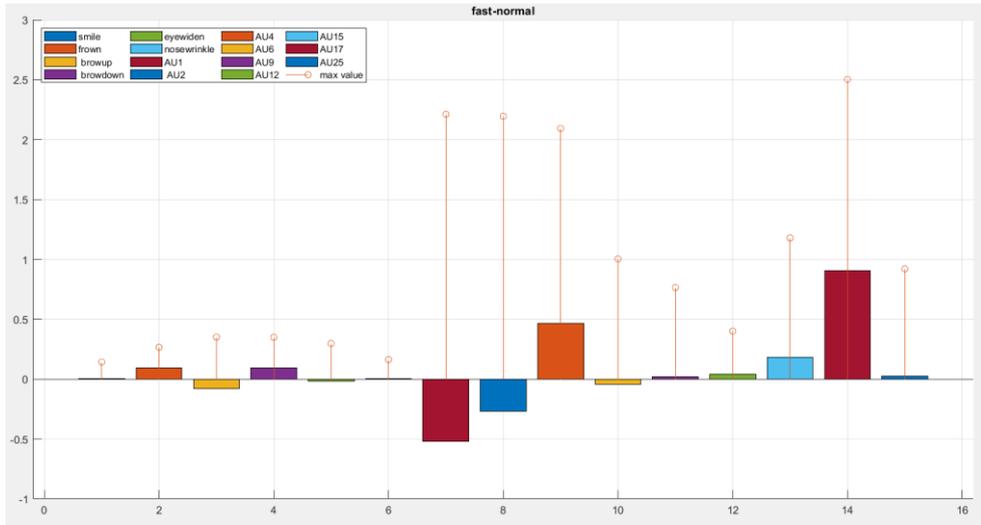


Figure 36: Exp1. High to Normal BR comparison

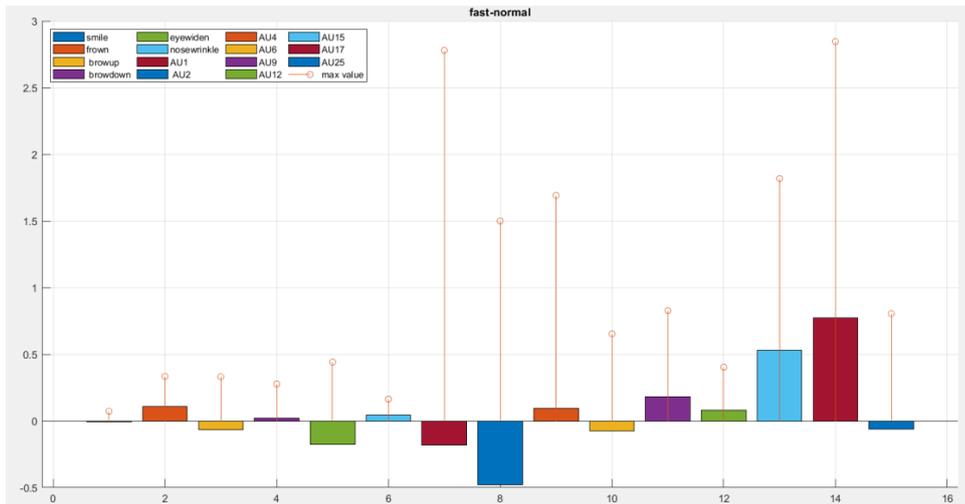


Figure 37: Exp2. High to Normal BR comparison

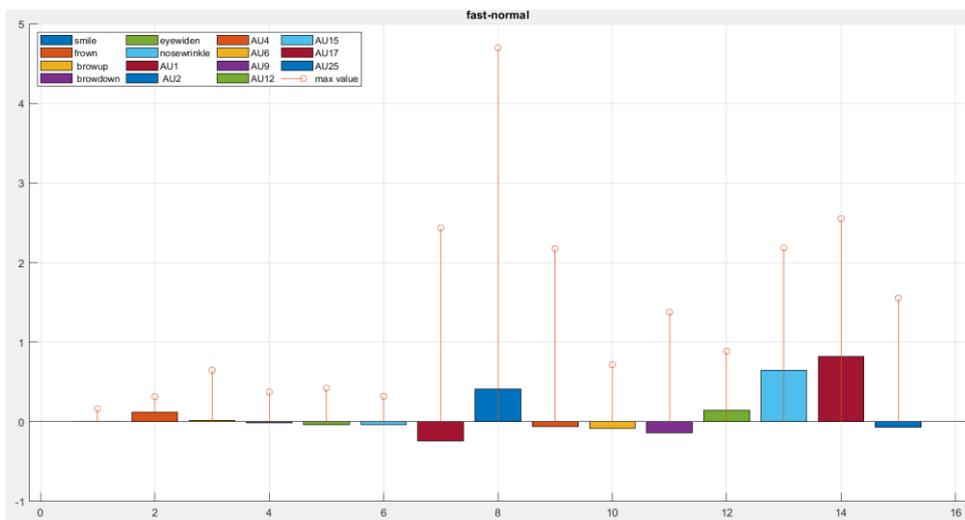


Figure 38: Exp3. High to Normal BR comparison

Again, the dependency between BR and AU15 and AU17 can be seen. As expected an increase in BR induces a greater stress of the eyebrow muscles which then determines greater variation of AU1, AU2 and AU4. These values cannot be directly traced back to BR because their variation is random both in module and in sign.

%NormalBR-HighBR	Exp1	Exp2	Exp3
AU15	38.25	28.02	29.22
AU17	47.74	17.88	18.36

Table 7: AU15/17 Normal to High BR

Finally, it can be noted that, as expected, a higher BR induces 'noise' on the AU connected to the eyebrows but the percentage variation of AU15 and AU17 is less than in the case of LowBR. This phenomenon can be explained as follows: the reduction of BR induces a slight prolonged contraction of the eyelid which in turn generates greater influence on chin and lips corners. On the contrary, a HighBR produces a more frequent and intense but short-lived muscle effort. This leads to a greater stress in the area near the eyes (upper face) but it hardly propagates in the lower face.

6.3.1 Outlier rejection

The outlier rejection method is therefore defined tanks to these experiments. Outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error. This can happen if the confidence has a drop so the Landmark detection is not correctly performed. A confidence level higher than 80% means that the landmarks are properly detected so they follow the face contractions and an outlier cannot occur. The minimum confidence threshold accepted is therefore set at 80%. Some geometric errors of contraction intensity may occur for a confidence between 80% and 90% but, as it will be explained in the OTM experiments chapter, the data that are over 80% are then filtered and mediated with a sliding window that allows to mitigate small detection inaccuracies.

In conclusion, data rejection thresholds are:

- Operator's distance from the camera between 45cm and 100cm
- Maximum rotation around X axis <20°
- Maximum rotation around Z axis <35°

As previously mentioned it is difficult that these thresholds are exceeded in fact in all the OTM experiments they have never been exceeded.

7 Experiments

The human-factors engineering research was conducted in the Royal Melbourne Institute of Technology (RMIT) Aerospace Intelligent and Autonomous Systems (AIAS) lab. Research at RMIT AIAS lab is primarily focused on the design and development of military and civil aerospace HMI² to enable safer and more efficient operations. Operational and technological evolutions in the aerospace field impose an increasing cognitive demand on the operator that must be mitigated with an adaptation of the level of automation, this auto-calibration of the system, however, must have an excellent level of reliability in order not to compromise safety. The experiments conducted therefore aim to analyze the trusted autonomy of a system that uses psycho-physiological parameters to adapt the level of automation. A trusted autonomy system must be able to adapt to both the task requirements as well as the human user's needs.

Informed consent for both study participation and publication of pictures was obtained from all the subjects after the explanation of the study. The experiment was conducted following the principles outlined in the Declaration of Helsinki of 1975, as revised in 2000.

The study protocol received the favourable opinion and approval by RMIT's University College Human Ethics Advisory Network (CHEAN) (ref: CHEAN A 201710-02-17). All respondents consented to participate by completing the online study. Only aggregate information has been released while no individual information was or will be diffused in any form.

The adopted protocol for the experiments that will be presented was defined by a team that started a research before this thesis project became part of the main project (ASEHAPP 72-16, ASEHAPP 111-16). I contributed to the team activity on Terminal Manoeuvring Area Scenario ATCo-in-the-loop experiment collaborating with THALES Australia.[76, 77]

The analyses performed also aim to define a new protocol for future experiments in which FE analysis will be introduced in the sensor network.

7.1 OTM experiments: Bushfire-fighting

As widely explained in the previous chapters, physiological sensing allows to evaluate the cognitive state of the operator in order to optimize the cognitive HMI²(CHMI²) which is based on Concept of Operations (CONOPS) addressing the one-to-many mode of operations for surveillance and bushfire-fighting. [78]

Unmanned Aircraft Systems (UAS) provide new opportunities for cost-efficient, persistent airborne surveillance for a wide field of application ranging from military reconnaissance a monitoring and industrial inspection.

This research therefore aims to analyze the possibility of managing multiple UAVs with a single operator through a system in which the trusted autonomy is based on a bio-sensing network. This CHMI²-based network operates on a single operator Ground Control Station(GCS) that has been prototyped and tested in a number of experimental studies. Sensors and simulators usually came as commercial-off-the-shelf products, requiring the development of suitable interfaces to support their integration within the laboratory network.

The data monitored by the sensors are collected and processed by a centralised server (HFE-Lab server). Each sensor has a dedicated client, which performs the data pre-processing before sending the processed data to the server. A key aspect for the monitoring of psycho-physiological parameters is the synchronization of data so the data server synchronizes incoming data from the different clients being hosted and facilitates the exchange of data between these clients.

Australia is affected every year by numerous bushfires that have devastating impacts on the environment and community. A UAS monitoring system could prevent and contain them informing fire fighters if a new bushfire is detected or provide location and size of bushfires already active flying over regions that would otherwise be unsafe for manned aircraft. For this reason a UAS simulator has been developed in order to analyze a scenario of applicability of an CHMI2 system based on CONOPS for One-to-Many operations according to the Australian Incident Management Team which identified several UAV applications, which included fire perimeter mapping information, night exploring, as well as improving communications and the safety of ground personnel. The main goal is to support both ground and airborne fire-fighting teams so the UAS operator (UASo), managing multiple UAV teams, can play the role of tactical coordinator to detect, monitor, localize and characterize fires in the operator’s Area of Responsibility (AOR). The UAS operator task is to detect and monitor fires providing geographic location, size, intensity, rate of spread and distance from objects of interest(roads, other fires, buildings) to both ground and airborne fire-fighting elements.

Relevant tasks and their coordination are illustrated in Figure 39.

The UASo has to plan and monitor the status of each UAV team. Mission planning entails assigning specific tasks to each UAV team and conducting path planning to maximize task performance. Tasks can be primary or secondary including area search, fire perimeter mapping, monitor available fuel and performance of on-board Control, Navigation and Surveillance (C/N/S) systems. The UASo is also responsible to process data from UAV teams and provide them to fire-fighters such as: fire presence, perimeter location, size, front velocity and spread location, and fire threat level(from 1 to 5 in function of fire intensity and the fire’s proximity to objects of interest)

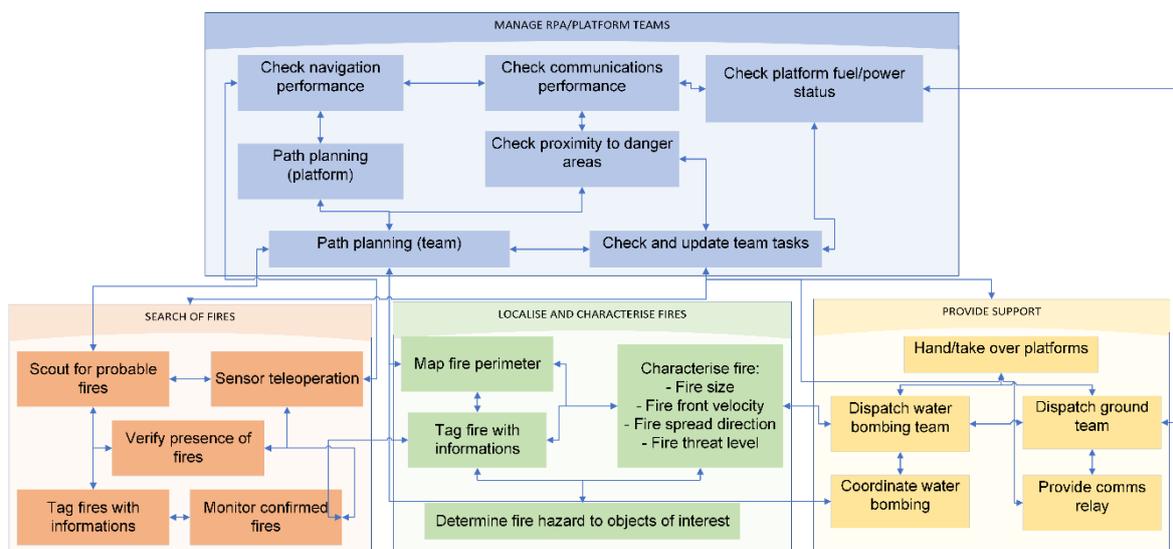


Figure 39: Task flowchart for the tactical coordination

7.1.1 GCS interface

The UASo performs supervisory tasks and higher-level decision making through GCS which implements aviation, navigation, communication and managing functionalities. The UASo manages the UAV teams by allocating tasks in the investigated AOR and monitors the task performance and system status of the team.

Figure 3 shows the six interfaces of the GCS display on which the UASo has to operate. The main feature are now described:

- the *Tactical Map* in the center of GCS display provides layers of geo-spatial information and allows the AUSo to select different UAV, makes modifications to team boundaries or platform trajectories, reviews sensor data, or adds information tags.
- The *Team Management Panel* allows to select teams and display relevant information on team assets in the form of glyphs. The UASo can select the team assets which can be active(actively managed by the human operator) or passive(outside of the human operator's command authority) and choose available task perform.
- The *Platform Management Panel* provides information on individual UAV for a selected team platforms, such as fuel and health status. The of review past sensor feed or the activation of different sensors for tele-operation can be managed.
- The *Task Planning Panel* provides support for automated path planning. The UASo can adjust some parameters to accomplish a task which are fed into a path planning algorithm which automatically computes feasible paths for each platform in the UAV team.
- The *Messaging Interface* lows the UASo to communicate with fire-fighters and in general all agents involved in the mission.
- The *Task Management Panel* is used to coordinate tasks between different teams, each bar represents a team task, tasks can have different level of urgency, a task color-coded in amber requires more attention from the UASo.

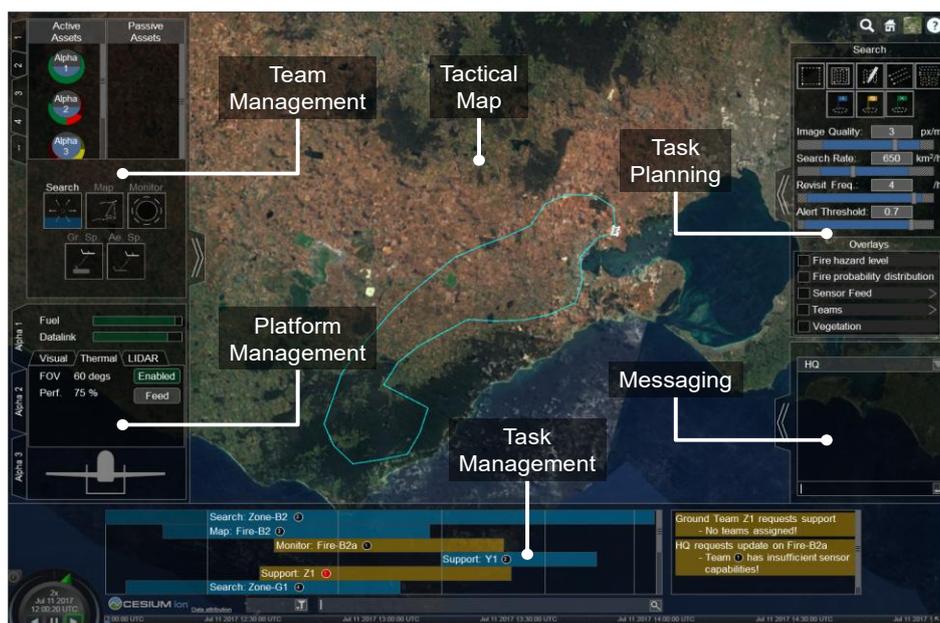


Figure 40: GCS Interface

The HMI2 system has to regulate the level of automation according to four aspects: information, task, team and path planning so relative mechanisms of adaptation are defined:

- The *Information Management adaptation* helps the UASo to maintain an appropriate situational awareness on UAV teams and relative tasks.
- The *Task Management adaptation* tracks progress and performances of all the tasks to help the operator identify the more relevant ones and to prioritize them to maximize the mission performance.
- The *Team Management adaptation* helps the operator to develop the optimal UAV team configuration, tasks assignment or re-allocation of UAVs and tasks.
- The *Path Planning adaptation* supports to generate or modify the optimal UAV team path to improve the efficiency of tasks achieving.

7.1.2 Test Scenario

Test participants assume the role of UASo of the system described above. The primary task of the mission is to find and localize any bushfires within the Area of Responsibility (AOR), whereas the secondary tasks are to maximize the search area coverage, to ensure that the UAV fuel levels, as well as navigation and communication performance are within serviceable range. The main AOR can be divided into smaller search regions (Team Areas) as shown in Figure 41. So-called because each Area is assigned to a UAV Team.

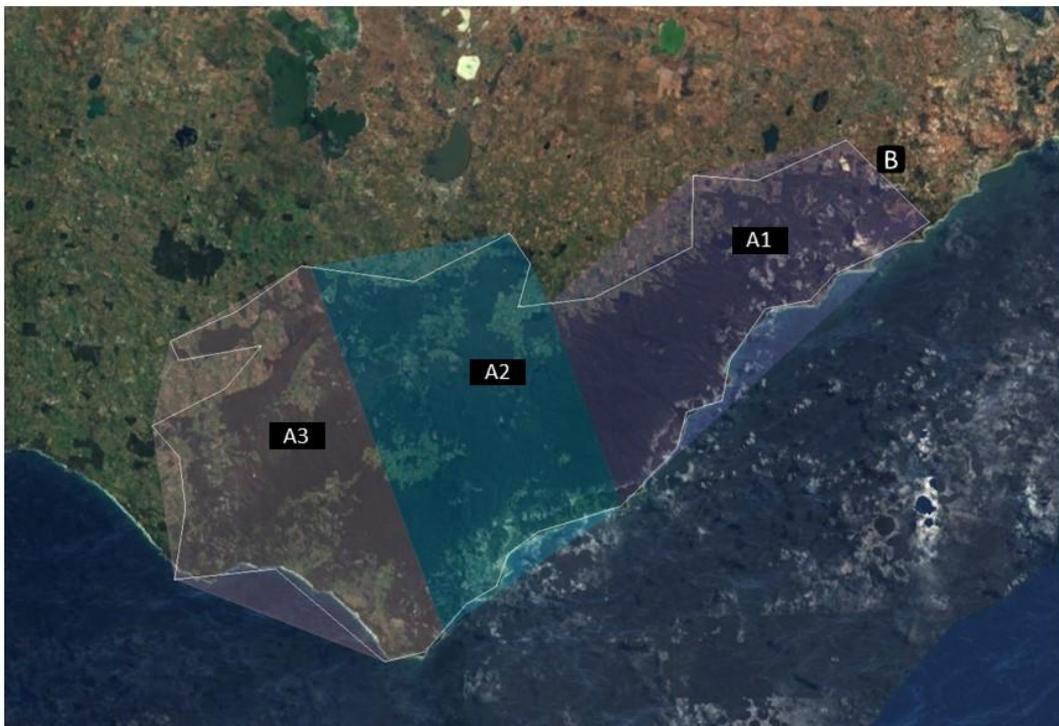


Figure 41: AOR partitioning

The UAV payload could be composed of a sensor (lidar) and/or a passive sensor (IR). The first one provides excellent range but a narrow field of view, it must be fired towards a ground receiver to measure the CO₂ concentration of the surrounding

atmosphere. The number of receiver in the AOR is limited so the operator has to manage when it is usable. To cover instead a larger field of view the Infra Red (IR) can be used, it doesn't require the use of a ground receiver but it has a smaller range.

The different tasks are now described and defined:

Primary Tasks

- Bushfires detection and tracking
 - At the beginning of the scenario bushfires are initialised and some spot fires can be created during the mission. The fire propagation is in function of the settled environmental conditions.
 - Maintain a constant visual coverage on the fire front

Secondary Tasks

- Sensor coverage maximization
 - Sensor coverage over the AOR is tracked in terms of revisit time, each scenario lasts 30 minutes so a revisit time longer than 30 minutes means that the area was no covered at all. Three covering measures can be related to the active sensor, the passive sensor or both.
 - Maintain a serviceable level of navigation and communication performance
- Navigation level and communication performance
 - These two parameters have to be maintained at a serviceable level
- Serviceable level of fuel
 - UAVs with low fuel need to be sent back to base, where they will undergo refuelling

Figure 42 shows the secondary tasks boundaries classification.

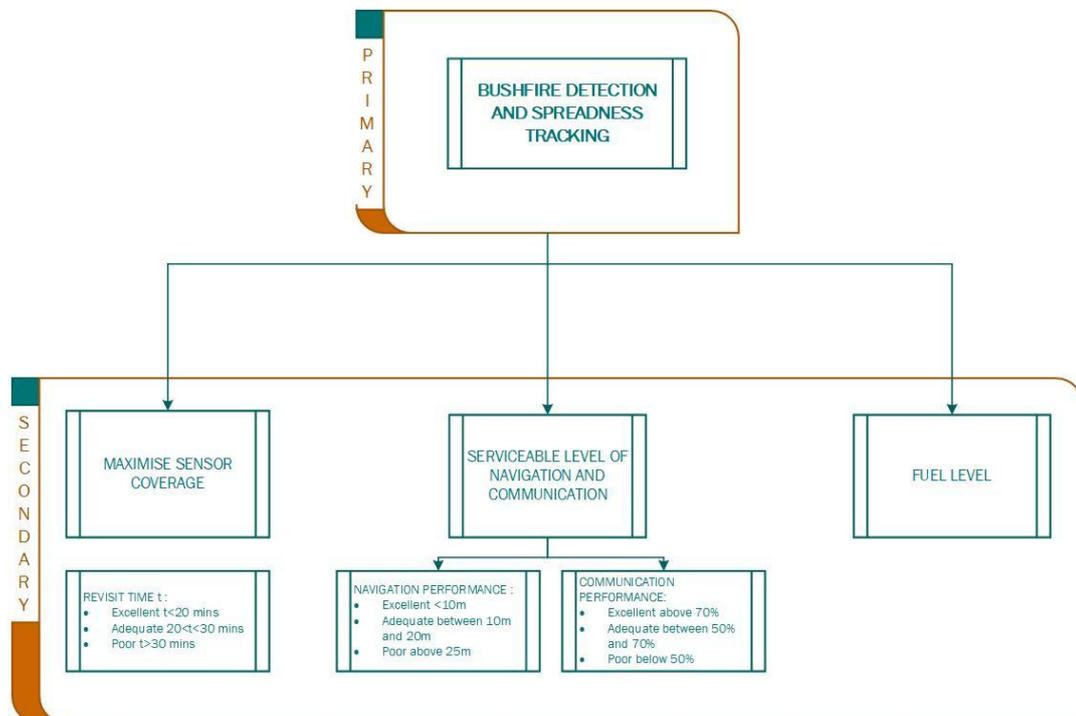


Figure 42: Secondary Tasks boundary classification

As described in *Cognitive state and Workload* chapter the most reliable and commonly used parameter to provide an objective and continuous measure of Workload is the Taskload. Secondary tasks are not critical to the detection of fires, but they are strictly related to mission performance. Secondary tasks are used because primary tasks are general and difficult to quantify. As the number of UAVs increases, the information that needs to be managed, monitored and planned increases which are mostly dictated by secondary tasks. Moreover the management of these tasks can easily refer to Sperandio's model for Workload evaluation because each operator chooses his own work method that determines feedback loops 1 and 2 as shown in Figure 17: Sperandio's model.

The Taskload is then defined as a weighted count of the number of pending secondary tasks from the UAV flight logs (i.e., system maintenance tasks). Each pending task for each UAV is assigned a score that can vary from 0 (best situation where the operator doesn't have to do adjustments) to 6 (performances are poor so the operator has to manage all the secondary tasks).

Pending secondary tasks	Score
Poor navigation performance (accuracy above 25m):	+1
Adequate navigation performance (accuracy between 10m and 25m):	+0.5
Excellent navigation performance (accuracy below 10m)	+0
Poor communication performance (comm strength below 50%):	+1
Adequate communication performance (comm strength between 50% to 70%):	+0.5
Excellent communication performance (comm above 70%):	+0
Critically low fuel (fuel needed to return to base less than 1.5x of fuel on board):	+1
Low fuel (fuel needed to return to base between 1.5x and 2x of fuel on board):	+0.5
Adequate fuel (fuel needed to return to base more than 2x of fuel on board):	+0
Autopilot mode in hold:	+1
Autopilot mode off:	+0
UAV not assigned into a team:	+1
UAV is assigned into a team:	+0
UAV does not have any sensor active:	+1
UAV does have any sensor active:	+0

Table 8: Secondary tasks rating

Table 9 shows the various phases of which the experiment is composed: calibration phase, phase 1, phase 2, phase 3 and final debriefing. During the calibration phase the monitored participant rest for 5 minutes, this phase is used to collect data that will be used for the calibrating the CHMI² baseline for phase 1.

The test case for each participant comprises a single scenario of 30 minutes composed by three phases of 10 minutes.

In Phase 1, 3 UAVs monitor the Team Area(TA) closest to the base so TA 1, after the Area has been searched, or when the mission transits to Phase 2 (whichever occurs first) the 3 UAVs are directed towards TA 2 and TA 3 new UAVs can monitor TA 1

After Area 2 has been searched, the operator repeats the same strategy with Area 3, moving 6 UAVs (3 in TA 1 and 3 in TA 2) to Area 2 and Area 3 so Area 1 can be monitored by 3 new UAVs. To allow the system to exploit some automated features described above each UAVs Team should be assigned to the corresponding Area (i.e., Team 1 for TA 1, Team 2 for TA 2, etc.). During phase 2 and phase 3, a change in environmental conditions would affect the navigation and communications performance of the UAVs, possibly requiring additional re-planning to satisfy operational requirements. UAVs can be selected out of a common pool comprising 4 UAVs with an active sensor, 4 UAVs with a passive sensor and 4 UAVs with a combination of active and passive sensors.

Pre-experiment logging	Phase 1	Phase 2	Phase 3	Post-experiment logging
5 mins	10 mins	10 mins	10 mins	5 mins
Rest condition	Cognitive load stimulation	Cognitive load stimulation	Cognitive load stimulation	Rest condition
Relaxed posture with open or closed eyes	3 UAV	Adding 3 UAV 6 UAV managed Env change	Adding 3 UAV 9 UAV managed Env change	Relaxed posture with open or closed eyes

Table 9: Test procedure for each participant.

7.1.3 Assumptions

The definition of movement patterns and variation of drone parameters has been modelled with assumptions that are not expected to affect the validity of the proposed offline training and online adaptation techniques. The purpose of the test cases is to verify the CHMI2 system so these assumptions have been made so that the tests are predictable and repeatable. The system has been developed with a modular architecture so that in case the potential validity of the FE monitoring is demonstrated also this sensor can be inserted in the sensor network.

The assumptions and simplified model in the test scenario are:

- *Kinematic model.* Lateral/vertical acceleration and yaw rate over time are integrated to propagate the UAVs position based on their groundspeed, climb rate and heading The pitch and roll are not accounted for in this kinematic model.
- *Power/fuel usage model.* A fixed endurance is assumed (e.g, 10 hrs), the power consumption is calculated on the time-in-flight. The refuel time is fixed (e.g, 1hr). Payload power usage is settled as a percentage of the total power. Aircraft which run out of fuel while in flight are considered ‘lost’.
- *Fire and CO2 model propagation.* Th area is divided in cells and for each one the rate and direction of fire spread is derived from models in the literature which account for vegetation type, terrain slope, wind speed,

temperature and humidity. The CO₂ concentration (to be detected by lidar) in each cell is assumed to vary between a range of values. The CO₂ concentration is assumed constant in each cell. The IR camera is able to detect if a given cell is burning, not burning or burnt.

