Co-funded by the Horizon 2020
Framework Programme of the European Union

HORIZON2020 FRAMEWORK PROGRAMME

ICT – 21 -2014

Advanced digital gaming/gamification technologies

ProsocialLearn

Gamification of Prosocial Learning

for Increased Youth Inclusion and Academic Achievement

D3.1

User data acquisition and mapping in game environments

Document Control Page

WP/Task	WP3 / T3.1
Title	D3.1 User data acquisition and mapping in game environments
Due date	30/09/2015
Submission date	21/10/2015
Abstract	This document describes algorithms for person-dependent affect analysis using gameplay interaction features, video, text analysis and evaluation on already existing datasets.
Author(s)	Konstantinos Apostolakis (CERTH)
Contributor(s)	Spyridon Thermos (CERTH), Kyriaki Kaza (CERTH), Athanasios Psaltis (CERTH), Kyriakos Stefanidis (CERTH), Kosmas Dimitropoulos (CERTH), Petros Daras (CERTH), Lee Middleton (ITINNOV), Stefano Modafferi (ITINNOV), Christopher Peters (KTH), Laura Vuillier(UCAM)
Reviewer(s)	Francesco D'Andria (ATOS)
Dissemination level	□ internal ☒ public □ confidential

Document Control Page

Version	Date	Modified by	Comments
0.1	19/08/2015	Kosmas Dimitropoulos	First draft, Table of Contents
0.2	25/08/2015	Lee Middleton	TOC expansion
0.3	28/08/2015	Konstantinos Apostolakis	First circulation draft
0.4	11/09/2015	Konstantinos Apostolakis	Integration of partner input
0.5	23/09/2015	Konstantinos Apostolakis	Pre-final version Implementation of comments received from partners
0.6	20/10/2015	Konstantinos Apostolakis	Formal review final version

List of Abbreviations

Abbreviation	Description
ACE	Automatic Contrast Enhancement
API	Application Programming Interface
ASM	Active Shape Model
AU	Action Unit
FACS	Facial Action Coding System
FERA	Facial Expression Recognition and Analysis
HCI	Human-Computer Interaction
KKCQ	Kitty King's Candy Quest (game)
MFCC	Mel-frequency cepstral coefficients
NUI	Natural User Interface
PLO	Prosocial Learning Objective
PoT	Path of Trust (game)
ROI	Region of Interest
SDK	Software Development Kit

Executive summary

The current deliverable is one of the outputs of work package **WP3: Player Modelling and Prosocial Affect**. It is a public document focusing on the **Deliverable 3.1: User data acquisition and mapping in game environments**.

In this report, user data acquisition channels are distinct into three categories, namely the sensory observation channels, In-game data logs and static data. The first comprises the full-range of player behavioral data, which is directly observable through any number of audio/visual and motion sensors. Each observation channel is examined with regard to player affect recognition, and where applicable, hints towards engagement measuring are presented. Sufficient scientific literature works are presented to justify the choice of feature descriptors per observation channel. These features are further processed to recognize affective properties per channel. Algorithms employed for affect recognition are presented, and where possible, evaluated on already existing datasets. The second observation channel concerns in-game data logs and player interaction patterns with the game itself (using traditional input devices such as the mouse or keyboard) as well as other players (via chat message sentiment analysis software to be incorporated onto the data acquisition platform). The final observation channel covers player psychological profiles and immediate gaming area contextual information to provide a complete description of the conditions in which a prosocial game session takes place.

In order to demonstrate and verify applicability of the observations acquisition channels described above, two prototype games have been developed by members of the consortium, with intent on collecting player input data for evaluation of the proposed algorithms in real-life scenarios and preliminary work on a generic framework for sensor connectivity is carried out. A demonstrative mapping of sensors, modalities, features and target output data is presented per game.

Index

List of Figures

List of Tables

1 Introduction

This section provides detailed information about the purpose of WP3 in general, placing of this Task and accompanying deliverable, as well as the scope and structure of the document, which set the tone for the intended audience and interested readers.

The aim of WP3 is to develop data fusion algorithms for the analysis of prosocial affect as well as define player models, all of which are intended for both offline and online adaptation of prosocial game content and the assessment of learning outcomes with regard to the Prosocial Learning Objectives (PLOs) set beforehand by teachers/parents/caregivers for student game players. Player models will encapsulate a summary of key characteristics influencing the students' potential to achieve prosocial learning outcomes, and ultimately will serve as a prosocial abstraction of any given player. Each player will have a unique player profile which will be adapted over time according to this particular player's gameplay interaction patterns and ability to reach a number of assigned PLOs. Adaptation of the player profile will depend on a persistence mechanism, which will keep track of player historical profiles, as well as online data fusion taking place during gameplay sessions, in which player interaction patterns are fused with subtle audio/visual affective cues. The results of this fusion are mapped to a multiple-axis prosocial affective space in order to determine whether adjustments need to be made to the game levels of difficulty (perceived challenge) and game graphical content to ensure high levels of player engagement. In turn, maintaining this balance between player skills and game challenge will produce the desired amount of engagement, which will maximize the potential of students in terms of achieving their assigned PLOs by completing certain tasks in the game environment.

Future deliverables D3.2, D3.3 and D3.4 will outline the continuous work on both prosocial affect fusion algorithms and player modelling. This document will focus on the core user data acquisition by monitoring gameplay interaction patterns and behavioral response signals captured using a reliable and scalable framework of sensory apparatus using a multitude of state of the art techniques.

1.1 Purpose of the document and positioning of Task 3.1

The purpose of Task 3.1, which concludes with the delivery of this document, is to identify and analyze all player input data coming from various incoming signals as well as player background knowledge, explore game input modalities beyond the scope of game interactions and process raw observations to provide input to the affect fusion algorithms (Task 3.2). Throughout the timeline of this Task, partners involved with player modelling and prosocial affect input modalities have identified prosocial observation acquisition channels in the following behavioral signals:

- Visual information coming from facial analysis, including facial expressions and gaze analysis.
- Visual information coming from body motion analysis, including full body motion, as well as head and hand movements, with the use of consumer-grade motion sensing gaming hardware.
- Audio information coming from speech emotion analysis.
- Gameplay data, including player interactions with the environment and transactions with other players (i.e. via chat messages).
- Contextual information regarding the conditions of the actual location where students are playing prosocial games and are being monitored by the multi-sensor platform.
- Player background/psychological profile data.

Each of these observation acquisition channels is comprised of a multi-layered set of features which have been acquired by studying already existing datasets on children interaction with serious games monitoring, or collected as part of prosocial studies (WP7). All features extracted as part of the work conducted in this Task are well-defined in the related scientific literature on cognitive and emotional state analysis. Evaluation schemes on existing datasets and testbeds for observation data local processing, i.e. turning low-level features to higher-level concepts of affect have also been investigated during this Task.

In this deliverable, we will describe the algorithms and processes used for person-dependent affect analysis using multi-channel observation acquisition, ranging from gameplay interaction features to audio/visual cues. This document will present in detail, a complete set of features extracted from every observation acquisition channel, and where applicable, elaborate on the evaluation process of the implemented algorithms on already existing datasets. A thorough scientific background on each modality's contribution to the affective state and related references on feature extraction techniques and algorithms are provided in text to solidify our approach with regards to each channel.

1.2 Scope and Audience of the document

The dissemination level of this document is public. This document will be made available on the project website for external parties interested in user data acquisition via behavioral signal cues, in-logging and player profiling. It is hoped that the report will assist interested parties in understanding why and how each modality is linked to player current affect and engagement states, and provide a base towards mapping this type of input data in prosocial game environments. We present this mapping to aid future game developers undertaking the task of creating games for the ProsocialLearn platform in incorporating sensory hardware onto their games. Towards this end, a number of features is thoroughly presented, that can be collected in real-time and post-session to feed forward to the prosocial affect fusion modules.

1.3 Structure of the Document

This document contains the following key sections, conveniently detailed in the list below:

Section 1: Introduction – an introductory section, i.e. this present section, which describes the WP as a whole, as well as the main purpose of the Task that generated this document.

Section 2: Sensory Observation Channels – this section will present and detail all the different features extracted through audio/visual sensory observation techniques and related to player subtle behavioral response cues during gameplay. An overview of all audio/visual sensors will be delivered. Each modality presented will include the total number of features detected, the techniques used for detection and feature extraction as well as evaluation on either existing datasets or data collected as part of prosocial studies (WP7).

Section 3: In-game Data Logs – this section describes the data logging process that will be supported by the ProsocialLearn platform, as well as propose a summary of tools and features collected online and in-game, with regards to student gameplay interaction patterns using keyboard and mouse interfaces as well as sentiment analysis performed on non-verbal communication (i.e. chat messages, emoticons, etc.) with other players.

Section 4: Static Player Data – this Section delves into the data remaining constant throughout the duration of the gameplay session. Information stemming from player profiles, persistence mechanisms and contextual information regarding the immediate area surrounding the physical space in which students are gaming will be explored in this Section.

Section 5: Multi-modal/Multi-sensor Capturing Setup – this section describes a generic approach for sensor/device connectivity and in-game observation data flow, according to the specifications of the overall ProsocialLearn requirements and architecture presented in D2.3. The aim is to present initial developments that will ensure input signal processing is available and robust for integration into the 1st ProsocialLearn prototype platform.

Section 6: Conclusion – this section presents the conclusion of the document.

2 Sensory Observation Channels

In this Section, the sum of all input modalities will be described in accordance to audio/visual signals retrieved from a variety of sensory apparatus plugged into the data acquisition platform. In the remainder of this Section, we will focus on describing the different sensors contributing to the data acquisition process (2.1), the extraction of audio/speech features (2.2), an overview of all visual data acquired for both facial expression and gaze analysis (2.3) and finally an insight on body motion analysis data collected using consumer-grade motion sensing techniques (2.4).

2.1 Capturing sensors

Below is a full summary of all the sensory hardware used in the ProsocialLearn data acquisition framework. Sensors can be used together, where applicable in order to provide multi-modal user capturing, although in some cases, the use of one sensor explicitly excludes the use of another (i.e. Kinect/LEAP Motion). An overview of a generic framework for sensor connectivity is provided in Section 5.

2.1.1 Microphone

A microphone is a sensor that converts sound into electrical signals. As it is the only device capable of capturing human speech it is commonly used in speech recognition and analysis. In ProsocialLearn the built-in microphones on modern computing devices (laptops, tablets) or on cameras are employed to extract raw audio data. Due to the proximity to the individual the microphones are able to extract voice information well. The voice signal needs to be cleaned and analyzed as described in Section 2.2.

2.1.2 Camera

Probably the most common sensor device used for monitoring user behavior during a multitude of Human-Computer interaction (HCI) activities is the standard (web-) camera. HCI activities vary in the scientific literature from implicit tagging [***Apostolakis & Daras, 2014***] to affective video viewing [***Soleymani et al, 2012a***] and monitoring of gameplay behavior [***Shaker et al, 2011***]. Cameras are particularly useful for close-up monitoring of the user's face, or in several cases the user's upper body [***Gunes & Piccardi, 2009***]. Studies vary in the use of only a single or multiple cameras, according to the accuracy required and cost-efficiency targeted by each respective proposed framework, and potential application uses. It is customary to place the sensor(s) close to the computer monitor, in an attempt to record user direct interaction with the viewed content. In this respect, the camera(s) is(are) placed either on top or in front of the user's monitor, depending on the intended best viewing angle per application.

In the ProsocialLearn data acquisition platform, a single standard web camera is used for acquiring a continuous flow of user facial images. We impose no special specification in terms of camera type or model, other than the capacity to record High Definition (HD) video. We find this specification non-limiting; as generally, most high- and mid-range laptops on the market at the time of writing this document are shipped with an onboard camera right on top of the built-in monitor's viewing area. In cases where a standard desktop PC is to be used, a standard-issue HD webcam is mounted on top of the desktop monitor. Raw camera frames obtained from the sensor will serve to capture a clear image of the student's face and eyes, which will be used for the extraction of facial expression and gaze analysis features, as described later in this Section.

2.1.3 Kinect

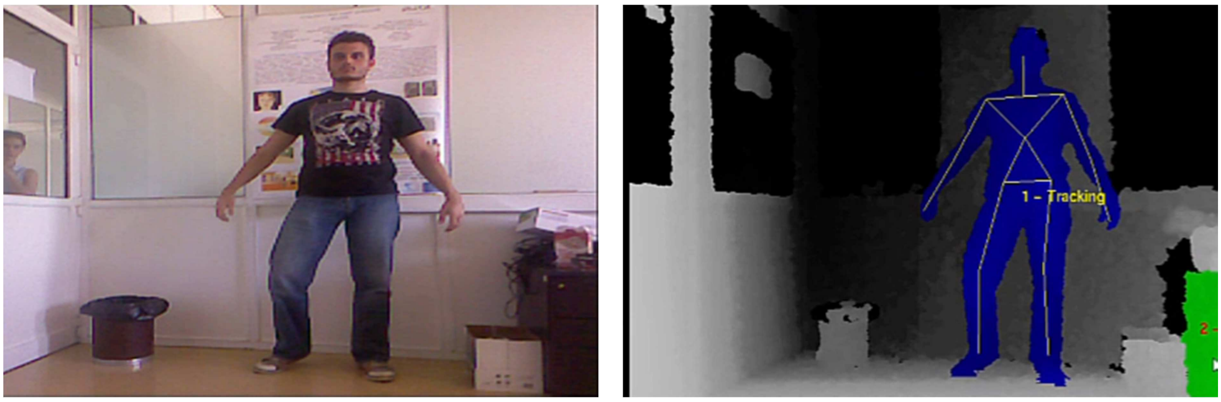
Late in 2010, developer Microsoft attempted to revolutionize the way game players interacted with their games and gaming hardware (Xbox 360[1]) by introducing the Kinect, the first in a line of motion sensing input devices intended to replace the traditional controller. The original device featured an RGB camera, a depth sensor and a multi-array microphone running proprietary software, and builds on software technology developed internally and range camera technology developed by the now defunct Israeli company PrimeSense. Instantly, the device broke the Guinness World Record for fastest selling consumer electronics device, and was positively accepted by the homebrew software and scientific community [***Zhang, 2012***]. In 2012, Microsoft introduced a Windows version of the device along with an official SDK to provide developers with the Kinect capabilities in order to build applications with C++, C# and Visual Basic using Microsoft Visual Studio 2010 and beyond. An upgraded version marketed with the company's latest gaming hardware (the Xbox One[2]) was released late in 2013.

Since the original Kinect for Xbox 360 sensor broke ground in 2010, several research topics surrounding low-cost motion capturing systems emerged [***Berger et al, 2011***]. More specifically, research work capitalized upon full-body skeleton tracking capabilities offered for the device through a number of software development kits created for the sensor, most notably by Kinect project partner PrimeSense (*OpenNI*), until the release of the official SDK by Microsoft [***Shotton et al, 2012***]. Shortly, research works on rehabilitation based on bodily activity patterns during gameplay emerged, capitalizing on the sensor's low cost and firmly established affiliation with the games industry [***Chang et al, 2011***] [***Lange et al, 2011***]. The notion was subsequently extended to include affective studies based on the extraction of full body motion features [***Piana et al, 2013***].

In the ProsocialLearn data acquisition platform, we utilize both Kinect hardware for the analysis and extraction of features related to the players' 3D full body motion. In this respect, we monitor and process data incoming from every visual component provided by, and the SDKs built for the sensor.

[1] Microsoft XBOX 360 console, http://www.xbox.com/en-US/xbox-360?xr=shellnav

[2] Microsoft XBOX One console, http://www.xbox.com/en-US/xbox-one?xr=shellnav



(a) (b)

Figure 1 - Sample data captured by the Kinect for Xbox 360 sensor

Left image (a) shows RGB input, while the right image (b) shows depth and skeleton tracking results obtained from the OpenNI library

More specifically, features are extracted with respect to the RGB camera frame, the 16-bit depth map as well as the skeletal joint locations and orientations provided by the OpenNI[3] and Kinect SDK packages for the original Kinect for Xbox 360 and Kinect for Xbox One respectively. Figure 1 demonstrates the raw data input of our feature extraction algorithms for both sensors.

2.1.4 LEAP Motion Sensor

The LEAP Motion controller is a computer hardware sensor device that supports contact-less hand and finger motions as input, allowing users to control applications via hand gestures and motions. The device is meant to be placed on the tabletop, and uses infra-red cameras and LEDs to generate a hemispherical, pattern-less IR light which is able to detect hand movement to a distance of about 600 millimeters [***Guna et al, 2014***]. According to the developer specification, user input is detected and analyzed at about 300 frames per second by the controller's built-in software, which uses complex maths to generate high-precision 3D hand position and joint data. Much like the Kinect, the LEAP Motion controller uses depth information to extract skeleton data, however, it's smaller observation area and higher resolution make it ideal for hand and finger tracking, whereas the Kinect is more suitable for full body tracking. The device was publicly made available in mid-2013, and has since been embraced by the Virtual and Augmented reality application development community.

Several obvious potential use cases of the sensor for the scientific community have since emerged, involving hand gesture recognition [***Schmidt et al, 2015***], sign language interpretation [***Potter et al, 2013***] and more recently, rehabilitation [***Grubišić et al, 2015***]. These studies display the controller's efficiency with hand tracking, concluding however that further development on the device software

[3] PrimeSense shutdown the original OpenNI project on which some of the capturing modules used in our data acquisition framework are linking to, when it was acquired by Apple on November 24, 2013. The modules retain their operability through the latest legacy version of the library (1.5.4.0 as of May 7, 2012).

API is required. At the time of writing these lines however, we are not aware of any studies that incorporate the LEAP Motion controller into an affect-sensing framework. Seeing how hand movement is considered to have a strong connection with the expression of emotion [***Pollick et al, 2001***] [***Gunes et al, 2015***], we aim at capitalizing on the high precision provided by the sensor to extract hand-specific features believed to be related to the expression of emotion [***Kessous et al, 2010***]. As the LEAP Motion controller has a limited range and is required to be set on top of an actual desktop, it is an ideal candidate for fusion with the facial expression and gaze modalities retrieved by positioning a camera on top of the user's monitor.

2.1.5 Standard input devices

We consider traditional HCI interfaces such as the keyboard, mouse or even the occasional game controller as standard input devices. Although these barely classify as "sensors", they provide the most recognizable means of interfacing with a game environment. Within ProsocialLearn, monitoring of user interaction patterns may implicitly reveal cues in prosocial affect and engagement through affect recognition [***Salmeron-Majadas et al, 2014***], or the inference of user interest via mouse or keyboard activity [***Claypool et al, 2001***]. We will elaborate on the sum of features collected through this interface in Section 3.3.

2.2 Audio data acquisition

In normal conversation the voice conveys a wide range of emotions. This emotion gives an insight in to the emotional state of the individual. This effect has been widely studied by psychologists mostly due to its role in diagnosis and treatment of a number of psychiatric illnesses [***APA 2013***]. More recently, the pattern analysis community has become interested in the problem due to the usefulness in general speech understanding problems and HCI. In ProsocialLearn we are using speech data as another channel from which to analyze the emotion of the children playing the games. This section provides an analysis of the problem and some specific techniques by which it can be achieved.

2.2.1 Selection of an appropriate dataset

In the literature there is a vast number of different datasets collected for extracting emotion from speech. These datasets can be analyzed via a number of discrete categories. These are:

- Natural versus Acted
- Level of classification
- Language
- Adults or children

Natural and acted datasets are produced via either a natural conversation or an actor following a script. Acted conversations are incredibly rare in datasets of children, for obvious reasons. Additionally, acted datasets are considered to be very different from natural speech and classifiers for acted speech often perform poorly on real speech [***Williams & Stevens, 1972***]. However, it is much easier to collect acted data than natural. The level of classification refers to which level the utterances have been classified with a specific emotion. Typically human experts perform this either at the sentence or word level. Generally, we would like emotions as they evolves so would like our classifications based on sound samples rather than, at a minimum, entire words. The final two categories should be self-evident and serve to provide some useful contextual information for our system. In Table 1 – Some common emotion from speech databases lists a number of different databases used in recent literature classified according to these categorizations.

Name	Natural/Acted	Level	Language	Adults/ Children
FAU Aibo [***Steidl, 2009***]	natural	words	german	children
Enterface [***Martin et al, 2006***]	acted	sentences	english	adults
AVEC [***Schuller et al, 2011***]	natural	words	english	adults
Recola [***Ringeval et al, 2013***]	natural	words	french/italian/german	adults
SES [***Montero et al, 1998***]	acted	sentence	spanish	adults

Table 1 – Some common emotion from speech databases (chosen one highlighted in green)

As the target audience is children the FAU Aibo dataset was used to build the classifiers. In this decision we are asserting that, within the context of emotion that adults and children are distinct [***Potamianos & Narayanan, 2007***], that emotion is independent of language spoken [***Bhatti et al, 2008***], and that words are sufficient to build our models.