- *Communication model.* the strength of the communications link between a UAV and the GCS is calculated by a logistic regression model (from 90% at 25km at GCS distance to 10% at 55km)

7.2 Experiment activities

The test activities took place over the course of 2 weeks, starting from 12 December 2019 and involved 6 participants (5 male; 1 female). The participants were Aerospace students at RMIT University and were selected based on their prior experience in aviation and aerospace engineering. Owing to a lack of familiarity with the HMI functions and the bushfire fighting scenario, participants had to undertake 2 hours familiarization training. During the familiarization training a pre-monitoring was made to evaluate the possible HMI2 adaptation methods. After the familiarization the participants were asked to wear two physiological sensors: the Bioharness strap, EEG cap. Then the calibration of remote sensors such as GP3 eye tracker and FE system took place and the experiment can start. Participants initially rested for 5 minutes (to collect rest-state data) before starting the scenario, which was a 30-minute exercise comprising 3 back-to-back phases of increasing difficulty. When Phase 3 is over Participants are asked to relax for 5 minutes while sensors keep logging so two comparable resting state are obtained, one pre-experiment and another post-experiment. Subsequently, participants provided subjective ratings for their workload and situational awareness in each of the three phases.

One of the participants could not fit the Cardiorespiratory sensor.

8 Online and offline analysis

This chapter presents the adopted mathematical methods in the data analysis and the considerations made in pre-processing.

First of all it is necessary to consider that during these experiments the operator wore various sensing devices and was aware of being monitored also by remote sensors, these factors can influence the naturalness of the subject's behavior. Monitoring was carried out on subjects with a different cultural background: Italian, Indian, Chinese, Thai, Australian so this analysis also evaluates the independence of the biometrical features from the cultural background. The subject's age vary between 24 and 36.

First and foremost, it is essential to define that two main types of errors can occur during tests: Type I and Type II. Type I errors or *false positives* occur when a condition is detected that does not actually appear in the test, an example in computer science is an antivirus that mistakenly considers a harmless program harmful, generating a false alarm. A Type II or *false negative* is an error in which a test result improperly indicates no presence of a condition (the result is *negative*), when in reality it is present. In statistical hypothesis testing a type I error is the rejection of a true null hypothesis, while a type II error is the non-rejection of a false null hypothesis.

The *null hypothesis* H_0 is a general statement or default position that indicates that no particular variations have occurred within the analyzed data, or for example there are no relationships between two phenomena or variables. The null hypothesis is generally assumed to be true until evidence. If the null hypothesis is rejected an alternative hypothesis H_1 is accepted which usually represents all other assumptions about the parameter not specified by the null assumption.

Initially a data analysis was carried out in relation to emotions as the only research available in the literature studied the relationship between groups of AUs and emotional states. This analysis was carried out by analyzing the AUs in half second time windows to highlight potential relationships with the Taskload and other sensors. Following numerous tests it was found that this approach did not provide satisfactory and repeatable results especially because the amplitude or frequency of the AUs was analyzed as an event. An AUs event represents an emotional state that involves the contraction of two or more AUs simultaneously. However, this contraction is too variable from person to person and is greatly influenced by various factors such as sleep and environment. In fact, each person feels the working environment in a different way and this can lead to a variation in the amount of contractions. For example, a person who is subject to stress at work or lives in a more serious environment may be unconsciously led to reduce the amplitude of FE than at home. In addition, a different mood (see chapter FE) influences the FE response to stimuli. For these reasons this approach has been abandoned and another method has been evaluated. The underlying pre-contraction of facial muscles in a time window up to 120 seconds. This data analysis method is now reported for OTM and ATM experiments.

The frequency of facial micro-expressions was also analyzed. A relationship was sought between the number of contractions that occur in a few seconds' time window and the workload. Also in this case the results are very noisy and although in some cases they show potential relationships in general they are not reliable and do not give results with good repeatability. For this reason also this method of data analysis

has been abandoned demonstrating that it is not a viable way using the open-source software available today and the knowledge you have about facial micro-expressions.

8.1 Filtering

The measured data by Open Face are filtered with a low-pass filter whose cut-off frequency has been determined based on studies available in the literature.

The carried out studies on the determination of the time window size necessary to detect an Action Unit. These tolerance windows represent the temporal precision with which action units are comprehensively coded. The parameter $kappa$ k is evaluated varying the window, which is the proportion of agreement above what would be expected to occur by chance (Cohen, 1960; Fleiss, 1981). Coefficients of 0.60 to about 0.75 indicate good, or adequate reliability instead $k > 0.75$ indicate excellent reliability. Figure 43 shows the k variation in function of the tolerance window for some AU.

AU	Facial muscle	Description of muscle movement	Frames	Tolerance window (seconds)			
				1/30th	1/6th	1/3rd	1/2
0	Not Applicable	Neutral, baseline expression	5124	0.73	0.79	0.81	0.83
1	Frontalis, pars medialis	Inner corner of eyebrow raised	386	0.66	0.71	0.74	0.76
2	Frontalis, pars lateralis	Outer corner of eyebrow raised	323	0.58	0.64	0.67	0.70
4	Corrugator supercilii, Depressor supercilii	Eyebrows drawn medially and down	480	0.68	0.76	0.79	0.82
5	Levator palpebrae superioris	Eyes widened	<48	—	—	—	—
6	Orbicularis oculi, pars orbitalis	Cheeks raised; eyes narrowed	201	0.72	0.78	0.82	0.85
7	Orbicularis oculi, pars palpebralis	Lower eyelid raised and drawn medially	136	0.44	0.49	0.53	0.56
9	Levator labii superioris alaeque nasi	Upper lip raised and inverted; superior part of the nasolabial furrow deepened; nostril dilated by the medial slip of the muscle	48	0.67	0.76	0.81	0.83
10	Levator labii superioris	Upper lip raised; nasolabial furrow deepened producing square-like furrows around nostrils	96	0.69	0.76	0.79	0.81
11	Levator anguli oris (a.k.a. Caninus)	Lower to medial part of the nasolabial furrow deepened	<48	—	—	—	—
12	Zygomaticus major	Lip corners pulled up and laterally	800	0.67	0.71	0.74	0.76

Figure 43: Kappa coefficient for single Action Unit [29]

It is easy to see how the reliability improve significantly between the 1/30th second and 1/6th second frame tolerance windows. k doesn't show significant variations between 1/6 and 1/3 so a tolerance window of 1/6th second provides adequate latitude for temporal agreement. [29] [24]

Thanks to these considerations a cut-off frequency of 6 Hz was chosen.

This value is also correct because the adopted FPS by the software are a function of the lighting level of the camera and they can reach a minimum of 7 FPS so it is insured that the cut-off frequency is lower than the data sampling.

In generally speaking not all AU's are detectable with the same reliability because some can be a mix of several AUs and therefore one is distorted or confused, for example AU23 is often confused with AU24. For this reason AUs that are generally more easily detectable are evaluated.

It is essential to note that these studies were carried out in 2001 when camera and software capabilities were significantly lower than nowadays. Finally Open Face

has a very high detection accuracy that is usually between 88 and 95, values below 80% were discarded and on average in an experiment lasting half an hour (59000 samples), lead to not consider $\sim 2.5\%$ of the values (1450 samples).

8.2 Prominence

The fundamental parameter for analyzing facial expressions is the amplitude of the facial expressions and therefore the sudden contraction of facial muscles. A Taskload variation can induce a muscle(AU) contraction which could be already pre-contracted so the AU prominence is evaluated. This is because during the performance of tasks in a wide period of time, tending to be more than 5 minutes, the face can contract in a constant way, like an underlying contraction that we call pre-contraction. In fact a new stimulus(Taskload increase) can induce a contraction and when the stimulus is finished the individual can maintain a pre-contraction even for minutes if the Taskload is still high, if in this time the operator undergoes a new stimulus the peak of muscle contraction will start from the pre-contraction value.

This concept is shown in Figure 44 using as an example the AU9 trend during an experiment phase.

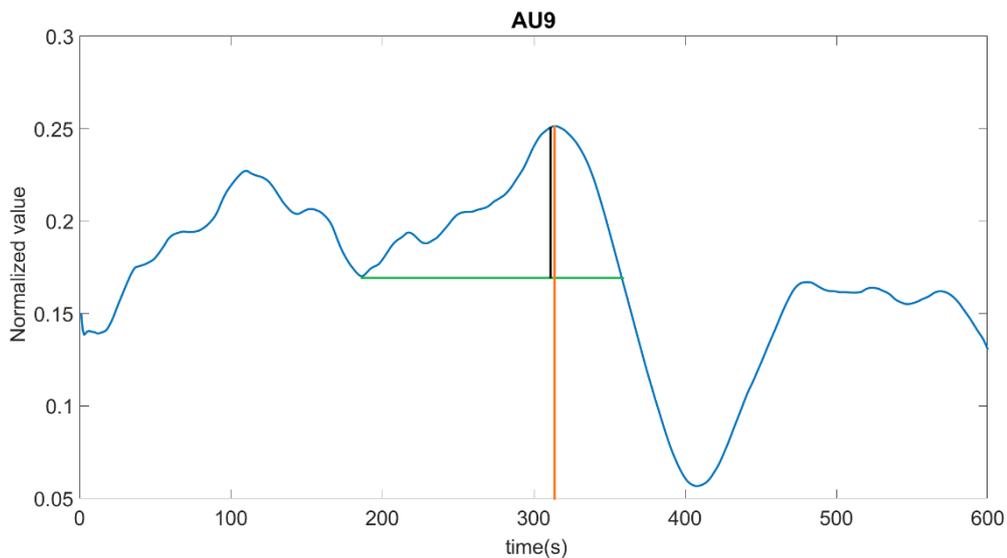


Figure 44: Peaks evaluation

The AU9 value up to 390s remains very far from the null value so to evaluate the variations of the parameter between an instant t and the instant $t+1$ it would not be correct to assume that the variation is the intensity of the peak (red line) but the prominence(black line). Let's suppose that at instant t the Taskload assumes a generic value of 6 and at instant $t+1$ it reaches a value of 8. For example if the operator has perceived a variation of Taskload equal to 2, in the same way the variation of AU related to this phenomenon should be considered as the difference between the pre-contraction at instant t and the value at $t+1$. The methodology of prominence determination is now presented.

The *prominence* of a peak measures how much the peak stands out due to its intrinsic height and its location relative to other peaks. An isolated peak may have a higher prominence than a peak of equal elevation located between two peaks. The prominence comes through the following steps:

1. Place a marker on the peak.
2. Extend a horizontal line from the peak to the left and right until the line does one of the following:
 - Crosses the signal because there is a higher peak
 - Reaches the left or right end of the signal
3. Find the minimum of the signal in each of the two intervals defined in Step 2. This point is either a valley or one of the signal endpoints.

The higher of the two interval minima specifies the reference level. The height of the peak above this level is its prominence as Figure 45 shows.

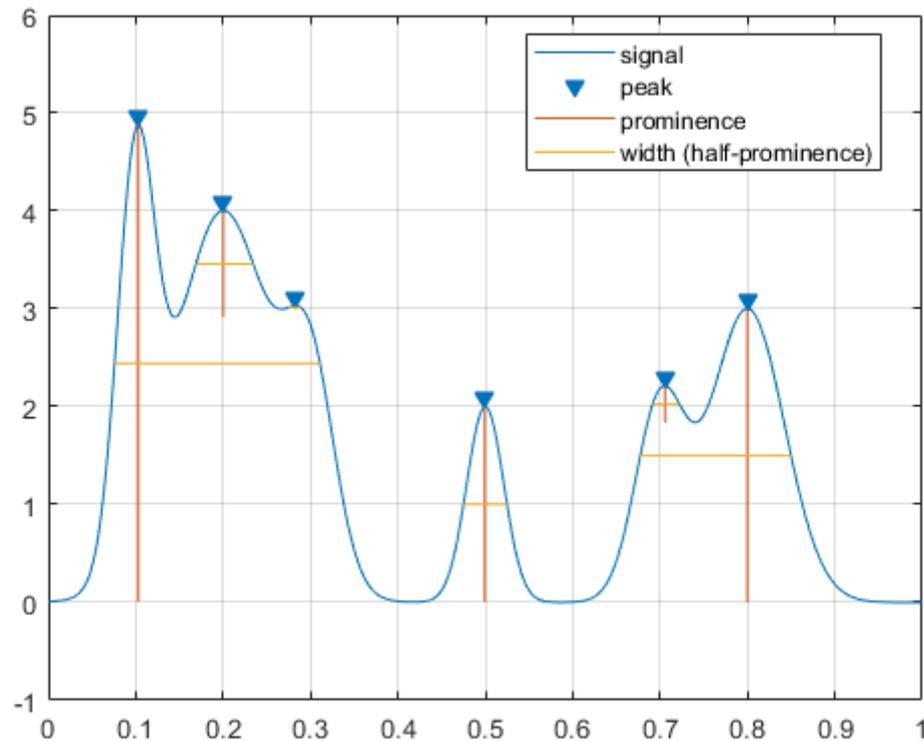


Figure 45: Prominence determination

The Matlab function *findpeaks* is used to determine the prominence of peaks, that makes no assumption about the behavior of the signal beyond its endpoints, whatever their height. This is reflected in Steps 2 and 4 and often affects the value of the reference level.

8.3 Variance and Covariance

The variance of a variable X , defined as $\sigma^2(X)$ or $Var(X)$, is a function that indicates the variability of the values assumed by variable X . It is defined as:

$$\sigma_X^2 = \frac{\sum(x_i - \bar{x})^2}{N} \quad (8.1)$$

It then indicates the square deviation of the variable from the arithmetic mean \bar{x} , N is the number of data variables. The AU trend is very variable so the mean is not evaluated for the whole data pool but is a variable average calculated with a sliding window.

The variance is the square of the standard deviation σ which is a measure of the amount of variation or dispersion of a set of values.

The covariance of two statistical variables X and Y is a function that allows to assess how much they vary together, i.e. their dependence. Covariance is defined as:

$$Cov(X, Y) = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{N} \quad (8.2)$$

Some of its properties are:

$$Cov(X, Y) = Cov(Y, X) \quad (8.3)$$

$$Cov(X, a) = 0 \quad (8.4)$$

$$Cov(aX, bY) = abCov(X, Y) \quad (8.5)$$

$$Cov(aX, bY) = abCov(X, Y) \quad (8.6)$$

$$Cov(X + a, Y + b) = Cov(X, Y) \quad (8.7)$$

Where a and b are constants.

If the two variables are identical, the variance is obtained. In fact the variance is a particular case of the covariance:

$$Cov(X, X) = var(X) \equiv \sigma^2(X) \equiv \sigma_X^2 \quad (8.8)$$

If the covariance between two variables is zero means they are uncorrelated.

These parameters are necessary to define the correlation coefficient used in subsequent data analysis.

8.4 Correlation coefficient

The adopted numerical measure relate AUs to other parameters from other sensors or Taskload is the correlation coefficient.

The Correlation Coefficient (CC) is a statistical measure that calculates the strength of the relationship between the relative movements of two variables.

The CC can assume values in a range between -1 and 1, a correlation of 1 means a perfect positive correlation so the two variables have the same trend. On the other side a correlation of -1 means a perfect negative correlation so when a variable

increase the other decrease in the same way. 0 indicates no correlation. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement.

There are several types of correlation coefficients, the Pearson Correlation Coefficient is adopted. It is a measure of the linear correlation between two variables X and Y, more precisely is defined as the covariance of the two variables divided by the product of their standard deviations.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad (8.9)$$

This measures the strength and direction of the linear relationship between two variables.

σ_X and σ_Y are defined as:

$$\sigma_X = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (8.10)$$

The Correlation coefficient cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

The two fundamental CC's properties are:

- It is symmetric: $corr(X,Y) = corr(Y,X)$
- It is invariant under separate changes in location and scale in the two variables, this means that even if variables X and Y are transformed into $a + bX$ and $c + dY$, where $a, b, c,$ and d are constants with $b, d > 0$, the CC remains unchanged.

In the engineering field, when measuring variables with a very high required precision, a $\rho_{X,Y} > 0.9$ is usually required, but in the biometrical measurement field parameters have a very low predictability, high variability and measurements themselves can be affected by numerous errors so the statistical Evans(1996) scale is assumed as basic reference:

- 0.0-0.19 *very weak*
- 0.2-0.39 *weak*
- 0.4-0.59 *moderate*
- 0.6-0.79 *strong*
- 0.8-1.0 *very strong*

According to this scale a minimum threshold of 0.5 is assumed as minimum correlation coefficient and data with a CC greater than 0.6 are considered potentially meaningful.

9 Data analysis

9.1 Objective parameters: Secondary tasks

Each experiment has a different task trend obviously according to the criteria adopted by the operator in carrying out the task. A method to process the given task has therefore been developed.

Since we want to consider the variation of the tasks over a wide time interval and the minimum variations are usually not perceived by the operator from a cognitive point of view:

- task variations of less than 5% are neglected
- the task curve is made smoother to facilitate correlation analysis

The smoothment is done by applying a local average through a sliding window using Matlab's *movmean* function. The width of the window is defined according to the criterion now described.

Please note that we do not want to evaluate the small tasks variations because the ultimate goal is to adapt the automation of the system and this does not have to be adapted to every single variation but it tends to be considered an adaptive automation hysteresis with variations greater than 30%/40%. To be conservative a variation of 25% has been considered. Analyzing the prominence of the task trend peaks, the peaks with prominence greater than 25% have been selected and it has been determined that the time duration varies from 25 to 50 seconds. At this point it was possible to determine the width of the sliding window considering the two cases of maximum and minimum duration of a peak.

$$Task_{m25} = movmean(Task, [25 0]) = \sum_{k=1}^{k=length(Task)-25} \frac{(\sum_{i=k}^{k+25} Task_i)}{25} \quad (9.1)$$

$$Task_{m50} = movmean(Task, [50 0]) = \sum_{k=1}^{k=length(Task)-50} \frac{(\sum_{i=k}^{k+50} Task_i)}{50} \quad (9.2)$$

The formulas show that the average is made with respect to the previous values so that it can be applied in real time. Subsequently the correlation analysis carried out for both dimensions of the sliding window will be reported. Generally it can be said that a window of 50s can be adopted when there are big variations of tasks while for smaller variations (but more than 25% of prominence) a 25s window is suggested. This approach can be easily integrated in the sensor network through an automatic window definition function with the criteria just described.

Figure 46 shows the adopted method with a window of 50 seconds. In conclusion, therefore, the yellow curve in the figure was used as a reference to evaluate the correlation with the AUs trends.

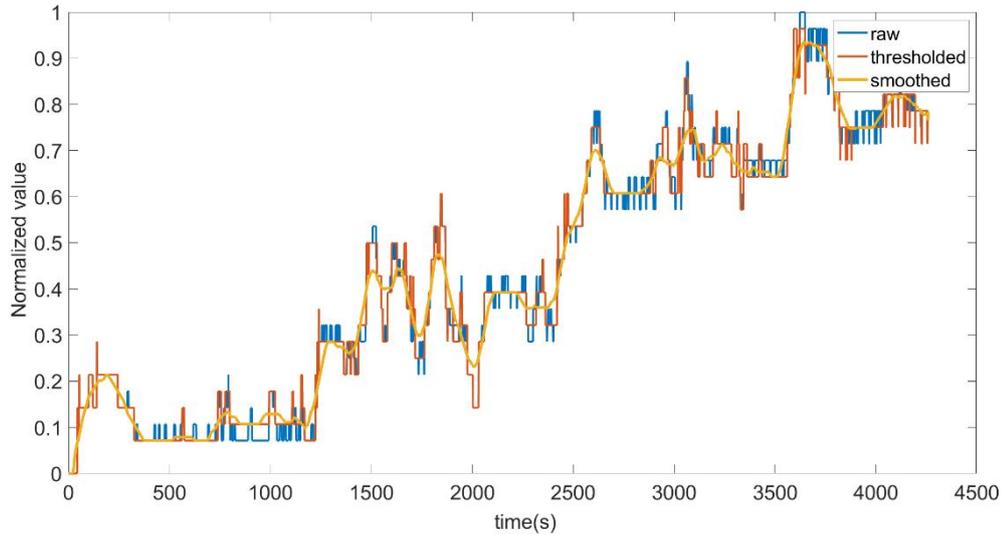


Figure 46: Task trend smoothing process

9.2 Sensor data smoothing

Numerous approaches have been tried in evaluating AUs and it has been noted that evaluating individual AUs in short time periods(0.5s - 2s) is very difficult or almost impossible especially because this approach is used in psychology but it is a human being who reads AUs and can therefore interpret them according to the context, so a subjective evaluation. Wanting instead to evaluate in an objective way the AUs to develop a mathematical model it is necessary to evaluate them in a greater temporal amplitude focusing on the baseline contractions of the face and evaluating the trend of contractions instead of the single contraction event.

Figure 47 shows how the raw data that the software provides which need to be pre-processed.

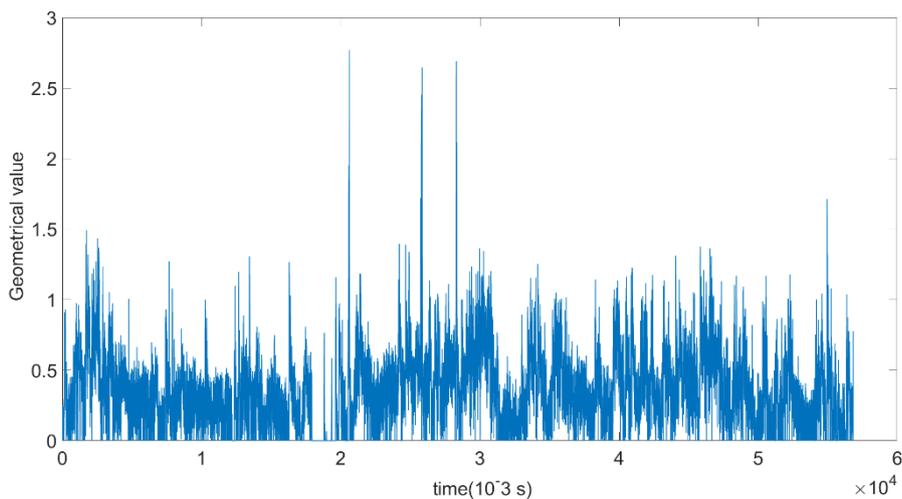


Figure 47: AU9 Raw

It has already been explained in the previous chapter how the first filtering is performed and the criterion for setting the cut-off frequency to 6 Hz. At this point a

smoothing is performed by averaging the values through a sliding window as for tasks and using the matlab function *smooth*. It is not used only *movmean* because this can involve an excessive reduction of the amplitude of the peaks and therefore a trend of the amplitude of the AU and their derivative that does not reflect the reality. For this reason *movmean* is used to reduce the noise and small oscillations that the filtering has not removed and then the function that uses the following equation is used:

$$y_s(i) = \frac{1}{2N+1} (y(i+N) + y(i+N-1) + \dots + y(i-N)) \quad (9.3)$$

where $y_s(i)$ is the smoothed value for the i th data point, N is the number of neighboring data points on either side of $y(i)$, and $2N+1$ is the span.

In this case, however, it is difficult to determine the size of the sliding window a priori because each individual has his own facial expression with different response times and contraction entities. The smaller the sliding window the greater the possibility of capturing significant variations but at the same time increases the risk of having false positives. In fact, small peaks may occur, even when the number of tasks is constant, due for example to itching, movements on the chair or others especially if the Taskload is low because the person tends to be more distracted. In general, therefore, it can be said that adopting a large window has the advantage to eliminate false positives and reducing the noise that has not been filtered with the passband filter, but may not point out variations that can be significant. For this reason different time *windowAU* of 25s, 50s, 100s, and 200s have been evaluated. The adopted approach is to adapt the window size in function of the Taskload variation over time. The variance of the Task trend is evaluated and the greater the variance of the Taskload the more it means that the operator is affected by cognitive load variations and so the size of the *AUwindow* is reduced to analyze them in more detail. This variation in *windowAU* have been conducted manually to verify whether this method is functioning so it can be implemented in a future online self-calibration system. This method and the self-calibration model are described in *Protocol*.

9.3 Correlation coefficient

The correlation coefficient is used to evaluate the relationship between the task and the AU trends. Not only the AU amplitude was analyzed, but also the peak prominence and frequency. As already explained the correlation coefficient is defined as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{N}}{\sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} * \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2}} \quad (9.4)$$

In the evaluation of biometric parameters the average reference value cannot be calculated as the simple average of the values of the vector x and y because as shown in Figure 48 defining an average value for the whole data pool in an experiment of 2100 seconds (plot every half second) would not provide a correct relation parameter because during the phases the average value of pre-contraction of the facial muscles is very different from the average value of the whole pool.

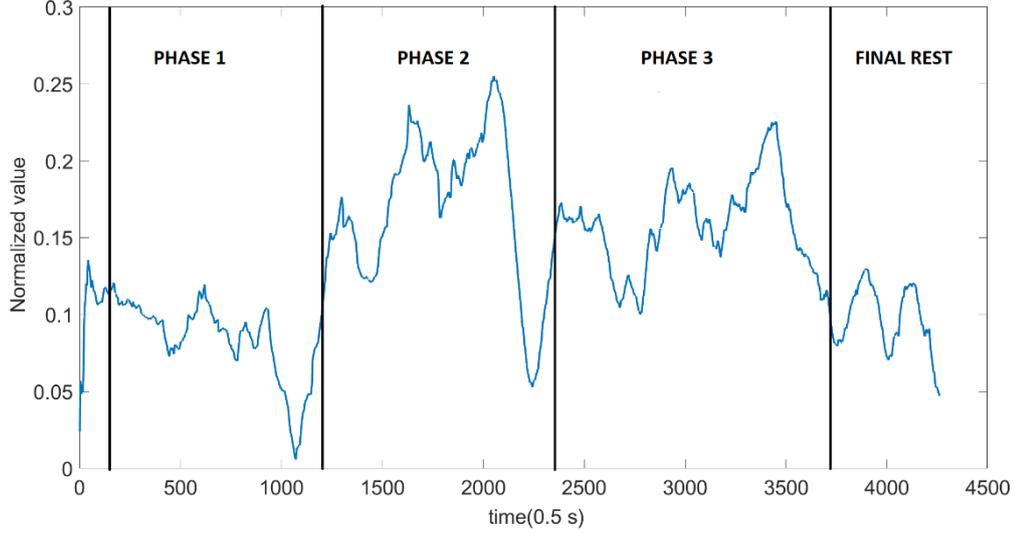


Figure 48: AU9 trend

For this reason, the data were analysed by dividing them into phases. For each phase k was calculated the average of the values that is used as reference for the values of that phase. The sampling rate is every half second, each phase lasts 600 sec=1200 values, so defining a parameter k that indicates the number of values on which the average is calculated, we have $k_1 = \{50:1250\}$, $k_2 = \{1251:2451\}$, $k_3 = \{2452:3652\}$. The first 50 seconds are for the initial rest and setting of the sensors, the last 609 values (4261-3652) are due to the final phase of rest at the end of the experiment to verify the rest values and any anomalies.

The correlation coefficient formula is then adapted by calculating a reference average for each phase, for the first phase for example:

$$mean(phase\ values) = mean(x_{i-k}) = \frac{\sum_{k_1(1)}^{k_1(end)} x_{k1}}{k_1} \quad (9.5)$$

This aspect is fundamental because each subject has his or her own facial expression, level of contraction intensity and response time. The adopted formula therefore becomes:

$$\rho_{X,Y} = \frac{\sum(x_i - mean(x_{i-k})) * (y_i - mean(y_{i-k}))}{N \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - mean(x_{i-k}))^2 * \frac{1}{N} \sum_{i=1}^N (y_i - mean(y_{i-k}))^2}} \quad (9.6)$$

The objective of these analyses is to highlight potential relationships between Taskload and AUs so the correlation coefficient is adopted as a indicator even if in some cases it can be an overly restrictive parameter to evaluate trend similarities. In fact sometimes two curves show a good similarity even if the correlation coefficient is around 50%, for this reason the AUs that show a correlation coefficient greater than 50% are analyzed in detail and considered potentially usable those with ρ greater than 60%.

Now the relationship with the Taskload trend is evaluated, the data of the Participant with more experience in flight control are reported. Participant 1 has a training as a flight controller that determines that during the formulation of the

solution his cognitive process is more structured and according to a precise logic. It is curious to note that Participant1 also gives the best correlation results, so it can be deduced that experience plays a positive influence on facial expression response because the operator has more awareness of the situation and how to solve it. This awareness affects the physiological response that also follows a trend outlined and not random.

On the contrary, it has been noticed that inexperienced participants show a less regular and more random AU trend especially when the Taskload increases. When the Taskload exceeds a certain level, these operators are affected by confusion, which obviously affects the facial expressions response which shows confusing contractions hardly interpretable.

The first phase does not show particular variations because the Taskload is contained and the individual is getting familiar with the interface so the second and third phases are reported.