2.2.2 The FAU Aibo dataset

The collection of the data was performed in an experiment with children where they interacted with a Sony Aibo robot. There were two specific types of interactions. A short one where the child guided the Aibo to one of several feeding bowls and a longer one where they directed the Aibo around a course asking it to perform certain activates at various stages. The children were told to treat it as if it were a real pet and to praise or reprimand based on the exhibited behavior.

The data was stored as mono WAV files sampled at 16kHz. The recordings were carried out at two different schools in Germany. Further details about the data are described in Table 2 – Simple metrics about the datasetTable 2 – Simple metrics about the dataset

Demographics	Data	Sound
51 children 30 female, 21 male	18216 files school 1 = 9959, number school 2 = 8257	~9 hours Median word length = 1.62s

Table 2 – Simple metrics about the dataset

A histogram illustrating the breakdown of the dataset in terms of sample (or word) length is shown in Figure 2 – Distribution of sample sizes.

The key feature to note is that most of the words are short (less than a few seconds). The specific range of word lengths is from 0.11s to 24.54s.

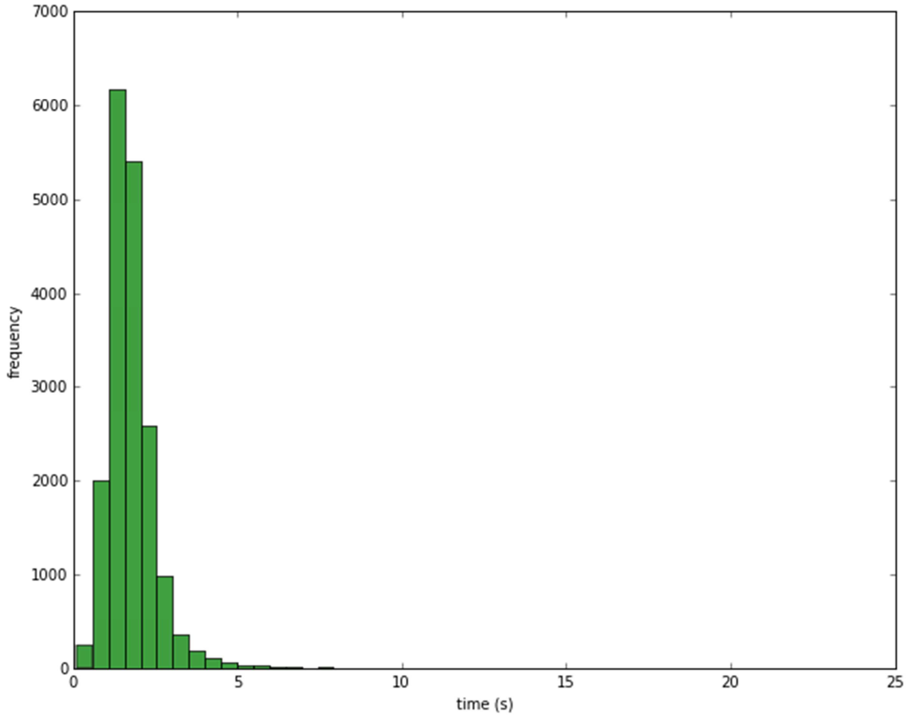


Figure 2 – Distribution of sample sizes

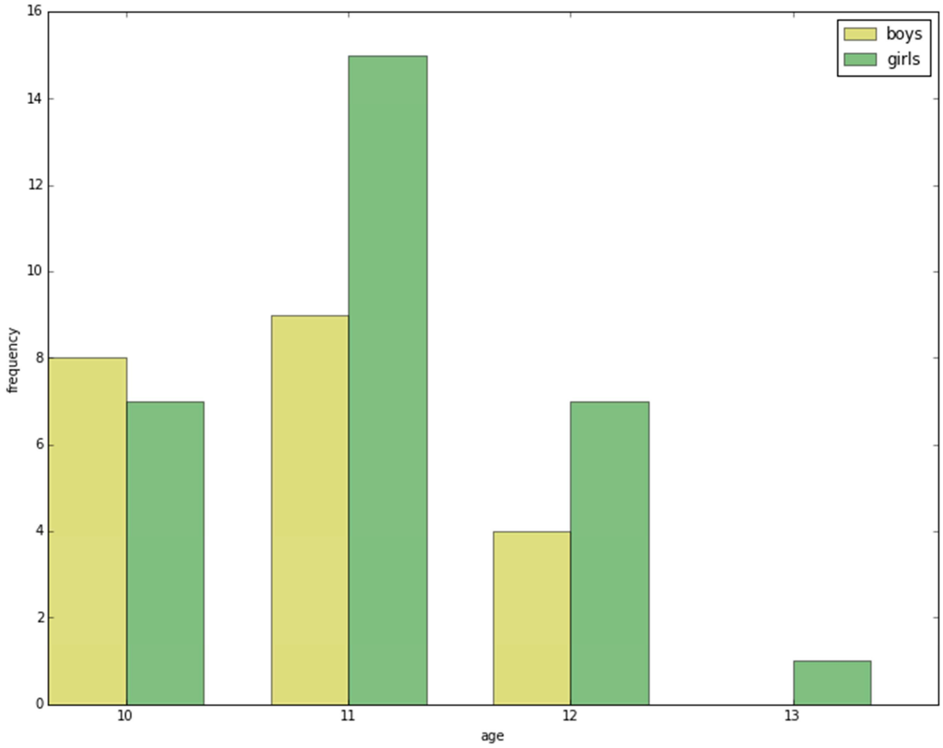


Figure 3 – Distribution of the ages of the children

The age distribution by gender of the dataset is illustrated in Figure 3 – Distribution of the ages of the children

This illustrates a median age of 11 years old with a bias towards girls.

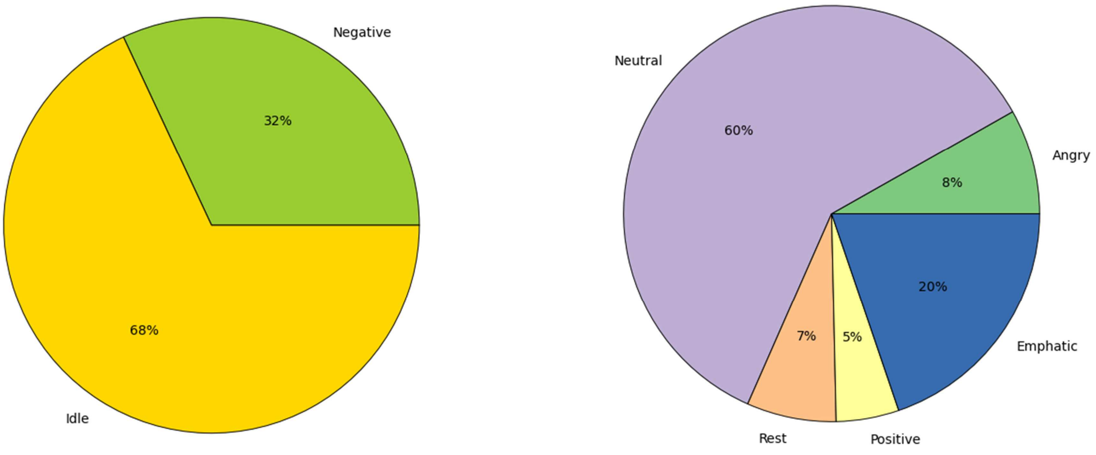
As the dataset is labelled we can examine the specific makeup of the labels. There are multiple sets of labels provided. The specific labels are described in Table 3 – The specific labels in the dataset

and a breakdown of the labels is in Figure 4 - Distribution of data labels (a) 2-class (b) 5-class

Notice that in both sets of labels the class representing no emotion (for 2-class this is Idle and for 5-class this is Neutral) is approximately two thirds of the label set. Thus the classes are exhibiting a large imbalance which must be corrected at the classifier stage.

2-class	5-class
Negative (NEG) Idle (IDL)	Angry Emphatic Neutral Positive Rest

Table 3 – The specific labels in the dataset



(a) (b)

Figure 4 - Distribution of data labels (a) 2-class (b) 5-class

2.2.3 Speech feature extraction

The work described in this section follows that of the INTERSPEECH 2009 (IS2009) emotional recognition challenge [***Schuller et al, 2009***]. This was the first attempt to bring together all the work on automated emotion recognition and perform a standard comparison on standard data. The aim was to provide a baseline which new algorithms and features could be evaluated against.

Additionally a set of standard profiles were also provided for feature extraction. This profile used OpenSMILE [***Eyben et al, 2013***], which is a tool which allows you to build batch oriented pipelines for the analysis of sound.

The basic flow of analysis of the speech for a single chunk of sound is illustrated in Figure 5 – Feature extraction from speech

Raw sound files enter the pipeline on the left and proceed through all the various processing steps until reaching the far right.

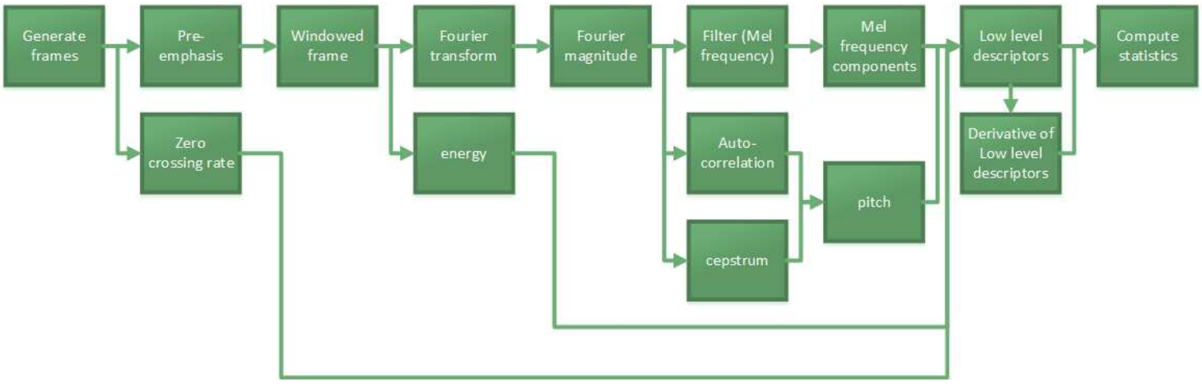
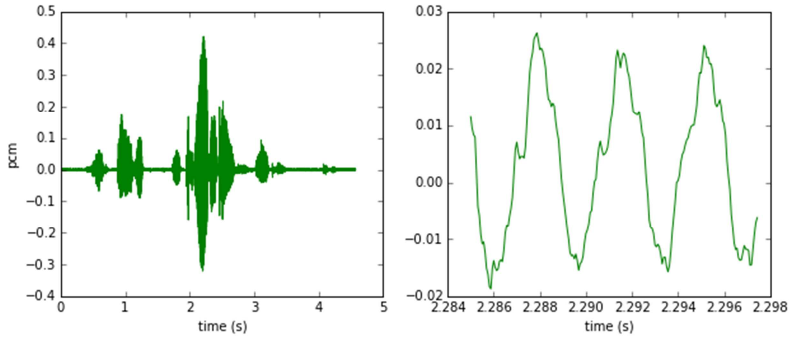


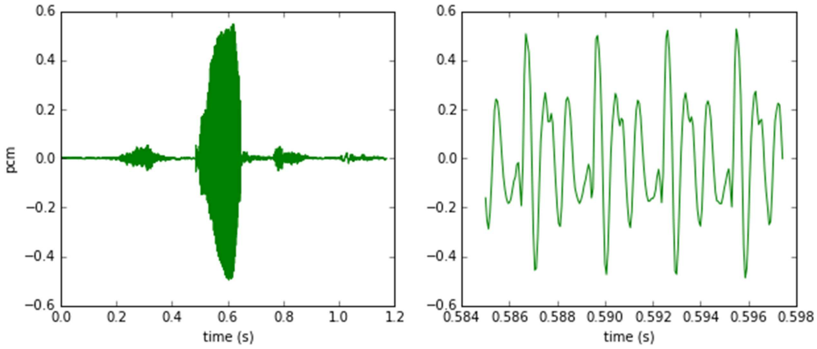
Figure 5 – Feature extraction from speech

To aid in understanding the various steps in the figure a pair of example sound samples will be employed. These are illustrated in Figure 6 – Two input waveforms (a) Idle (b) Negative

There are notable visual differences between the two different waveforms. This is illustrated more clearly via the zoomed version of the waveforms on the right of each sub-figure.



(a)



(b)

Figure 6 – Two input waveforms (a) Idle (b) Negative

In Figure 5 – Feature extraction from speech the next two steps serve to flatten the magnitude spectrum and balance out the high and low frequencies. The zero crossing rate is the rate of sign changes along the time axis of a signal. It occurs when the signal transitions from positive to negative or negative to positive. An example of the zero crossing rate for the idle waveform is pictured inFigure 7.

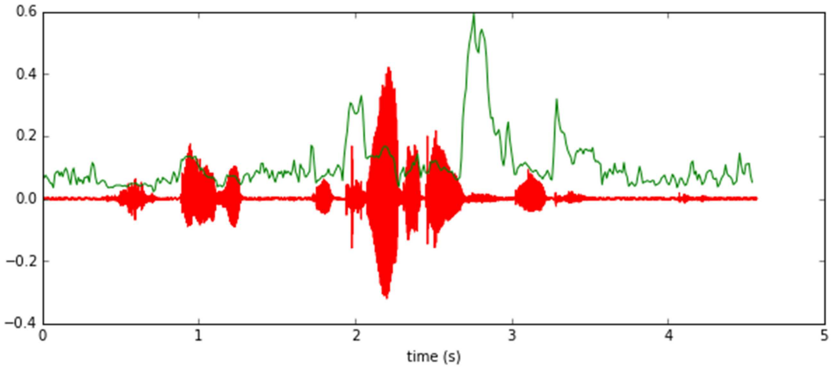


Figure 7 -Original waveform compared with the zero crossing rate (green)

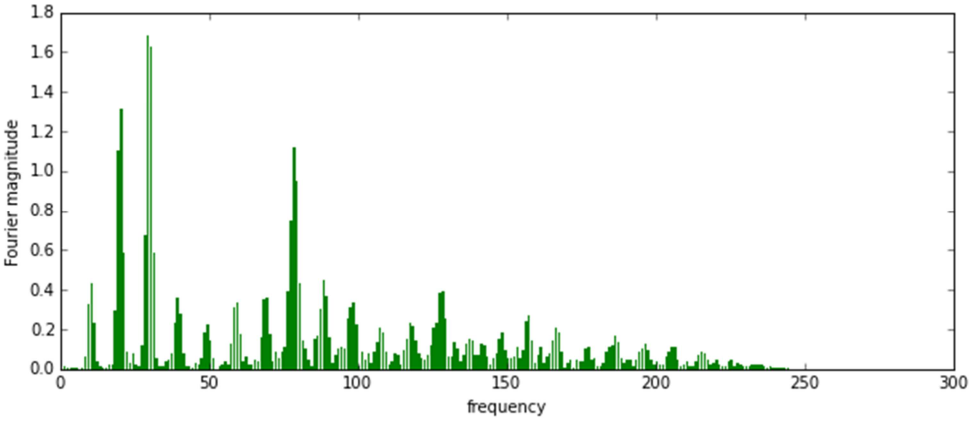


Figure 8 - Fourier magnitude for the 235 time window in the "Idle" signal

The Fourier transform is computed for each specific time window. A visualization of the "Idle" waveform for approximately 2.35 seconds is illustrated in Figure 8. Only the magnitude is shown as this is used in subsequent stages of the processing chain.

The energy within each time window is now computed. A figure showing a comparison of the original signal with the energy is shown in Figure 9. Note that peaks in the energy correspond to the portions of the signal with rapid transitions. These transitions have been shown to be indicative of emotions within the speech.

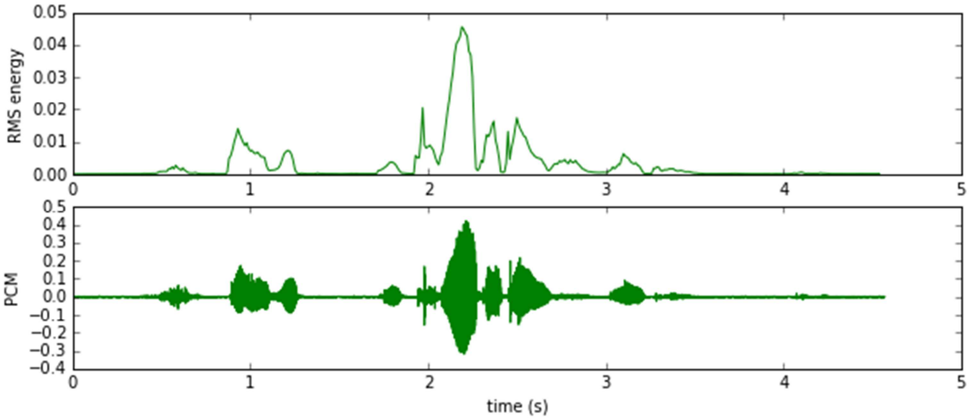


Figure 9 -Energy of the "Idle" waveform

The Mel-frequency cepstral coefficients (MFCC) are derived from the Fourier transform magnitude of the signal. Like the Fourier transform magnitude this computes a number of values for each specific frame in the original waveform. Two consecutive frames are shown in Figure 10. Note that the coefficients are changing rapidly even within such a short time frame. The MFCCs have shown efficacy in measuring emotion from voice.

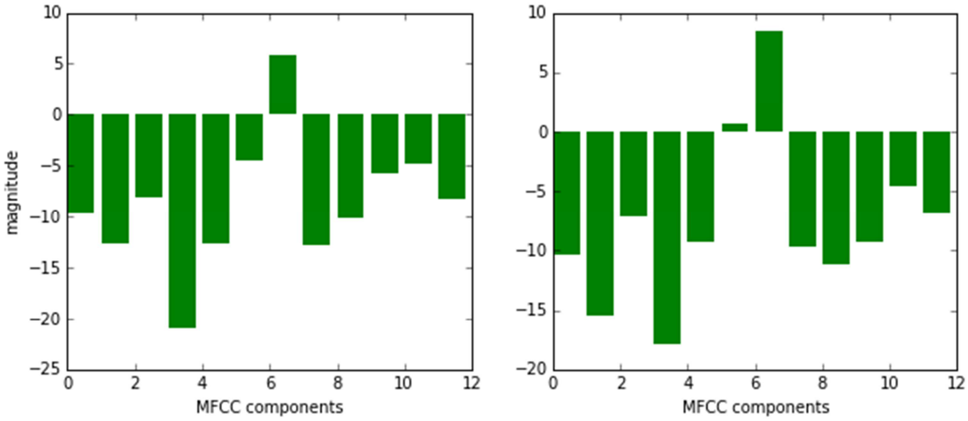


Figure 10 - MFCC components of two consecutive frames of the Idle waveform

The final of the low level descriptors that are computed are the pitch. Intuitively pitch is related to emotion as it changes when people are emotionally charged. For a negative emotion the track of the fundamental frequency (F_0) and the original waveform are shown in Figure 11.

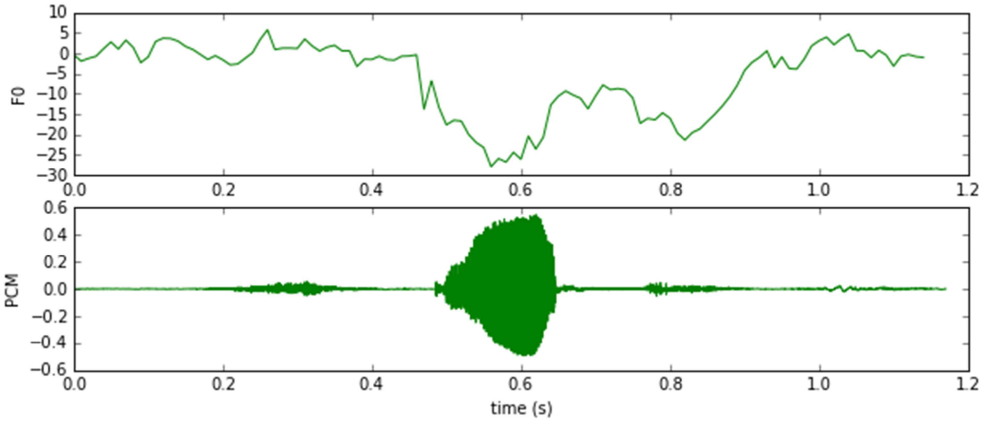


Figure 11 - Track of the fundamental frequency for a negative emotion

Processing the data this way leads to a very large number of features for each sample within the original signal. The breakdown is shown in Table 4. In addition the first derivatives of all the quantities are also computed. So to compute all the features for the dataset (described in sub-Section 2.2.2) would involve between 542 and 132790 features.

Obviously, the large number of features generated is intractable for analysis. The solution is to reduce the feature set by modelling the feature values computed across a word. This has the secondary effect of reducing the impact of noise. The final reduced set of features is illustrated in Table 5. In the table there are a number of low level descriptors (and their first derivatives) and functionals (scalar valued functions which can be used to represent the low level descriptors).

Feature	Number of values
Zero crossing	1
Fourier Magnitude	256
Mel-frequency cepstral coefficients	12
Pitch	1
Voiced probability	1
total	**271**

Table 4 -Features computed each frame

Low level descriptors	Functionals
Zero crossing rate	Mean
RMS energy	Standard deviation
F_0	Kurtosis, skewness
Harmonic noise ratio	Extremes: value, relative position, range
MFCC	Linear regression: offset, slope, mean square error

Table 5 -Complete set of final features

The low level descriptors are chosen to be ones which have shown to be effective in detecting emotion. The functionals are computed for each descriptor. This leads to a total of 384 features for each word chunk.

2.2.4 Building a classifier for speech emotion extraction

Before training the classifiers the data needs to be partitioned. In the original INTERSPEECH emotion challenge the partitioning was performed by school. However, the size of the datasets this way is imbalanced (one school has a 1000 more samples than the other). Furthermore the presentation order is based on speaker order which may impact the training process. In order to ameliorate these issues the training and test set are created from the complete set of all data. To achieve a good performance without overfitting the split is set to approximately 75% of the available data. Furthermore, the data is randomized in the process. The range on each of the specific features in the original dataset is high. In order that one feature does not get treated in preference to another the data was initially normalized to remove the mean and contain the spread within a single standard deviation. In essence this constrains all the individual features to lie within (-1, 1) and have 0 mean.

The choice of classifier is huge but in this work support vector machines were chosen as they have been shown to perform well on this problem. As the dataset is of high dimensionality and the separations may be complex kernel functions based on radial basis functions were chosen. When training a cross validation approach was employed. This means that the training set was partitioned and training performed on a single partition. Then evaluation was performed against the other partitions and the outliers used as the basis for on-going training. This approach has been shown to be more effective especially with unbalanced datasets. Specifically, a cross validation ratio of 5 was employed.

The classifier performance is measured in terms of 3 different measures. These are precision, recall, and F_1-score. These all provide different ways of evaluating the effectiveness of the classifier. *Precision* is the proportion of the true positive results in the total set of true and false positives. It measures how likely a random guess is likely to be correct. *Recall* is the ratio of true positives to the set of true positives and false negatives. This is the sensitivity of the classifier in predicting a correct outcome for a specific class. F_1-score is the weighted harmonic mean of the precision and recall.

As described in section 2.2.2 there are two sets of labels provided as part of the FAU Aibo dataset. These are the 2-class and the 5-class labels. This section will look at both of these in turn.

2.2.4.1 The two-class problem

The break down by class for the training and testing set is provided in Table 6.

Class	Training	Testing
IDL	9295	3098
NEG	4367	1456
total	**13662**	**4554**

Table 6 - Breakdown of the dataset into training and testing sets

The results are provided in Table 7. The overall results are slightly better than the baseline results for the INTERSPEECH emotion challenge[4] though not dramatically so.

	precision	recall	F_1-score
Idle	0.81	0.89	0.85
Negative	0.70	0.57	0.63
Average	**0.78**	**0.78**	**0.78**

Table 7 - Results on the two class problem

The table shows expected results given the class imbalance. The Idle class is the best represented. Overall the classifier will work well to track the presence of negative emotion in speech versus no emotion.

2.2.4.2 The five-class problem

For the five class problem Table 8 Illustrates the breakdown of the training and testing sets by class.

Class	Training	Testing
Angry	1132	360
Emphatic	2674	927
Neutral	8196	2771
Positive	680	209
Rest	980	287
total	**13662**	**4554**

Table 8 - Numbers of each class in the training and testing sets

The specific results are demonstrated in Table 9. As with the two class case the average results are comparable to the results from the baseline INTERSPEECH results[5].

	precision	recall	F_1-score
Angry	0.58	0.35	0.43
Emphatic	0.56	0.38	0.45
Neutral	0.70	0.91	0.80

[4] The INTERSPEECH emotion challenge achieved a precision of 72% and a recall of 70% for the two class problem.

[5] For the 5 class problem a precision of 65% and a recall of 57% was achieved.

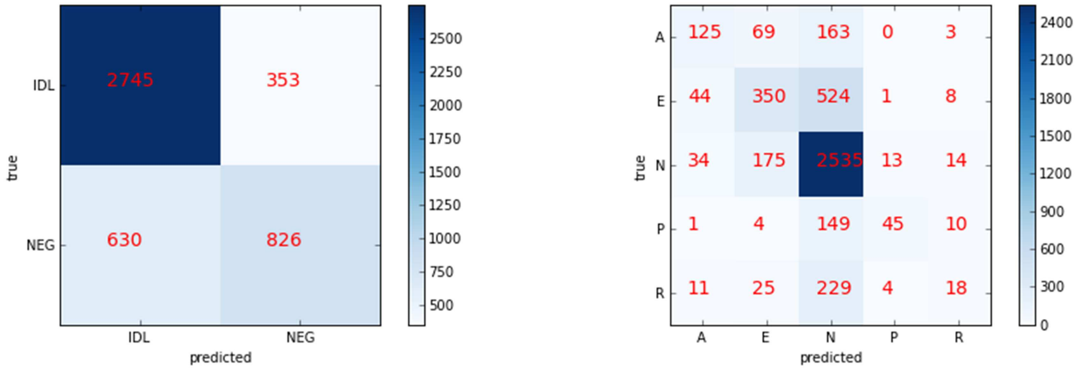
Positive	0.71	0.22	0.33
Rest	0.34	0.06	0.11
Average	**0.64**	**0.67**	**0.63**

Table 9 - Results for the 5 class problem

The average results perform well with all but one class being over the random guess possibility (20%). The worst performing class is the rest class. However as the rest class is the class of no emotion so it is allowable that it has poor performance for our application.

2.2.5 Outlook

The work covered here described initial work on building an emotion classifier on a known dataset. The results are promising and show good alignment with that of the INTERSPEECH emotion challenge. In order to get a feeling for what needs to be improved we will first examine the classification in each class by way of a confusion matrix which illustrates the classification for all the members of each class. These are shown in Figure 12.

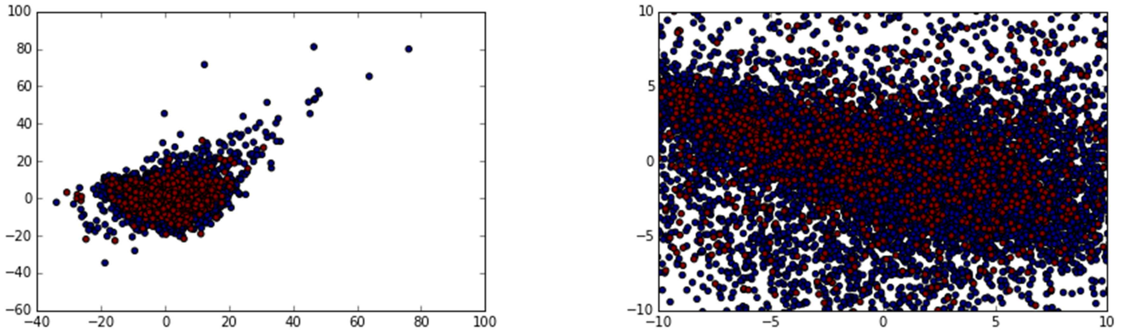


(a) (b)

Figure 12 - Confusion matrices for the (a) 2 and (b) 5 class problems

The first observation to make here is that the imbalance of the initial training set problem was the greatest contributor for the results. This can be observed via the ratios of the individual classes to the total of all classes (the diagonal). In the 5-class problem the biggest cause of performance degradation is from the Neutral class. A quick check via the use of PCA is shown in Figure 13 – PCA performed on the set of features for Rest and Neutral (a) entire set (b) zoom of the central region.

It indicates our intuition in that the features are not well separated. In order to overcome this issue, we will focus our efforts on trying and improve the feature set (by using some sort of feature selection method) or on modifying the feature sets.



(a) (b)

Figure 13 – PCA performed on the set of features for Rest and Neutral (a) entire set (b) zoom of the central region

Examination of the results from the two- and 5-class classifiers shows that we could achieve benefit by fusing them together. This would lead to an improvement in the performance. As another alternative we could move to a differing style of classifier.

Finally, the work here is based on word boundaries within the sequence. This means that before feeding to the classifier the sound needs to be chopped into word chunks. The next phase of work will be involved with moving the architecture to this. This will involve initially decomposing the features and training on smaller time windows. This work will be reported in D3.2 1st Prosocial affect fusion and player modelling.

2.3 Visual data acquisition

It is said that one image is worth a thousand words. Truthfully, humans display a variety of non-verbal, mainly visual signals to communicate with their peers as well as during human-computer interaction. While audio communication serves as the main channel for transferring raw information, it is by decoding visual signals such as facial expressions, gaze and body gestures one can truly decipher the true meaning of the words being spoken, and how the communicated message is intended to be perceived [***Massaro, 1998***]. Furthermore, it is through the observation of spontaneous reactions and behavior one can understand another's state of mind when no words are being spoken. Through continuous subconscious training during everyday life and the experience of a multitude of emotional episodes, humans are able to understand and identify certain emotional characteristics in their transactions with others i.e. we can tell if another person is happy, sad, scared or surprised without needing to be explicitly told so. In terms of the ProsocialLearn project, visual observation of students playing prosocial games holds the key to draw important conclusions about the player's experience during gameplay. Visual cues can tell us whether players are decently challenged in the game, whether they are engaged or bored and whether the entire experience is felt pleasant or irritating. As described in Section 0, our aim is to use this information along with player patterns and historical profiling in order to ensure the proper challenge is being injected to the game, and that content is adapted such that engagement levels during gameplay are kept high. Seeing this condition met, we can expect players to reach their maximum potential with regards to achieving an assigned PLO.

We refer to Visual data as every piece of information that can be acquired through the use of camera-like sensors, and by means of, mainly image processing and visual tracking. As is the case in human everyday life, many of the algorithms used to detect features useful for gaining insight on a human's inner emotional experience rely on an offline training procedure using large datasets, frequently containing both positive and negative examples [***Kanade et al, 2000***] [***Pantic & Rothkrantz, 2003***]. In the remainder of this sub-Section, we will describe the full range of features obtained through all visual sensors described in a previous paragraph. These include features related to the user's face and eye gaze in particular.

2.3.1 Facial Expression Analysis

Facial expressions are probably the most well-studied emotion expression channel, being one of the most natural means of emotional communication. Therefore, automatic analysis of facial expressions is among the most interesting topics in the scientific community. It is almost impossible to cover all the published work, different approaches and multitude of automatic systems implemented in the scientific literature, since the field first broke ground in the early nineties. Therefore, a number of surveys on the state of the art in facial expression analysis have been published, covering various timelines and research trends throughout the years [***Pantic & Rothkrantz, 2000***] [***Bettadapura, 2009***].

In the interest of extracting specific features to feed to the online fusion algorithms for extracting prosocial affect, output of the visual data acquisition techniques for facial expression analysis has been classified into low-, mid-, and high-level abstraction layers, based on the amount of information (core data and meta-data) being encapsulated in the extracted signal. This effectively means that features extracted by applying facial expression analysis techniques can range from simply geo-locating and calculating actual anthropometric measurements, to summarizing an entire group of feature-group elements under a single emotional category, such as *happiness* or *surprise*. Using sophisticated and well-trained shape and landmark tracking techniques, specific facial feature points can be identified and located for every consecutive frame obtained by a camera-like sensor. We obtain this information by fitting an Active Shape Model (ASM) [***Cootes et al, 1995***] onto each consecutive facial image. Early forms of low level data processing can then be applied to identify and track specific muscles and muscle groups' displacement over time, theorized to be involved in the formulation of specific facial expressions. The Facial Action Coding System (FACS) proposed by Ekman & Friesen [***Ekman & Friesen, 1978***] provides a useful tool for describing such a mapping, and additionally coding detected muscle activity into specific Action Units (AUs). These AUs can be seen as a form of mid-level representation of the raw data obtained by fitting the ASM, effectively describing mere observations into meaningful muscular activity occurring at any given input frame. In turn, specific facial expressions and combinations of AUs can lead to the detection of a specific emotion via expression classification [***Pantic & Rothkrantz, 2000***]. This form of representation is usually considered the final step in the automatic emotion recognition using a standard facial expression analysis pipeline, as presented in the majority of the scientific literature. However, studies show that the fusion with other visual and non-visual modalities can enhance the confidence rates by a significant amount [***Pantic & Rothkrantz, 2003***] [***Busso et al, 2004***] [***Soleymani et al, 2012a***].

In this report we will outline all of the extracted features and signals related to our facial expression analysis data acquisition module. We will explain the transitional phases by which low-level features are assigned into mid-level AUs, and how groups of the latter are further associated with high-level emotional labels. Prosocial affect fusion algorithms described in later reporting periods of the project will select which level of feature abstraction is deemed sufficient according to the availability of sensory input information (i.e. how many sensors are used for a single gameplay session) and the density of the input data (i.e. the total number of features that can be acquired throughout the duration of the session), for improved robustness and reliability of the overall output. Figure 14 represents the overall structure of facial expression analysis features extracted by our data acquisition platform.

2.3.1.1 Low level facial features

At the start of every automatic facial features extractor lies a face detection algorithm. This process is able to locate a Region of Interest (ROI) in the image where a human face is located and returns its coordinates. The process greatly enhances performance of any tracking scheme, as it significantly reduces the search area for the algorithm to cover. In the case of the ProsocialLearn facial features acquisition platform, an ASM is fitted onto the ROI returned by a standard Viola-Jones Haar-like features classifier cascade employed for face detection [***Viola & Jones, 2001***]. The latter use is rather straightforward, as both the detector and classifiers are publicly available through large-scale open source computer vision libraries such as *OpenCV*[6]. The ASM algorithm used for landmark tracking is built on top of this face detection scheme and described in [***Wei, 2009***], with an implementation being available in the form of the *ASMLibrary*[7] library.

In order to ensure a sufficient number of facial landmarks are pinpointed, we developed a dense-

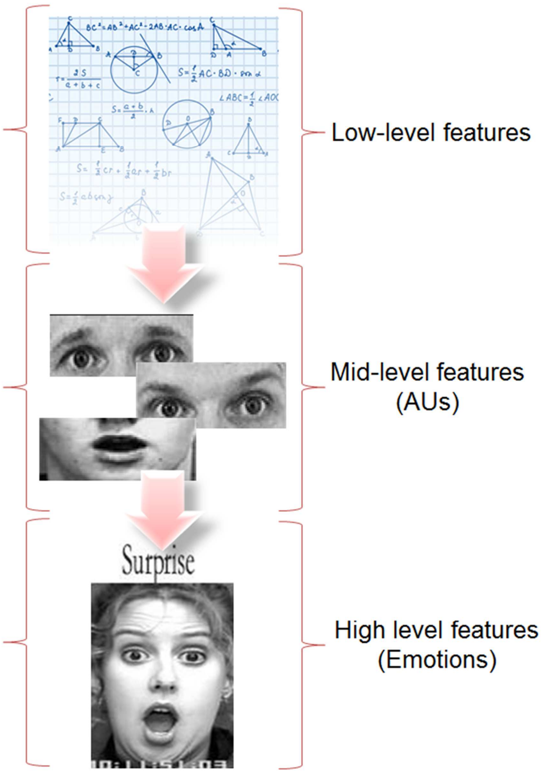


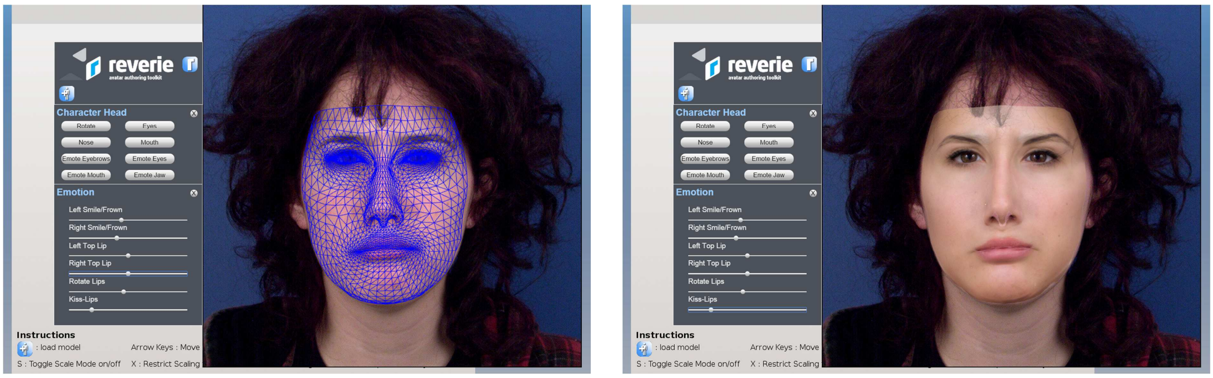
Figure 14 – Facial expression analysis features' structure.

[6] Open Source Computer Vision library http://opencv.org/, started at Intel in 1998, currently led by itseez (http://itseez.com/)

[7] ASMLibrary, ©2008-2011 by Yao Wei, available under the MIT licence at https://github.com/greatyao/asmlibrary

ASM shape representation comprised out of 1,761 points. The shape was trained using an in-house web-based landmarking application based on the *RAAT* library [***Apostolakis & Daras, 2013***]. The application allows a human annotator to overlay the entire shape model as a 2D mesh over a training image and make adjustments using a wide range of blendshapes created for the mesh representation. The application uses two distinct viewing modes to assist the annotator during the process. Namely, the *wireframe mode* allows the annotator to inspect where each and every one of the training feature point landmarks will map onto the image, assisting in approximating the shape of the depicted face. The *textured mode* on the other hand assists in outlining inner-face texture features such as the eyebrow shape and lip line. Screenshots of both modes in the application are depicted in Figure 15 – Dense – ASM Annotation application used for annotating images with 1,761 landmarks

Using this landmarking application, a database consisting of over 1,200 images of faces retrieved from a variety of publicly available face datasets [***Minear & Park, 2004***] [***Nordstrøm et al, 2004***] [***Aifanti et al, 2010***] [***Thomaz & Giraldi, 2010***] [***Shaker et al, 2011***], as well as other sources[8,9] was annotated. These images include both male and female subjects posing a variety of facial expressions, and due to the density of the ASM, are picked for training as a result of their large resolution or relatively large size of the depicted face. We demonstrate the effectiveness and robustness of the resulting fitting process throughout different subjects and posed facial expressions in Figure 16.



(a) (b)

Figure 15 – Dense – ASM Annotation application used for annotating images with 1,761 landmarks

(a) shows the application's wireframe mode used for aligning the shape and basic features (such as the eyes and nose), while (b) shows textured mode for approximating the innermost features such as the eyebrows, nostrils and lip line (training image source: [*Aifanti et al, 2010*]).