The contraction of the eyelids, monitored by the eye gaze sensor, determines the contraction of muscles in the areas surrounding the eyes (AU1,AU2,AU4,AU5,AU6), so these AU have been valued with less interest than those of the mouth (AU9, AU12, AU15, AU17, AU25) in order to use Open Face to evaluate parameters which are independent from those already evaluated by other sensors.

9.3.1 Cross correlation

Experiment Phase 2

Figure 49, Figure 50, Figure 51, Figure 52, Figure 53 show a close relation between the amplitude of the action units in the mouth area and the Taskload. As expected, the physiological response presents itself with a time delay.

This delay is the *Facial Expression Response Latency*(FE-RL) τ_d that occur between a stimulus and the related facial muscles contraction. This latency is not constant but it can increase or decrease in time so it can't be simply used the cross-correlation method by adopting a constant τ_{delay} but a *Dynamic Time Warping* (DTW) has to be adopted. The cross-correlation analysis is performed to determine the range of variation of τ_d and its average value. Subsequently this analysis is used to define the time warping strategy.

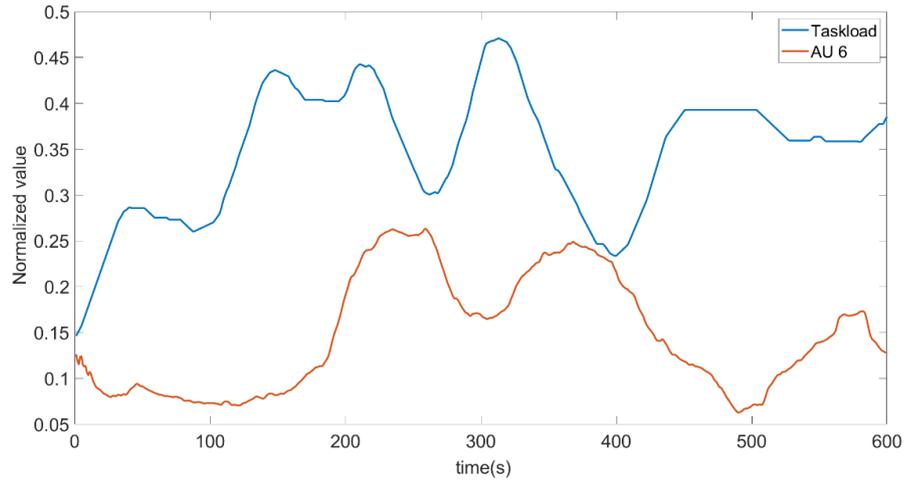


Figure 49: AU6 FE-RL

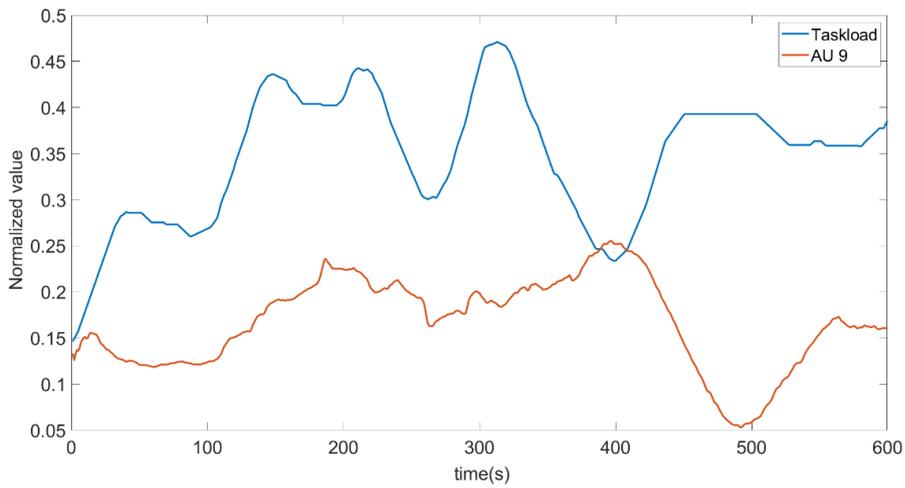


Figure 50: AU9 FE-RL

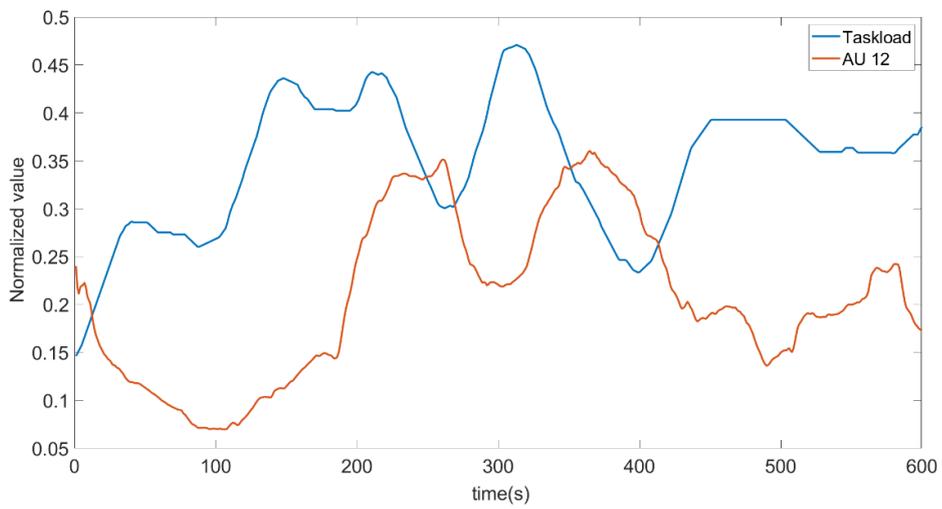


Figure 51: AU12 FE-RL

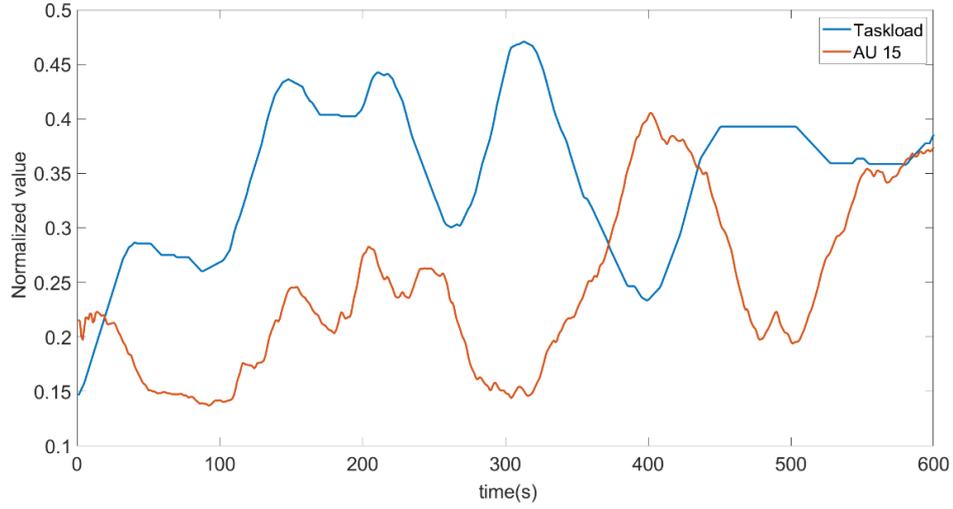


Figure 52: AU15 FE-RL

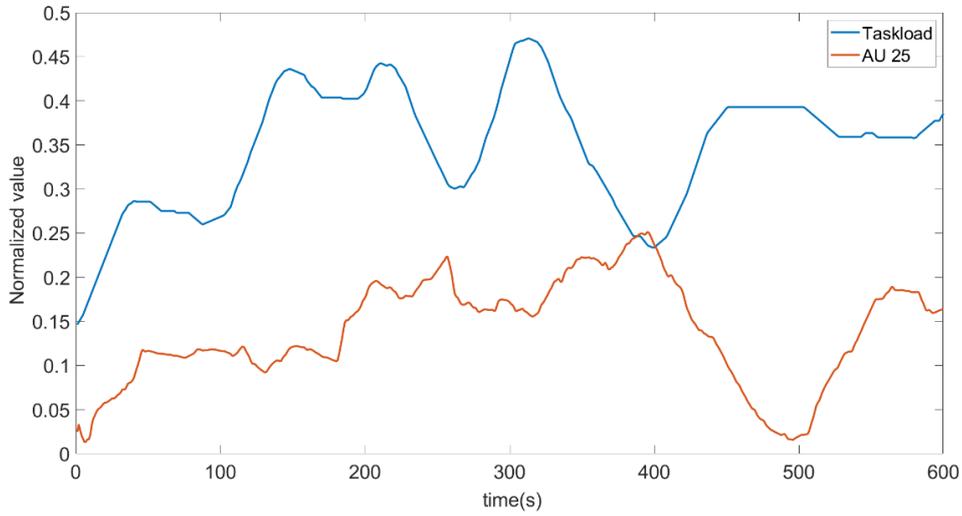


Figure 53: AU25 FE-RL

The cross-correlation is a measure of similarity of two series as a function of the displacement of one relative to the other. Considering two functions f e g differing only by an unknown shift along the x-axis. The cross-correlation can be used to determine what delay the g curve has with respect to f and therefore how much must be shifted along the x-axis to make it identical to f . This method slides g and calculates the integral of the product $f * g$ at each position, the functions match when the integral of $f * g$ is maximized. This is because when peaks (positive areas) are aligned, they make a large contribution to the integral. The cross-correlation is defined as:

$$(f * g)(\tau) = \int_{-\infty}^{+\infty} \overline{f(t)} g(t + \tau_d) dt \quad (9.7)$$

$\overline{f(t)}$ is the complex conjugate of $f(t)$ and τ_d is the time delay or lag.

In particular to assess τ_d the cross covariance has been evaluated which is the cross-correlation function of f and g with their means removed:

The cross-covariance is the cross-correlation function of two sequences with their means removed:

$$C(m) = E\{(f(n+m) - \text{mean}f) * \text{conj}(g(n) - \text{mean}g)\} \quad (9.8)$$

Where E is the expected value operator , $\text{mean}f$ and $\text{mean}g$ are the means of f and g respectively for each phase.

This function allows to obtain a vector in which C is the covariance between two shifting vectors. Once the *crosscov* vector is obtained, the phase shift value that maximize the covariance is determined and the AU curve is anticipated of that value. Consistently with the studies carried out and the available literature, delay values up to 180s are acceptable. Obviously the mathematical operators provide a different τ_d for each curve but they are very similar as shown in Table 10 :

	AU6	AU9	AU12	AU15	AU25
τ_d (s)	78	90.5	72	89.5	89.5

Table 10: phase shifting AU-Tasks Experiment Phase2

The average τ_d for Phase 2 is $\tau_{dmean} = 81.2 \text{ sec}$.

As mentioned before the size of the sliding windows *movmeanTA* and *movmeanAU* is varied according to the variance of the Taskload and the participants expressivity and FE-RL, for future analysis it will be necessary to define a protocol of adaptability of the smoothing to the participant that is reported in Protocol. τ_{dmean} may change if the window size is changed but in general it can be said that for Phase 2 $\tau_{dmean} = 81 \pm 15\text{sec}$ for all the participants. It is not necessary to calculate an exact value of τ_{dmean} but this analysis is necessary to estimate it so later the dynamic time warping range can be settled.

Figure 54, Figure 55, Figure 56, Figure 57 show the AU curve anticipated of τ_d . Some AUs like AU6 are aligned very well to the task curve for the whole duration of the Phase because the physiological response keeps about the same delay but other AUs like AU12 have a τ_d that is not constant so even if the τ_d maximizes the covariance there is a good relationship between the curves only from 300s, while before AU12 are too early. For this it cannot be assumed a τ_d constant to determine the correlation coefficient but the DTW is necessary.

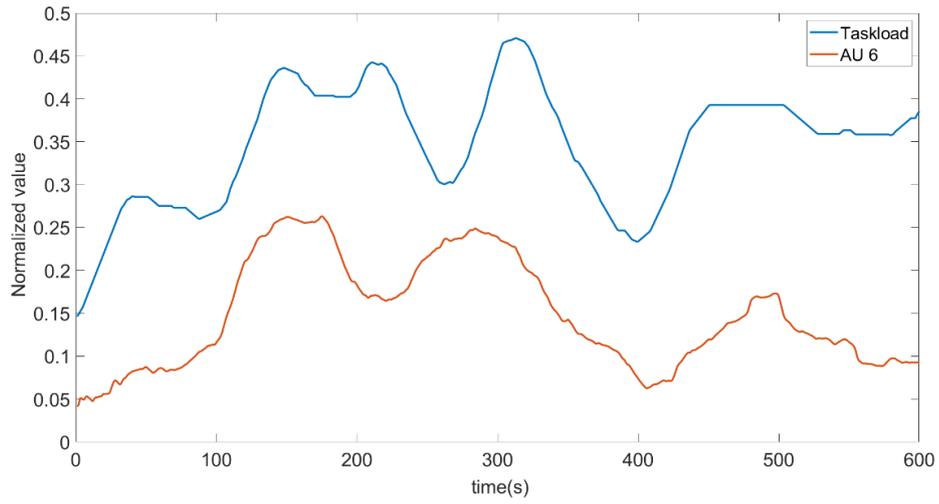


Figure 54: Phase correction AU6

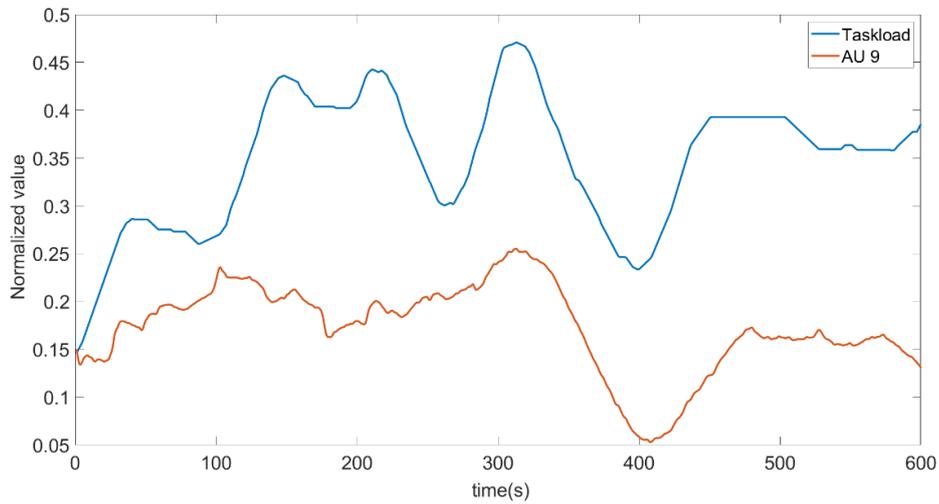


Figure 55: Phase correction AU9

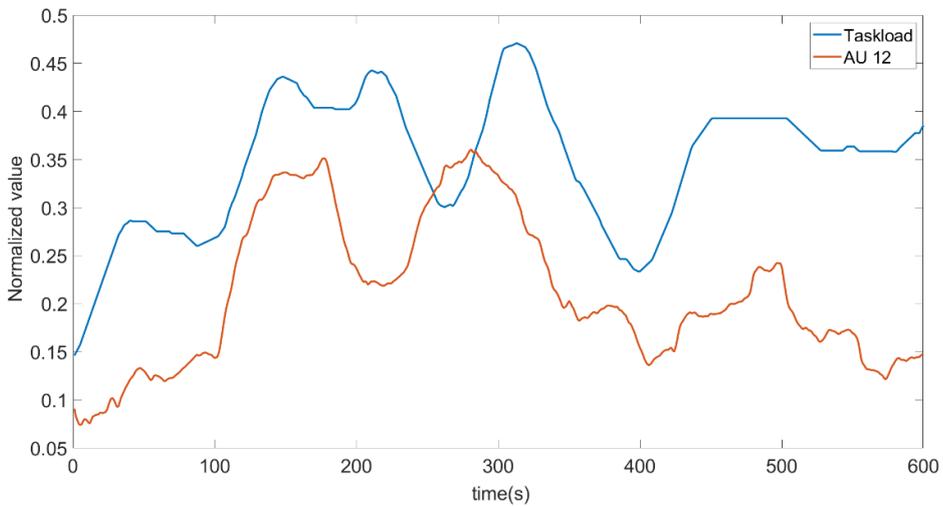


Figure 56: Phase correction AU12

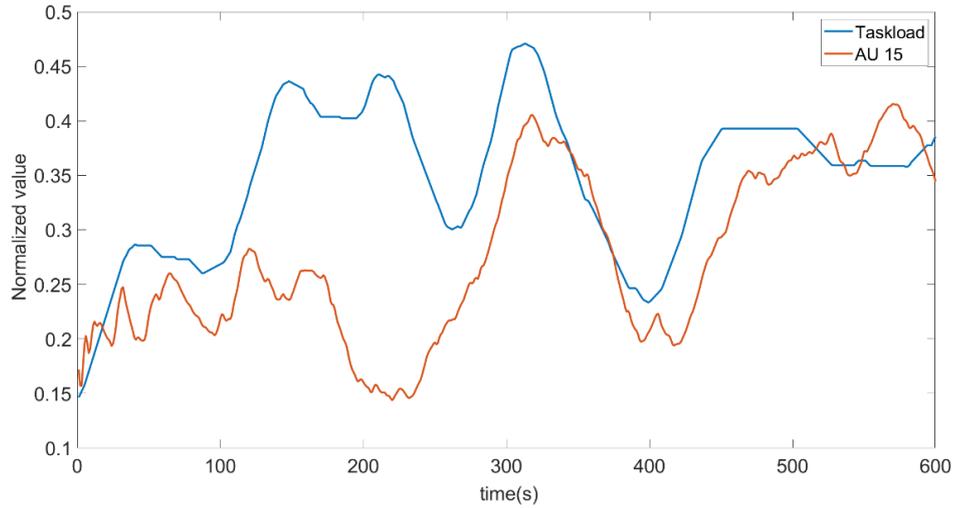


Figure 57: Phase correction AU15

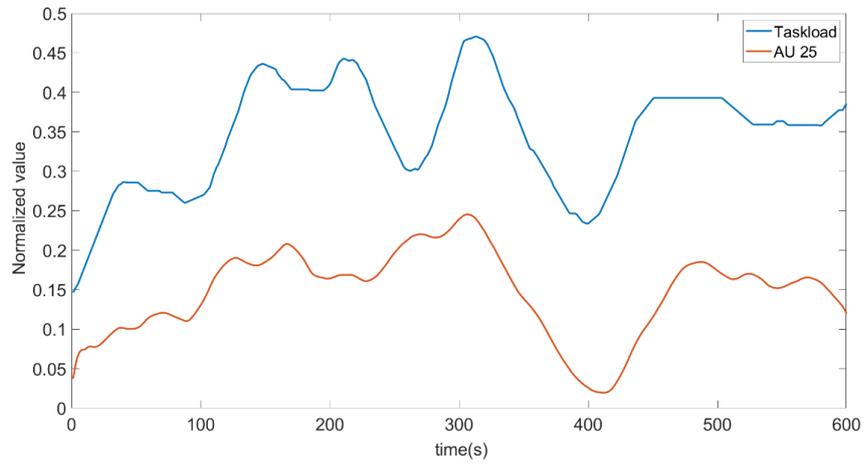


Figure 58: Phase correction AU25

As all the figures show, a generic cross-correlation cannot be adopted because, as expected, the physiological response does not have a constant delay time.

Experiment Phase 3

The AU6, 9, 12, 15 and 25 trends are now evaluated for phase 3. In this phase the delay time is increased as shown in Table 11:

	AU6	AU9	AU12	AU15	AU25
τ_d	234	264	229	305	253

Table 11: phase shifting AU-Tasks Experiment Phase3

The average τ_d for Phase 3 is $\tau_{dmean} = 128.5 \text{ sec}$.

Figure 59, Figure 60, Figure 61 show the AUs curves with the phase shift applied.

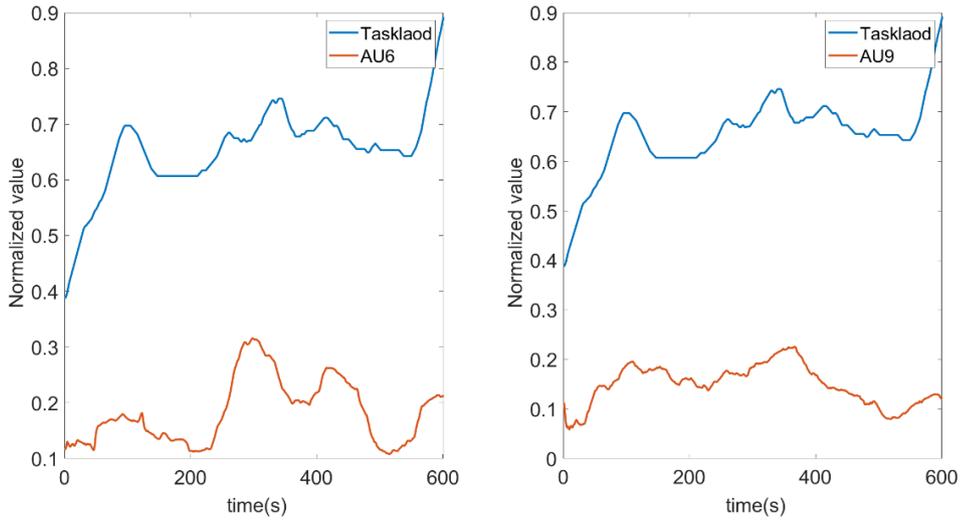


Figure 59: Phase correction AU6 AU9

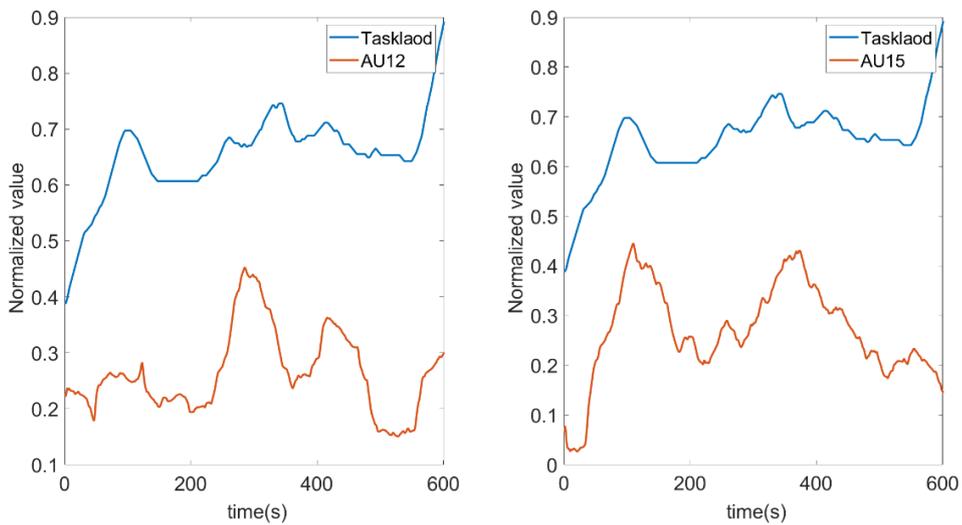


Figure 60: Phase correction AU12 AU15

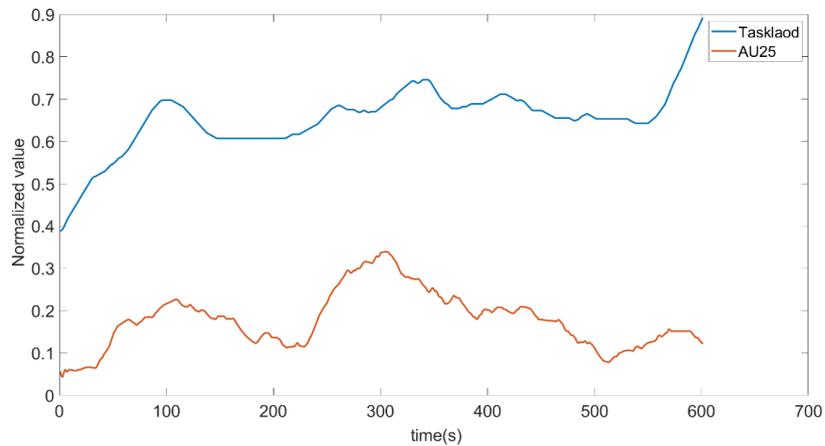


Figure 61: Phase correction AU25

In all the other experiments very similar values are obtained, in some cases the AUs are few seconds in advance but this is a problem of the experiment because there may be delays in the interface between the sensors and also the calculation of

secondary tasks can be done with a slight delay because the operator in some cases starts a cognitive process that stimulates his MFE and then performs a physical task resolution operation that is then detected by the system as a variation of the Taskload. In general, therefore, AUs variations are also accepted in advance of the related Taskload variation but only for a few seconds.

It is not possible to define a criterion of variation of τ_d because each individual has his own response. For example Figure 62 shows the response of the participant 4 during the third phase, the first peak is delayed while the second is almost in phase.

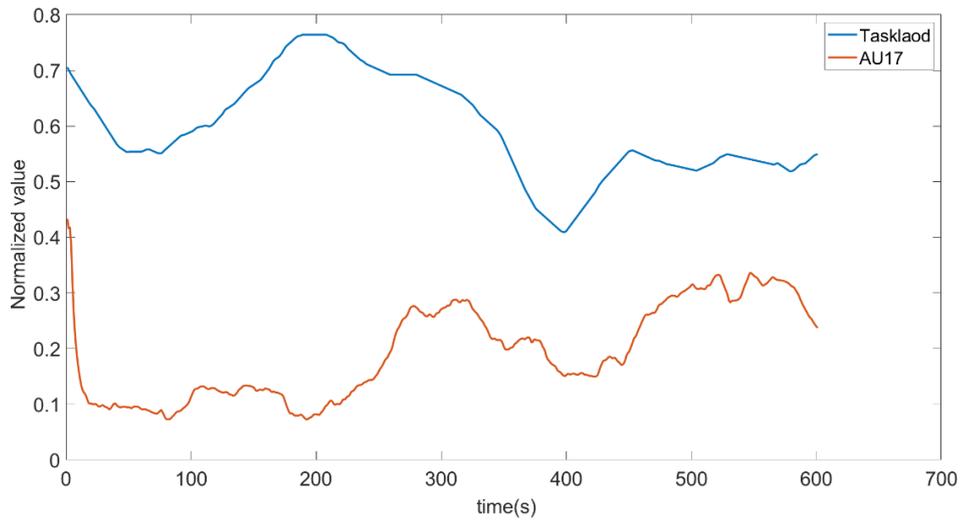


Figure 62: Participant 4 MFE AU17 Phase 3

In conclusion, considering the data of all the conducted experiments, it has been determined that the pre-contraction time delay of the facial muscles after a stimulus can be between 0s and 150s.

9.3.2 Dynamic time warping

Experiment Phase 2 is again evaluated to explain the DTW adopted method. The correlation coefficient between AU and Taskload is shown in Table 12.

	AU6	AU9	AU12	AU15	AU25
Corr-coef	0.7031	0.3627	0.6944	0.2987	0.6869

Table 12: Correlation coefficient Phase 2 mouth AUs

The table shows that the correlation coefficient (CC) is low for AU9 and AU15 but visually analyzing Figure 55 and Figure 57 a close relationship between the curves can be noted. This happens because the CC evaluates the similarity between the curves point by point and this leads to a low value because the AU curve is more variable and sometimes has an opposite slope for short stretches. In addition, some peaks are delayed with a variable time delay as described in the previous section. To overcome these two problems the *Dynamic Time Warping* is adopted. Through this

technique the small local variations that lead to a CC reduction are reduced and the relationship between the curves can be evaluated without having to fix a constant τ_d for the whole Phase. The DTW in fact manipulates the curves by applying a variable τ_d and also changes the general trend of the curves. Figure 63 and Figure 64 show the AUs that have undergone the slightest and greatest variation as a result of time warping.

It is easy to see that the problem of phase shift is solved but in the case of AU15 the curve is too distorted. For this reason it is necessary to set a curve modification threshold in order to find the right compromise between a meaningful correlation without losing the physical meaning of the data.