[8] Ten24 3D Scan Store ©2012, Head Scans, RAW Individual Expressions, product image gallery http://www.3dscanstore.com/

[9] Iranian women 2D face set, 369 images, 34 women, mostly with smile and neutral in each of five orientations. http://pics.stir.ac.uk/2D_face_sets.htm

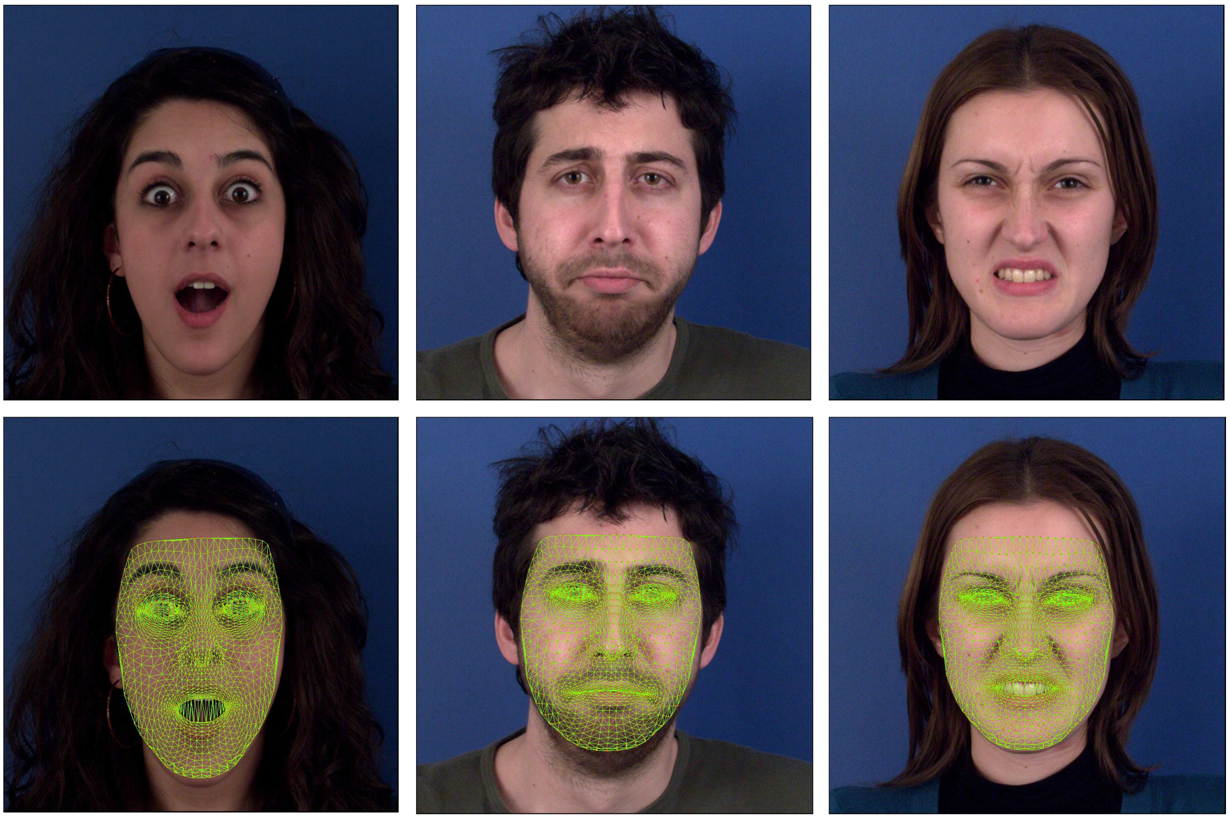


Figure 16 – ASM Fitting results for a variety of facial expressions.

(Test images source: [Aifanti et al, 2010]).

We consider the location and tracking of a number of specifically selected facial landmarks over time as the lowest level of facial expression analysis feature possible. Therefore 1,761 features can be collected per frame in a raw manner. Of course, some level of processing has to be applied in order to generate somewhat meaningful data rather than just reporting the location of each landmark at any given time to the fusion algorithms. In this respect we have followed the approach described in [***Soleymani et al, 2012b***], in which landmark processing is leading to low-level facial features describing the three most expressive regions of the human face: the upper component consisting of the forehead and eyebrows, the middle component comprised of the eyes and cheekbones and the lower component containing the nose, mouth and chin [***Ekman & Friesen, 1978***]. In total, 20 low-level features are extracted from the eyes, eyebrows and mouth areas, as presented in Table 10. The following paragraphs describe these measurements in more detail.

Facial Feature Area / Total Number of Features	Extracted low-level features
Eyes	*Distances between outer eyes' corner and upper eyelids, distances between outer eyes' corner and lower eyelids, distances between inner eyes' corner and upper eyelids, distances between inner eyes' corner and lower eyelids, vertical distances between upper-lower eyelids.*

Eyebrows	*Angles between horizontal line connecting the inner corners of the eyes and the line that connects inner and outer corner of the eyebrow, vertical distances connecting the outer eyebrows to the line that connects the inner corner of the eyes.*
Mouth	*Distances between mouth corners and upper lip, distances between mouth corners and lower lip, distance between mouth corners, vertical distance between upper-lower lip.*

Table 10 -Summary of Facial Expression Analysis low-level features.

Eyes

It is said that the eyes are the most expressive feature of the entire human face, a “window to the soul”. Studies however put the claim to the test· nowadays it is generally accepted that the eyes (and particularly eye gaze) are involved with the experience of emotion, although the eyebrows and mouth tend to be more salient facial features [***Sadrô et al, 2003***] [***Calvo & Fernández-Martín, 2013***]. We treat eye gaze as a separate signal comprised of multiple low-level features on its own right later in this document. In this paragraph we strictly focus on measurements related to the activity of the eyelids, as complementary information to the observations made on the eyebrows and mouth regions.

More specifically, our feature extracting pipeline is able to detect landmarks on the inner/outer eye corners, as well as top/bottom eyelid centers. We then measure a total of five Euclidean distances defined by these points:

1. Outer eye corner to upper eyelid d_{out-up}.
2. Outer eye corner to lower eyelid $d_{out-low}$.
3. Inner eye corner to upper eyelid d_{in-up}.
4. Inner eye corner to lower eyelid d_{in-low}.
5. Upper eye lid to lower eyelid d_{up-low}.

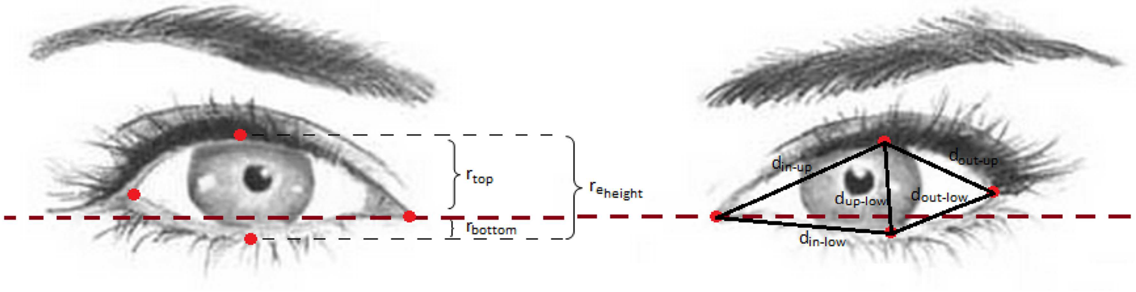


Figure 17 – Visual representation of low level eye features

The bold black lines indicate low-level distance metrics features. Mid-level features for AU extraction are also depicted (see Section 2.3.1.2).

The actual number of features is doubled through the consideration of both eyes. A visual representation of these features is provided in Figure 17 – Visual representation of low level eye features.

Eyebrows

The importance of eyebrows in the domain of emotional expression as well as nonverbal communication in general has been acknowledged throughout years of psychological research [***Sadrô et al, 2003]***. The eyebrows are generally considered to be able to communicate the extremes of aggression and fear, as well as the entire range of human emotion in coordination with other facial movements, playing a key role in the expression of happiness, surprise and anger [***Ekman & Friesen, 1978***].

In order to extract low-level eyebrow features for our facial expression analysis data acquisition pipeline, we first define the imaginary line connecting the inner eye corners as a reference line. This particular line is selected due to the fact that the inner eye corners are considered stable features meaning, that their relative positions inside the face area remains constant throughout the display of any possible facial expression. After this line is drawn, we are able to calculate angular and Euclidean measurements which correspond to the following features:

1. Angle between the line connecting the inner eye corners and the line connecting selected feature points on they eyebrow contour (inner/outer eyebrow) θ_{brow}.
2. Vertical distance between outer eyebrow landmark and the line connecting the inner eye corners $r_{b_{outer}}$.

As was the case in the extraction of eye region features, the total number of features for the eyebrow region is the sum of all measurements for both eyebrows. A visual representation can be seen in Figure 18 – Visual representation of low level eyebrow features

Mouth

A person's mouth is the most prominent element in the lower facial regional component which also contains the nose and chin. The visual saliency of the mouth region, especially in the case of a smile, and subsequent role as a conveyer of emotion has been shown to overpower other expressive facial elements, such as the eyes [***Calvo & Fernández-Martín, 2013***].

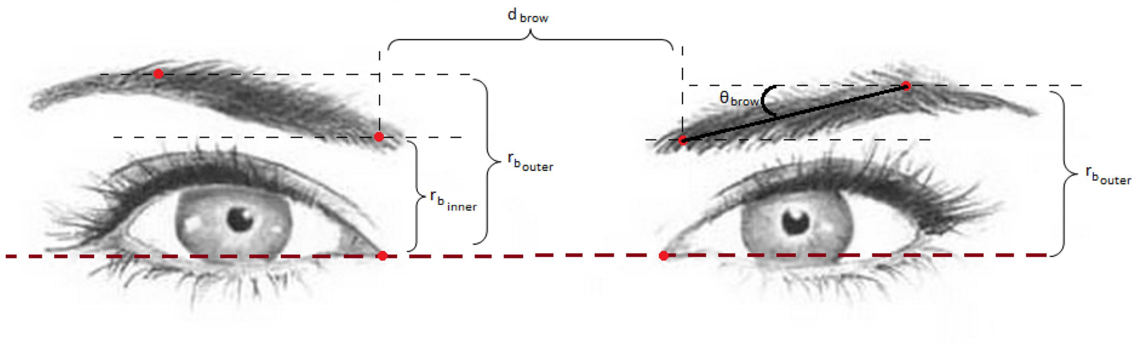


Figure 18 – Visual representation of low level eyebrow features

The bold black lines indicate low-level distance metrics features. Mid-level features for AU extraction are also depicted (see Section 2.3.1.2).

This effectively means that people are more likely to correctly associate emotions such as happiness, sadness and anger when presented with facial images depicting an expressive mouth, even more so than being shown only the eyes. For the acquisition of low level mouth features, we rely upon measuring several key Euclidean distances between strategically selected feature points on the ASM. These include:

1. Distance between each mouth corner and the upper lip d_{upper}.
2. Distance between each mouth corner and the lower lip d_{lower}.
3. Distance between mouth corners r_{width}.
4. Vertical distance between upper and lower lip d_{height}.

A visual representation of these features is given in Figure 19 – Visual representation of low level mouth features

.

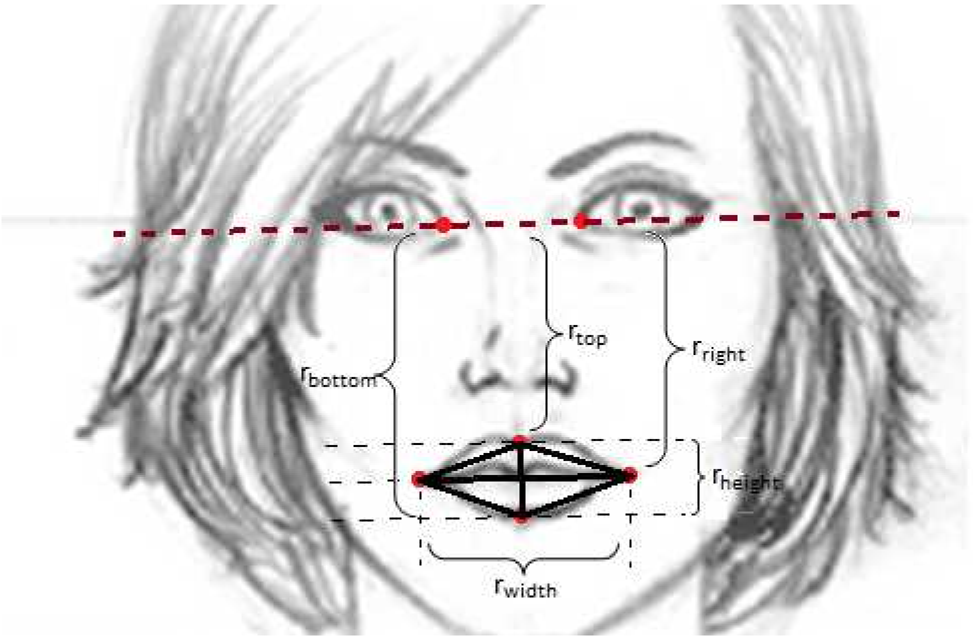


Figure 19 – Visual representation of low level mouth features

The bold black lines indicate low-level distance metrics features. Mid-level features for AU extraction are also depicted (see Section 2.3.1.2).

2.3.1.2 Facial Action Coding System - Action Units (FACS – AUs)

As mentioned in the previous paragraphs, in addition to extracting low-level features describing geometrically the movement flow of specific feature points during the display of a facial expression, specific coding units have been developed encapsulate these measurements into a more meaningful representation. The FACS is one such mapping system, proposed by [***Ekman & Friesen, 1978***], and still serving as one of the most frequently used tool in the related literature on automatic extraction of facial features and emotion recognition. The FACS is a manual of all possible muscle and muscle group movements involved in the formulation of a facial expression. Every single possible movement of such a construct is described as an Action Unit (AU). The detection and identification of AUs is one of the most interesting challenges in human expression recognition [***Valstar et al, 2011***]. Recently, researchers have concluded that intensity and frequency of facial expression AUs can predict moment-by-moment engagement and frustration during learning, efficiently inferring tutoring outcomes [***Grafsgaard et al, 2013***].

As traditional research work on facial expression analysis and AU recognition would have it, we distinct our extracted AU features in two categories, mainly *upper face* and *lower face* AUs [***Tian et al, 2001***], [***Valstar et al, 2012***]. A summary of AUs extracted in each category is given in Table 11.

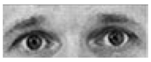
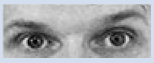
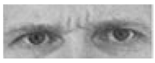
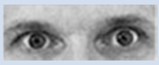
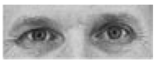
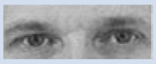
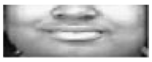
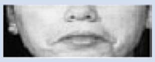
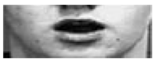
AU facial region	Extracted AU features		
	AU Code	***AU Description***	***Visual representation***
Upper face	*AU1*	*Inner brow raiser*	
	AU2	*Outer brow raiser*	
	AU4	*Brow lowerer*	
	AU5	*Upper lid raiser*	
	AU6	*Cheek raiser*	
	AU7	*Lid tightener*	
Lower face	*AU12*	*Lip corner puller*	
	AU15	*Lip corner depressor*	
	AU26	*Jaw drop*	

Table 11 -Summary of extracted AU features.

In order to extract the aforementioned set of AUs, we follow a similar approach as [***Tian et al, 2001***], which incorporates the feature tracking capabilities offered by our dense-ASM tracking framework. More specifically, two three-layer neural networks with one hidden layer to recognize AUs through a number of parameters defined by low-level features extracted for the upper and lower face regions by standard back-propagation are employed. Our neural networks are trained to recognize the corresponding AUs from Table 11 as well as a “neutral” category which represents the case where no AUs are visible on the current face. We use 13 low-level parameters as the input layer for the upper face AU recognition network, depicted in Table 12.

Left/right facial features			Other features
Inner brow motion $r_{b_{inner}}$	Outer brow motion $r_{b_{outer}}$	Eye height $r_{e_{height}}$	Brow distance d_{brow}

Eye top lid motion r_{top}	Eye bottom lid motion r_{bottom}	Cheek motion (angle) r_{cheek}	-

Table 12 -Upper face AU recognition network input layer parameters.

Similarly, 6 parameters depicted in Table 13 are used as input for the lower face AU neural network input layer. Parameters shown are naturally calculated in a similar manner to the low-level features described in the previous paragraph, by measuring distances and angles of selected landmarks on the ASM (refer to Figure 17 – Visual representation of low level eye features, Figure 18 – Visual representation of low level eyebrow features and Figure 19 – Visual representation of low level mouth features for a visual reference of the networks' input parameters). Through continuous experimentation, we set the number of hidden units in each network to be 12, as our networks demonstrated much more accurate results. Figure 20 shows the structure of both networks for the recognition of AUs in the upper and lower face regions.

Left/right facial features	Other features	
Lip corner motion $r_{left/right}$	Lip height r_{height}	Lip width r_{width}
-	Top lip motion r_{top}	Bottom lip motion r_{bottom}

Table 13 -Lower face AU recognition network input layer parameters.

The networks are trained on a collection of AU-coded frames obtained from the Cohn-Kanade AU-Coded Face Expression Image Database [***Kanade et al, 2000***]. This particular database contains training samples showing AUs occurring singly or in combination. Training samples were gathered per AU occurrence (i.e. every image sequence depicting at least one of the AUs displayed in Table 11 is considered in the training set). We use the first and final frame of each sequence to gather our training data. Then the set is doubled by flipping all the images around the Y-axis, resulting in 848 total training frames. The neutral images are used to train a neutral class for each network, to recognize cases where no AUs are being displayed. The training is performed as such, so that when AUs occur in combination, multiple output nodes on the neural networks are returned.

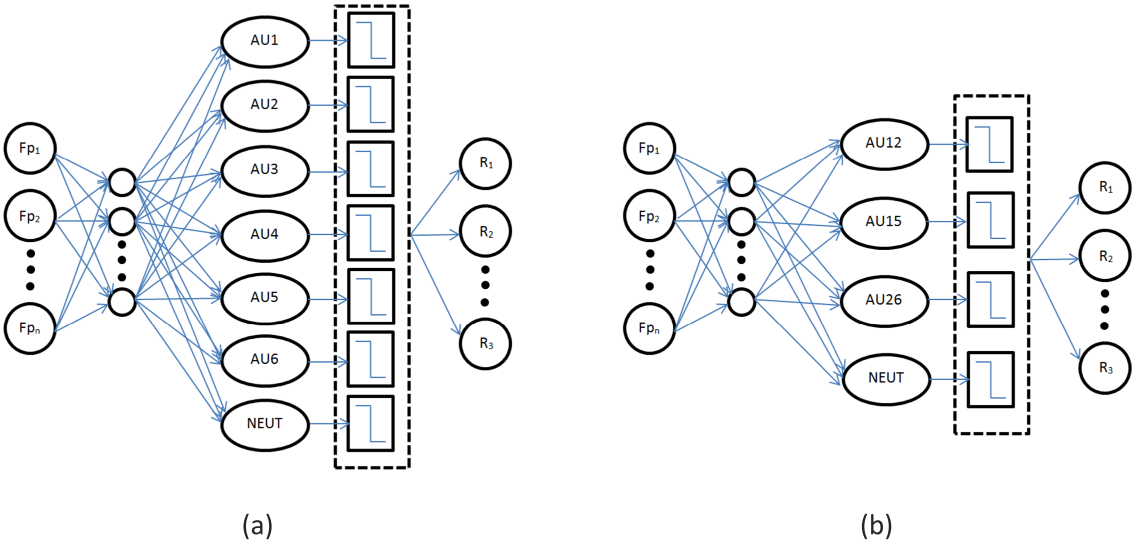


Figure 20 – Neural networks for AU recognition of upper face region (a) and lower face region (b).

2.3.1.3 Emotions

The ultimate goal of identifying and extracting AUs is to ultimately classify expressions under a certain emotion category [***Pantic & Rothkrantz, 2000***]. According to the FACS creators themselves, several AU combinations are associated with emotion [***Ekman et al, 2002***]. We proceed to follow their emotion predictions based on the occurrence of prototypic or major variants of AU combinations to extract emotion. We only utilize the criteria formulated through the use of the AUs recognized by our neural network recognition systems, as described in the previous paragraph. Therefore, the emotion label "*Disgust*" is not part of our recognition results. Instead, we define a "neutral" category to classify all frames where no criteria apply that indicate any of the distinct remaining emotion classes. A summary of the criteria for each emotion class is presented in Table 14.

All of our observations above are provided with caution in terms of the following reasons:

- Evidence: Universal evidence for de facto association of any of the AU combinations in Table 14 with its corresponding emotion class does not exist. The above observations may differ among cultures and according to psychological and physiological conditions of the observed participant and experimental conditions.
- Conversational signals: Several of the aforementioned AUs may be observed during speech and may not necessarily be tied to the experience of emotion. Conversational patterns and manners of speech also differ among participants. At any rate, referring to a combination of AUs as a sign of emotion does not necessarily indicate that emotion is actually being experienced, much like someone may refer to an emotion by name without actually experiencing the said emotion [***Ekman et al, 2002***].

Therefore, our entire facial expression analysis feature extraction pipeline enables the extraction of features more appropriate for each use case as well as each user group.

Emotion	AU Criteria - Prototypes	AU Criteria – Major Variants	
Surprise	**1+2+5+26**	1+2+5	1+2+26
	-	5+26	-
Fear	-	1+2+4+5	1+2+5
	-	1+2+5+26	-
Happiness	**6+12**	-	-
	12	-	-
Sadness	**1+4+15**	1+4+15+26	6+15+26
	6+15	-	-
Anger	**4+5+7**	4+5+7+26	5+7(+26)
	-	4+7(+26)	4+5(+26)

Table 14 - Summary of emotion predictions and corresponding AU criteria (prototypic/major variations).