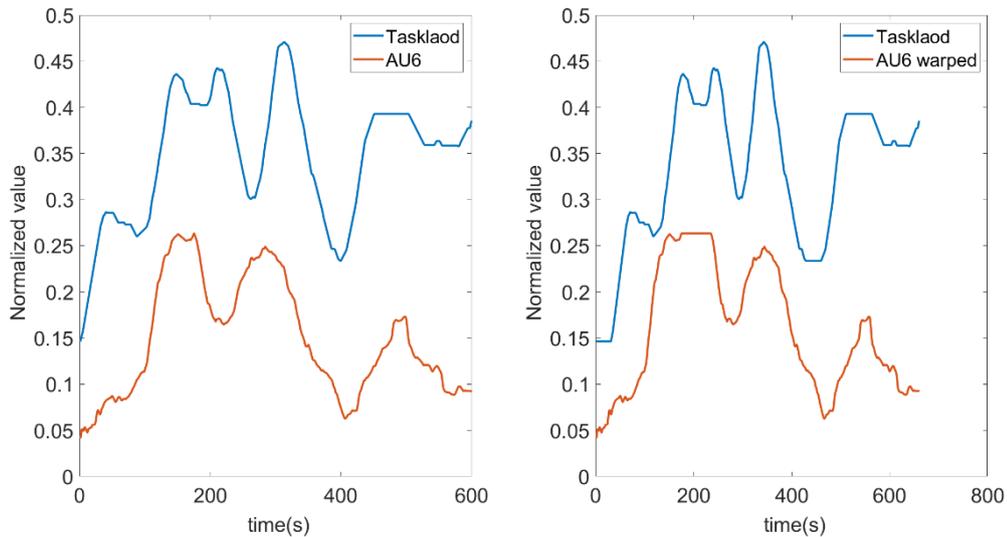


Figure 63: AU6 Warped

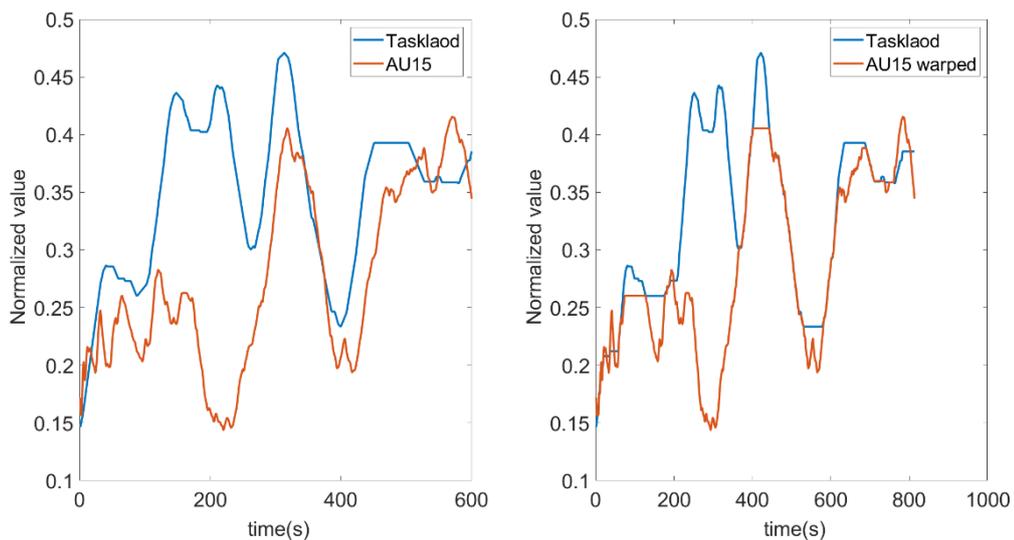


Figure 64: AU15 Warped

The *Matlab dtw* function has been used and Figure 65 shows how this algorithm works.

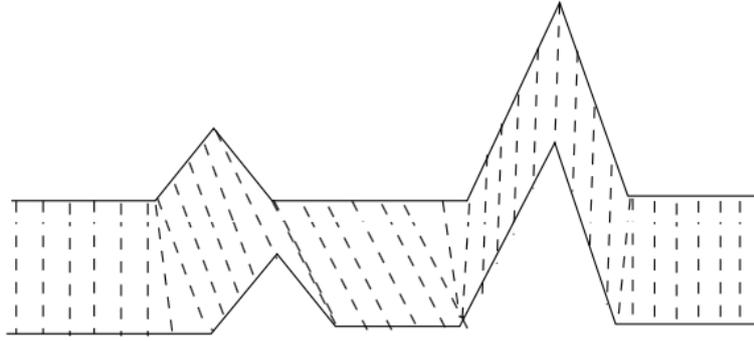


Figure 65: Dynamic Time Warping path

It computes the Euclidean distance between the samples of vectors Taskload and AU. It returns the warping path, IX and IY, that minimizes the total Euclidean distance between Taskload(IX) and AU(IY). This algorithm works very well for similar sinusoidal signals such as audio and video but as for the data analyzed in this research it is necessary to impose constraints on the algorithm. Figure 66 shows the original signal (blue) and the warped signal (red), the warped signal is modified too much (up to $\tau_d = 400s$) which is not acceptable because it would mean that the MFE response has a delay of 400s. Thanks to the cross correlation analysis it has been discovered that the τ_d is between 0s and 150s so a maximum delay of two minutes (120s) is set as the warping limit, which leads to the curve on the bottom left in Figure 66, Figure 67, Figure 68, Figure 69.

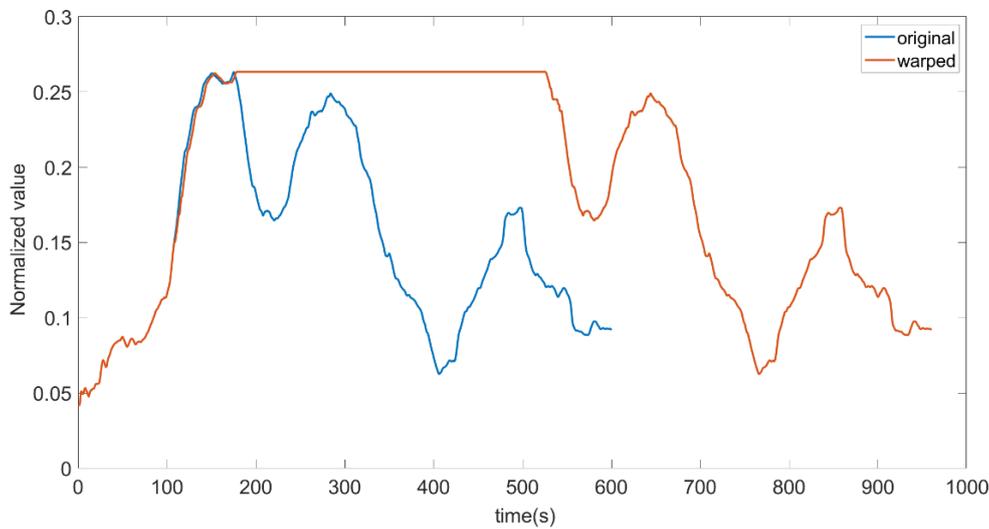


Figure 66: DTW without constraints

At this point the last problem has to be solved, the DTW function shifting algorithm. The figures at the bottom left shows how the curves, now with a physically meaningful delay, sometimes still are in delay and this leads to a low correlation coefficient. This happens because the function uses the Euclidean distance between the i -th point of curve 1 and the $i-j$ -th point of curve 2 to determine the warping path. The Euclidean distance is the distance between two points, i.e. the measure of the segment having at its extremes the two points x_{ik} and y_{ik} .

$$d(i) = \sqrt{\sum_{k=1}^n (x_{ik} - y_{ik})^2} \quad (9.9)$$

Unfortunately sometimes the minimizations of the Euclidean distance between curves does not lead to a reduction of the phase shift. This problem is solved by analyzing the slope of the curves, i.e. making the DTW of the derivatives of the Taskload and AU curves obtaining the warping path IX_der and IY_der.

```
AU_der= diff(AU)./diff(time);
Task_der=diff(task)./diff(time);
[Dist_der,ix_der,iy_der]=dtw(AU_der,task_der,240)
```

The warping paths obtained are then used for the phase shifting of the amplitude of the curves.

So the graphs at the bottom right of Figure 67, Figure 68 and Figure 69 are obtained.

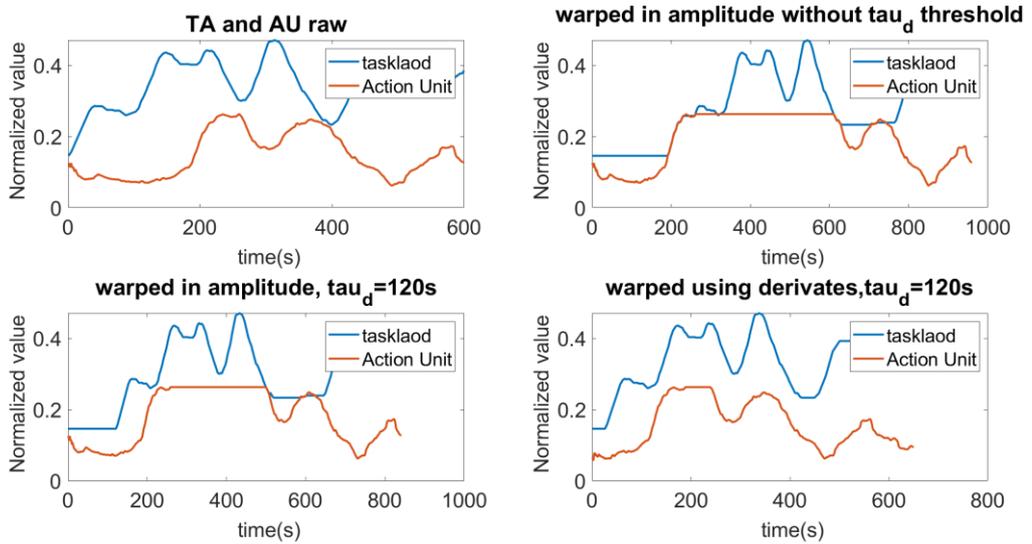


Figure 67: AU6 derivate warping

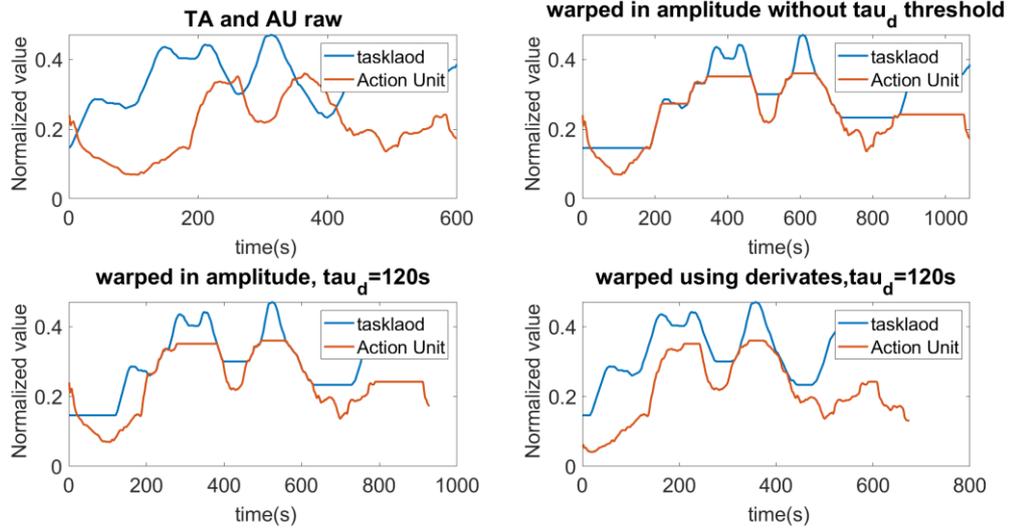


Figure 68: AU12 derivate warping

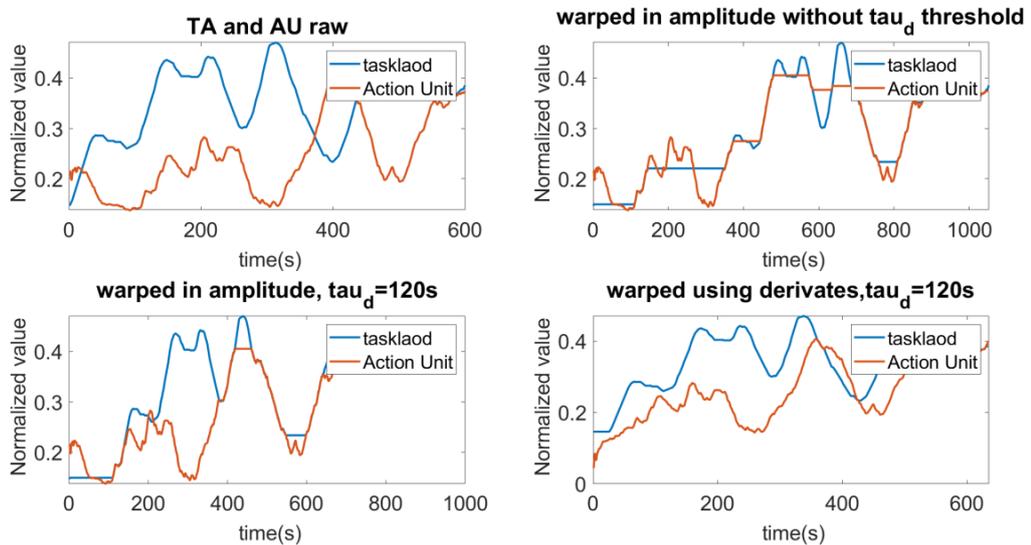


Figure 69: AU15 derivate warping

This method gives good results if the curves have a relation of at least 35%, otherwise it generates excessive deformations but it is not a problem because the minimum correlation threshold accepted in this analysis is 60%.

As has been explained in the previous sections the Taskload and AU are mediated with a sliding window, the first is set to 50s instead the second must be variable because each individual has its own facial expression and muscle contraction magnitude. The correlation coefficient between the two curves is evaluated for a $window_{AU}$ of variable dimension with values from half of the minimum $window_{TA}(25s)=12.5s$ to the double of the $window_{TA}(50s)$ that is 100s. The size of $window_{AU}$ is chosen in function of the variance of $window_{TA}$, if $window_{TA}$ has a high variance in time a smaller $window_{AU}$ is adopted, otherwise a bigger one. The correlation coefficient of conducted analyses on the six participants are shown in Table 13. In the table are reported the AUs that show the greater correlation with the Taskload for the three phases and the six analyzed participants. Data are presented for all three phases because it was found that the stimulation of the facial muscles of an individual can change over time so for example the AU6 can be an

excellent indicator for phase 1 and phase 2 but in phase 3 the AU15 shows a better correlation.

Participant 1	AU6	AU9	AU12	AU15	AU17	AU25
PHASE 1	0.83	0.86	0.81	0.77	0.94	0.77
PHASE 2	0.55	0.60	0.85	0.94	0.22	0.67
PHASE 3	0.45	0.26	0.72	0.56	0.32	0.65
Participant 2	AU6	AU9	AU12	AU15	AU17	AU25
PHASE 1	0.77	0.70	0.88	0.22	0.35	0.87
PHASE 2	0.69	0.57	0.24	0.87	0.86	0.65
PHASE 3	0.68	0.23	0.71	0.68	0.39	0.67
Participant 3	AU6	AU9	AU12	AU15	AU17	AU25
PHASE 1	0.68	0.17	0.83	0.75	0.20	0.69
PHASE 2	0.46	0.65	0.92	0.67	0.24	0.82
PHASE 3	0.71	0.78	0.72	0.73	0.85	0.87
Participant 4	AU6	AU9	AU12	AU15	AU17	AU25
PHASE 1	0.65	0.68	0.25	0.45	0.68	0.66
PHASE 2	0.74	0.73	0.65	0.23	0.57	0.72
PHASE 3	0.59	0.59	0.47	0.13	0.83	0.69
Participant 5	AU6	AU9	AU12	AU15	AU17	AU25
PHASE 1	0.97	0.94	0.95	0.87	0.75	0.92
PHASE 2	0.45	-0.17	-0.55	0.56	0.53	-0.67
PHASE 3	0.77	0.70	0.28	0.67	0.78	0.68
Participant 6	AU6	AU9	AU12	AU15	AU17	AU25
PHASE 1	0.45	0.68	0.69	0.66	0.23	0.74
PHASE 2	0.67	0.87	0.67	0.81	0.45	0.81
PHASE 3	0.53	0.55	0.45	0.64	0.33	0.58

Table 13: Correlation coefficient AUs-Taskload

Table 13 shows that AU6, AU15 and AU25 are the most reliable with a correlation bigger than 60% in the 72% of cases.

These experiments show that one of the goal of the research as been achieved, mouth AUs can be used as a Workload indicator so integrated in the sensor network.

It is fundamental to note that, analyzing the graphs, in some cases a good correlations is visible even for correlation coefficient between 50% and 60%, this happen because the developed algorithm needs improvements to be adapted to the individual. For this reason subsequently the definition of a protocol for the adaptation of the software to the operator will be analyzed in Protocol chapter.

The relationship between FEs and other psycho-physiological parameters is now analysed in order to assess whether FEs can provide an independent evaluation parameter that can be added to the others to confirm the workload estimation.

10 Psychophysiological parameters

10.1 Eye features

The relationship between eye features and cognitive states has been analyzed in numerous studies that have shown a close relationship between these psychophysiological parameters and the cognitive state. [4, 9, 40, 79-82]

Blink Rate(BR) and Visual Entropy(VE) are the two main parameters that are used. As anticipated in the previous chapters in this analysis the facial expressions of the mouth are evaluated because they are not involved in blinking and so they can provide an independent evaluation parameter to prove a possible variation of the cognitive state detected by the Eye sensor.

The blink rate represents the variation over time of the number of blinks. It has been demonstrated that a reduction in the blink rate means an increase in attention that in extreme cases can lead to attention tunnelling, characterized by very low BR. The inverse of the BR (IBR) is then analysed, so an increase of IBR means an increase in attention and concentration. Phase 2 and Phase 3 are characterized by a major workload increase so are now analyzed. Figure 70 and Figure 71 show the normalized Blink Rate and the AUs that have shown a higher correlation with Taskload: AU6, AU15 and AU25. They show that after each Taskload a peak of IBR and AUs appear with the FE-PD discussed in the previous chapter.

It has been noted that, in most cases, during an increasing effort in the cognitive process to solve a task, the blink rate is reduced while the underlying contraction of the muscles at the sides of the mouth increases. Both Figure 70 and Figure 71 show how AU15 is closely related to the blink rate.

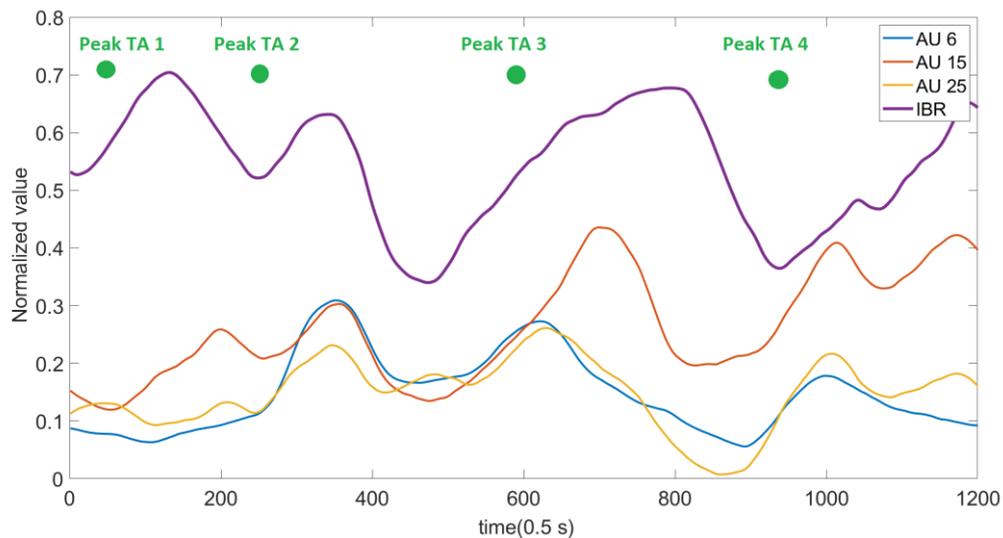


Figure 70: Blink Rate Ph2

In Phase 2, AU15 has a trend very similar to IBR and AU25 (lower lip muscles, *Depressor Labii*) presents a remarkable similarity but smoother, this is due to the fact that the contraction of the muscles linked to AU25 leads to open the mouth, a phenomenon that in facial mimicry has less magnitude unless it is voluntary or due to a high effort. In fact, the contraction of the lower lip that leads to the opening of

the mouth is due to greater cognitive effort and it is essential to point out that the participants were in a professional environment and knew they were being monitored, so this may have reduced their facial expressivity.

AU25 therefore presents peaks when a cognitive effort is prolonged over time and this is clearly visible in the first two peaks of phase 2.

Figure 46 shows how the first Taskload peak (Peak TA1 in green in Figure 70) is generated by a considerable increase of Taskload whereas the second peak (PeakTA2) appear at an already high Taskload level and therefore a prolonged effort that generates an AU25 increase. The same reasoning can be done for AU6 because also the increase in cheeks contraction occurs when the Participant's increase in concentration lasts over time.

AU6 (*Cheek Raiser* or *Orbicularis oculi, pars orbitalis*) can be considered as a parameter that identifies a lasting effort over time because a reduction in the Blink Rate leads to a greater underlying contraction of the eyelid muscles, which in turn is transmitted to the cheeks by contracting them. AU6 and AU25 can be therefore considered long time response parameters. It can be therefore concluded that AU15 and AU25 work in synergy during the increase in concentration, AU15 contracts more quickly giving the first indication of increase in Workload and if the effort is persistent over time even AU25 increases whereas AU6 is the consequence of ocular focalization so can be used to prove a detected Workload variation by the Blink Rate.

In addition, the lowering of the lower lip(AU25) leads to widen the lowering of the sides of the mouth (*Depressor anguli oris* or *Lip corner depressor*) amplifying the magnitude of AU15.

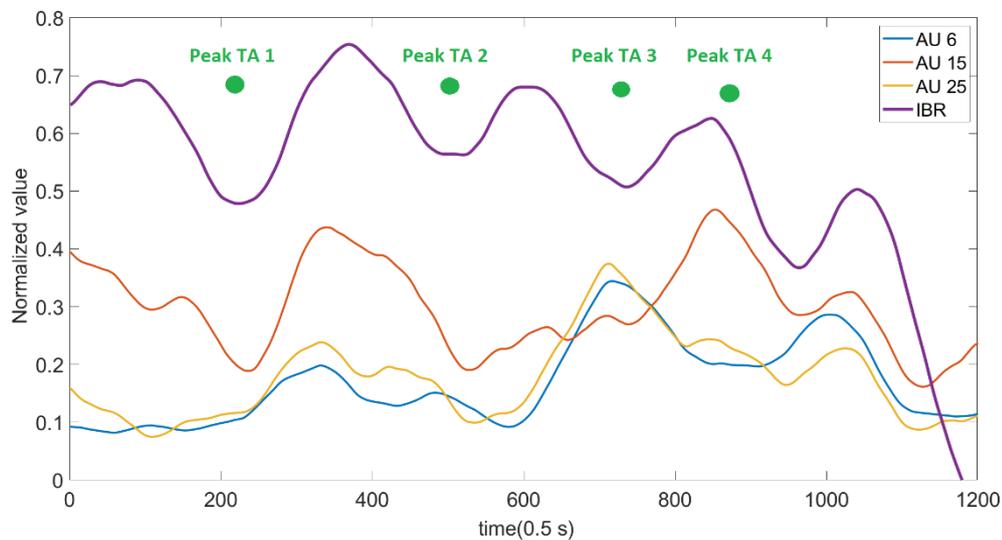


Figure 71: Blink Rate Ph3

The same phenomenon can be seen in Figure 71 for phase 3 where each PeakTA is followed by a peak of IBR and AUs. Again analyzing Figure 46 it can be seen how Peak TA2,3 and 4 occur during an enduring cognitive effort increase, AU6 and AU25 do not show the first PeakTA2 which instead AU15 shows, even if small, but they increase reaching a peak for PeakTA3.

It can be therefore concluded that AU15, AU25 and AU6 work in synergy during the increase in concentration, AU15 contracts more quickly giving the first indication of increase in Workload and if the effort is persistent over time even AU25 and AU6 increase with a certain delay obviously variable depending on the individual.

10.1.1 Visual Entropy

Visual Entropy (VE or H) has been defined in the layer chapter and can be broadly defined as the randomness of fixations.

The value of entropy is high when randomness in scanning patterns is high, whereas, the value will be reduced when less fixation transitions are employed as well as steady fixation patterns are formed together with systemically gaze patterns are also observed. A low entropy can be interpreted as higher attention level of the operator, as well as high workload is suggested by an increases of dwell time and reduces in fixations, since higher workload demands trigger a restricted eye movement with narrower range of fixate areas and more time is required to focus on area with higher priority task. [83]

Again Figure 72 and Figure 73 show relationships especially with AU15, in this case the relationship is a little less marked but the curves have the same trend. We expected to find this relationship because although BR and VE are two different parameters, they are very close to each other because when an operator focuses on something both IBR and IVE increase, obviously with different intensity but with a similar general trend.

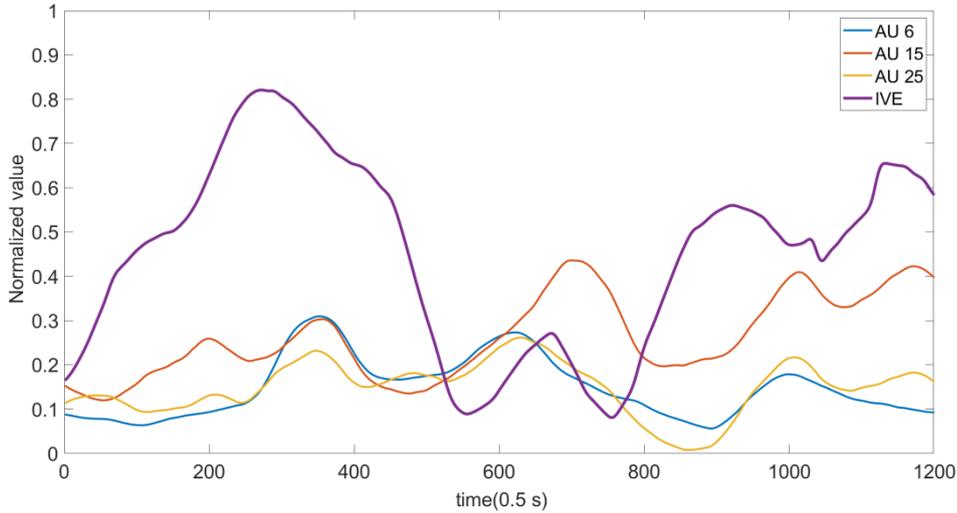


Figure 72: VE Phase 2

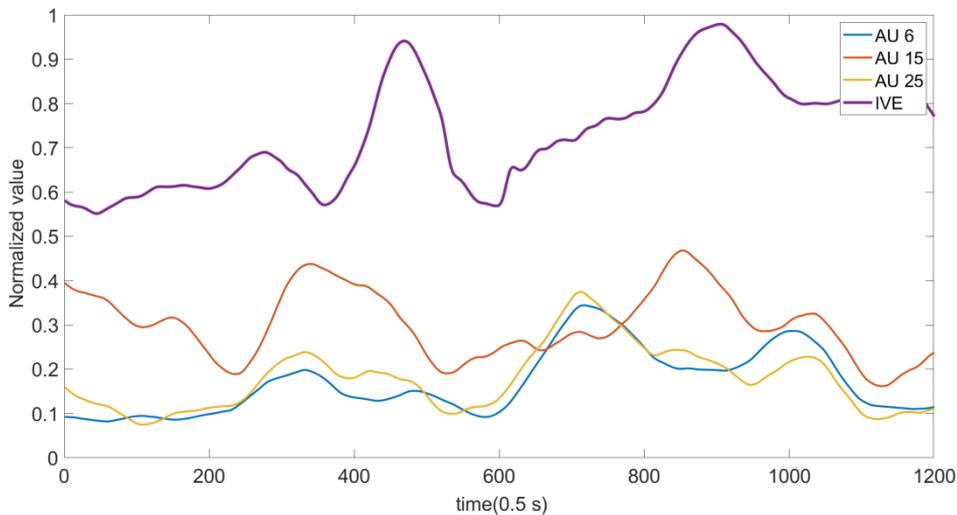


Figure 73: VE Phase 3

The same considerations made for the BR can be extended for the VE but analyzing all the experiments it has been noticed that the relationship between VE and AU is less generalizable and less accentuated compared to the BR. In spite of this anyway there is a relationship between VE and AUs to determine the cognitive state. AU15 in Figure 72 and Figure 73 shows a similar general trend with some phase shifts of the peaks due to the FE-RL. In this case a maximum FE-RL of 150s is not accepted because both VE (or BR) and AU are physiological responses delayed respect to the stimulus (Taskload). In fact, the highest phase shifting (60s) occurs between the first peaks of AU15 and IVE. For all participants the phase shifting between eye sensor and AUs values does not exceed 70s. In general VE has a less clear and harder relatable trend to the Taskload and AUs than BR.