2.3.1.4 Evaluation

Seeing how low-level features in our facial expression analysis data acquisition pipeline are well-established results of the ASM fitting algorithm and our emotion recognition results stem from criteria based on AU combination detection, our approach was evaluated in terms of AU recognition. We conducted two experiments to evaluate the performance of our system. Both cases consider the detection of both single and AU combinations. In the first case, we measure the accuracy of our method using the same data for generating the test set. This is formally referred to as the *person-specific* case, in which subjects appearing in the test set may also be present in the training set [***Valstar et al, 2011***]. In the second case, we measure the performance of our system using a completely different dataset. A summary of our experimental setup for AU detection can be seen in Table 15 – Action Units include in the AU detection.

As mentioned previously in Section 2.3.1.2, and can be seen in Table 15 - Action Units include in the AU detection, we used 848 the AU-coded frames obtained from the Cohn-Kanade AU-Coded Face Expression Image Database [***Kanade et al, 2000***] for training of our neural networks. A total of 231 test frames obtained from the Cohn-Kanade database were used for recognition of AUs in both upper and lower face regions. We evaluated our system by testing separately for every single AU. The test videos were fed as input to our neural network which detected a sub-set of the nine desired AUs. For every frame of the video every AU's value was compared to the ground truth frame. The results of the AU detection on the designated test set were measured in F1-measure, according to Equation (1):

		COHN - KANADE	COHN - KANADE	GEMEP - FERA
AU	**Description**	**Training Frames**	**Test Frames**	**Test Videos**
1	Inner Brow Raiser	256	37	10
2	Outer Brow Raiser	169	56	11
4	Brow Lowerer	268	26	7
5	Upper Lid Raiser	136	71	5
6	Cheek Raiser	202	27	10
7	Lid Tightener	188	25	9
12	Lip Corner Puller	206	50	17
15	Lip Corner Depressor	142	20	4
26	Jaw Drop	98	35	4
Overall		**848**	**231**	**22**

Table 15 – Action Units include in the AU detection

Training Frames are from Cohn – Kanade dataset. Test Frames denote seen subjects from Cohn – Kanade dataset and Test Videos denote unseen subjects from GEMEP-FERA 2011 dataset [Valstar et al, 2011].

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (1)$$

where $precision$ is the number of correct positive results divided by the number of all positive results and $recall$ is the number of correct positive results divided by the number of positive results that should have been returned. The overall results for both classifiers per AU, as well as the combined average for complete AU recognition are shown in Table 16 – F1 measure for Action Unit detection results on the test set for the person specific participation of the Cohn – Kanade dataset

COHN - KANADE			
Upper face AU	**F1-measure**	**Lower face AU**	**F1-measure**
1	0.727	12	0.933
2	0.843	15	0.553
4	0.778	26	0.806
5	0.864	-	-
6	0.763	-	-
7	0.566	-	-
Avg.	**0.756833333**	Avg.	**0.764**
Avg.			**0.759222222**

Table 16 – F1 measure for Action Unit detection results on the test set for the person specific participation of the Cohn – Kanade dataset

The first and third column show the specific Action Unit currently measured for each neural network recognizer.

We then proceeded to evaluate our algorithms for feature extraction via facial expression analysis by following the baseline set for the Facial Expression Recognition and Analysis (FERA) challenge [***Valstar et al, 2011***], measuring the detection rates of AUs on the designated subset of the GEMEP database [***Bänziger et al, 2010***], made publicly available as part of the challenge[10]. We formalize our comparative results against the baseline method set for the FERA 2011 challenge, focusing on the person-independent partition, seeing how our neural networks were trained on an entirely different dataset. This partition was designed specifically to showcase the ability of AU detection systems partaking in the challenge to generalize to unseen subjects, as the test data are not present in the training data.

Table 17 – F1 measure for Action Unit detection results on the test set for the person independent partition of the GEMEP-FERA 2011 dataset shows the results of the AU detection measured in F1-measure (Eq. 1), for direct comparison of our approach against the FERA 2011 baseline method and the corresponding reported results of a naïve AU detector. It is also noted that both methods results reported in [***Valstar et al, 2011***] do not detect AU5 (upper lid raiser) at all, and therefore their average F1-measures correspond to the set of remaining AUs only. Our approach' F1-measure was instead averaged by considering all of our detected AUs, as well as the AUs remaining after excluding AU5. As can be seen by the summary of results shown in Table 17 – F1 measure for Action Unit detection results on the test set for the person independent partition of the GEMEP-FERA 2011 dataset, our method outperforms both the baseline as well as the naïve classifier in both cases by a considerable amount. At this point we should note that the scope of this report is not to achieve competition-level results, as we do not consider participation on the next FERA challenge [***Valstar et al, 2015***], but rather demonstrate the effectiveness of the proposed methods and support our claim for integrating these techniques onto our data acquisition framework for ProsocialLearn.

AU	[*Valstar et al, 2011*]	Random	Our method
1	0.634	0.506	0.612
2	0.675	0.477	0.682
4	0.133	0.567	0.728
5	N/A	N/A	0.576
6	0.536	0.626	0.738
7	0.493	0.619	0.589
12	0.769	0.739	0.779
15	0.082	0.182	0.352
26	0.371	0.495	0.408
Avg.	**0.461625**	**0.526375**	**0.607111111**
Avg. (excluding AU 5)			**0.611**

Table 17 – F1 measure for Action Unit detection results on the test set for the person independent partition of the GEMEP-FERA 2011 dataset

[10] http://sspnet.eu/2011/05/gemep-fera/

We display our results in comparison to the baseline method of the FERA 2011 challenge, as well as the corresponding reported results of a naïve classifier on the overall test set.

2.3.2 Gaze Analysis

Eye gaze is one of the most expressive signals in human non-verbal communication, and the way gaze shifts, averts or focuses on perceived stimuli can hold useful information towards the beholder's inner emotional state, explaining liking and disliking motives. Gaze supports information gathering, signaling interest and emotional state, and the regulation of conversations by managing turn-taking between participants [***Argyle and Cook, 1976***] [***Kendon, 1990***]. Gaze communicates information to a viewer about internal states, attitudes, attentions and intentions, expresses intimacy, exercises social control and also functions as a means of mobile sensory investigation [***Kleinke, 1986***]. Gaze also supports non-verbal feedback behaviors, such as glances towards and away from others, mediating flow in conversational situations, indicating the addressee, paying attention, displaying attentiveness, effecting turn transitions and signaling requests for backchannels [***Heylen, 2006***].

Psychological research also indicates that gaze can be explained as post-hoc reaction to changes in an individual's core affect state, which manifest themselves with a shift in attention towards the objects attributed as being the cause of that change [***Russell, 2003***]. Eye gaze has been particularly studied as part of implicit human-centered tagging (IHCT) experiments, in an attempt to "read" users' minds during video and image browsing by monitoring gaze patterns [***Vrochidis et al, 2011***] [***Hajimirza et al, 2012***] [***Apostolakis & Daras, 2014***]. In terms of automatic emotion recognition explored in the scientific literature, gaze has only recently been considered as a separate modality for the extraction of emotion [***Soleymani et al, 2012a***]. It has also been proven that combining the information retrieved through more traditional modalities such as facial expression analysis with the appropriate gaze features, extracted either using specialized, wearable hardware, or remote eye center/pupil detection schemes via single or multiple cameras can greatly enhance the performance of the classification schemes [***Soleymani et al, 2012b***]. For the remainder of this document, we will focus our feature extraction procedures on remote eye tracking techniques, efficiently utilizing the same camera-like devices used for facial expression analysis (see Section 4 for more details on hardware specifications). Currently, wearable eye trackers are deemed too expensive for rapid deployment in school environments and too sensitive for use by children in the age group 7-10, which constitute the target audience of the project's prosocial games platform. However, many of the features described in the next paragraphs can be extracted by using wearable devices and are in fact derivations of related work in implicit sentiment tagging through multi-modal monitoring of annotators which used a Tobii[11] wearable eye tracker during its experiments [***Soleymani et al, 2012b***].

As was the case with facial expressions, multiple levels of feature descriptors are extracted, with regards to raw gaze pattern measurements as well as indications on higher level cognitive processes, such as engagement and attention. In a similar way to the Facial Expression Analysis features described in the previous sub-Section, the fusion algorithms will gain access to either level of feature description, and utilize the input in the most efficient and practical way possible in order to reach a decision on the player's prosocial affective state.

2.3.2.1 Low level gaze features

Our decision to focus on remote eye tracking techniques stems from the fact that we already deploy a robust and reliable ASM to track feature points in the eye area. We therefore utilize the same

[11] Tobii AB Group, http://www.tobii.com/

scheme to achieve feature extraction from facial expression analysis. We employ the ASM to locate the actual eye areas on the face which in turn provide us with a second ROI in which our pupil center extraction algorithms operate. We identify as low-level gaze features the actual measurements that can be categorized under four distinct eye gaze activity classes, namely the user's *gaze distance*, *gaze locations on the computer screen*, *pupillary measurements* and *blinking patterns*. We employ several techniques proposed in the scientific literature, adjusted to accommodate a re-imagined remote eye tracker provided by project partner CERTH [***Apostolakis & Daras, 2014***] [***Papadopoulos et al, 2014***] in order to obtain these features. Table 18 contains a summative description of the features extracted in each of the aforementioned categories.

Eye gaze activity / Total Number of Features	Extracted low-level features
Gaze Distance	*Approach time ratio, avoidance time ratio, approach rate, average approach time.*
Gaze location on screen	*Standard deviation, skewness, kurtosis, average fixation time, average scan path length, number of fixation zones, average and standard deviation of the SD of gaze coordinates in each fixation zone.*
Pupil diameter	*Average, standard deviation*.
Blinking	*Blink depth*, *blink rate*, *length of longest blink*, *time spent with eyes closed.*

Table 18 -Summary of Gaze Analysis low-level features.

Gaze Distance

We refer to gaze distance, as the result of approaching or withdrawing from the sensor, which is strategically placed on top (in case of a standard camera), or in front of the user's monitor (in case of a body motion sensor), indicating an act of approach or withdrawal from the presented game content. Psychological reviews argue that such information can be related to action regulation due to liking or disliking motives [***Russell, 2003***], or more specifically, with the experience of positive and negative affect [***Davidson et al, 1990***]. The distance between the user and the screen can also provide valuable information on the user's posture during the activity [***Soleymani et al, 2012b***]. At any rate, gaze distance features extracted by our data acquisition framework are closely related to the player's global positioning in relation to the capturing sensor/gameplay monitor and are thus easily calculated by using the information provided by the ASM or face detection algorithm.

Due to perspective projection, users seen approaching the sensor will appear larger than users who move away. We are able to detect such changes by measuring the scale of the ASM model (by means of the shape's width and height in relation to the input frame size) and detecting subsequent changes. We demonstrate an example in Figure 21 – ASM scale variations comparison with initial measurement (a) with regard to avoidance (b) and approach (c) activity

Once we are able to determine whether the distance between the user and the sensor has increased or decreased, we extract gaze distance features in relation to gameplay time. More specifically we extract:

1. The amount of time spent getting close to the screen (approach time ratio) $t_{approach}$.
2. The amount of time spent getting away from the screen (avoidance time ratio) t_{avoid}.

3. Frequency of users' movement towards the screen (approach rate) $r_{approach}$.
4. Average frequency of users' movement towards the screen (average approach rate) $\overline{r_{approach}}$.

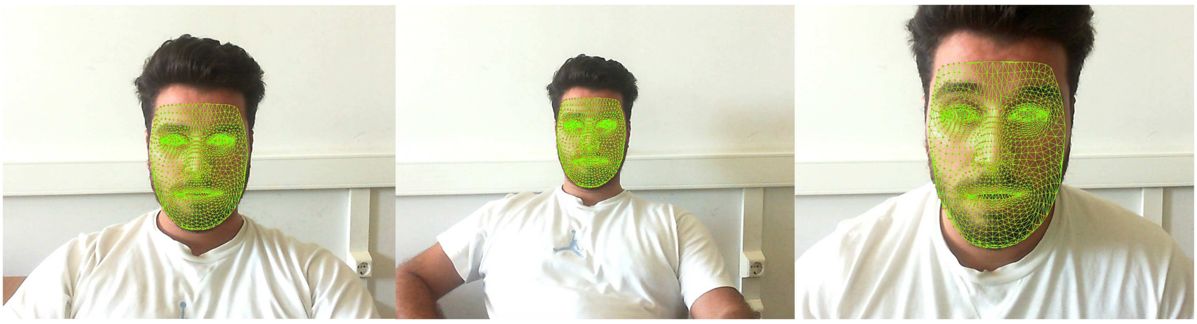


Figure 21 – ASM scale variations comparison with initial measurement (a) with regard to avoidance (b) and approach (c) activity

These features can prove useful for both on-line (measuring gaze distance activity in relation to the sum of previous activity during a gameplay session) as well as off-line prosocial affect fusion (by obtaining gaze distance metrics for an entire session after gameplay has ceased).

Gaze Location on Screen

Observing eye gaze patterns and analyzing scan paths during gameplay may hold information related to the player's engagement and/or frustration during gameplay [***El-Nasr & Yan, 2006***], or indicate players focusing certain amounts of attention towards specific objects in the game world [***Sunstedt et al, 2013***]. In order to generate a user's gaze location on the screen, distinct facial features, such as the eyes corners and pupil centers need to be detected. Then mapping functions can be used to relate gaze parameters to screen coordinates, after an off-line calibration procedure is performed. Utilizing the ASM fitting algorithm, we are able to extract the user's eye corners' coordinates. These landmarks are used to generate a ROI image that isolates each eye, and further contribute to the localization of the user's pupil center, by applying an adaptive version of the Otsu histogram shape-based thresholding algorithm. The process involves the detection of the darker iris/pupil area against the lighter-toned sclera. After the contour is extracted, the pupil center is estimated by calculating the median of points on the contour.

To locate user gaze point on the screen we adopt linear 2D mapping of eye corner-to-pupil center vectors to a corresponding pair of screen coordinates [***Zhu & Yang, 2002***]. This procedure associates eye corner $E(\chi,\psi)$ to pupil center $P(\chi,\psi)$ vectors for each eye to a set of eight known calibration points, which are successively displayed on the screen boundaries during the tracker calibration process, as shown in Figure 22. The process requires an eye corner to pupil center vector $U_i(\chi_i,\psi_i), i \in [1,8]$ to be stored for each of the eight calibration points $D_i(x_i,y_i)$. To extract the top-left/bottom-right screen rectangular area used for the linear 2D mapping, the following equations apply with respect to the calibration points displayed in Figure 22:

$$\chi_{left} = \frac{\chi_1 + \chi_4 + \chi_6}{3} \qquad (2)$$

$$\chi_{right} = \frac{\chi_2 + \chi_5 + \chi_8}{3} \qquad (3)$$

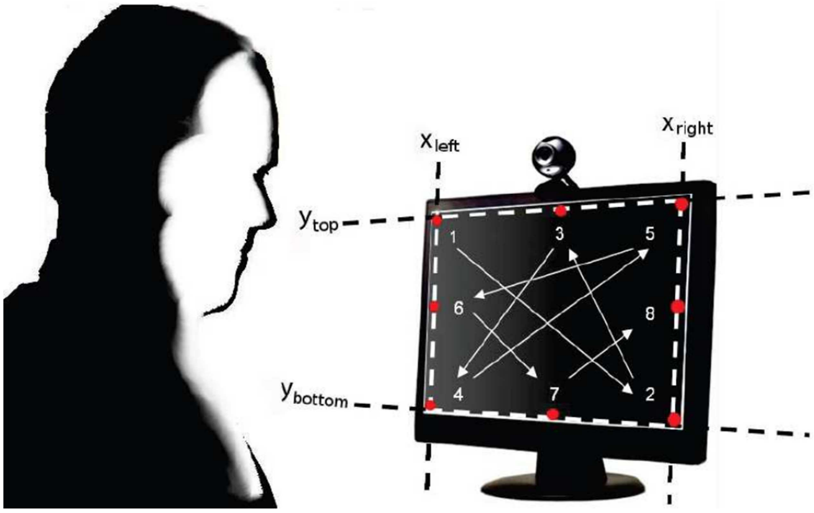


Figure 22 –Gaze tracker calibration process and point display procedure.

$$\psi_{top} = \frac{\psi_1 + \psi_3 + \psi_5}{3} \qquad (4)$$

$$\psi_{bottom} = \frac{\psi_2 + \psi_4 + \psi_7}{3} \qquad (5)$$

In this way, the set of camera image plane coordinates $(\chi_{left}, \chi_{right}, \psi_{top}, \psi_{bottom})$ acquired through Equations 1-4 can be mapped to the set of screen coordinates $(x_{left}, x_{right}, y_{top}, y_{bottom})$ which corresponds to the eight calibration points. After the calibration process is complete, each new eye corner to pupil center vector $U(\chi, \psi)$ can be mapped to its corresponding gaze location on the screen $D(x, y)$ according to Equations 5 and 6:

$$x = x_{left} + \frac{\chi - \chi_{left}}{\chi_{right} - \chi_{left}} \cdot (x_{right} - x_{left}) \qquad (6)$$

$$y = y_{top} + \frac{\psi - \psi_{top}}{\psi_{bottom} - \psi_{top}} \cdot (y_{bottom} - y_{top}) \quad (7)$$

Pupil Diameter

According to recent psycho-physiological studies, pupil diameter has been known to change in different emotional states, reflecting emotional arousal associated with increased sympathetic activity [***Bradley et al, 2008***]. Therefore, monitoring the student's pupillary reactions to the viewed content during gameplay of a prosocial game, might help indicate, along with other features, changes in emotional arousal felt during the experience of both positive as well as negatively associated game elements (such as graphics or perceived challenge).

To remotely measure pupil diameter and changes over time, we employ a real-time pupil contour extraction algorithm based on the description in [***Radu et al, 2011***], built within the framework of the

CERTH Eye Tracker. More specifically, after the ASM is fitted onto the input image frame, the eye regions are located using strategically designated landmarks around the eyes. A ROI image of the eye alone is obtained and then decomposed to its RGB channel components, before histogram equalization is applied to adjust the contrast on the Red channel image, in order to enhance the boundary separation of the iris with the sclera on the resulting ACE (Automatic Contrast Enhancement) image. The ACE image is then passed through a threshold, and an ellipse is fitted around the extracted iris contour. The latter elliptic shape is then used to construct a binary mask which is applied to “crop” the ACE image to the iris, and apply a second histogram equalization to enhance the boundary of the darker pupil region. Again, thresholding is applied and a new ellipse is fitted around the resulting pupil contour. The ellipses for both the iris and pupil sections are computed after interpolating ellipse centers and axes with the ellipses found in the previous frame. An example of the procedure is shown in Figure 23 – Real time pupil contour extraction algorithm integrated onto the CERTH eye tracker

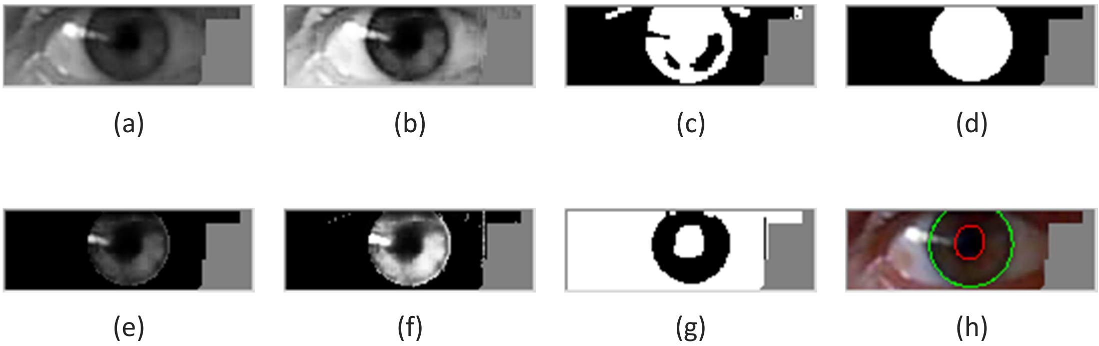


Figure 23 – Real-time pupil contour extraction algorithm integrated onto the CERTH eye tracker

(a) Red channel image; (b) ACE image #1; (c) threshold binary image #1; (d) ellipse fitting mask image; (e) Masked ACE image #1 (f) ACE image #2; (g) threshold binary image #2; (h) final output with elliptic contours for iris (green) and pupil (red) extracted. Images are shown in their original size (subject very close to the camera).

A downside to this method is that, due to remote tracking, the subject’s eye image is usually too small for the algorithm to produce good results. One way to overcome this issue is to enlarge the ROI eye image, although noise is bound to be introduced through scaling. A second option is to employ full HD 1080p or 4K webcams during tracking to obtain large eye ROI images. A second issue stems from the fact that the thresholds need to be specified at the beginning of the session by hand. Enhancement of the described algorithm and automatization of the thresholding procedure in order to decouple the framework from human experiment coordinator effort will be explored as part of work described in D3.2. In its’ current state, sample results of the pupil ellipse estimate axis sizes in pixels measured on consecutive frames of the input video feed (at 720p, enlarged x1.5 times) are presented in Figure 24 – Iris (green) and pupil (red) contour extraction in 8 consecutive frames.

Blinking

Eye blinks can be categorized as spontaneous, voluntary or reflexive and each has different associated dynamics [***Van der Werf et al, 2003***]. While the frequency of spontaneous blinks has been linked to cognitive state and activity [***Stern et al, 1984***], blink rates typically appear to be highly variable during natural interactions, such as conversation [***Doughtym, 2001***].

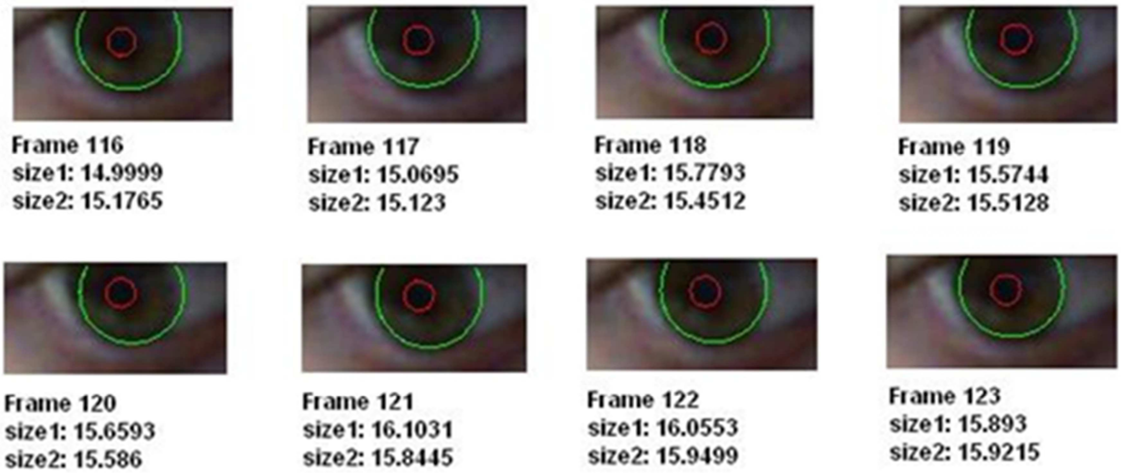


Figure 24 – Iris (green) and pupil (red) contour extraction in 8 consecutive frames

Frame counter and pupillary ellipse axis sizes (in pixels) shown below each image.

Spontaneous eye blinking serves a critical physiological function, but also interrupts incoming visual information. Yet, studies have shown that humans will spontaneously inhibit eye blinks in an attempt to minimize this loss, particularly when viewers perceive that information to be important. Therefore, inhibition of eye blinking during natural viewing can be used as a quantifiable metric of viewers' moment-by-moment engagement with the visual content [***Shultz et al, 2011***]. We can effectively infer that a similar eye blinking pattern will occur during user engagement with active gameplay content. Therefore, eye blinking activity is broken down into several low-level features, which hold information on the frequency and length of player blinks.

To detect blinks in our system, we use two parameters, one for each eye. The first parameter is the projection of the upper eyelid to the imaginary line the connects the inner eye corner points, which, as explained in Section 2.3.1.1, is considered a good reference, as these points' relative positions inside the face area is accepted to remain constant throughout the display of any possible facial expression. The second parameter is the projection of the lower eyelid to the same line. When both metrics are found to be below a certain threshold, we confirm can confirm that the eyes are indeed closed, and therefore report a single blink that lasts as long as the threshold conditions for each parameter hold. Such metrics as time spent with eyes closed and length of each blink can be recorded in real time throughout the duration of the session. Others, such as blink rate and length of the longest blink are accumulated at the end of the session.

2.3.2.2 Visual Attention

In the previous break-down of low-level features acquired through gaze analysis, we have frequently mentioned terms such as arousal and engagement. These notions are closely tied to visual attention. Scientific research has shown that the interest of a person towards a web page, multimedia presentation, video clip or any other form of electronic document is the degree of engagement or interest towards the computer screen it is shown on [***Asteriadis et al, 2009***]. Determining head pose as well as the direction of a user's gaze are a vital part of this kind of feedback. In ProsocialLearn, we extend this notion towards the gaming medium, by employing our algorithms for gaze feature in the context of HCI to extract the degree of interest and engagement of students playing games on a computer screen. In this respect, we can use the position and movement of prominent points around the eyes (see Section 2.3.1.1) and the position of the irises (previously mentioned in this Section) to reconstruct vectors which illustrate the direction of gaze and head pose. These vectors will be used as an indication of whether the user is currently *attentive*, i.e. looking into the screen or not and, in conjunction with our gaze tracking system, whether the users' eyes are fixed at a particular spot for long periods of time. This information will then be used to determine the behavioral state of the user towards the gaming medium. More specifically, the level of interest and attention will be extracted, based on concepts from the Theory of Mind [***Baron-Cohen, 1995***], as well as annotation from experts obtained through crowdsourcing schemes. We will utilize these annotations to turn high-level visual attention concepts described in the Theory of Mind (such as 'distracted') to features detectable with computer vision techniques. Several suggestive features derived from study of visual attention on e-learning environments for the assessment of children's' reading performance [***Asteriadis et al, 2009***] are presented in Table 19 – Visual evidence attention features.

Visual Evidence	Possible values	Related low-level feature
Eyes looking at the screen	*Yes/No*	*Gaze location on screen (eye gaze vector).*
Eyes wide open	*Strong, above normal, normal, below normal, reduced*	*ASM, distance between points around the eye (Section 2.3.1.1).*
Head is moving	*Yes/No*	*ASM, Gaze Distance.*
Head is moving (direction)	*None, forward, backward, up, down, left, right*	*ASM, Gaze Distance.*
Head is moving (speed)	*None, fast, normal, slow*	*ASM, Gaze Distance.*
Eyes blinking	*Yes/No*	*ASM, Blinking.*

Table 19 – Visual evidence attention features

Proposed in [*Asteriadis et al, 2009*] and corresponding related low-level gaze features in this report.

We will use our findings for relating the process of directing one's gaze towards the screen to the level of interest on the prosocial game, as well as the process of staring away from the screen to distraction or lack of interest. These measurements will be collected depending on the time span calculated for each process. Measurements with respect to time will provide useful information towards game adaptation mechanics (Task 4.1). In this respect, we can correlate sudden and abrupt, as well as repeating movements to nervousness and frustration. If the user is not looking at the screen, the game needs to reinstate user interest, for example, by playing a sound file to turn the user's attention back at the screen. In case of consistent distraction, adaptation should further configure game mechanics and content (such as graphics) in order to minimize the observed gaze aversion times. We aim to report on our findings using data collected as part of prosocial studies (WP7) in future deliverables.

2.3.2.3 Evaluation

Evaluation of the gaze features analysis techniques reported in this Section for both low-level as well as visual attention evidence detection involves mainly the evaluation of the employed gaze tracker, i.e. the spatial accuracy and temporal coherence of the tracker. We consider the cases of gaze distance, blinking and pupil contour extraction to be certifiably robust as they constitute implementations of works reported in the scientific literature (ASM, [***Radu et al, 2011***]). Since changes have been made to the overall framework of the employed eye tracker, we employ a similar evaluation experiment as is reported in [***Papadopoulos et al, 2014***], which reported an average measured accuracy (angular error) approximately equal to 0.83 degrees. Gaze accuracy in general, describes the angular average distance from the actual gaze point to the one measured by the eye tracker. Gaze accuracy is measured in degrees of visual angle. One degree accuracy corresponds to an average error of 11 mm (0.45”) on a screen at a distance of 65 cm (26”)[12].

The experiment for measuring accuracy of the tracker involves the display of a red circle following a circular trajectory which the experiment subjects are asked to follow with their gaze. Accuracy is defined as the mean gaze angle deviation that corresponds to the distance of the estimated gaze location on the screen from the center of the trajectorized circle. We replicated the reported experiment with five subjects, as described in the aforementioned scientific paper. Every subject was granted several experimental tries before recording of the results to get acquainted with the tracker. Three-out-of-five subjects had never used remote eye tracking systems before. All of the specifications reported were accounted for, except for the use of webcam and tracking resolution (720p in contrast to the reported 844x448). All experimental data was logged per participant, by recording the eye tracker result gaze point location on the screen (in pixels) as well as the target location (again, in pixels). The average accuracy of the tracker per participant as well as the overall average angular error is presented in Table 20 – Mean gaze accuracy of the employed remote gaze tracker measured in degrees.

The reported results suggest a slight improvement was observed over the original framework, and solidifies our choice to employ the tracker for collecting gaze location on screen features.

Participant	Mean gaze accuracy
Participant #1	0.6723626˚
Participant #2	0.7892300˚
Participant #3	0.8213406˚
Participant #4	0.7574546˚
Participant #5	0.7124208˚
Avg.	**0.75056172˚**

Table 20 – Mean gaze accuracy of the employed remote gaze tracker measured in degrees

Results are shown per participant, where mean gaze accuracy over all experimental runs is reported. A mean average angular error for all participants is reported as well.

2.4 Body Motion Analysis

Most state of the art emotion recognition frameworks capitalize only on facial expression or voice analysis; however recent studies have shown that the movements involving the entire human body can also be used to infer user affective state [***Gunes et al, 2015***]. Seeing how ProsocialLearn aims to

[12] Specification of Gaze Accuracy and Gaze Precision, Tobii X2-60 Eye Tracker, retrieved at http://www.tobii.com/Global/Analysis/Downloads/Product_Descriptions/Tobii_X2-60_Eye_Tracker_Technical_Specification.pdf

achieve a variety of playing styles and through the support of Natural User Interface (NUI) controllers using gesture-driven and engagement-based interactions with cameras and/or depth sensors (WP4), we set out to extract a set of body motion analysis features that can be fused along with the information from the audio/visual channels towards identifying prosocial affect. Furthermore, bodily expression provides a means for recognition of affect from a distance [***de Gelder, 2009***], and therefore motion analysis data are crucial in generating multi-modal data in gameplay environments where players' facial analysis data is either too remote (i.e. players interacting with Kinect sensor – see subSection 2.1.3) or partially obstructed (i.e. children wearing glasses, hats or other headwear). Additionally, the inclusion of bodily expression as an additional channel for affect communication can help resolve ambiguity observed in the identification of certain basic mental states, such as anger and fear [***Gunes et al, 2015***].

Most of the features described in the next paragraphs comprising the body motion modality are extracted through joint-oriented skeleton tracking using depth and RGB information. Additionally, these data can be collected and processed without compromising anonymity of the players, which plays a crucial part in the projects' ethics board, mainly in terms of collecting experimental data as part of prosocial studies (WP7). In the next paragraphs, we will discriminate between body motion features extracted by analyzing the whole body skeleton and features gathered by studying the motion and posture of the head and the hands. A complete summary of all features in this modality is provided in Table 21.

Body motion category / Total Number of Features	Extracted features
3D Body	*Kinetic energy, contraction index, density, smoothness, symmetry, forwards/backwards leaning of the upper body and relative positions.*
Head Motion	*Yaw, pitch, roll.*
Hand Motion	*Velocity, acceleration, fluidity of hand barycenter.*

Table 21 -Summary of Body Motion Analysis features.

2.4.1 3D body features

Indications in the movement of the entire body are related to specific emotions according to experimental psychology literature, for example, contracting the body as an attempt to appear as small as possible is shown to be a strong indicator of fear [***Boone & Cunningham, 1998***]. Accurate full-body motion capturing has been an expensive privilege owned by few, specialized movie and game industry studios, which use sophisticated tracking techniques based on wearable markers. As mentioned earlier in this Section, the introduction of consumer-grade gaming hardware utilizing RGB and depth information to remotely track the users' "skeleton" joint trajectories in 2010, has since revolutionized the way researchers and game developers manufacture low-cost motion capturing frameworks for natural interaction and emotion recognition during gameplay. In this report, the specified set of 3D body features was first defined as part of FP7 ASC-Inclusion[13] project by [***Piana et al, 2013***] and is deeply inspired by psychological literature, to be related to the inference of emotion.

[13] http://asc-inclusion.eu/

2.4.1.1 Kinetic energy

Kinetic energy provides an estimate of the overall energy spent by the user during movement. The amount of movement activity has been shown to be relevantly important for differentiating emotions [***Camurri et al, 2003***]. We employ Kinect skeleton tracking libraries to obtain 3D user joint tracking information. Then, the kinetic energy can be measured as the total amount of displacement in all of the tracked joints, providing an approximation of the user’s body real kinematic energy. The kinetic energy is proportional to the square of velocity. We ignore the mass term in kinetic energy as it is not relevant. The velocity can be approximated in our case by considering finite differences of position divided by the sampling time interval ΔT [***Junjie Shan et al, 2014***]. So the proportional amount of the kinetic energy of each joint K_i is calculated as:

$$K_i = \frac{1}{2} v_i^2 \qquad (8)$$

Then, the kinetic energy of the entire body is calculated as the weighted sum of all joints’ kinetic energies. We demonstrate an example real-time measurement in Figure 25 – Kinetic energy data measurement using the Kinect sensor during “Path of Trust” gameplay.

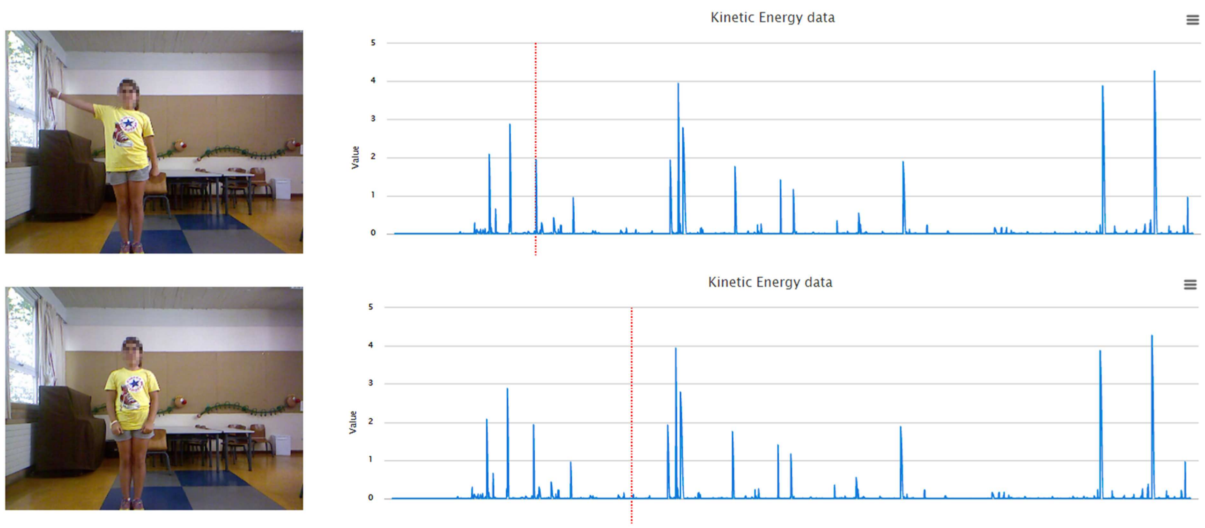


Figure 25 – Kinetic energy data measurement using the Kinect sensor during “Path of Trust” game play

(see Section 5.2.1) acquired as part of small experimental studies. The continuous blue line indicates Kinetic energy measurements over time throughout the entire session. The dotted red vertical line indicates the current frame. The top image shows kinetic energy measurement during gameplay gesture. Spike length on the graph is representative of the movement intensity. The image below shows the corresponding kinetic energy measurement during a player rest/immobile period.

2.4.1.2 Contraction index

The contraction index is measured as an indication of the users’ body spatial extent and is related to the definition of ones’ “personal space” [***Piana et al, 2013***]. It is an estimate of how the body occupies the 3D space surrounding it. According to research in experimental psychology, the contraction index can be used to infer specific emotional states; people are considered to usually spread out when they are happy, angry or surprised, and similarly reduce their size when in fear [***Boone & Cunningham, 1998***]. Contraction index in 3D is therefore defined as the normalized bounding volume containing the user's body. Given the 3D positions of the user's limbs' end effectors we can approximate this volume as the minimum parallelepiped surrounding the user's body. The 3D contraction index is then calculated by comparing this bounding volume and an approximation of the volume of the density (DI) of the 3D coordinates calculated as follows:

$$DI = \tfrac{3}{4}\pi \cdot DI_x \cdot DI_y \cdot DI_z \qquad (9)$$

where DI_x, DI_y, DI_z are the approximated density indices calculated respectively on x, y and z axes as described in the following Equations:

$$DI_x = \frac{1}{n}\sum_{i=1}^{n} dx_i \qquad (10)$$

$$DI_y = \frac{1}{n}\sum_{i=1}^{n} dy_i \qquad (11)$$

$$DI_z = \frac{1}{n}\sum_{i=1}^{n} dz_i \qquad (12)$$

in which dx_i, dy_i and dz_i are the distances between the center of mass and the i^{th} joint.

The 3D Contraction Index is then calculated as the normalized ratio between DI and the Bounding Volume. If the limbs of the user are fully stretched and not lying along the body, the 3D contraction index CI will be low, while if the limbs are kept tightly nearby the body, it will be near to 1.0.

2.4.1.3 Density

A different measurement of body spatial extent is represented by the density index. Given the center of mass of the user's tracked skeleton C, the density index is calculated as the average sum of Euclidean distances of all tracked joints from C:

$$DEI = \frac{1}{n}\sum_{i=1}^{n} d_{Ci} \qquad (13)$$

A graphical representation of this index is shown in Figure 26 – Density index measurement using the Kinect sensor during "Path of Trust" gameplay.

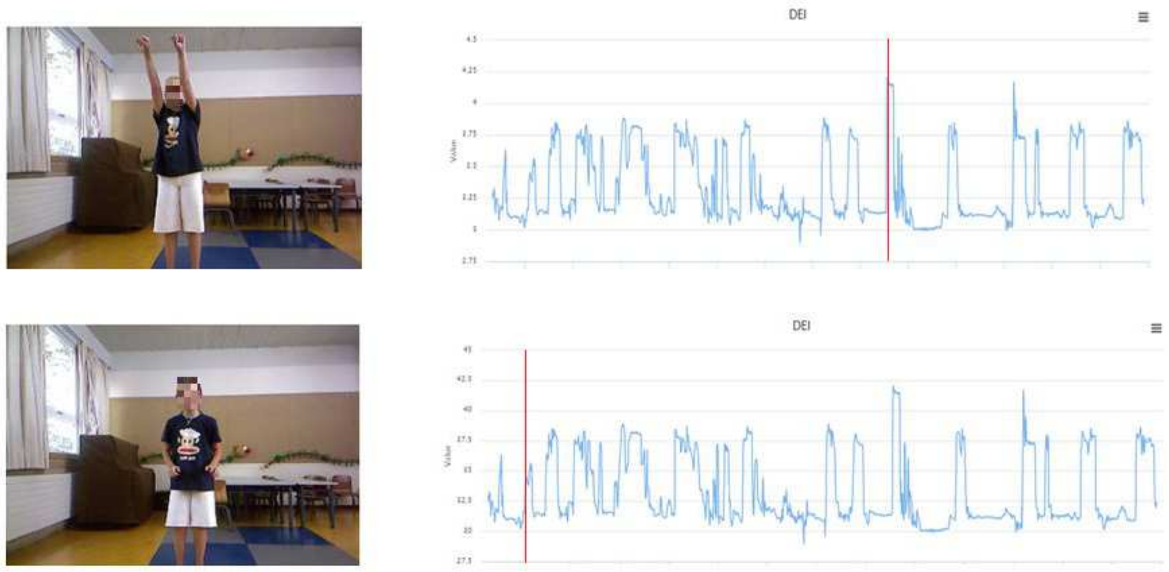


Figure 26 – Density index measurement using the Kinect sensor during "Path of Trust" gameplay

(see Section 5.2.1) acquired as part of small experimental studies. The continuous blue line indicates density measurements over time throughout the entire session. The dotted red vertical line indicates the current frame. Top image depicts density calculation when user body spatial extent is increased through the extension of the hands. Bottom image shows the corresponding measurement when the student's body is contracted.