The Taskload and the Eye features analysis have led to a potential applicable interpretation of the Face response, AU15 varies readily after a stimulus (Taskload increase) with a maximum delay of a few seconds while AU6 and AU25 significantly increase if the cognitive effort lasts over time and they are characterized by a maximum FE-RL of 130s.

This concept is now applied on the relationship between AU and Breathing Rate and Heart Rate.

10.2 Bioharness

The data from the cardiorespiratory sensor are now analyzed. In the Sensors section it has been explained that this sensor provides three main parameters: the Heart Rate(HR), the Short term HRV characteristic (SD1) and the Breathing Rate(BHR). Numerous researches have demonstrated the relationship between these parameters and the mental workload. [4, 6, 7, 9, 41, 83-88]

An increase in workload leads to an increase in Heart Rate, SD1 and a Breathing Rate(Breath/min) decrease, for this reason the analyzed parameter is the inverse of the BHR (IB). In fact, when a person concentrates to solve a task tends to block the breath or slow it down, an increment of workload determines then an increment of IB. The Heart Rate is a parameter that generally varies with the WL but is strictly dependent on the subject because a trained person to have a lower HR and a slower variation over time. In these experiments the training level of the participants has not been evaluated so it is accepted that the response can be very different in terms of amplitude and Breathing Rate Response Latency(BHR-RL) from individual to individual. The BHR-RL can reach few minutes whereas the FE response is usually prompter so phase shifts between the two curves are expected. Phase 2 and phase 3 for Participant 1 are again analyzed. Figure 74 show that the Heart Rate (beats/min) has a very smooth trend and is therefore difficult to relate it to the AUs but on the contrary IB shows interesting correlations.

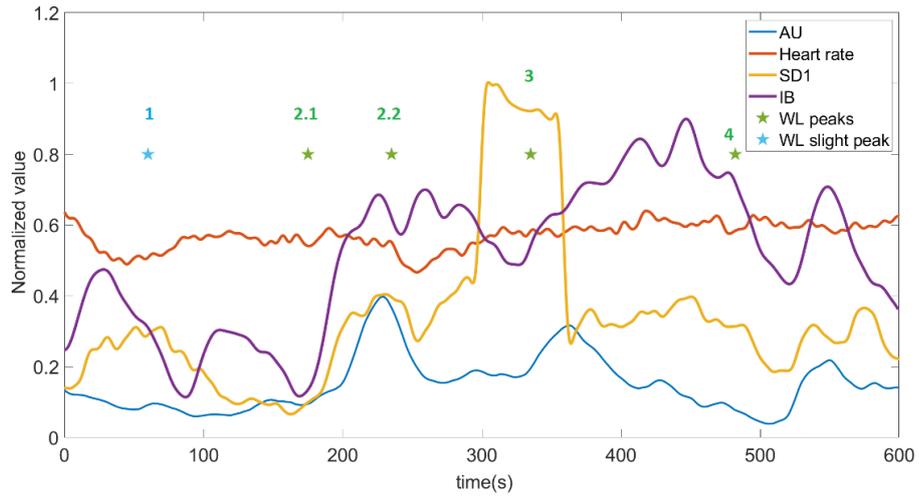


Figure 74: Participant 1 Phase 2 Bioharness AU6

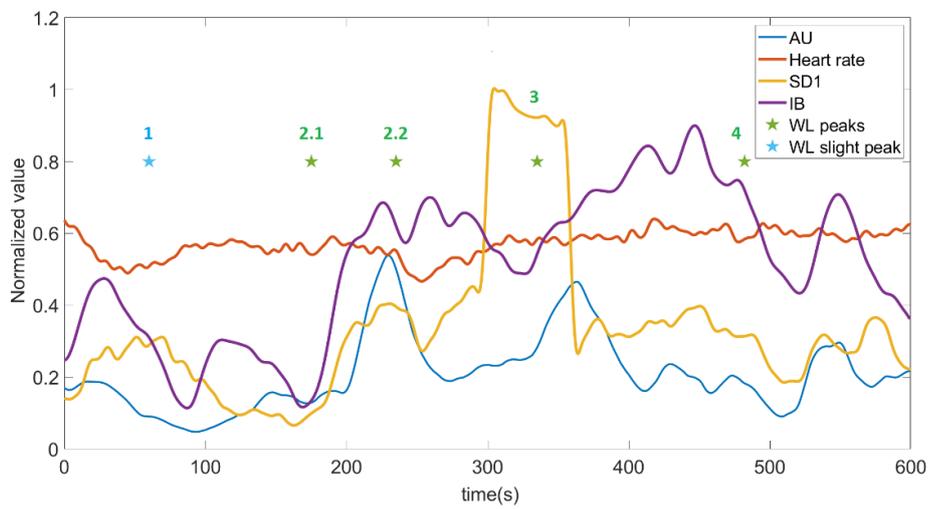


Figure 75: Participant 1 Phase 2 Bioharness AU12

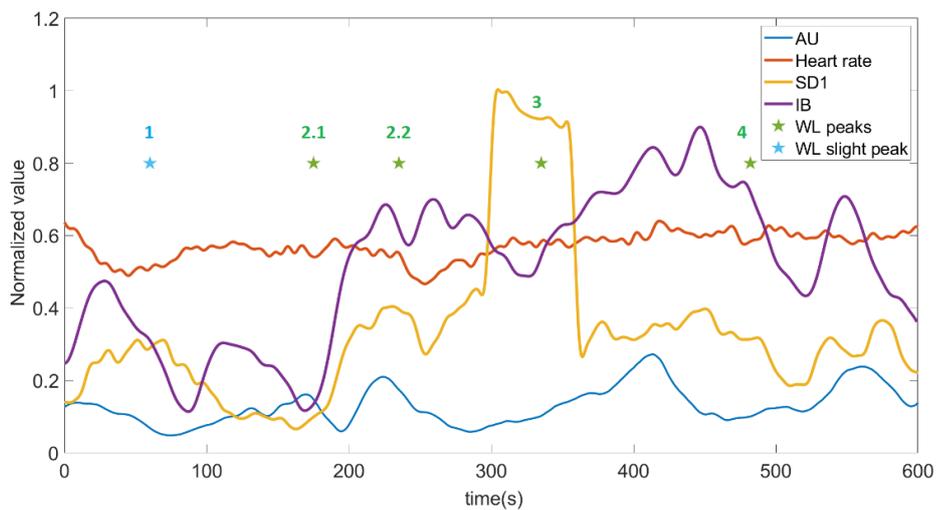


Figure 76: Participant 1 Phase 2 Bioharness AU15

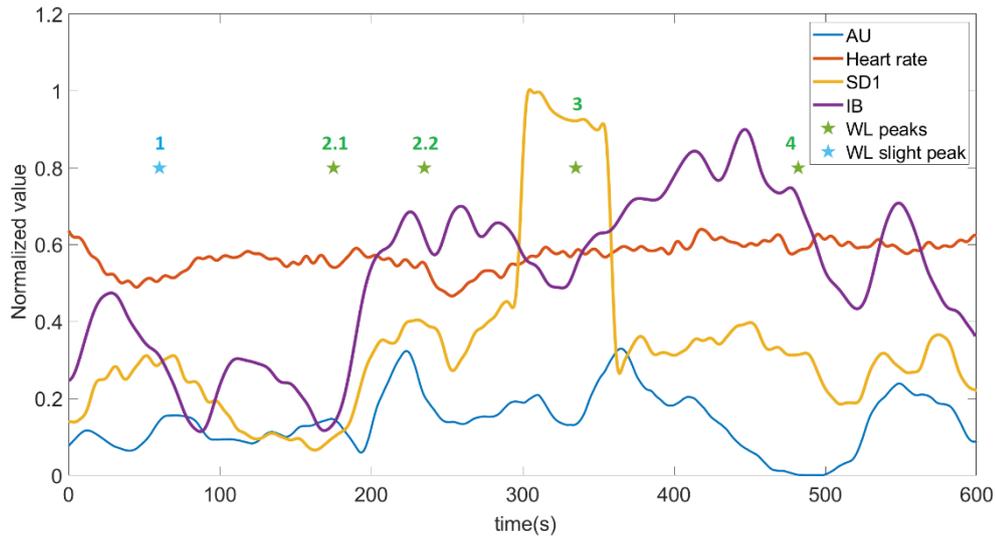


Figure 77: Participant 1 Phase 2 Bioharness AU25

The stars in the figure above represent the WL peaks, the first blue star (1) identifies a slight peak while the others (2.1-2.2,3,4) represent the main peaks. The first peak on the left of IB is relative to the previous phase while the second is relative to the peak of WL 1. This peak is not highlighted by AU6,AU12 and AU15 instead AU25 has a slight peak at 80s in conjunction with the blue star. The other main peaks(green) are instead highlighted by all the AU shown in the figures. The first two green peaks(2.1 and 2.2) are part of a lasting increase of WL that induces a single psycho-physiological response. The WL 3 peak induces a promptly response of AU6,AU12,AU25 and SD1 and then also AU15 and IB show the related peaks, AU15 about 40s later whereas IB about 100s later. Finally, the last WL peak(4) induces an increase of all parameters with about the same response latency.

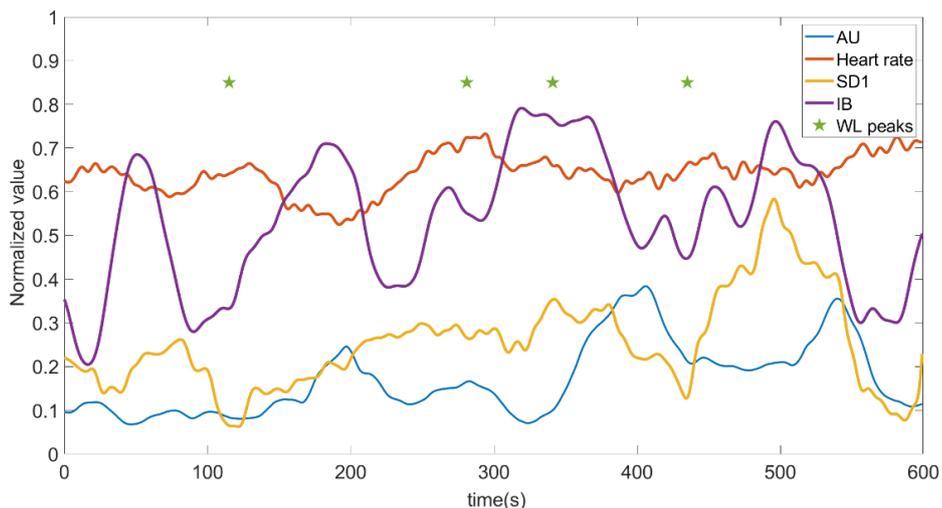


Figure 78: Participant 1 Phase 3 Bioharness AU6

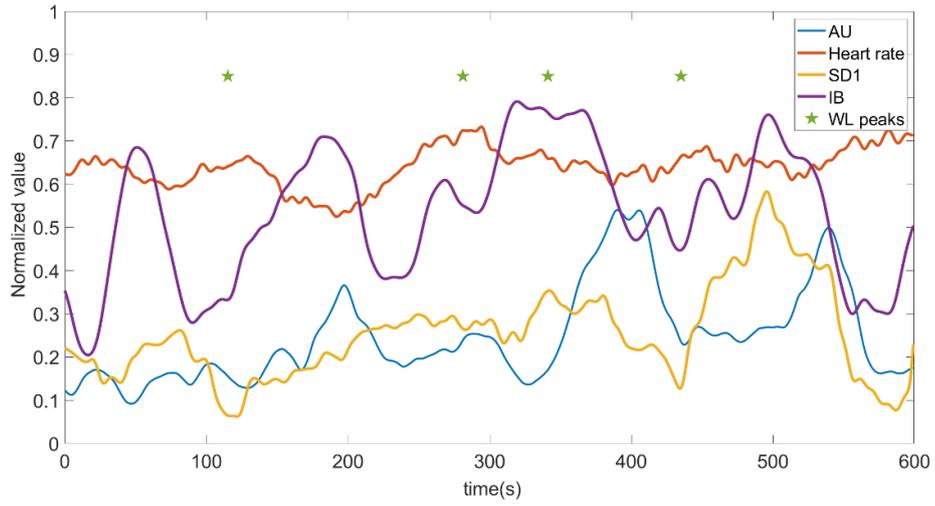


Figure 79: Participant 1 Phase 3 Bioharness AU12

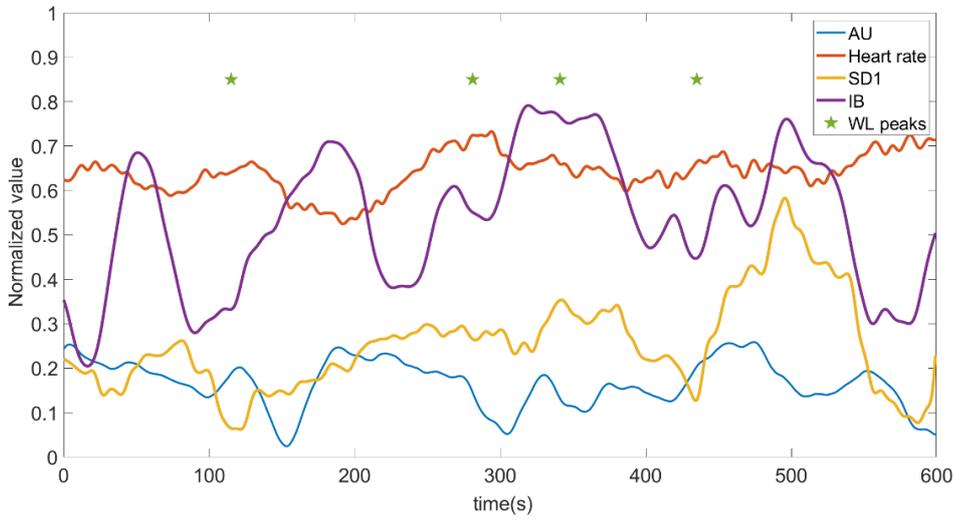


Figure 80: Participant 1 Phase 3 Bioharness AU15

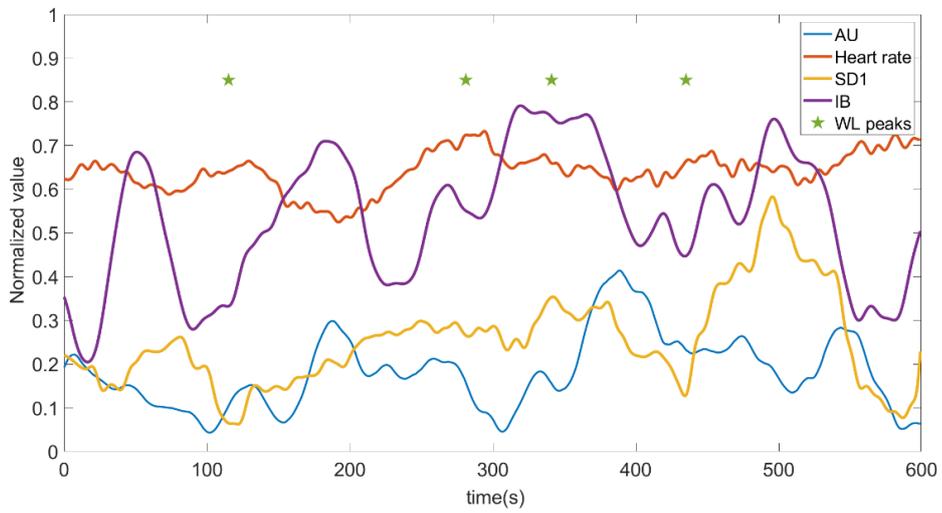


Figure 81: Participant 1 Phase 3 Bioharness AU25

Figure 75 shows again that the psycho-physiological response of IB,AU6,AU12 and AU25 is strictly related to WL peaks, but in this case SD1 does not show all the peaks in fact the first WL peak induce the second IB peak and the first of the AUs but SD1 doesn't show a remarkable change.

A comparison between phase 2 and phase 3 of the participant 1 shows that in all the phase 3 the FE-RL is greater than the IB-RL,

However, this phenomenon does not occur for the other participants, each has its own FE-RL which in some cases remains approximately constant throughout the experiment whereas in others it varies from phase to phase. The maximum lag between the facial response and the breath response in participant 1 is:

$$\Delta_{RL} = FE_{RL} - IB_{RL} = 69s \quad (10.1)$$

This variability of the FE-RL can be related to the WL derivate, the higher the WL variation the faster the FE-RL is. In fact the last 3 peaks of Figure 75 are very gently WL variations.

Further research will demonstrate this interpretation.

The 80% of the participants show a clear correlation between IB and AU25, obviously in some it is more marked whereas in others less but the correlation coefficient is generally higher than 0.65. AU6 and AU12 often have peaks at the IB peaks but in general their variation show more noise and lower muscle contractions than AU25. AU15 presents partial correlations in 70% of the participants and further analysis could be prove its potential usefulness. SD1 in general describes WL well but in some cases it does not detect some WL peaks as in the case of Figure 75.

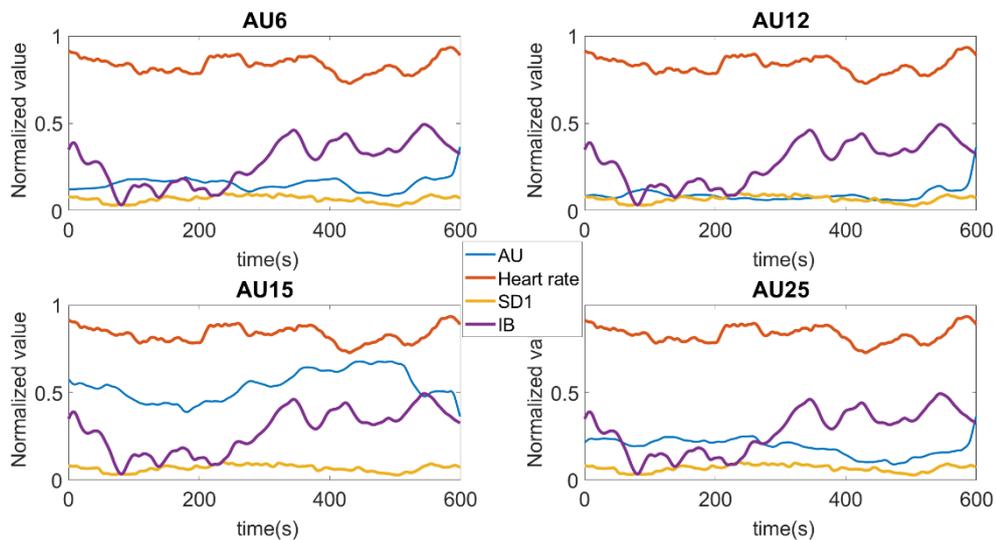


Figure 82: Participant 2 Phase 3 Bioharness

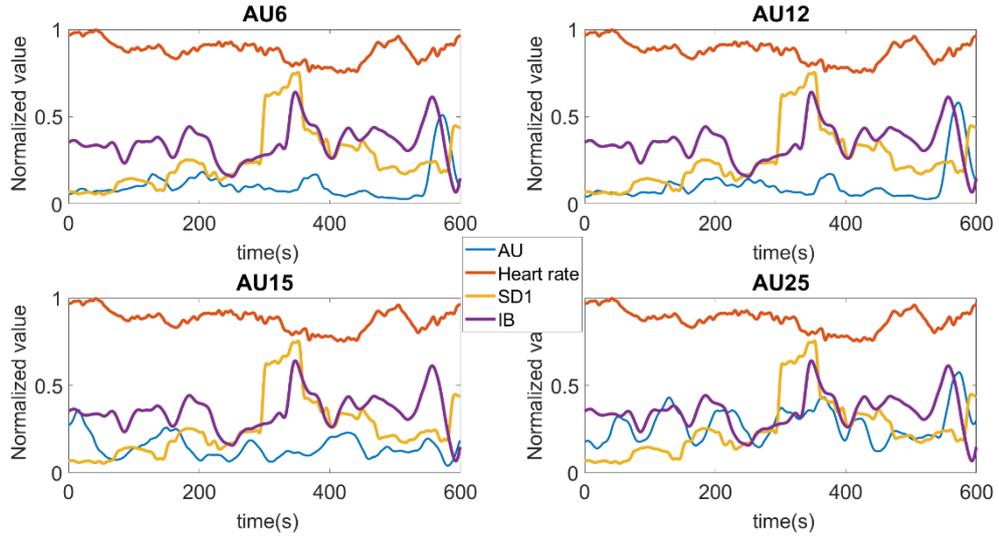


Figure 83: Participant 4 Phase 3 Bioharness

Figure 82 and Figure 83 show the data of participants 2 and 4 for phase 3 which is the one with the highest WL, Figure 82 shows the case in which there was less correlation whereas Figure 83, as for participant 1, shows an evident correlation between AU6,AU12,AU15 and AU25 with obviously with different FE-RL and IB-RL.

It is interesting to note that the fundamental aspect is that sensors are able to detect when there is a significant peak of WL, instead minor WL variations have a secondary importance. All substantial variations of WL have been detected by the FE sensor and the most critical cases where the WL suddenly increases the FE show such variation, sometimes even faster than SD1 and IB.

It can therefore be concluded that again in this evaluation AU25 is the most potentially promising FE parameter and given the similarities with the monitored features by the Bioharness sensor, the FE monitoring could be introduced in the sensor network. AU25 can be used as a WL evaluation parameter whereas AU6, AU15 and AU12 can be used as secondary parameters to confirm the WL estimation made byAU25, Breathing Rate and eye-sensing.

10.3 Electroencephalography

Numerous studies have demonstrated the relationship between the brain waves monitored by the EEG and the workload. [4, 10, 86, 88-93] In particular, the parameter evaluated is the EEG index explained in the sensors chapter:

$$EEG\ index = \frac{\theta_{F4+C4}}{\alpha_{O1+O2}} = \frac{\int_{4Hz}^{8Hz} f(\lambda)d\lambda}{\int_{8Hz}^{12Hz} f(\lambda)d\lambda} \quad (10.2)$$

The obtained pre-processed values are then related to the AUs making them smoother with the two matlab functions *movmean* and *smooth*. The first function averages the values in a sliding window of 20 seconds for the EEG index and 30 seconds for AUs. The *smooth* function uses a moving average which, unlike *movmean*, does not smooth the prominence of the original curve so *movmean* is used to reduce noise while *smooth* is used to extrapolate the general curves trend. Smoothing span for both EEG index and AUs is settled at 30 seconds. Finally, a

dynamic time warping of only 20 seconds was applied to correct any phase shift due to different *Brain Waves Response Latency* (BW-RL) and FE-RL.

Table 14 shows the correlation coefficient between the processed curves as mentioned above. Participant 5 monitoring presented some setting problems in fact it can be seen how the data show low CC values and phase 1 was reported instead of phase 3 because in the latter the sensor presented anomalies.

Participant 1	AU6	AU9	AU12	AU15	AU17	AU25
PHASE 2	0.41	0.70	0.17	0.68	-0.31	0.63
PHASE 3	0.69	0.85	0.61	0.70	0.18	0.80
Participant 2	AU6	AU9	AU12	AU15	AU17	AU25
PHASE 2	0.46	0.44	0.47	-0.10	0.039	0.54
PHASE 3	0.73	0.60	0.65	0.62	0.53	0.70
Participant 3	AU6	AU9	AU12	AU15	AU17	AU25
PHASE 2	0.43	0.61	0.25	0.71	0.57	0.68
PHASE 3	0.48	0.61	-0.45	0.70	0.65	-0.41
Participant 4	AU6	AU9	AU12	AU15	AU17	AU25
PHASE 2	0.78	0.63	0.79	0.26	0.31	0.69
PHASE 3	0.85	0.78	0.89	0.23	0.12	0.82
Participant 5	AU6	AU9	AU12	AU15	AU17	AU25
PHASE 1	0.24	0.37	0.46	0.41	0.69	0.0049
PHASE 2	0.72	0.71	0.68	0.73	0.82	0.66
Participant 6	AU6	AU9	AU12	AU15	AU17	AU25
PHASE 2	0.26	0.32	0.35	0.78	0.79	0.75
PHASE 3	0.44	0.75	0.31	0.77	0.66	0.80
CC>60%	42%	75%	48%	66%	33%	75%

Table 14: Correlation coefficient between EEG index and AUs

Again AU25 is the one that shows the best correlation and the table shows how the correlation between EEG and AUs increases as the WL increases, in fact higher correlation coefficient values have been calculated for phase 3.

This phenomenon has also been noticed for other sensors and shows an important aspect. AUs monitoring increases his reliability on evaluating workload level as the Workload increases. This may be due to the fact that when the workload is low the person can get distracted and have different behaviours but when the operator increases his concentration and cognitive effort this leads to more determined and quantifiable psycho-physiological responses.

The last row of the table shows the percentage of phases in which the CC was greater than 60%. In this case also AU9 showed interesting behaviours that will be further analysed in the ATM experiment to verify its potential relationship with EEG. Figure 84 and Figure 85 show the AU9 and AU25 for the participant 1 before

the dynamic time warping is applied. Although there are still normal phase shifts between the curves, it is possible to notice a close relationship between the trends.

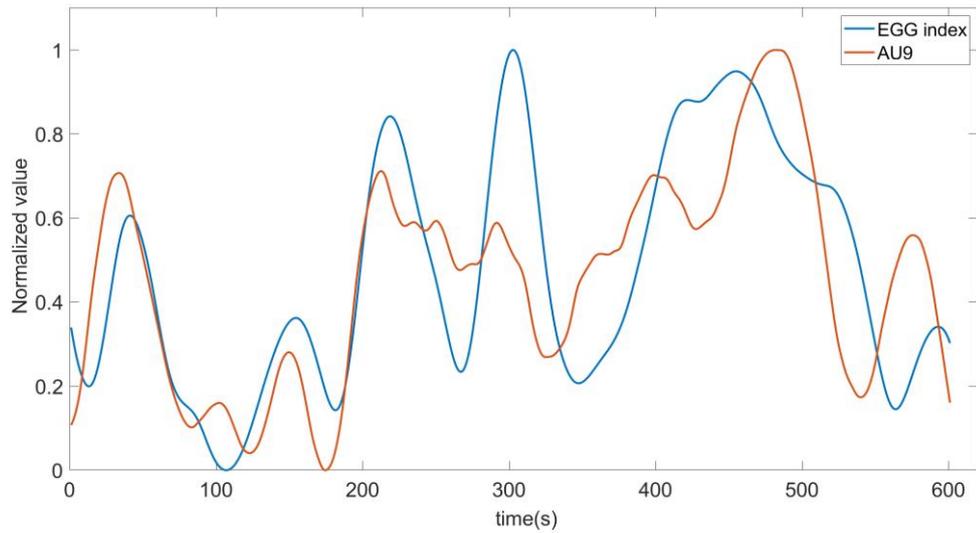


Figure 84: Participant 1 Phase 3 AU9

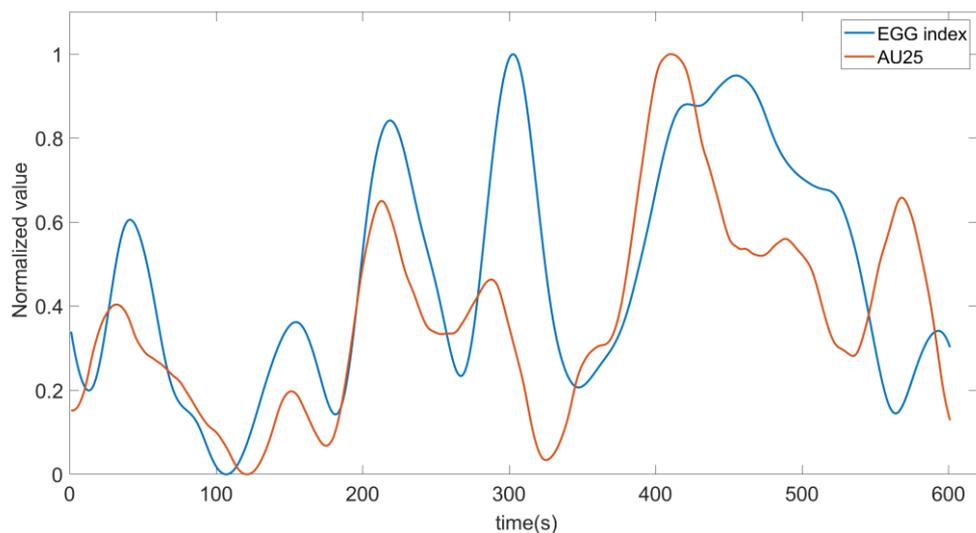


Figure 85: Participant 1 Phase 3 AU25

Participant 2 shows close relationships between the curves but a smaller *movmean* window had to be adopted than the other participants because the magnitude of the facial expressions for this participant is very small compared to the others and is therefore reduced smoothing. This again shows the need to develop a protocol to define a data pre-processing adaptation model on the each specific operator.

11 ATM experiment

The OTM experiments allowed us to have greater knowledge about physiological facial reactions following stimuli in a given environment so another experiment was conducted in collaboration with THALES Australia to verify the considerations previously made. This experiment concerns Air Traffic Management (ATM) in the Terminal Manouvring Area (TMA). The software and hardware components of the AIAS lab used in this research are now presented. Figure 86 shows in the foreground the two radar workstations with industry-grade Esterline MDP-471/4 LCD tactical situation displays. The ATM station in the background comprises an immersive 270° control tower simulator. The external visuals for the 270 degrees Air Traffic Control (ATC) Tower Simulator can be supplied by either Lockheed-Martin's Prepar3D v3.4 or X-Plane 10. The simulators are commercial-off-the-shelf products so C++/Python/Javascript and Matlab were used to develop of suitable interfaces to integrate them within the laboratory network. The public available adopted traffic generator software are openScope, Hoekstra's BlueSky, Albatross Display, Eurocontrol's e-DEP, Euroscope.



Figure 86: ATM simulation environment

The CHMI² system modules are presented in figure. The collection, processing and logging of sensor and simulation data is supported by a centralised server(HFE-Lab server). Testing the full CHMI² system required further stages of HMI design and development.

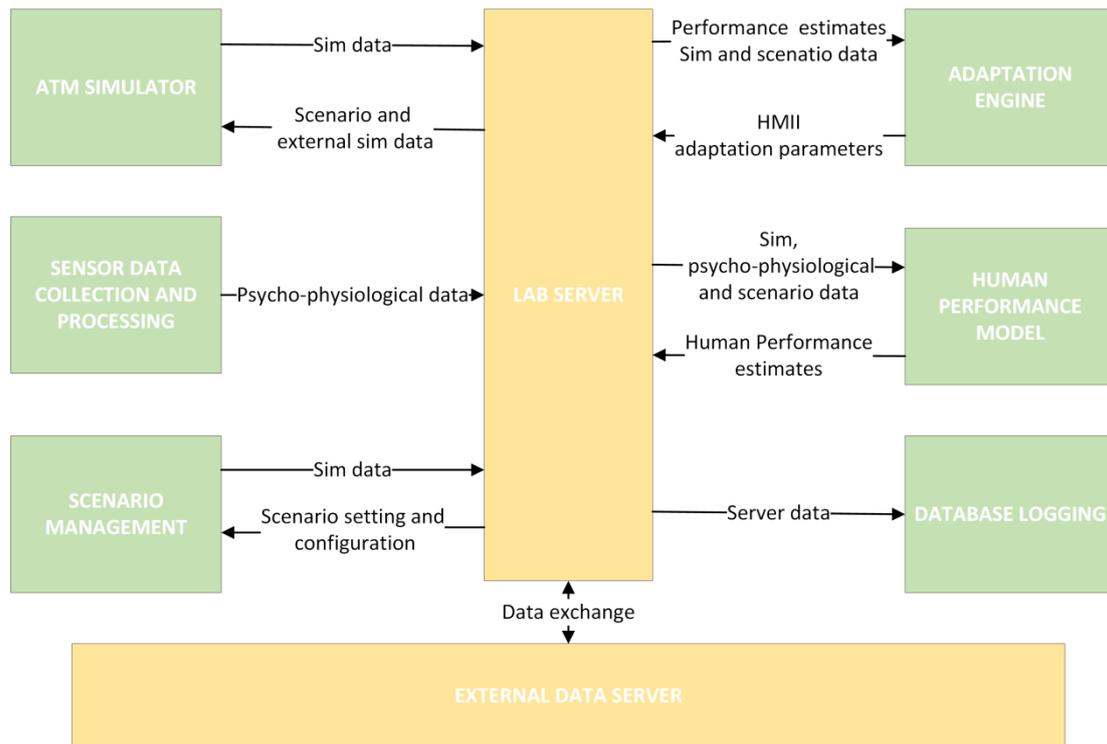


Figure 87: Hardware architecture

The modular architecture provides significant flexibility in the integration of different types of bio-sensors developing the client-side interface as well as the server-side thread and buffer. Modifications can then be made to other threads to read data from the new buffer.

The experiment lasts 45 minutes and is performed administered by a formally trained Air Traffic Controller (ATCo). The TMA ATM duties involved arrival traffic sequencing and spacing by mean of hybrid procedural and radar vectoring control, as well as deconfliction of arrival and departure streams. Only voice-based interactions with the simulated aircraft are considered. The Taskload is controlled by manipulating arrival and departure traffic throughputs and traffic density. In the middle of the scenario, there is a substantial increase of traffic amount (and complexity), which was devised to intentionally induce a noticeable change in physiological parameters. During the TMA ATM session, the participant must manually accept new arrival or departure traffic from upstream ATM sectors. When aircraft reach the prescribed handover conditions, the participant can release the aircraft to next sector. Table 15 shows TMA ATM tasks and subtasks considered in our study.

Task	Sub-task
Assigning deconfliction instructions to conflicting flights	<ol style="list-style-type: none"> 1. Predicting the conflicts in terms of both lateral and longitudinal separation 2. Evaluating the deconfliction resolutions in terms of level changes, path stretching or speed adjustments 3. Calculating path stretching geometry or speed adjustment magnitude and the associated time impacts 4. Assigning new speeds and altitudes
Maintaining separation	<ol style="list-style-type: none"> 1. Assign proper altitude considering vertical constraints and terrain 2. Assign heading and speed for general approach
Dynamic airspace / stream management	<ol style="list-style-type: none"> 1. Evaluate runway configuration and arrival/departure throughputs as defined by the airport tower and/or local restrictions 2. Evaluate weather trends and traffic inflows/outflows 3. Configure subsectors as departure-only, arrival-only or mixed (this determines which departure and arrival procedures will be allocated)

Table 15: Key tasks analysis for TMA ATM([17])

At this point the number of aircraft has been chosen as objective parameter to use as workload index because studies on ATC and ATM related to workload evaluation both through subjective evaluations (NASA-TLX) and evaluation through psycho-physiological parameters (EEG,Heart Rate,...) have shown a close relationship between traffic density and workload. [8, 13, 15, 17, 62]. The traffic density is defined as the number of aircraft divided by sector volume:

$$Traffic\ density = \frac{number\ of\ aircraft}{sector\ volume} \quad (11.1)$$

Simulations have shown that in some cases the operator has to perform a different number of tasks in equal number of aircraft because each aircraft could require slightly different tasks. It has also been decided to consider a parameter linked to the physical actions that the ATM operator performs. This parameter is the number of control input (operator mouse click) and assumes a lower weight in the definition of the objective workload but must be considered as will be explained below.

11.1 Objective parameters: number of aircrafts and control inputs

In this chapter the AUs that have shown the best correlations in ATM experiments are analyzed to verify the interpretations and concepts elaborated in the previous chapter.

In this experiment a smaller magnitude FE trend is expected, smoother and less noisy because the participant is an experienced ATM operator and therefore its physiological response tends to be more controlled and with less pronounced peaks because the individual is familiar with the interface and quick workload variations. As will be seen later this expectation is verified.

The same data pre-processing methodology is adopted. The objective is to verify that the AU25 is the best AU to estimate Workload and other Aus(AU9) can be used as secondary parameters to confirm what is captured by the other sensors and AU25.

In this case the objective parameter for workload estimation are no longer secondary tasks but the number of aircraft and control input (mouse click).

Researches in the ATM field have already demonstrated that the number of aircraft is closely related to workload. [57, 89, 94]

The experiment lasted 1969 seconds. The OTM experiments have shown that the greater the workload the greater the relationship between AUs trend and workload and the other psycho-physiological parameters so a time window between 550 and 1750 seconds is evaluated to exclude the first minutes when the number of aircraft is low and the operator is familiarizing with air traffic initial conditions. We do not consider the last 200 seconds because the number of aircraft drastically decreased and the operator was finishing the test session so the psycho-physiological parameters are not meaningful.

In some situations the number of airplanes increases but the operator has no more tasks to perform because the behaviour of the latter does not require additional work whereas in other situations there can be a lower number of airplanes but the ATM operator has more tasks to perform because there are problems to solve. For this reason we cannot say that the workload is directly proportional to only the number of aircrafts and therefore the number of control input has been considered. This parameter allows to consider so the number of actions that the controller physically carries out in the workstation.

Figure 88 shows the trends in the number of aircraft and control input. The two trends are very similar but sometimes the control input is in advance (the controller performs tasks before the number of aircrafts increase, points 1,2,3,4) or delay (the number of aircraft has decreased but the controller still has to perform tasks related to past aircraft, point 9).

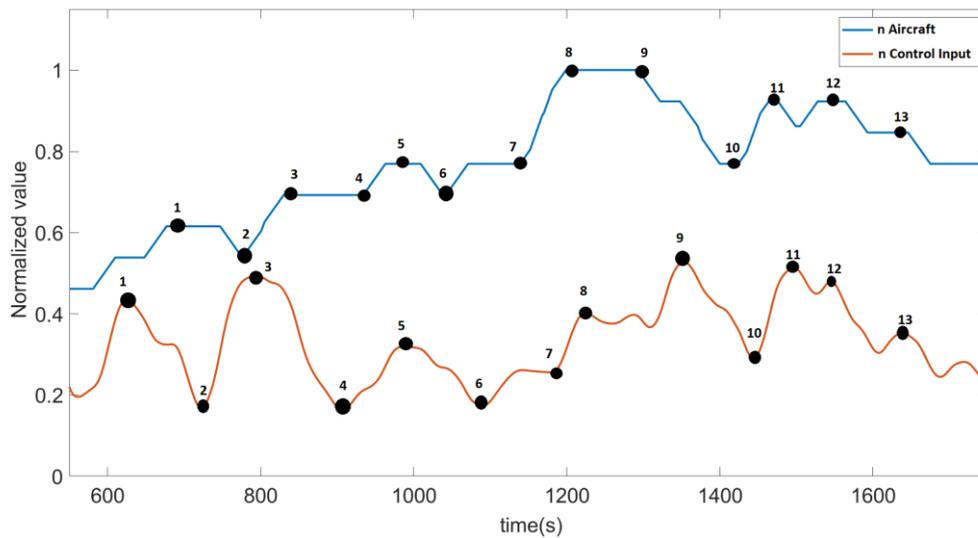


Figure 88: n aircraft and control input trends

For these reasons the workload has been defined as a weighted sum of number of aircraft ($nacft$) and number of control input(nci) as follow:

$$WL = 0.7 * nacft + 0.3 * nci \quad (11.2)$$

The workload trend in Figure 89 is therefore obtained. In this case a double time phase(1200 s) is analyzed compared to the OTM experiments(each phase lasts 600 s) so the maximum dynamic time warping is settled to 240 seconds, anyway less than 5 minutes which is the maximum accepted threshold usually used as Response Latency of psycho-physiological features.

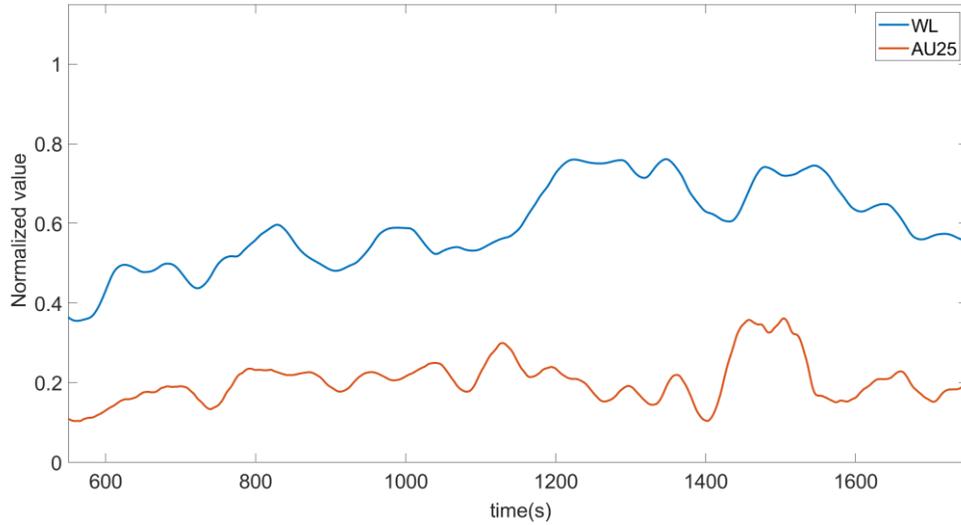


Figure 89: WL and AU25 trends

Table 16 shows the Correlation Coefficient between AUs and Workload.

	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU25
CC	-0.36	-0.52	-0.44	0.36	0.13	0.51	0.36	0.1	0.2	0.63

Table 16: Correlation Coefficient WL-AUs

In general in this experiment CCs are not very high but again the mouth AUs show the best relationship and the highest is for AU25.

It can be therefore concluded that lips parting is the main indicator of workload, it has been demonstrated in 7 experiments (6 OTM and 1 ATM). Future researches will be able to study in more depth the relationship between AU25 and WL by defining a prediction model of WL according to lips parting. The FE-RL for AU25 is 29s, which is a coherent value compared to what was obtained in OTM experiments.

AU9 also shows interesting relationships that have also been highlighted in the OTM experiment (Table 13). The data of the other sensors are now analyzed.

11.2 Eye features

Blink Rate and Visual Entropy of the participant are now evaluated in the time window 550-1750 seconds. As explained in the previous chapters we use the inverse of BR and VE. In the evaluation of the eye features of the OTM experiment we analyzed the time BR_RL and FE_RL deducing that AU6 has a response time greater than AU25. Performing the cross-covariance analysis the same result was calculated

in this experiment. In fact AU6 has a Δ_{RL} of 15 seconds while AU25 has 3 seconds with respect to the Blink Rate. Δ_{RL} is defined as follow:

$$\Delta_{RL} = FE_{RL} - IBR_{RL} \quad (11.3)$$

This results in a new potential discovery that further research may prove.

Figure 90 shows the relationship between IBR and AU25, the workload was also graphed to highlight the similarity between the curves and the different response delays of BR and FE.

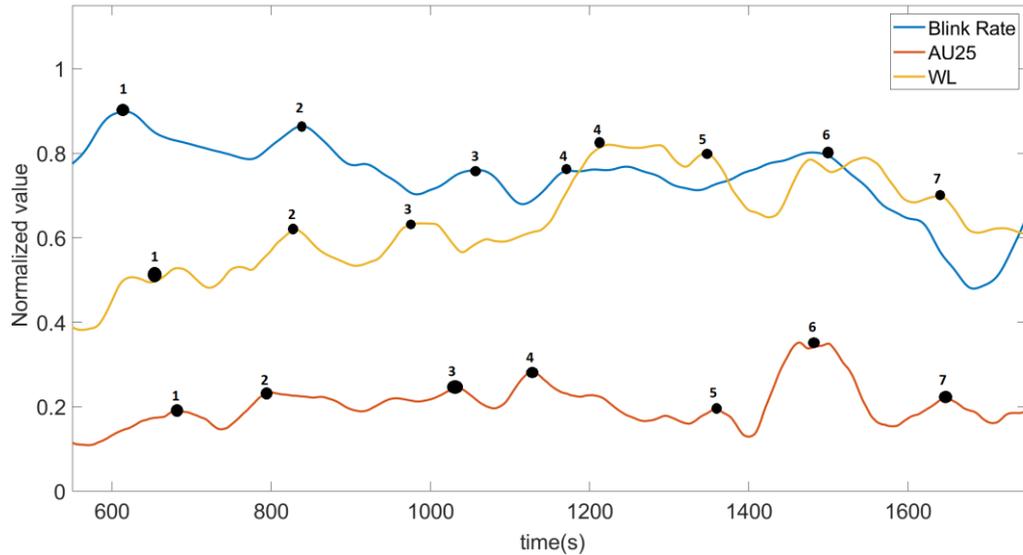


Figure 90: AU25- Blink Rate comparison

The figure shows how WL peaks are highlighted by both BR and FE, minor peaks 2 and 3 are not highlighted much by AU25 but after the drop after peak 3 it's easy to see how AU25 shows an increasing workload. After peak 4 the workload remains approximately constant and the BR shows a similar trend. AU25 on the other hand due to the physiognomy of the facial muscles as explained above tends to decrease. This is because the facial muscles of the corners of the mouth (*Depressor Labii*) tend to relax after a sudden contraction if the cognitive effort remains at the same level. Subsequently, however, the WL presents another peak(5) which is highlighted by AU25. Finally, peak 6 is clearly shown by both AU25 and BR.

Figure 91 shows that AU6 also has a very similar trend to BR but very smooth. It is important to note that it shows in a clear way the two major peaks of WL(4-5,6) consistently with BR so also in this experiment the relationship between AU6,AU25 and BR has been demonstrated.

In this experiment AU15 does not show particular relationships, it shows the first major peak (4) and the second (6) but not in a clear way.

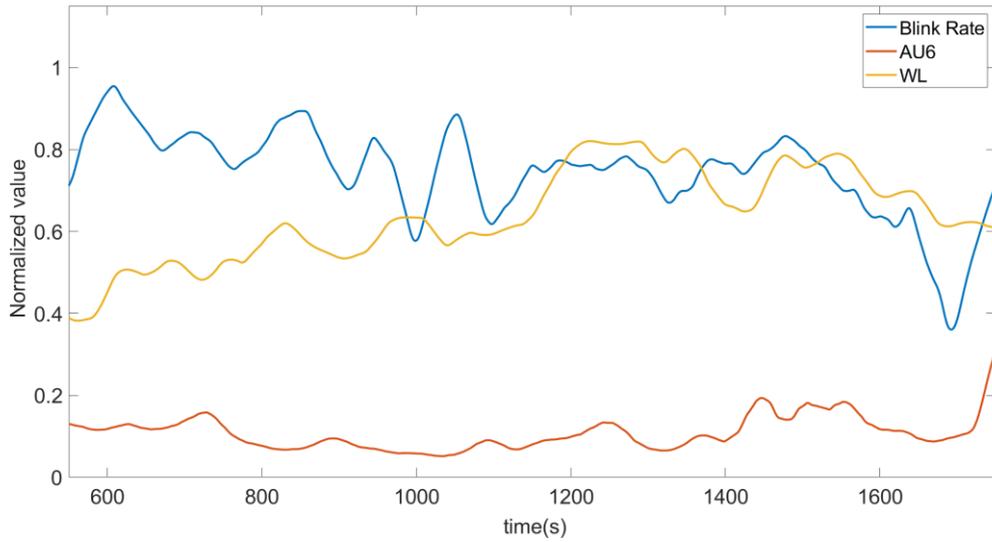


Figure 91: AU6- Blink Rate comparison

11.2.1 Visual Entropy

The inverse of the VE is now evaluated. Figure 92 shows how once again AU25 presents the highest correlation although not for each peak. It is important that there are relationships for the main peak because it is the one that during the online implementation will give the main feedback of adaptation of the automation level and it is highlighted.

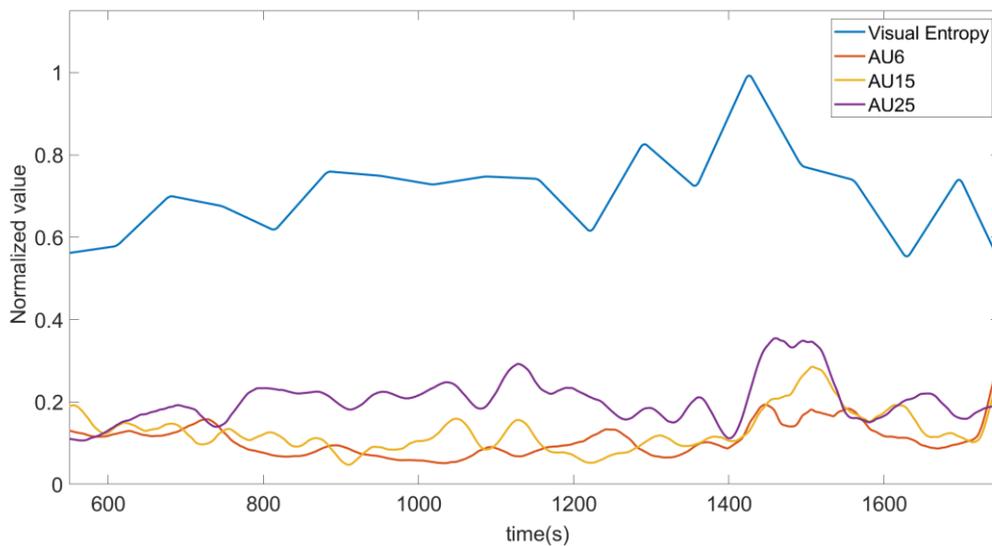


Figure 92: AUs-VE comparison

In this case the phase shifting between FE and VE, for the major peak, is about 35s so slightly longer than BR but consistent with the operator's RLs calculated for the other sensors.

11.3 Cardiorespiratory features

By relating the various parameters SD,HR and BR (inverse of BR) a correlation was found only with HR because BR.

Table 17 shows the correlation coefficient with the Aus, the time warping is carried out with the same modalities of the OTM experiment.

	AU6	AU9	AU12	AU15	AU17	AU25
Heart Rate	0.61	0.65	0.65	0.58	0.72	0.81
Breathing Rate	0.59	0.62	0.60	0.75	0.48	0.36

Table 17: Correlation Coefficient AUs-Bioharness

The figures show how the breathing rate (IB) is smoother and shows a significant and basically constant increase from 1200s when the workload presents the first major peak. Given the very smooth IB trend, it is difficult to find direct relationship between IB and AUs, but both show an increase in workload after 1200s. In the OTM experiment AU6,AU12,AU15 and AU25 were the ones with the highest correlation and again AU15 and AU25 show a close correlation with IB. AU6 and AU12 had presented partial reports and therefore they were proposed as indicators of secondary importance, in this experiment again the relationship between these AUs and the bioharness parameters is not clear and noisy therefore they are not considered reliable. AU15 instead had shown interesting reports that in this experiment happen again.

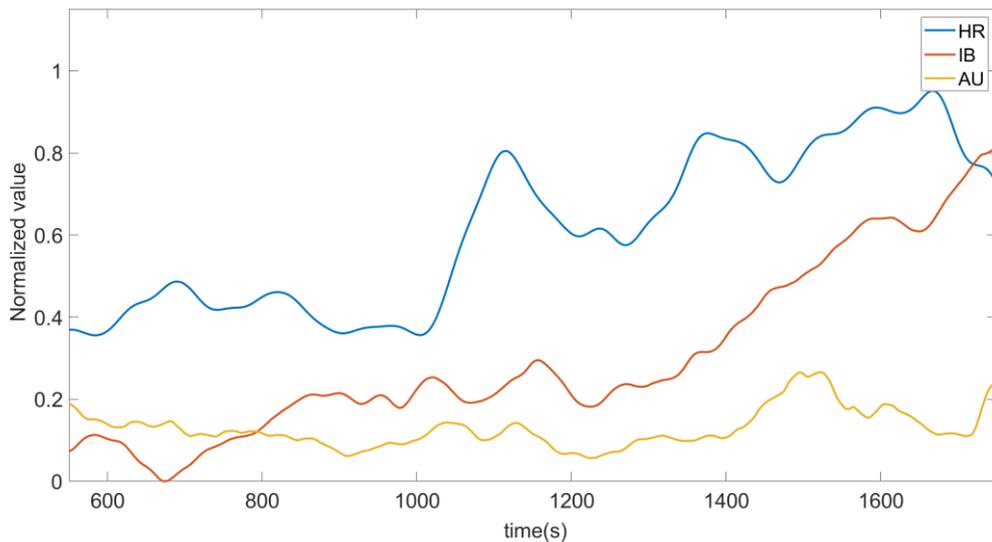


Figure 93: AU15 trend

Figure 93 shows that AU15 trend is very similar to IB, both curves grow around 1200s but obviously after a certain contraction the same phenomenon found in the OTM experiments occurs, i.e. the muscles return to decontract because a high and long lasting contraction is not physiologically possible unless it is voluntary. For this

reason AU15 reduces again after a maximum contraction period between 1460s and 1500s. The same phenomenon occurs for AU25.

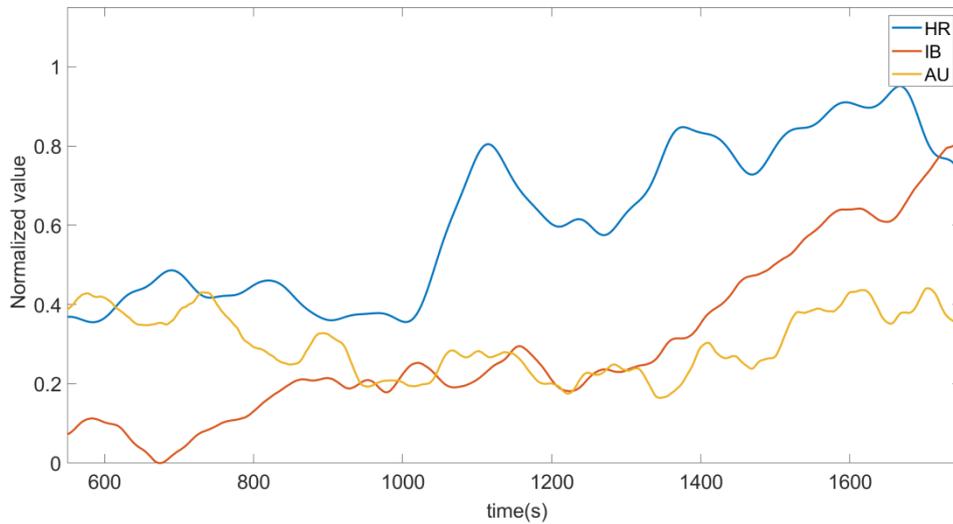


Figure 94: AU17 trend

Obviously each individual has its own typical facial expressiveness and tends to use more certain muscles, EEG data showed a relationship with the AU17 and is therefore also analyzed in relation to bioharness to analyze the possibility that each individual privileges the contraction of certain facial muscles induced by a cognitive load increase. AU17 presents the same peaks of HR but the first two occur with a delay of 45s motivated by the fact that the workload is low and therefore the FE_{RL} is greater. It is interesting to note that the peaks at 1400s and 1600s have about the same shape of HR whereas the lower peak at 1240s that is highlighted little by HR is evident in the AU17 trend.

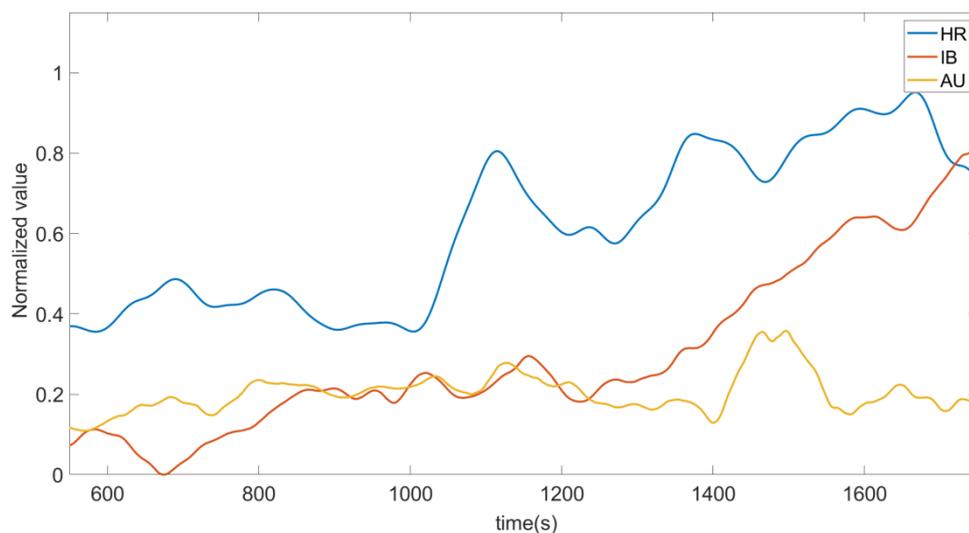


Figure 95: AU25 trend

AU25 initially shows a slight increase but the trend is very smooth up to 1100s and then has three main peaks. Analyzing again the prominence of the peaks as the oscillations are small (except for the peak at about 1500s) three main peaks are identified. Table 18 shows the time of occurrence of AU25 and HR peaks.