2.4.1.4 Smoothness

Wallbott, in his analysis of qualitative aspects of psychiatric patients' hand movements, noticed that movements judged as smooth "*are characterized distally by large circumference, long wavelength, high mean velocity, but not abrupt changes in velocity or acceleration (standard deviations of velocity and acceleration). Thus, smooth movements seem to be large in terms of space and exhibit a high but even velocity*" [***Wallbot, 1998***]. Based on Wallbott's statements on the qualitative dimensions of under-constrained arm movements, we use hands trajectories curvature to identify trajectories' smoothness. *Curvature* (k) measures the rate at which a tangent vector turns as a trajectory bends. A hand trajectory following the contour of a small circle will bend sharply, and hence will have higher curvature; by contrast, a point trajectory following a straight line will have zero curvature. The curvature is computed for each point trajectory as follows:

$$k_i = \frac{\bar{\dot{r}}_i \times \bar{\ddot{r}}_i}{|\bar{\dot{r}}_i|^3} \qquad (14)$$

where $\bar{\dot{r}}_i$ is the velocity of the trajectory of the i-th point and $\bar{\ddot{r}}_i$ is its acceleration. Based on the above formula, the smoothness index for three dimensional curvatures is computed as follows:

$$k_i = \frac{\sqrt{(\dot{x}_i \cdot \ddot{y}_i - \dot{y}_i \cdot \ddot{x}_i)^2 + (\dot{z}_i \cdot \ddot{x}_i - \dot{x}_i \cdot \ddot{z}_i)^2 + (\dot{y}_i \cdot \ddot{z}_i - \dot{z}_i \cdot \ddot{y}_i)^2}}{(\dot{x}_i^2 + \dot{y}_i^2 + \dot{z}_i^2)^{\frac{3}{2}}} \qquad (15)$$

The estimation of smoothness is depicted in Figure 27 – Curvature index measurement using the Kinect sensor during "Path of Trust" gameplay.

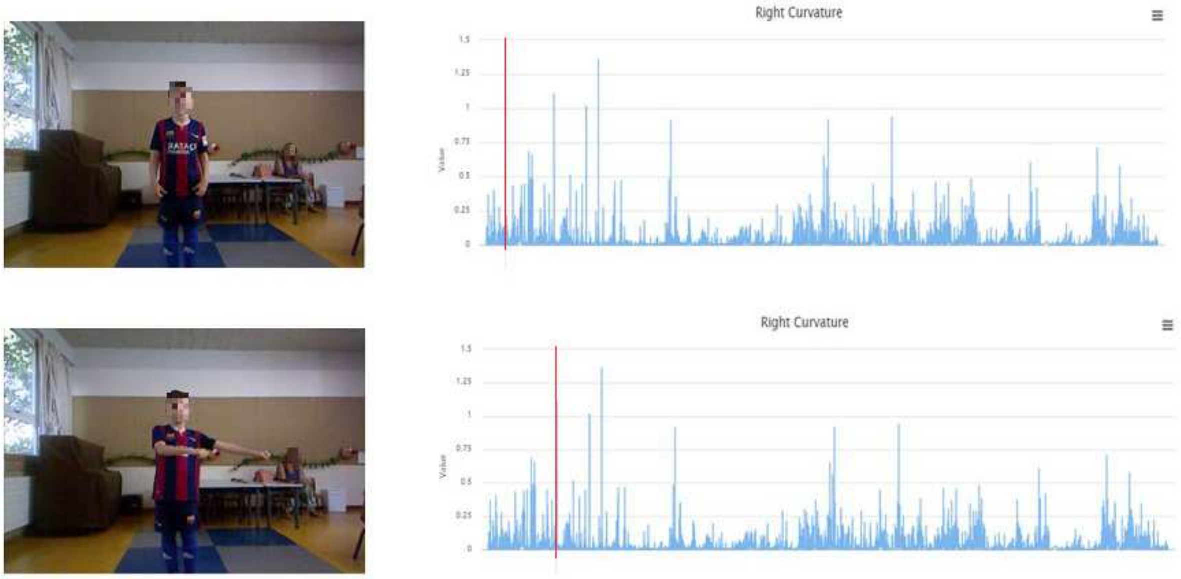


Figure 27 – Curvature index measurement using the Kinect sensor during "Path of Trust" gameplay

(see Section 5.2.1) acquired as part of small experimental studies. The continuous blue line indicates curvature measurements over time throughout the entire session. The dotted red vertical line indicates the current frame. Top image depicts curvature calculation during player immobile/rest period. Bottom image shows the corresponding measurement during gameplay gesture.

2.4.1.5 Symmetry

A study on human gait demonstrated that lateral asymmetries exist not only in face expressions, but also in human emotional full-body movement [***Roether et al, 2008***]. Twenty-four actors (with an equal number of right and left-handed subjects) were recorded by using a motion capture system during neutral walking and emotionally expressive walking (anger, happiness, sadness). For all three emotions, the experiments showed that the left body side moves with significantly higher amplitude and energy. Taking into account the role of asymmetry as indicator of behavioral and affective features, we address the symmetry of gestures and its relation with emotional expression. It is measured evaluating limbs spatial symmetry with respect to the body computing symmetry on each of the available dimensions. Each symmetry (SI_x, SI_y, SI_z) is computed from the position of the barycenter and the left and right joints (e.g., wrists, shoulders, feet, knees) as described below:

$$SI_{Xi} = \frac{(x_B - x_{Li}) - (x_B - x_{Ri})}{x_{Ri} - x_{Li}} \qquad (16)$$

$$SI_{Yi} = \frac{(y_B - y_{Li}) - (y_B - y_{Ri})}{y_{Ri} - y_{Li}} \qquad (17)$$

$$SI_{Zi} = \frac{(z_B - z_{Li}) - (z_B - z_{Ri})}{z_{Ri} - z_{Li}} \qquad (18)$$

Where x_B, y_B, z_B are the coordinates of the center of mass, x_{Li}, y_{Li}, z_{Li} are the coordinates of a left joint i (e.g., left hand, left shoulder, left foot, etc.) and, x_{Ri}, y_{Ri}, z_{Ri} are the coordinates of a right joint (e.g., right hand, right shoulder, right foot, etc). The three partial indices are then combined in a normalized index that expresses the overall estimated symmetry:

$$SI = \frac{SI_{Xi} + SI_{Yi} + SI_{Zi}}{3} \qquad (19)$$

A graphical representation is shown in Figure 28 – Symmetry data measurement using the Kinect sensor during “Path of Trust” gameplay.

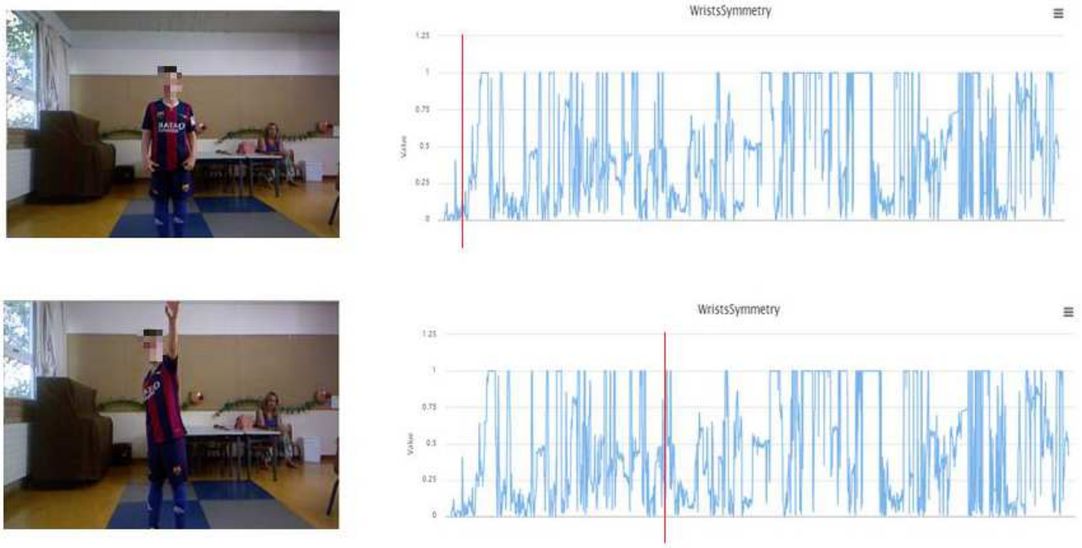


Figure 28 – Symmetry data measurement using the Kinect sensor during “Path of Trust” gameplay

(see Section 5.2.1) acquired as part of small experimental studies. The continuous blue line indicates user wrist Symmetry measurement over time throughout the entire session. The vertical red line indicates the current frame. The top image shows symmetry measurement during player rest/immobile period. The image below demonstrates user symmetry index during gameplay gesture.

2.4.1.6 Forward/backward leaning of upper body and relative positions

Head and body movement and positions are relied on as an important feature for distinguishing between various emotional expressions [***Schowstra & Hoogstraten, 1995***]. The amount of forward and backward leaning of a joint is measured by the velocity of the joint's displacement along its z component (depth) respective to the body position and orientation, as follows:

$$L_i = \dot{z}_i \qquad (20)$$

A graphical representation is shown in Figure 29 – Upper body leaning data measurement using the Kinect sensor during “Path of Trust” gameplay.

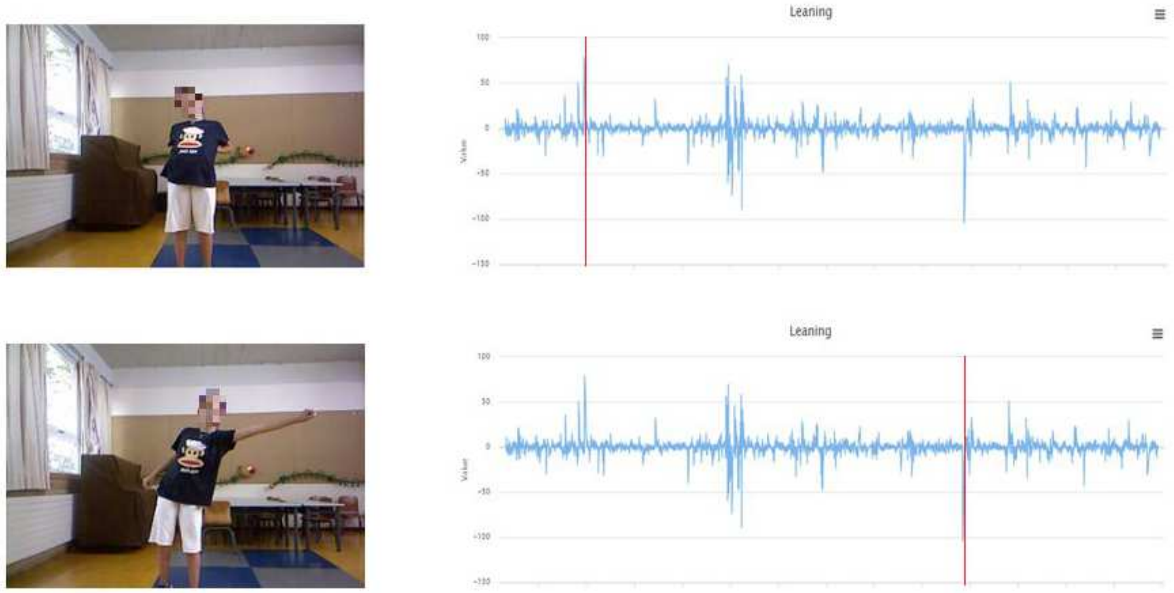


Figure 29 – Upper body leaning data measurement using the Kinect sensor during "Path of Trust" gameplay

(see Section 5.2.1) acquired as part of small experimental studies. The continuous blue line indicates user wrist Symmetry measurement over time throughout the entire session. The vertical red line indicates the current frame. Images show user backwards (top) and forwards (bottom) leaning examples during gameplay.

2.4.2 Head Motion Analysis

Postural configurations of the head [***Kleinsmith & Bianchi-Berthouze, 2007***], movements [***Cohn et al, 2004***] and gestures (for example, head nods and shakes) [***Cowie et al, 2010***] are some of the more representative bodily cues related to affect recognition studies. We therefore use the representation the users' head pose as a sum of its angles in three dimensions (*yaw*, *pitch*, and *roll*) and consider them as additional features for the multi-modal fusion algorithms. This information is *a priori* available through the internal head joint tracking built into the Kinect sensor. As our capturing framework for gathering full body motion data for affect analysis already capitalizes on the sensor's skeleton tracking capabilities, we find ourselves gaining a reliable measurement of the user's 3D head pose to add to our collection of body features.

2.4.3 Hand Motion Analysis

Analysis of arm movements has shown that, considering a dimensional emotional space represented by measures for valence and arousal, the velocity, acceleration, and jerk of the hand movement is highly correlated with the arousal component [***Pollick et al, 2001***]. Seeing how the multi-modal capturing setup presented previously in this Section considers the use of not only full-body skeleton tracking sensors such as the Kinect, but also hand motion sensing technology in the form of the LEAP Motion Controller, we additionally measure and extract features related to the user's motion of the hands. More specifically, the features proposed in [***Kessous et al, 2010***], are extracted. These features consider the hand as a singular point in 3D space, represented by its barycenter. Therefore, hand *velocity* and *acceleration* are directly related to the trajectory of the hand barycenter. *Fluidity*, on the other hand provides a measure of the uniformity of motion. Fluidity is therefore considered maximum when the acceleration of the hand during movement between two specific points in 3D space is zero. These features are gathered for both hands in case of a full-body movement tracking environment (such as when using a Kinect sensor), resulting in a total of 6 features, or for a single hand (i.e. 3 features) in case of hand motion sensing interfacing with the prosocial game (e.g. LEAP Motion).

2.4.4 Outlook

Due to the nature of the features extracted as part of the body motion analysis data acquisition framework we presented in this Section, we are currently unable to map skeleton data acquired through the use of a Kinect sensor onto a certain emotion label. Though several databases correlating motion capture data with emotion classes exist, as shown in Table 22 – Overview of existing databases on emotion from body data, we are unaware of any work that solely maps emotional labels to user tracked skeleton movements, thus providing ground truth on which to train and choose classification schemes on. Therefore, as part of the ProsocialLearn project, we aim to create a database to complement work presented in Table 22 – Overview of existing databases on emotion from body data.

Name	Nr. of emotions	Format	Modalities	Mode	Actors
FABO [***Gunes & Piccardi, 2006***]	10	Video	Face, Body	Elicited	1,900
GEMEP [***Bänziger et al, 2012***]	17	Audio, Video	Face, Body, Speech	Elicited	1,260
[***Savva et al, 2012***]	4	Motion capture	Body	Natural	161
[***Sneddon et al, 2012***]	8	Audio, Video	Face, Body, Speech	Natural	1,400
[***Volkova et al, 2014***]	11	Motion capture	Body	Elicited, Natural	1,447

Table 22 – Overview of existing databases on emotion from body data

We will mainly target an audience of game players, preferably within the age groups defined for our prosocial studies, and proceed to record Kinect data (RGB + depth channels, skeleton joint tracking data) during gameplay sessions. Each session will then be annotated per frame with emotional labels through crowdsourcing schemes. We will then proceed to generate baseline emotion classification methods utilizing features presented in this report, while hopefully being able to break ground on the proposition of new 3D body features as cues on user emotional state. We will then proceed to publish our annotated datasets as part of the ProsocialLearn project, hopefully contributing to the scientific community while also striving to excel beyond the scope of Task 3.1. Further details on the developments concerning this endeavor will be reported in future deliverables. Samples of the database showcasing acted emotions being captured by a Kinect for Xbox One device are shown in Figure 30 – Samples of the database work in progress on emotional body motion for feature extraction and analysis.

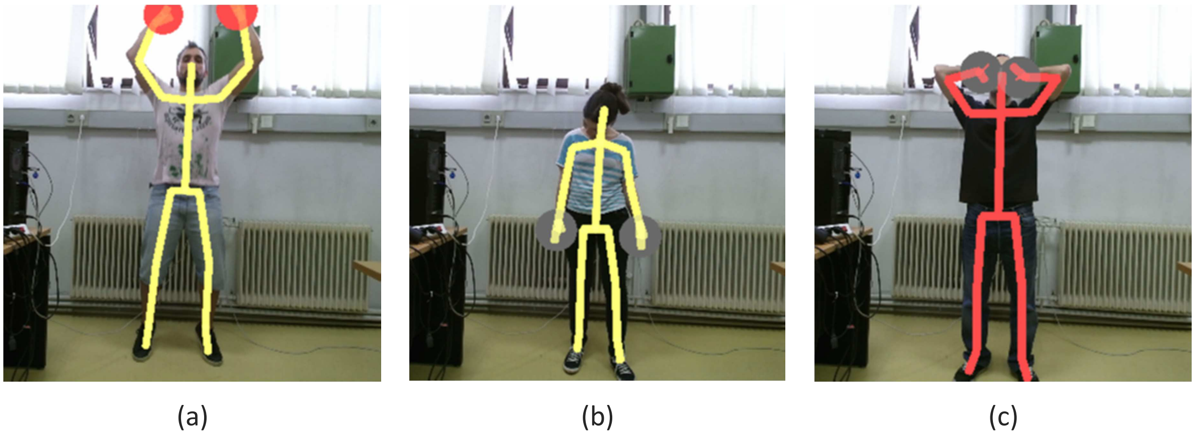


Figure 30 – Samples of the database work in progress on emotional body motion for feature extraction and analysis

Frames are being shown for *happiness* (a), *sadness* (b) and *surprise* (c). Full-body skeleton joint tracking and hand motion patterns are shown highlighted, in accordance to the features presented in this Section.

3 In-game Data Logs

In this Section, the acquisition of in-game data logs will be described, along with proposed format based on current popular online multiplayer games. We categorize in-game logging channels into low-level behavioral observations and high-level game mechanics, which will comprise the input to the game logging service. Additionally, in this Section, behavioral cues stemming from chat message sentimental analysis as well as interaction patterns with standard I/O devices such as the mouse and keyboard will be explored. A thorough survey on proposed features in the related literature will be provided, with the intent on detecting users' affective states from their keyboard and mouse interaction features and chat messages, which is aimed to enrich the multimodal approach we are following for automatic prosocial affect and engagement detection through the sensory observation channels described in Section 2.

3.1 In-game logging channels

In parallel with sensor data acquired during at run-time, ProsocialLearn games are also expected to generate 'in-game' data describing important logical behaviors related to game interactions, mechanics and transactions. ProsocialLearn game logging is used to support the capture of data that is a) generated directly from within the game logic (rather than by a separate and de-coupled sensor acquisition process) and b) intended specifically for use in down-stream WP3 related fusion processes. In-game logging is therefore categorized in two main types:

- Behavioral observations (for down-stream emotion analysis)
    - Mouse input signals
    - Text input via keyboard
- High-level game mechanics (for down-stream prosocial state management)
    - High-level game constructs and events
    - Game transactions

3.1.1 Behavioral observation data logging

Behavioral observation data logging focusses on the relatively low-level data related to conventional input events taken from the keyboard and mouse. Streamed data representing mouse movement and button signals are candidates for inferring further evidence of user emotion and engagement at run-time. Conversational interactions between players can also be re-constructed through the aggregation of player messages; this then leads on to the possibility of inferring additional evidence of emotion and engagement through sentiment analysis. A brief overview on these subjects will be presented in Section 3.3.

3.1.2 High-level game mechanics logging

High level game mechanics message data will carry common game constructs and events, such as:

- Game title and instance identifiers
- Game player and avatar identifiers
- Common game events (i.e., the beginning & ending of games, levels or scenarios)

In addition to these, specific interactions between players, expressed using a well-defined vocabulary defined for ProsocialLearn, will also be used to log a series of game events that collectively represent a game transaction. These data, along with classifications of emotional responses from sensor data, will be used later to calculate prosocial state changes in a player, downstream from the fusion process (described in D2.3 and to be discussed further in D3.2).

3.2 Acquisition of in-game logging data

The actual route any particular game log message takes will depend on whether it is game play or game mechanic data. Messages destined for emotion classification can either be routed via the game logging service (as illustrated in Figure 31) or be transmitted using another network protocol and third party service, if required by the game server. Game mechanics data are expected to be handled directly by the game logging service.