	Peak1	Peak2	Peak3
HR	1115s	1380s	1670s
AU25	1125s	1460s	1650s

Table 18: AU25 and HR peaks occurrence

The phase shifting between the two curves is between 10s and 80s, consistent with the data obtained in the OTM experiments. As already described in the previous analysis AU25 does not show particular relationships in this experiment in which AU15 seems to be more reliable, but both show the second peak of Workload around 1500s.

We can conclude that AU25 could still be considered as the main AU of reference but individual tests on the operator are necessary to understand which muscles privilege the contraction which in this case are the muscles related to AU15 and AU17.

11.4 Electroencephalography

The OTM experiments have shown a close relationship between AU9 and AU25 which is analysed for this experiment. Table 19 show the correlation coefficient between AUs and EEG.

	AU6	AU9	AU12	AU15	AU17	AU25
CC	0.41	0.71	0.76	0.42	0.55	0.1

Table 19: Correlation Coefficient AUs-EEG

Unfortunately in this experiment the relationship between AU25 and EEG is not clear. On the contrary, Figure 96 and Figure 97 show how AU9, AU12 and AU17 have a close relationship with the *EEG index*. AU9 between 1100s and 1400s shows a clear and constant increase in muscle contraction which is also shown by EEG if the general trend of the mean value is evaluated. In general the AUs are not in phase with EEG except AU17.

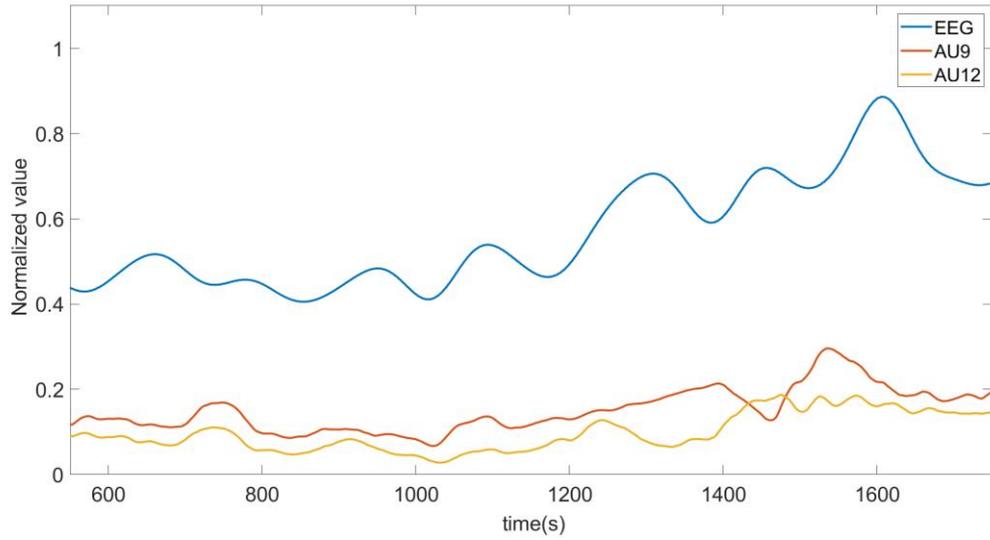


Figure 96: AU9-AU15 trends

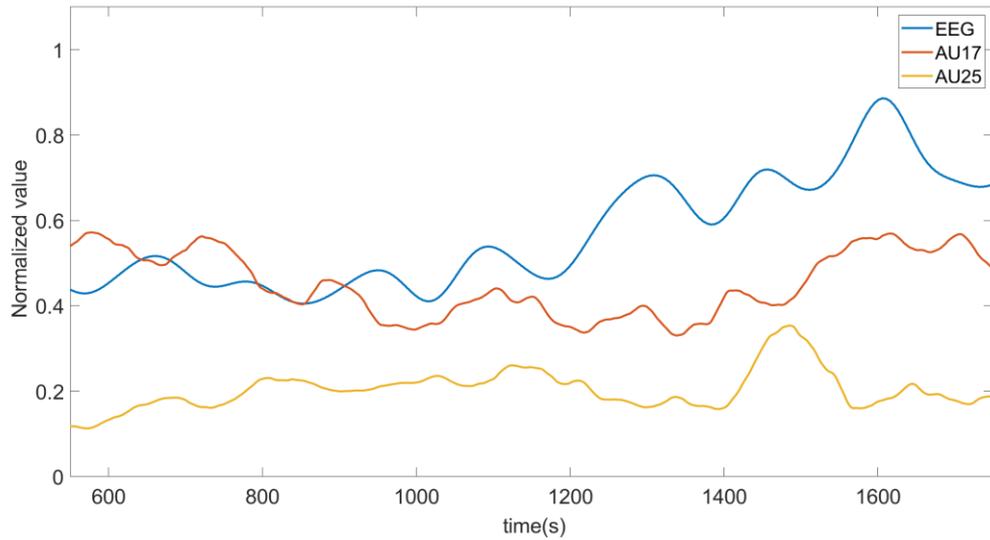


Figure 97: AU17-AU25 trends

AU17 in fact has peaks about in phase with EEG when the workload increases. In this case study, therefore, a particular correlation with AU25 was not detected but the hypothesized relationship with AU9 in the OTM experiments can be confirmed. In addition, can be confirmed as well the relationship between Workload and contraction of the mouth muscles because AU12(Lip Corner Puller, *Zygomatic Major*) and AU17(Chin Raiser, *Mentalis*) both act on lip corners.

12 Protocol

The conducted experiments have demonstrated the relationship between some AUs and the cognitive state and that each individual shows his or her own FE response following Workload variations. For this reason it is necessary to develop an offline calibration protocol in order to adapt the given pre-processing model to each operator. For each individual it is therefore necessary to make calibration experiments that are used to define the pre-processing data method and saved in a database. Following the creation of this database each operator before starting the experiment must select his profile so that the CHMI2 can process the monitored data consistently with the model defined with the calibration tests. In general, the pre-processing data consists essentially of the following steps:

1. Data rejection if the confidence of the monitored data is less than 80%.
2. Lowpass filtering at 6Hz
3. Smoothing of small oscillations by averaging the values in a sliding window (*movmean* function)
4. Smoothing to highlight the general trend of the curve using the *smooth* function

Steps 1 and 2 must always be applied whereas the last two steps vary according to the participant because each one shows a different amount of contraction of the muscles with equal Workload variation. *Movmean* and *smooth* entity are defined through calibration experiments.

It should be noted that this protocol is aimed at operators who already have experience or at least familiarity with the interface because, as has been demonstrated by the conducted analysis, newbies individuals have more noisy physiological responses and can more easily loom in emotional states such as confusion and anxiety that lead to be the emotional state that plays the most important role and not the cognitive state. An experienced individual, instead, is in comfort zones in performing tasks so does not have a high variation of the emotional state that therefore does not distort the data, or rather the emotional component in the cognitive state equation remains at low values.

Performance evaluation chapter describes the limits of the software in terms of the position of the face in relation to the camera. Participants during both calibration tests and experiments must remain within the thresholds listed in Performance evaluation but more stringent thresholds are required in calibration tests to optimize the system calibration:

- Operator's distance from the camera between 50cm and 90cm
- Maximum rotation around X axis <10°
- Maximum rotation around Z axis <15°

Reference system showed in Figure 27.

After installing the Open Face software on the computer, simply connect the camera to it and start recording, data are saved to an excel file as described in chapter The Software: Open Face.

The calibration tests must consist of the following steps, presented in Figure 98:

- Pre-test: 5 minutes of pre-test in which the operator starts to perform some tasks with a very low workload. This phase serves to immerse the operator in the environment to distract him from any thoughts not consistent with the experiment that could induce emotional reactions and thus generate bias.
- Variable workload experiment
- Post-test: 5 minutes at low workload equal to the first pre-test phase (e.g. same number of aircraft or UAVs to manage) to compare the facial response with the pre-test phase to check that there are no substantial variations. In this phase are accepted background contractions lower than the pre-test (up to 20%) due to possible tiredness and a more relaxed emotional state because the operator has finished the test.

The central part of the experiment represents the main evaluation phase and is in turn divided into sub-phases. It is recommended not to screen the clock or a timer because the operator, so aware of when workload variations occur, may self-induce changes in physiological response.

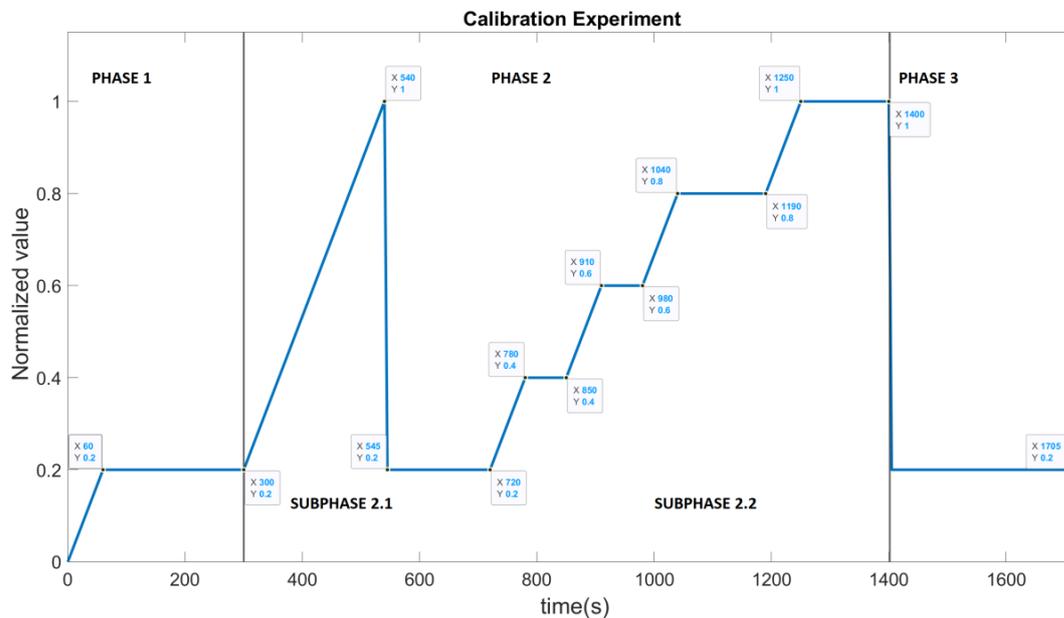


Figure 98: Calibration experiment

Figure 98 shows the workload trend of the experiment, the workload indicates an objective parameter that can be, as for the experiments conducted, the number of aircraft and control input or secondary tasks. It is not necessary but it is recommended to use the same parameter both in calibration tests and in the main experiments. In figure the normalized value is presented and the maximum value equal to 1 represents the maximum WL value that an ATMo or OTMo can experiment during a working day. The workload rise rate is always equal to a 20% increase every minute. The experiment develops as follows:

Phase 1

1. increasing WL from 0 to 20% of the maximum value in one minute.
2. WL plateau of 4 minutes to allow the operator to adapt to the context and a low workload

Phase 2

1. Sub-Phase 2.1: increase the workload by 0.2 every minute to the maximum value and then quickly (5 seconds) back to 0.2.
2. Sub-Phase 2.2: step trend with increments of 0.2 per minute and 4 plateaus, the first two by 70s and the last two by 150s.

Phase 3

1. WL plateau of 5 minutes to allow the operator to re-establish his background contraction and then compare it with *Phase 1* values.

In *Phase 3* contractions are expected to be higher than in *Phase 1* for the first 2/3 minutes or so due to the effort made in *Phase 2* (contraction inertia and FE-RL) and then reach lower values due to fatigue and relaxation of the participant.

Two trends are generated in *Phase 2* because experiments have shown that after a WL plateau above 10s muscle contractions tend to reduce as if it were an adaptation of the face and therefore of the cognitive state to that particular WL value. Moreover, even if the cognitive effort remains constant, especially for high efforts, the muscles do not remain very contracted for a long period. For this reason, a steadily increasing WL is initially analysed to assess the relationship between the magnitude of contraction and the relative WL value and the maximum reachable contraction of the muscles (maximum geometrical variation). The second sub-phase (2.2), on the other hand, is used to determine the response of facial muscles following WL plateau of more than 10s duration. The analysis of the experiments showed that the Constant Workload Reduction in Contraction (CWRC) during a plateau depends on the level of workload, at lower workloads the CWRC occurs faster while at higher workloads the CWRC occurs more slowly. At high WL there is a greater contraction that in 10s/20s is reduced, but the variations may last up to 30s with possible re-contractions, so for WL over 60 the plateaus last longer (150s).

It is recommended to repeat the experiment on the same subject at least 2 times at the same time of day. These experiments therefore make it possible to determine the specific parameters of each individual:

- Maximum muscle contraction (extent of facial expression)
- Change in contraction due to change in WL
- Response times: Facial Expression Response Latency (FE-RL)
- Constant Workload Reduction in Contraction (CWRC)

These parameters are used to define two fundamental aspects of data processing:

- Size of the sliding window *movmean*
- Acceptability range of FE-RL

In the previous chapters it has been explained how the smoothing of the values carried out through the *movmean* functions should be modified from person to person, the size of the mean sliding window can vary from about 40s to 200s. The data

obtained from calibration experiments are used to evaluate offline the size of this window so that it can be used for online analysis.

The FE-RL also varies from individual to individual and can reach a maximum of 150s. For this reason is the two high WL plateaus were chosen to last 150s so the physiological response happen during the plateau and can not happen that it appears when the plateau is already finished. If it exceeds this value it can no longer be accepted that a variation of WL has induced such a variation of FE. The offline analysis allows to determine the late parameter and its variation when the workload varies so that outlier rejection thresholds can be established. For example an individual with FE-RL=70s implies that during the comparison of the parameters evaluated by other sensors all the values of the psycho-physiological parameters monitored in a 70s window are evaluated and if all or most of these present a consistent variation within the window then a variation of WL is detected.

The other fundamental aspect in calibration experiments is the determination of dominant contractions. In fact it has been seen that for the analyzed subjects AU25 represents the most reliable for OTM and AU17 for ATM. This difference can be due to the fact that the two experiments stimulate the participants in different ways so the ATM operator could be made to perform the OTM experiments to make a further analysis of dependence of the dominant AU with the performed experiment. The other fundamental aspect is that it has been shown that each individual prefers the contraction of some muscles rather than others. Thus two different individuals undergoing the same stimulus could show different dominant AUs, albeit in the mouth region, such as AU12, AU15, AU17 or AU25. Thanks to the calibration test, therefore, it is possible to adopt a 'boolean' approach in which it is detected which AU is the dominant one and others are used as secondary confirmation of cognitive load variation detection as proposed in the previous chapters.

Some final notes for the experiment are:

- Constant artificial lighting from above.
- Camera centred with monitors and positioned above it in front of the participant.

The FE monitoring is then modelled to be inserted into the sensor network. The development of a model is needed for the online human-system adaptation during the mission using databases which are defined in the calibration experiments. The relationship between cognitive states, evaluated with the psycho-physiological monitored parameters, and the workload has been modelled through Fuzzy systems. This relationship is not linear and is different between individuals so it is necessary to use continuous self-calibration systems. For this reason Fuzzy systems were chosen because they allow to create a loosely relationships between physiological features, cognitive states and mission performance through membership functions. Considering that psycho-physiological measurements are very noisy, Fuzzy membership functions give the possibility to estimate the degree of truth of a classification value based on suitably modelled statistical errors. The evaluation of workload and cognitive state is an inexact science that does not yet allow to precisely define a value of the latter on a universally recognized scale and for the purposes of adaptation of the level of automation it is initially necessary just to divide it into categories such as *high*, *medium*, *low*. Therefore the Fuzzy *if-then rules* or *fuzzy conditional statements* represents a good solution because allow to capture inaccurate modes of reasoning typical of humans decision making environment.[95, 96] These rules are used in a Fuzzy inference system that allows to adapt the definition of membership function based on historical data and calibration data.[40, 97-99]

An offline training is therefore necessary to define the Fuzzy model, membership functions, clustering and other parameters that will be later explained. Neural-Fuzzy System(NFS)[100] which integrate aspects of neural networks and fuzzy inference systems present a good repeatability and software maturity, so they had been investigated as the most promising offline training technique. NFS reproduces the brain structure which is formed by distributed network of nodes, each performing simple functions. Each connection is assigned a weight which is defined during the training session. The offline training allows to create a database with all *if-then* rules that is used during the online training (incremental learning). In our evaluations the input are FEs while the output is the workload objectively evaluated through secondary Taskload for OTM and number of aircraft and control input for ATM. Fuzzy logics provide a simple structure to the classifier and support a greater degree of result interpretability when compared against other machine learning approaches such as deep learning. These logics reflect very much the human reasoning method and operate in its ambiguity and subjectivity. The main aspect is that the characterization is based on the level of uncertainty of an information that is the degree of truth of that event occurring. It should not be confused with the likelihood of an event's occurrence. Uncertainty is the basis of clustering which represents the first level of the Neural Network. The categories of *high*, *medium* and *low* can be expressed by fuzzy sets. Dividing the workload into three categories: high, medium and low, the concept of uncertainty can be expressed as the degree of truth of a piece of information considering that it is known that the data obtained from the sensors have uncertainties and noise. The Fuzzy rules have already been conceptually anticipated during the data analysis of the experiments as for example:

Rule i: IF AU25 is High AND Blink Rate is High THEN Workload is High

These rules represent common characteristics between each individual such as a relationship between AU17 and/or AU25 and the Blink Rate but each individual has a different definition of these categories or sets (set's centers and spreads). For example, assuming to use three categories, they can be divided as follows:

AU 25 normalized amplitude	Partecipant1	Partecipant2
Low	0.3	0.2
Medium	0.6	0.55
High	1	1

Table 20: Clustering rules

Once the rules describing the relationship between input and output are defined, different inference methods of NFS computation can be used.

Figure 99 shows the typical architecture of a neural-fuzzy network composed of 5 layers.

The first layer consists of all inputs that are passed to the second layer that uses the membership functions to *“fuzzify”* the inputs obtaining the fuzzy set membership for a given input parameter. Then the values are passed to the decision making layer which performs the inference operations on the rules. In the next layer all the results of each rules are compared through a fuzzy OR operator obtaining the membership value of the output parameter. Now the workload level and his evaluation uncertainty is obtained in accordance to the defined rules but these data are still linguistic values so they are *“defuzzified”* into a crisp output that represents the last layer.

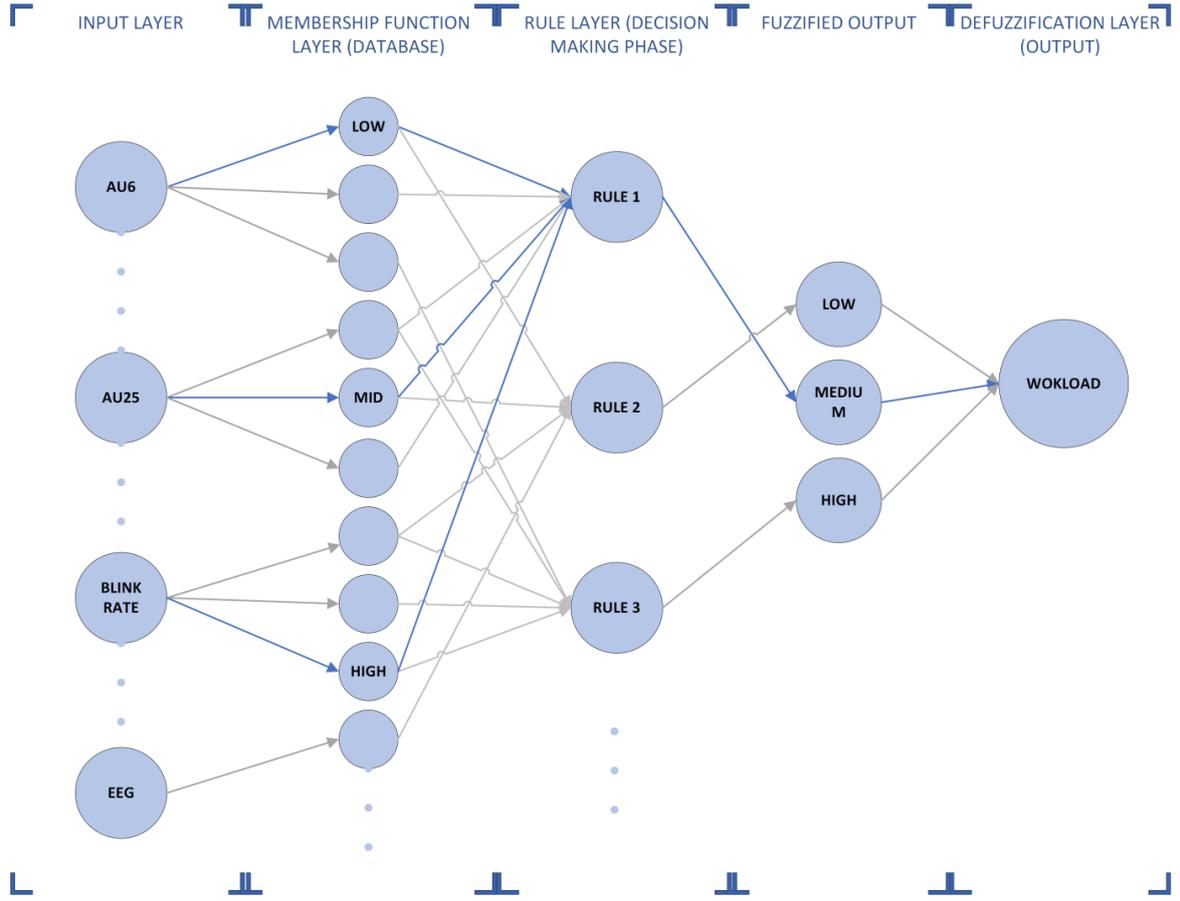


Figure 99: Neural Fuzzy Network architecture

Rules are defined thanks to the conducted experiments whereas the Membership function layer is defined through the calibration experiments. The calibration experiments are then used to perform offline calibration of the system for future evaluations and implementations for real-time HMI² adaptation. The offline calibration of the CHMI² membership function(classification) layer is carried out through the use of the Adaptive Neuro-Fuzzy Inference System (ANFIS). The training is based on Fuzzy Inference System(FIS) which use a clustering algorithm to create the input x_i Gaussian membership functions and y_q Sugeno-type FIS output functions. The Sugeno-type FIS used in the workload classifier uses output membership functions that are a linear combination of the input values. The adopted approach assume the same number of membership functions(j^{th}) and clusters. The function equation is:

$$f(x_i, \sigma_{ij}, c_{ij}) = \exp\left(-\frac{(x_i - c_{ij})^2}{2\sigma_{ij}^2}\right) \quad (12.1)$$

Where c_{ij} is the centre of the Gaussian and σ_{ij} the spread.

Each y_q Sugeno output functions is defined as a linear combination of the input:

$$y_q(x_1, \dots, x_n) = p_{q0} + \sum_{i=1}^n p_{qi} \cdot x_{qi} \quad (12.2)$$

where $\{p_{q0}, \dots, p_{qn}\}$ are the parameters of the output membership function.

Different types of subtractive clustering algorithms can be used and the adopted one is the Fuzzy C-Means(FCM) clustering.[101]

This clustering technique requires to specify the number of clusters and in the case of CHMI2 it has an advantage because it can be defined according to the levels of automation to be implemented in the system and the physiological responses determined in calibration experiments. FCM divides the pool of n data into c fuzzy groups, each data is associated with a membership matrix U which denotes the fuzzy membership of data point x_i with respect to group j .

The membership grade of each data i in group j is defined by a membership grade u_{ij} , i.e. the degree of belonging to that fuzzy group. The sum of all membership grade across all groups must sum to unity:

$$\sum_{j=1}^c u_{ij} = 1, \quad \forall_i = 1, \dots, n \quad (12.3)$$

Each i^{th} point is at a certain distance from the j^{th} cluster centre which is calculated through the Euclidean distance:

$$d_{ij} = \|\mathbf{x}_i - \mathbf{c}_j\| \quad (12.4)$$

The degree of fuzziness of each data cluster is defined by m . A high values of m data points far away from the cluster center have a significant membership grade. It can be therefore defined a cost function as follow:

$$J(\mathbf{U}, \mathbf{c}_1, \dots, \mathbf{c}_c) = \sum_{j=1}^c \sum_i^n u_{ij}^m \cdot d_{ij}^2 \quad (12.5)$$

The above reported equation is differentiated with respect to all input arguments to find its minimum:

$$\begin{aligned} J(\mathbf{U}, \mathbf{c}_1, \dots, \mathbf{c}_c, \lambda_1, \dots, \lambda_n) &= J(\mathbf{U}, \mathbf{c}_1, \dots, \mathbf{c}_c) + \sum_i^n \lambda_i \cdot \left(\sum_{j=1}^c u_{ij} - 1 \right) = \\ & \sum_{j=1}^c \sum_i^n u_{ij}^m \cdot d_{ij}^2 + \sum_i^n \lambda_i \cdot \left(\sum_{j=1}^c u_{ij} - 1 \right) \end{aligned} \quad (12.6)$$

The necessary conditions are given by:

$$\mathbf{c}_i = \frac{\sum_{j=1}^c u_{ij}^m \cdot \mathbf{x}_i}{\sum_{j=1}^c u_{ij}^m} \quad (12.7)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}} \quad (12.8)$$

The main steps that the FCM carries out are now conceptually reported:

- Cluster initialization randomly defining the c cluster centers.
- Membership grade calculation.
- Effective cluster centre calculation.
- Membership grade and cluster centers are iteratively calculated until J satisfies a given threshold or until $\|U^{(k+1)} - U^{(k)}\|$ satisfies a termination criterion.

When the clustering phase is finished the neural fuzzy system (ANFIS) tunes the parameters of the generated FIS using a back-propagation method for the input membership functions whereas least square estimation is adopted to define the output membership function parameters.

The ANFIS workload classifier is trained using a subset of the participant dataset, Experiments have shown that the physiological response has a more regular trend when the workload is high, so the training is affected for a subset equal to 20% of the dataset during the high WL phase. This training can be tested during further experiments conducted on the same subjects performing the same type of experiments.

For this first evaluation of facial expressions in the sensor network, the relationship between the single AUs and the objective workload was evaluated. Future research will determine further rules to be able to tow ANFIS with more inputs from different sensors. The root mean-squared error of all classifiers was found to be 0.0471 ± 0.0253 .

Figure 100 and Figure 101 shows the Membership Functions AU25 clusters for the ATM experiment and participant 1 in the OTM experiment. The larger clusters are generated during training whereas the sharper ones represent clusters applied to all the dataset. They show that there is accordance between the clustering of training and testing, especially the ATM experiment shows how the higher is the workload the higher is the correlation between FE and objective workload(n aircraft and control input).

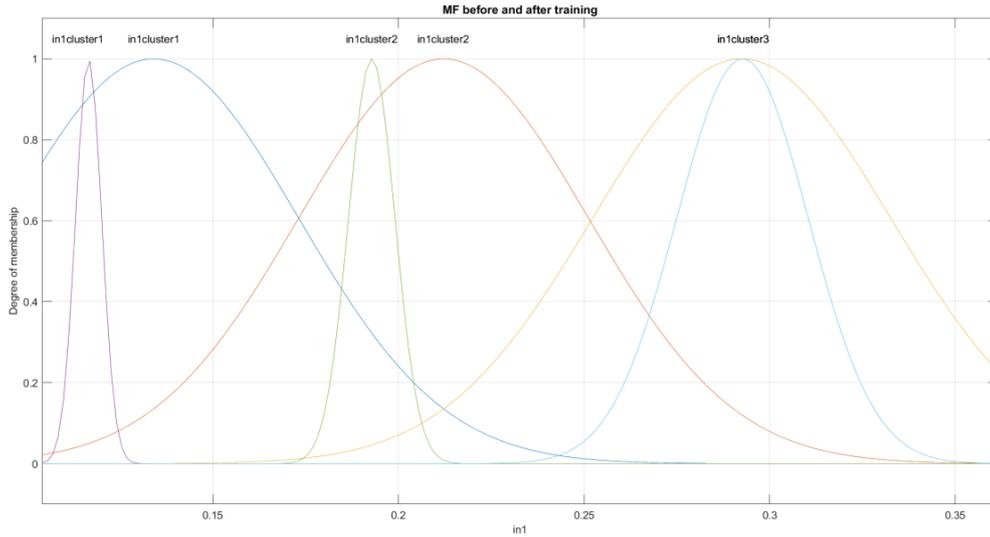


Figure 100: ATM experiment

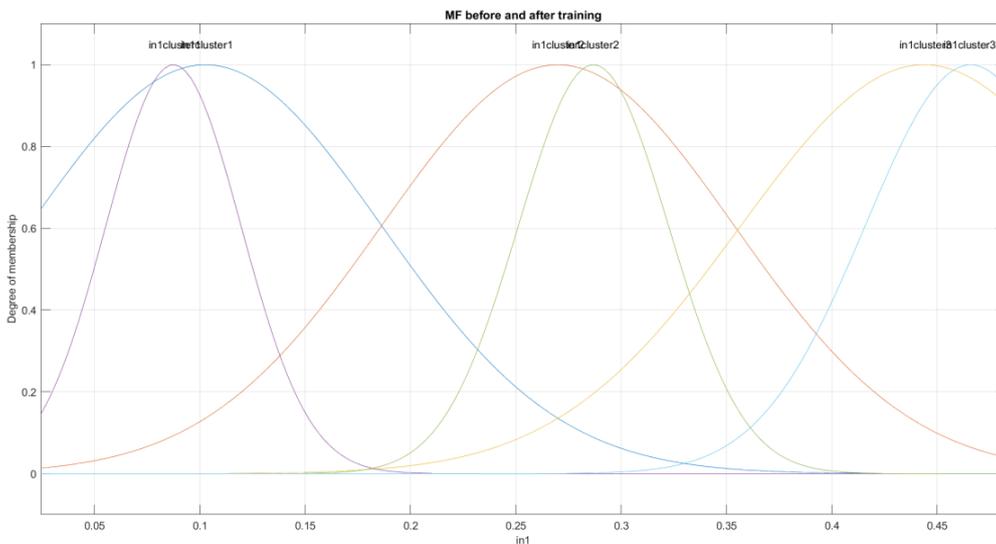


Figure 101: OTM Participant1

The parameter that indicates how close is the correlation between input and output is the Relative Fitting Error (RFE) defined as:

$$RFE = \frac{\text{mean}(|\text{fitting error}|)}{\text{max}(\text{original values}) - \text{min}(\text{original values})} \quad (12.9)$$

Where *original values* is the objective workload vector that has been set as ANFIS output. The fitting error is defined as follow:

$$\text{Fitting error} = \text{FIS data} - \text{original values} \quad (12.10)$$

FIS data are the input data(AU) trained by ANFIS as explained before. Table 21 shows the obtained RFE for all the conducted experiments.

	AU6	AU9	AU12	AU15	AU17	AU25
ATM	20.07	15.94	17.26	18.09	23.06	19.32
OTM						
Partecipant1	18.75	19.78	16.20	18.98	12.74	15.68
Partecipant2	23.17	14.98	19.95	17.60	21.62	24.91
Partecipant3	17.51	10.65	12.46	8.34	10.02	5.78
Partecipant4	17.54	14.61	19.61	22.55	21.0.3	19.42
Partecipant5	20.02	7.65	21.27	11.37	23.22	15.28
Partecipant6	11.75	16.12	14.74	13.76	16.47	12.98

Table 21: Relative fitting error ANFIS

Considering that these are the first analyses conducted on the FE in this project and the human body is very unpredictable, values below 20% are considered meaningful. Table 21 shows that most RFE are less than 20% and the minimum obtained values is 7.65%, which indicates an excellent correlation between AU and WL using a fuzzy logic inferred method. These analyses have demonstrated the validity of the adopted method and the potential usefulness of the AUs for workload determination. Further experiments are necessary to create a large ANFIS database that allows to further reduce the error and acquire a greater awareness of the physiological response. In the current test activity, the only physiological data that could be exchanged in real time was eye tracking data, while cardiorespiratory and EEG data were only analysed in the post-processing period like AUs. In future researches it will be possible to perform calibration tests with the previously described methods to train ANFIS and then carry out experiments such as those already performed to verify and validate this model in order to implement the online automation adaptation.

13 Conclusions

This research addresses the growing need to monitor humans in safety critical operations in which the number of tasks to be carried out is increasing. The role of sensor networks in cyber-physical aerospace system applications is increasingly important to manage more information by developing human-machine teaming and performance. This teaming is represented by the concept of Cognitive Human-Machine Interfaces and Interactions(CHMI²) systems which represents the core of the carried out study.

The conducted analyses led to the achievement of the research objectives initially set. A relationship between cognitive state and the contractions of the facial muscles has been demonstrated and will be analysed in more detail in future researches. The carried out studies have demonstrated a close relationship between mental workload and Taskload variation with mouth and cheekbones Action Units (AU12, AU15, AU17, AU25). This concept had never been highlighted in past research and represents the main contribution of this project to the research field.

In the *Performance evaluation* chapter the software and its applicability have been analyzed in order to highlight possible bias and the optimal approach to monitor Facial Expressions.

The subjectivity of FEs is the aspect that has the greatest impact on the meaning of recorded data so it has been analysed in detail to define a protocol. In *Protocol* chapter were reported all the aspects that have to be considered during FE monitoring. This protocol will be used to perform calibration tests in future research in order to highlight the subjective component and avoid misinterpretations of face muscles contractions. The relationship between FE and biometric parameters such as EEG, Eye movements, Heart Rate, Breath Rate has been also studied demonstrating the possibility to integrate Facial Expression sensing in a sensor network for the cognitive state evaluation of ATM and OTM operators.

CHMI² is conceived as a multi-sensor fusion system because each sensor is affected by bias and noise. In fact, this research confirms what expressed in other researches ([18]) for which the more sensors are used, the greater is the reliability and accuracy of the cognitive states estimation. This is because the change in one measure is usually related to few cognitive states.

The EEG sensor works at very low voltages and is therefore affected by background noise, the Cardio-respiratory sensor is affected by the training of the individual, a trained subject has a lower Heart Rate and Breath Rate than a not trained one, whereas the Eye Sensor may suffer disturbances due to the movement of the operator's head. All these bias can be overcome by integrating the various sensors in a single system to use them in a complementary way. This research has demonstrated the potential reliability of FE monitoring which can therefore contribute significantly to the complementarity of sensors in a CHMI² framework:

- Increasing the accuracy of Cognitive States estimation.
- Increasing the range of applicability.

In fact, Face Monitoring uses remote sensing: a simple camera. This allows it to be easily installed in different environments and can be directly connected with other sensors in a framework for real-time monitoring.

Consequently this work can be useful in both defence and aerospace fields and also for Single Pilot Operations (SPO).[102] Furthermore, some space applications currently mainly use sensor networks for medical monitoring purposes and it is expected that in the future CHMI²systems will also be adopted. Cognitive state estimation by FE monitoring can also be used in other fields such as automotive, psychology and all environments with the aim to monitor the human to increase safety and operational efficiency.

This research has provided useful data for the definition of the FE time response after a stimulus (FE-RL) and how AUs can be integrated in a parametric model that defines the mental workload.

It has been noted that the FE response is generally quicker if the workload variation is higher (high derivate) but further experiments are needed to prove this point.

Future developments will better analyze the relationship between workload and mouth AUs studying the maximum time of muscle contraction after which a relaxation appear even if the cognitive load remains high.

Future research activities will analyze the relationship with an additional psychophysiological parameter: the voice pattern.

Bibliography

- [1] J.-D. J. Yeong Heok Lee, and Youn-Chul Choi, "Air Traffic Controllers' Situation Awareness and Workload under Dynamic Air Traffic Situations," *Transportation Journal*, vol. Vol. 51, No. 3, pp. pp. 338-352, 2012.
- [2] J. Harrison *et al.*, "Cognitive Workload and Learning Assessment During the Implementation of a Next-Generation Air Traffic Control Technology Using Functional Near-Infrared Spectroscopy," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 4, pp. 429-440, 2014, doi: 10.1109/thms.2014.2319822.
- [3] M. Jipp and P. L. Ackerman, "The Impact of Higher Levels of Automation on Performance and Situation Awareness," *Journal of Cognitive Engineering and Decision Making*, vol. 10, no. 2, pp. 138-166, 2016, doi: 10.1177/1555343416637517.
- [4] G. F. Wilson, "An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures," *The International Journal of Aviation Psychology*, vol. 12, no. 1, pp. 3-18, 2009, doi: 10.1207/s15327108ijap1201_2.
- [5] L. X. Xin Liu, Zhiliang Wang, Dongmei Fu, "Cognitive-affective regulation process for micro-expressions in active field state space," *IEEE CCIS2012*, vol. Proceedings of IEEE CCIS2012, 2012.
- [6] M. A. Bonner and G. F. Wilson, "Heart Rate Measures of Flight Test and Evaluation," *The International Journal of Aviation Psychology*, vol. 12, no. 1, pp. 63-77, 2009, doi: 10.1207/s15327108ijap1201_6.
- [7] R. Castaldo, L. Montesinos, T. S. Wan, A. Serban, S. Massaro, and L. Pecchia, "Heart Rate Variability Analysis and Performance during a Repeated Mental Workload Task," in *Embec & Nbc 2017*, (IFMBE Proceedings, 2018, ch. Chapter 18, pp. 69-72.
- [8] R. L. Charles and J. Nixon, "Measuring mental workload using physiological measures: A systematic review," *Appl Ergon*, vol. 74, pp. 221-232, Jan 2019, doi: 10.1016/j.apergo.2018.08.028.
- [9] M. De Rivecourt, M. N. Kuperus, W. J. Post, and L. J. Mulder, "Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight," *Ergonomics*, vol. 51, no. 9, pp. 1295-319, Sep 2008, doi: 10.1080/00140130802120267.
- [10] F. Dehais *et al.*, "Monitoring Pilot's Cognitive Fatigue with Engagement Features in Simulated and Actual Flight Conditions Using an Hybrid fNIRS-EEG Passive BCI," presented at the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2018.
- [11] F. Di Nocera, M. Camilli, and M. Terenzi, "A Random Glance at the Flight Deck: Pilots' Scanning Strategies and the Real-Time Assessment of Mental Workload," *Journal of Cognitive Engineering and Decision Making*, vol. 1, no. 3, pp. 271-285, 2016, doi: 10.1518/155534307x255627.
- [12] T. Kistan, A. Gardi, and R. Sabatini, "Machine Learning and Cognitive Ergonomics in Air Traffic Management: Recent Developments and

- Considerations for Certification," *Aerospace*, vol. 5, no. 4, p. 103, 2018, doi: 10.3390/aerospace5040103.
- [13] S. Loft, P. Sanderson, A. Neal, and M. Mooij, "Modeling and predicting mental workload in en route air traffic control: critical review and broader implications," *Hum Factors*, vol. 49, no. 3, pp. 376-99, Jun 2007, doi: 10.1518/001872007X197017.
- [14] B. G. L. Marian Stewart, Ian Fasel, Javier R. Movellan, "Real Time Face Detection and Facial Expression Recognition," *IEEE Computer Society*, Proceedings of the 2003 Conference on Computer Vision and Pattern Recognition Workshop 2003.
- [15] S. C. Corver, D. Unger, and G. Grote, "Predicting Air Traffic Controller Workload: Trajectory Uncertainty as the Moderator of the Indirect Effect of Traffic Density on Controller Workload Through Traffic Conflict," *Hum Factors*, vol. 58, no. 4, pp. 560-73, Jun 2016, doi: 10.1177/0018720816639418.
- [16] Y. Lim, S. Ramasamy, A. Gardi, T. Kistan, and R. Sabatini, "Cognitive Human-Machine Interfaces and Interactions for Unmanned Aircraft," *Journal of Intelligent & Robotic Systems*, vol. 91, no. 3-4, pp. 755-774, 2017, doi: 10.1007/s10846-017-0648-9.
- [17] A. G. Nichakorn Pongsakornsathien, Roberto Sabatini, Trevor Kistan, "Cognitive Human-Machine Interfaces and Interaction for Terminal Manoeuvring Area Traffic Management."
- [18] N. Pongsakornsathien *et al.*, "Sensor Networks for Aerospace Human-Machine Systems," *Sensors*, vol. 19, no. 16, p. 3465, 2019.
- [19] K. R. Y. Lim, A. Gardi, N. Ezer, and R. Sabatini, "Human-Machine Interfaces and Interaction," Proceedings of the 31th Congress of the International Council of the Aeronautical Sciences 2018.
- [20] G. G. De la Torre, J. M. Mestre Navas, and R. Guil Bozal, "Neurocognitive performance using the Windows spaceflight cognitive assessment tool (WinSCAT) in human spaceflight simulations," *Aerospace Science and Technology*, vol. 35, pp. 87-92, 2014, doi: 10.1016/j.ast.2014.02.006.
- [21] R. Parasuraman and G. F. Wilson, "Putting the Brain to Work: Neuroergonomics Past, Present, and Future," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 50, no. 3, pp. 468-474, 2008, doi: 10.1518/001872008x288349.
- [22] P. Ekman, *I volti della menzogna. Gli indizi dell'inganno nei rapporti interpersonali*. Giunti Editore, 1995.
- [23] M. Balconi, *Neuropsicologia della comunicazione*. Springer, 2008.
- [24] J. F. C. Z. A. P. Ekman, *Observer-Based Measurement of Facial Expression With the Facial Action Coding System*.
- [25] C. L. L. a. D. J. Schiano, *automatic-facial-expression-interpretation-where-humancomputer-i-2000* (Pragmatics & Cognition Vol. 8). 2000.
- [26] J. A. R. José-Miguel Fernández-Dols, *The Science of Facial Expression* (Social Cognition and Social Neuroscience). 2017.
- [27] J. F. Grafsgaard, K. E. Boyer, and J. C. Lester, "Predicting Facial Indicators of Confusion with Hidden Markov Models," Springer Berlin Heidelberg, 2011, pp. 97-106.

- [28] S. Craig, A. Graesser, J. Sullins, and B. Gholson, "Affect and learning: An exploratory look into the role of affect in learning with AutoTutor," *Journal of Educational Media*, vol. 29, no. 3, pp. 241-250, 2004, doi: 10.1080/1358165042000283101.
- [29] M. A. Sayette, J. F. Cohn, J. M. Wertz, M. A. Perrott, and D. J. Parrott, *Journal of Nonverbal Behavior*, vol. 25, no. 3, pp. 167-185, 2001, doi: 10.1023/a:1010671109788.
- [30] B. D. M. McDaniel, Sidney King, Brandon et al., "Facial Features for Affective State Detection in Learning Environment," *Proceedings of the Annual Meeting of the Cognitive Science Society*, 29(29), 2007.
- [31] J. C. Hager, "A comparison of units for visually measuring facial actions," p. 19, 1985 Behavior Research Methods, Instruments, & Computers.
- [32] G. JJ., "Emotion regulation: affective, cognitive, and social consequence," *Psychophysiology*, 2002.
- [33] B. Sidney K. D'Mello, Lehman, Natalie Person, "Monitoring Affect States During Effortful Problem Solving Activities," *International Journal of Artificial Intelligence in Education Volume*, 2010, doi: 10.3233/JAI-2010-012.
- [34] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *International Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 724-736, 2007, doi: 10.1016/j.ijhcs.2007.02.003.
- [35] J. B. W. Joseph F. Grafsgaard, Kristy Elizabeth Boyer, Eric N. Wiebe, James C. Lester, "Automatically Recognizing Facial Expression_Predicting engagement and frustration."
- [36] K. Ihme, A. Unni, M. Zhang, J. W. Rieger, and M. Jipp, "Recognizing Frustration of Drivers From Face Video Recordings and Brain Activation Measurements With Functional Near-Infrared Spectroscopy," *Front Hum Neurosci*, vol. 12, p. 327, 2018, doi: 10.3389/fnhum.2018.00327.
- [37] S. D'Mello, B. Lehman, R. Pekrun, and A. Graesser, "Confusion can be beneficial for learning," *Learning and Instruction*, vol. 29, pp. 153-170, 2014, doi: 10.1016/j.learninstruc.2012.05.003.
- [38] G. C. Littlewort, M. S. Bartlett, L. P. Salamanca, and J. Reilly, "Automated measurement of children's facial expressions during problem solving tasks," 2011: IEEE, doi: 10.1109/fg.2011.5771418.
- [39] J. B. W. Joseph F. Grafsgaard, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester, "Automatically Recognizing Facial Indicators of Frustration."
- [40] Y. Lim, A. Gardi, N. Pongsakornsathien, R. Sabatini, N. Ezer, and T. Kistan, "Experimental characterisation of eye-tracking sensors for adaptive human-machine systems," *Measurement*, vol. 140, pp. 151-160, 2019, doi: 10.1016/j.measurement.2019.03.032.
- [41] P. Rani, J. Sims, R. Brackin, and N. Sarkar, "Online stress detection using psychophysiological signals for implicit human-robot cooperation," *Robotica*, vol. 20, no. 6, pp. 673-685, 2002, doi: 10.1017/s0263574702004484.
- [42] R. P. Evan A. Byrne, "Psychophysiology and adaptive automation."
- [43] B. L. Brandon Amos, Mahadev Satyanarayanan, "OpenFace A general purpose face recognition."

- [44] B. L. A. Mica R. Endsley, and Debra G. Jones, "Designing for Situation Awareness_ An Approach to User-Centered Design," *full book*, 2003.
- [45] D. J. G. Mica R. Endsley, "Situation Awareness Analysis and Measurement " *full book*, 2000.
- [46] S. R. Yixiang Lim, Alessandro Gardi, and Roberto Sabatini, "Aviation Human Factors Engineering."
- [47] M. R. Endsley, "Level of automation effects on performance, situation awareness and workload in a dynamic control task," *Ergonomics*, vol. 42, no. 3, pp. 462-492, 1999, doi: 10.1080/0014013991855595.
- [48] A. P. G. Martins, "A Review of Important Cognitive Concepts in Aviation," *Aviation*, vol. 20, no. 2, pp. 65-84, 2016, doi: 10.3846/16487788.2016.1196559.
- [49] M. Malik, "Heart rate variability standards of measurement, physiological interpretation, and clinical use," *Eur Heart J*, vol. 17, pp. 354-381, 1996.
- [50] P. Sauseng, W. Klimesch, M. Schabus, and M. Doppelmayr, "Fronto-parietal EEG coherence in theta and upper alpha reflect central executive functions of working memory," *International Journal of Psychophysiology*, vol. 57, no. 2, pp. 97-103, 2005.
- [51] S. Weiss and H. M. Mueller, "'Too many betas do not spoil the broth': the role of beta brain oscillations in language processing," *Frontiers in Psychology*, vol. 3, p. 201, 2012.
- [52] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," Proceedings of the 2000 symposium on Eye tracking research & applications, 2000.
- [53] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 symposium on Eye tracking research & applications*, 2000: ACM, pp. 71-78.
- [54] J. Gilland, *Driving, eye-tracking and visual entropy: Exploration of age and task effects*. ProQuest, 2008.
- [55] M. A. V. Pamela S. Tsang, "MENTAL WORKLOAD AND SITUATION AWARENESS," pp. 243-262, doi: 10.1002/9781118131350.
- [56] P. A. Hancock and G. Matthews, "Workload and Performance: Associations, Insensitivities, and Dissociations," *Hum Factors*, vol. 61, no. 3, pp. 374-392, May 2019, doi: 10.1177/0018720818809590.
- [57] G. B. Pietro Aricò, Ilenia Graziani, Jean-Paul Imbert, G´eraud Granger, et al.. "Air-traffic-controllers (ATCO): neurophysiological analysis of training and workload," *Italian Journal of Aerospace Medicine, Italian Society of Aerospace Medicine (AIMAS)*, pp35, 2015.
- [58] M. J. Kochenderfer, *Decision Making Under Uncertainty theory and application*.
- [59] L. Longo and M. C. Leva, "Human Mental Workload Models and Applications," 2019, doi: 10.1007/978-3-030-32423-0_13.
- [60] J. S. M. Christipher D. Wickens, "Applied Attention Theory," 2008.
- [61] N. Moray, "Mental Workload: Its Theory and Measurements," 1979, doi: 10.1007/978-1-4757-0884-4_6.

- [62] J.-C. Sperandio, "The Regulation of Working Methods as a Function of Workload among Air Traffic Controllers," vol. 21, no. 3, pp. 195-202, 1978, doi: 10.1080/00140137808931713.
- [63] S. L. E. William B. Rouse, John M. Hammer, "Modeling the Dynamics of Mental Workload and Human Performance in Complex Systems," *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS*, vol. 23, 6, 1993.
- [64] N. Pattyn, X. Neyt, D. Henderickx, and E. Soetens, "Psychophysiological investigation of vigilance decrement: boredom or cognitive fatigue?," *Physiol Behav*, vol. 93, no. 1-2, pp. 369-78, Jan 28 2008, doi: 10.1016/j.physbeh.2007.09.016.
- [65] R. F. a. R. W. Picard, "Automatic prediction of frustration," 1998.
- [66] Y. L. Tian, T. Kanade, and J. F. Cohn, "Recognizing Action Units for Facial Expression Analysis," *IEEE Trans Pattern Anal Mach Intell*, vol. 23, no. 2, pp. 97-115, Feb 2001, doi: 10.1109/34.908962.
- [67] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010 2015: IEEE, doi: 10.1109/cvprw.2010.5543262. [Online]. Available: <https://dx.doi.org/10.1109/CVPRW.2010.5543262>
- [68] B. L. Brandon Amos, y Mahadev Satyanarayanan, "OpenFace A general purpose face recognition," June 2016.
- [69] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," 2015: IEEE, doi: 10.1109/cvpr.2015.7298682.
- [70] P. R. Tadas Baltrušaitis, Louis-Philippe Morency, "OpenFace: an open source facial behavior analysis toolkit."
- [71] P. R. Tadas Baltrušaitis, Louis-Philippe Morency, "Constrained Local Neural Fields for robust facial landmark detection in the wild," *2013 IEEE International Conference on Computer Vision Workshops*, 2013, doi: 10.1109/iccvw.2013.54.
- [72] T. Baltrusaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic Action Unit detection," 2015: IEEE, doi: 10.1109/fg.2015.7284869.
- [73] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," 2018: IEEE, doi: 10.1109/fg.2018.00019.
- [74] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing Parts of Faces Using a Consensus of Exemplars," vol. 35, no. 12, pp. 2930-2940, 2013, doi: 10.1109/tpami.2013.23.
- [75] K. Sundaraj, "Investigation of Facial Artifacts on Face Biometrics."
- [76] P. N. Gardi A. , Sabatini R., Kistan T, "Development of a Cognitive HMI for Air Traffic Management Systems – HMI Specification document," RMIT/SENG/CPS/002-2019, Aug. 2019.
- [77] P. N. Gardi A., Planke L., Lim Y., Kistan T., Sabatini R., "Development of a Cognitive HMI for Air Traffic Management Systems – Report No. 4: CHMI2

- System-Level Implementation and Verification in a Representative Simulation Environment," RMIT/SENG/CPS/003-2019, Aug. 2019.
- [78] K. R. Yixiang Lim, Alessandro Gardi, Neta Ezer, Roberto Sabatini, "Human-Machine Interfaces and Interaction."
- [79] P. P. Caffier, U. Erdmann, and P. Ullsperger, "Experimental evaluation of eye-blink parameters as a drowsiness measure," *Eur J Appl Physiol*, vol. 89, no. 3-4, pp. 319-25, May 2003, doi: 10.1007/s00421-003-0807-5.
- [80] J. B. W. Joseph F. Grafsgaard, Kristy Elizabeth Boyer, Eric N. Wiebe, James C. Lester, "Automatically Recognizing Facial Expression: Predicting Engagement and Frustration."
- [81] M. W. Schurgin, J. Nelson, S. Iida, H. Ohira, J. Y. Chiao, and S. L. Franconeri, "Eye movements during emotion recognition in faces," *J Vis*, vol. 14, no. 13, p. 14, Nov 18 2014, doi: 10.1167/14.13.14.
- [82] M. Iwasaki and Y. Noguchi, "Hiding true emotions: micro-expressions in eyes retrospectively concealed by mouth movements," *Sci Rep*, vol. 6, p. 22049, Feb 26 2016, doi: 10.1038/srep22049.
- [83] E. Vlemincx, J. Taelman, S. De Peuter, I. Van Diest, and O. Van Den Bergh, "Sigh rate and respiratory variability during mental load and sustained attention," vol. 48, no. 1, pp. 117-120, 2011, doi: 10.1111/j.1469-8986.2010.01043.x.
- [84] J. P. K. Taija Lahtinen, Tomi Laitinen, Tuomo K Leino, "Heart rate and performance during combat missions in a flight simulator," *Aviation Space and Environmental Medicine* 78.
- [85] M. Grassmann, E. Vlemincx, A. von Leupoldt, J. M. Mittelstadt, and O. Van den Bergh, "Respiratory Changes in Response to Cognitive Load: A Systematic Review," *Neural Plast*, vol. 2016, p. 8146809, 2016, doi: 10.1155/2016/8146809.
- [86] C. K. L. Sun K. Yoo, "Changes-in-EEG-and-HRV-during-Event-Related-Attention," *International Journal of Biomedical and Biological Engineering*.
- [87] R. L. M. Burcu Cinaz, Bert Arnrich and Gerhard Tröster, "MONITORING OF MENTAL WORKLOAD LEVELS."
- [88] K. J. Jaquess *et al.*, "Empirical evidence for the relationship between cognitive workload and attentional reserve," *Int J Psychophysiol*, vol. 121, pp. 46-55, Nov 2017, doi: 10.1016/j.ijpsycho.2017.09.007.
- [89] P. Arico *et al.*, "Adaptive Automation Triggered by EEG-Based Mental Workload Index: A Passive Brain-Computer Interface Application in Realistic Air Traffic Control Environment," *Front Hum Neurosci*, vol. 10, p. 539, 2016, doi: 10.3389/fnhum.2016.00539.
- [90] M. C. L. Luca Longo, "Human Mental Workload Models and Applications.."
- [91] A. Kartali, M. M. Janković, I. Gligorijević, P. Mijović, B. Mijović, and M. C. Leva, "Real-Time Mental Workload Estimation Using EEG," in *Human Mental Workload: Models and Applications*, (Communications in Computer and Information Science, 2019, ch. Chapter 2, pp. 20-34.
- [92] L. Giraudet, J. P. Imbert, M. Berenger, S. Tremblay, and M. Causse, "The neuroergonomic evaluation of human machine interface design in air traffic control using behavioral and EEG/ERP measures," *Behav Brain Res*, vol. 294, pp. 246-53, Nov 1 2015, doi: 10.1016/j.bbr.2015.07.041.

- [93] D. Tao, H. Tan, H. Wang, X. Zhang, X. Qu, and T. Zhang, "A Systematic Review of Physiological Measures of Mental Workload," *International Journal of Environmental Research and Public Health*, vol. 16, no. 15, p. 2716, 2019, doi: 10.3390/ijerph16152716.
- [94] N. P. Yixiang Lim, Alessandro Gardi, Roberto Sabatini, "Dynamic_Airspace_Management_for_Enhanced."
- [95] A. Cuzzocrea, E. Mumolo, and G. M. Grasso, "An Effective and Efficient Genetic-Fuzzy Algorithm for Supporting Advanced Human-Machine Interfaces in Big Data Settings," *Algorithms*, vol. 13, no. 1, 2019, doi: 10.3390/a13010013.
- [96] J.-H. Zhang and J. Raisch, "Hybrid-Data-Driven Fuzzy Recognition of Operator Cognitive State in Human-Machine Cooperative Control System," *IFAC Proceedings Volumes*, vol. 46, no. 13, pp. 327-334, 2013, doi: 10.3182/20130708-3-cn-2036.00001.
- [97] S. S. a. S. N. D. S.N.Sivanandam, *Introduction_to_Fuzzy_Logic_using_MATLAB*. 2007.
- [98] R. J. Jyh-Shing, "Adaptive-Network-Based Fuzzy Inference System," *IEEE transaction on systems, man, and cybernetics*, vol. 23, 1993.
- [99] U. M. Rao, Y. R. Sood, and R. K. Jarial, "Subtractive Clustering Fuzzy Expert System for Engineering Applications," vol. 48, pp. 77-83, 2015, doi: 10.1016/j.procs.2015.04.153.
- [100] R. J. Jyh-Shing, "ANFIS_adaptive-network-based fuzzy," *IEEE Transaction on system, man, and cybernetics*, vol. 23, no. 3 1993.
- [101] B. N. a. H. S. B. Janmenjoy Nayak, "Fuzzy C-Means (FCM) Clustering Algorithm: A Decade Review from 2000 to 2014," *Computational Intelligence in Data Mining - Volume 2*, doi: 10.1007/978-81-322-2208-8_14.
- [102] J. Liu, A. Gardi, S. Ramasamy, Y. Lim, and R. Sabatini, "Cognitive pilot-aircraft interface for single-pilot operations," *Knowledge-Based Systems*, vol. 112, pp. 37-53, 2016, doi: 10.1016/j.knosys.2016.08.031.