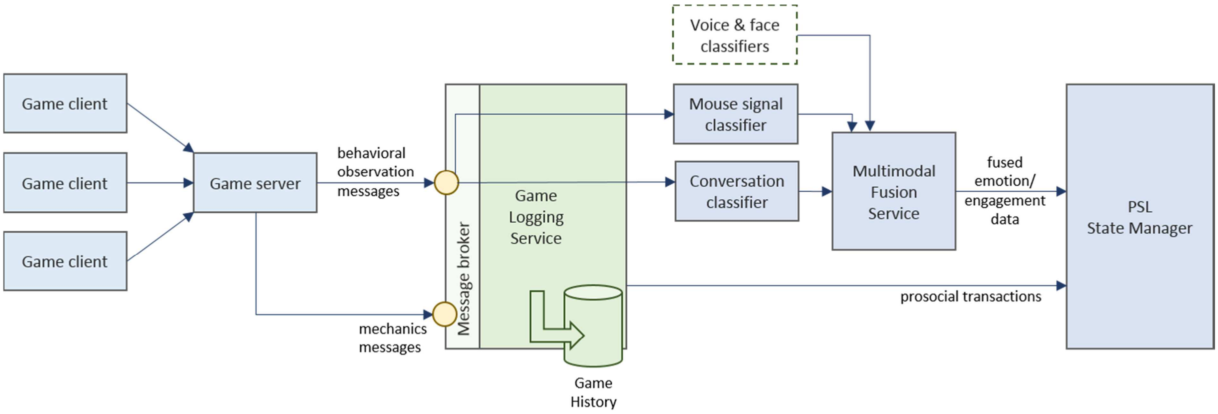
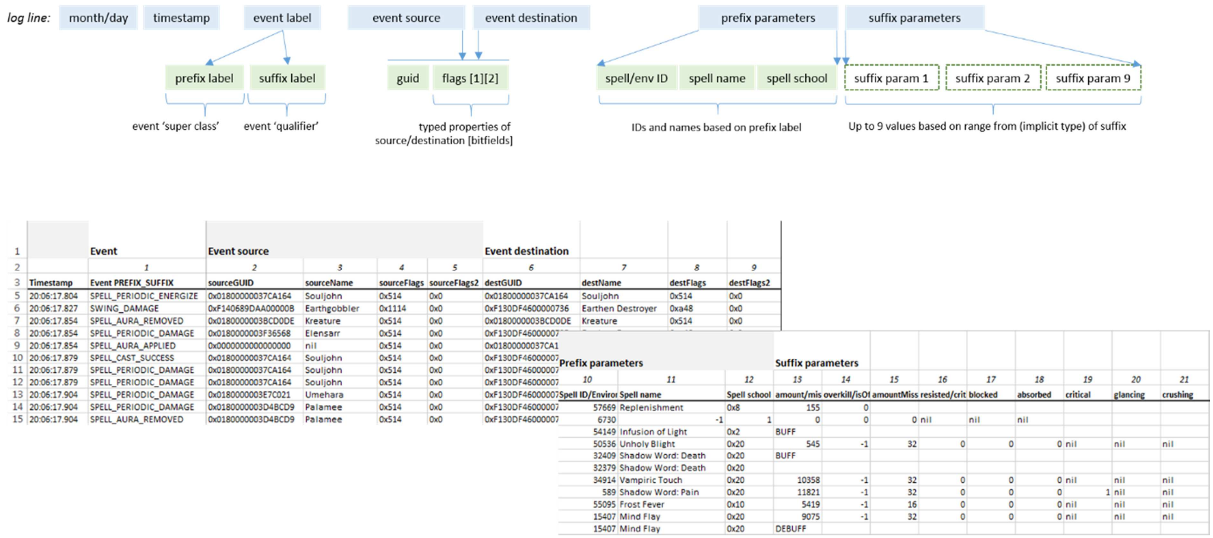


Figure 31 -The role of the game logging service in the fusion pipeline

In support of log message acquisition, the logging service will provide a message brokering service; any messages intended for emotion/engagement analysis will be routed on to the appropriate classifiers. Game mechanics data will be processed and persisted in a game history store. Aggregated messages that combined to form game transactions will be pushed forward to the PSL state manager. A common format for all logging messages will be used, based on a widely recognized formalism already in use in the game industry and exemplified by the “World of Warcraft” format, an example of which is shown in Figure 32.



1		Event	Event source				Event destination			
2		1	2	3	4	5	6	7	8	9
3	Timestamp	Event PREFIX_SUFFIX	sourceGUID	sourceName	sourceFlags	sourceFlags2	destGUID	destName	destFlags	destFlags2
5	20:06:17.804	SPELL_PERIODIC_ENERGIZE	0x01800000037CA164	Souljohn	0x514	0x0	0x01800000037CA164	Souljohn	0x514	0x0
6	20:06:17.827	SWING_DAMAGE	0xF140689DAA00000B	Earthgobbler	0x1114	0x0	0xF130DF4600000736	Earthen Destroyer	0xa48	0x0
7	20:06:17.854	SPELL_AURA_REMOVED	0x0180000003BCD0DE	Kreature	0x514	0x0	0x0180000003BCD0DE	Kreature	0x514	0x0
8	20:06:17.854	SPELL_PERIODIC_DAMAGE	0x0180000003F36568	Elensarr	0x514	0x0	0xF130DF46000007			
9	20:06:17.854	SPELL_AURA_APPLIED	0x0000000000000000	nil	0x514	0x0	0x01800000037CA1			
10	20:06:17.879	SPELL_CAST_SUCCESS	0x01800000037CA164	Souljohn	0x514	0x0	0xF130DF46000007			
11	20:06:17.879	SPELL_PERIODIC_DAMAGE	0x01800000037CA164	Souljohn	0x514	0x0	0xF130DF46000007			
12	20:06:17.879	SPELL_PERIODIC_DAMAGE	0x01800000037CA164	Souljohn	0x514	0x0	0xF130DF46000007			
13	20:06:17.904	SPELL_PERIODIC_DAMAGE	0x0180000003E7C021	Umehara	0x514	0x0	0xF130DF46000007			
14	20:06:17.904	SPELL_PERIODIC_DAMAGE	0x0180000003D4BCD9	Palamee	0x514	0x0	0xF130DF46000007			
15	20:06:17.904	SPELL_AURA_REMOVED	0x0180000003D4BCD9	Palamee	0x514	0x0	0xF130DF46000007			

Prefix parameters			Suffix parameters								
10	11	12	13	14	15	16	17	18	19	20	21
Spell ID/Enviror	Spell name	Spell school	amount/mis	overkill/isOf	amountMiss	resisted/crit	blocked	absorbed	critical	glancing	crushing
57669	Replenishment	0x8	155	0							
6730	-1	1	0	0	0	nil	nil	nil			
54149	Infusion of Light	0x2	BUFF								
50536	Unholy Blight	0x20	545	-1	32	0	0	0	nil	nil	nil
32409	Shadow Word: Death	0x20	BUFF								
32379	Shadow Word: Death	0x20									
34914	Vampiric Touch	0x20	10358	-1	32	0	0	0	nil	nil	nil
589	Shadow Word: Pain	0x20	11821	-1	32	0	0	0	1	nil	nil
55095	Frost Fever	0x10	5419	-1	16	0	0	0	nil	nil	nil
15407	Mind Flay	0x20	9075	-1	32	0	0	0	nil	nil	nil
15407	Mind Flay	0x20	DEBUFF								

Figure 32 -A well-known game logging format: World of Warcraft example.

A similar approach will be adopted in the ProsocialLearn project since game developers will be familiar with the approach; a full specification of the logging format and API will be provided in D3.2.

3.3 Behavioral observation in-game data acquisition

The following paragraphs provide a summative survey into the behavioral observation in-game data acquisition channels defined in Section 3.1. The aim is to lay a foundation on which ProsocialLearn will build on to generate gameplay data to complement the observation channels presented in Section 2. Additional details on how this data will be fused with the audio/visual modalities will be presented in D3.2.

3.3.1 Mouse/Keyboard input signals

As already described in Section 2, expressive, non-intrusive audio/visual methods for recognizing user affective states have primarily dominated the scientific literature on affective computing. However, interaction with the most common interfacing devices such as the keyboard and mouse, have also been extensively studied for obtaining affective indicators. Mouse or keyboard Input behavioral patterns have been studied in the scientific literature mainly as an indicator of interest [***Claypool et al, 2001***], interface design [***Sengupta & Jeng, 2003***] or as evaluation metrics on user experience [***Tullis & Albert, 2010***]. In this Section we focus on the affective computing element, which has studied the use of these devices extensively for recognizing user affective states from behavioral patterns [***Salmeron-Majadas et al, 2014***]. We present the results of this survey in the state of the art in the following paragraphs, focusing on mouse and keyboard input devices. More recent studies relating user affective states with other standardized methods of common device input arise with the increased popularity of touch-screen interfaces [***Gao et al, 2012***], as well as recent smartphone and tablet sensors for device environmental data and gyroscopes (e.g. device shaking, inclination etc.) [***Lee et al, 2012a***].

Affect extraction from mouse and keyboard interaction has been explored in several studies in the related literature. Mouse interactions are mainly studied in relation to movement, and movement rate has been shown to be related to arousal in an attempt to recognize the user's mood [***Sottilare & Proctor, 2012***]. More prominently, features such as average speed, inactivity, speed and orientation of mouse movements have been proposed for the detection of students' affective state detection during online lessons [***Tsoulouhas et al, 2011***]. Biometrics can be applied to these measurements when using specialized biometric mouse hardware, which are able to efficiently extract indicators such as hand shaking, temperature, humidity and pressing intensity [***Kaklauskas et al, 2009***]. Other popular unique features that relate to biometry and do not necessarily require the use of specialized hardware include the time between mouse button presses [***Tsai et al, 2012***] and analysis of mouse movements in terms of coordinates, distance, path etc. [***Lin et al, 2012***]. Biometrics is then used to identify users from these patterns.

Keyboard input has also been studied as an input signal for determining user affective states. More prominently, these studies relate to the typing of text rather monitoring keyboiard activity during gameplay sessions. Therefore, typical features extracted include typing speed, number of typed characters during set intervals, relative timing (total time taken for typing during a single session), number of errors (backspaces) and idle times [***Khanna & Sasikumar, 2010***] [***Felipe et al, 2012***] [***Bixler & D'Mello, 2013***]. These features can be used to recognize between the reported states of boredom, engagement and neutral [***Bixler & D'Mello, 2013***]. As is the case with mouse interactions, biometrics are also studied for keyboard interactions during short text (e.g. passwords) [***Karnan et al, 2011***] [***Bakelman et al, 2013***] or large text input [***Villani et al, 2006***] [***Monaco et al, 2012***]. Biometrics mainly focusing on keystroke time, interval, input rate, errors, key press durations, key pressure, text length, difficulty etc. is then mainly used to enrich security systems, recognizing and modeling users, rather than recognizing a particular user's affective state.

Multimodal input interactions using both mouse and keyboard input have also been subjected to studies in the related state of the art [***Zimmermann et al, 2006***] [***Lee et al, 2012b***] [***Salmeron-Majadas et al, 2014***]. The multi-modal approaches typically consider a wide range of features collected by both devices, and have recently been used in a similar manner to ProsocialLearn's aim of enriching multi-modal affective computing systems that utilize behavioral signals incoming from sensory units, most prominently physiological signals and facial expressions [***Salmeron-Majadas et al, 2014***]. A summative view on a large number of proposed features in the multi-modal approach scientific literature is presented in Table 23.

Within ProsocialLearn, we aim to proceed with appropriate features' selection for the fusion algorithms further down the player modelling and prosocial affect pipeline, according to availability of devices per game. Our selection schemes should also take into consideration, how each device is intended to be used for providing input (e.g. only measuring features during intervals when player input is expected). Additionally, feature reduction may be necessary due to high correlation between parameters. More details on this feature selection process will be given in future deliverables.

3.3.2 Sentiment analysis on chat messages

Sentiment analysis refers to the evaluation of a piece of text. This statement includes various aspects, as sentiment analysis has been used to detect the polarity of messages (positive, negative or neutral), the detection of objective or subjective sentences, detecting emotions such as joy, anger or fear and applying sentiment analysis in various domains such as healthcare, commerce and disaster management [***Kiritchenko et al, 2014***].

Work	Mouse features	Keyboard features
[*Zimmermann et al, 2006*]	*Total number of mouse clicks, single mouse clicks (multiple clicks counted*	*Number of keystrokes, median length of a keystroke, etc.*

	as one click), total distance of mouse pointer, mouse speed, median click time (time between pressing and releasing a mouse button), number of pauses in mouse movement, median distance of single mouse movement, mouse acceleration, angle and direction of mouse movements.	
[***Salmeron-Majadas et al, 2014***]	*Number of button presses (left/right/both), overall distance, distance covered between two button press events (1), distance covered between button press and following button release (2), distance covered between two button release events (3), distance covered between a button release and the following button press events (4), Euclidean distance in cases (1), (2), (3) and (4), difference between covered and Euclidean distance in cases (1), (2), (3) and (4), times elapsed between events in cases (1), (2), (3) and (4).*	***Individual keystroke indicators*** *Number of key press events, average time between key press events, average time per stroke, number of times a given key has been pressed, number of times a set of keys has been pressed.* ***Digraph/Trigraph*** *Time between down keys, duration of each key event, time between a key up and following key down, duration of digraph/trigraph, number of events in the key events combination*

Table 23 -Summary of Mouse & Keyboard multi-modal input behavioral features in the scientific literature.

Sentiment analysis generally aims at determining the writer's attitude (i.e. the emotional state of the writer at the time of writing), the emotional effect bestowed upon the written medium by the author or the overall contextual polarity of the written segment. Textual information takes on numerous forms (articles, blogs, SMS messages, chatrooms, tweets, etc.). In ProsocialLearn, we will focus on sentiment analysis performed on short, informal textual messages, which may be sent among players collaborating or competing within a shared game environment. These messages are limited in length, usually spanning one sentence or less. Some characteristics associated with these messages are the many occurrences of misspellings, slang terms, and shortened forms of words, as well as the occasional inclusion of special markers, such as hashtags and emoticons. Such markers may be used to facilitate search, but can also indicate a topic, trend or sentiment.

Due to an explosive amount of scientific literature concerning sentiment analysis, a large portion of which is beyond the scope of this document (or this project in general), we will point out the interested reader to some of the more prominent surveys on the topic [***Pang & Lee, 2008***] [***Kim et al, 2011***] [***Cambria et al, 2013***] [***Kiritchenko et al, 2014***]. In ProsocialLearn, we aim at utilizing available solutions in the form of Application Programming Interfaces (APIs) for integrating sentiment analysis on player chat messages, whenever chatting is available as an option in the prosocial game. The preferred APIs should generate sentimental values that concern both individual words as well as whole phrases, while also demonstrating robustness against the use of emoticons. Some early suggestions on state of the art APIs are presented in Table 24. Further down the player modelling and prosocial affect pipeline, heuristic approaches will be used to attempt an identification of prosocial signals. This topic will be further elaborated in future deliverables of the project.

API	License	Output
Stanford NLP[14]	*academic*	*5-scale Polarity values (Very Negative, Negative, Neutral, Positive, Very Positive)*
Alchemy API[15]	*commercial*	*Score-based Polarity value (Negative, Neutral, Positive)*
Semantria[16]	*commercial*	*Score-based Polarity value (Negative, Neutral, Positive)*
Sentiment140[17]	*commercial*	*3-scale Polarity values (Negative, Neutral, Positive)*

Table 24 -Sample academic and commercial APIs for online text sentiment analysis.

[14] http://nlp.stanford.edu/sentiment/

[15] http://www.alchemyapi.com/

[16] https://semantria.com/demo

[17] http://help.sentiment140.com/api

4 Static Player Data

All player data stemming from player profiles, parent/teacher/caregiver input as well as session data considered constant throughout the duration of the session (i.e. physical and contextual conditions in which a gameplay session is progressing), are categorized in the ProsocialLearn project as 'Static' player data. In this Section, we will elaborate on the nature and acquisition procedures for these particular factors, effectively differentiating from audio/visual player behavioral cues and game-related behavioral patterns and wrapping up our observation data acquisition pipeline features. This breakdown of player data into the distinct categories of sensory observations, in-game logs and static data presents a solid and consistent structure of observations, which will greatly assist in enhancing future work on player profile modelling (Task 3.3) and dynamic fusion of modalities (Task 3.2).

4.1 Contextual data acquisition

Data concerning relevant contextual issues is useful for strengthening our understanding of, and providing additional meaning to automatically acquired data. Contextual data relates to the multitude of aspects outside of the recorded audiovisual streams and game play interaction recordings that may be judged to be important factors in influencing behavioral interactions in the game (to be differentiated from 'Contextual factors affecting prosocial learning', as described in D2.1, which relates to player profile acquisition in this document – see Section 4.2). It is useful as a means for increasing confidence in inferences made from behavioral data, for example, that behaviors are a result of in-game interactions and not due to other causes, such as physical space constraints or disruptions in the environment. Contextual data is therefore of use to human annotators, whose awareness after the fact during playback sessions is sometimes limited due to the available sensor equipment and game footage, and to automated learning mechanisms.

Examples of contextual data categories include:

- Physical set-up: The presence of a window behind the main game screen may cause attention to be shifted from the game and help explain fixations outside of the game screen that may otherwise be interpreted as nervousness or a decrease in engagement.
- Physical space constraints: The use of the Kinect sensor in a constrained room may alter body motion strategies producing more contracted motions.
- Cohort and peer presence: Games played with peers may be more stressful than those played alone.
- Educational context: The use of the system for graded activities may result in different behaviors to usage as leisure activity.
- Disruption level: Events taking place outside the range of the audio and visual sensing equipment described in sub-Section 2.1 may cause distraction.

Typically, due to the complex and dynamic nature of contextual data, it may be gathered manually via an online questionnaire before or after the session and, for practical purposes, relate solely to issues considered to be salient or noteworthy by the experimenter/teacher in that particular gaming instance.

4.2 Player Profile data acquisition

A player profile is a repository of information about an individual, merging multiple sources of data from psychometric questionnaires (or subsets of questionnaires), previous interactions, and other sources from parents, teachers and caregivers (based on questionnaires assessing the value systems of schools, for example). The purpose of the profile is to assist in the estimation of an individual's prosocial disposition in order to inform adaptation mechanisms and provide a foundation for into which to place their likely future decisions and behaviors related to prosociality.

4.2.1 Psychometric questionnaires

Psychometrics is concerned with the objective measurement of skills and knowledge, abilities, attitudes, personality traits, and educational achievement. In relation to ProsocialLearn, as described in Deliverable D2.1, contextual factors affecting prosocial learning primarily concern temperament, personality, attachment style, demographics (age, gender, and income), culture, family and characteristics of the beneficiary. In principle, this data may be captured offline by means of questionnaires that are distributed to the individuals (questionnaires may be intended for the children, teachers or parents). An overview of the questionnaires is provided in Table 25. See Deliverable D2.1, Appendices 1 and 2 for a more detailed description of each of the questionnaires.

Questionnaire name	Participant	Languages	Time to complete	Age group
TMCQ: Temperament in middle childhood questionnaire	Child / Parent / Teacher	English, Italian	10-20 mins	7-10
EAQ: Emotional Awareness Questionnaire	Child	English, Italian	10-20 mins	9-16
CBS: Child behavior scale	Teacher	English	10-20 mins	6-13
BFQ-C: Big Five Questionnaire in late childhood (personality)	Child / Parent / Teacher	English, Italian	10-20 mins	7-13
ICID-s: Inventory for Child Individual differences	Child / Parent / Teacher	English	10-20 mins	2-15
The Junior Temperament and Character Inventory	Parent	English, Italian	10-20 mins	6-16
Parenting Styles	Parent	English	5-10 mins	-
ASCQ: attachment	Child	English	<5 mins	7-14

Table 25 -Summary of psychometric questionnaires for children, parents and teachers.

In practice, the approach of filling out questionnaires may not be feasible due to the time and organizational requirements involved, often in classroom environments. While the player profile could contain substantial information covering all aspects of these questionnaires, are more feasible approach involves the development of questionnaire subsets that are acquired directly within the game environment, ideally as part of the game scenario.

4.2.2 Other inputs from parents, teachers and caregivers

The player profile also contains other potentially relevant data elicited, for example, as feedback from teachers, parents and caregivers. For example, data relating to the value systems of schools and the ways in which prosociality is already being taught (see Deliverable D2.1, Section 4) is of relevance to developing the profile of individuals. Data sources are typically derived from questionnaires probing how valued various aspects of prosociality are in specific cultures, for example, asking teachers to rank empathy, fairness, cooperation concern for others, and so on. These data sources provide a valuable backdrop for better understanding individual profiles.

5 Multi-modal / Multi-sensor capturing setup

Player input data and processes necessary for affect fusion described in this document stem from the identification and exploration of all available game input modalities, which can be summarized in Section 2.1. Games developed for the ProsocialLearn project may embed use of these sensors in the game control, and should definitively include reasons in the game logic for using the sensors input (e.g. collaboration through voice channel) that are in general provided by the platform and not by the game. Sensor data collected can be then pre-processed and fused (Task 3.2) with the ultimate goal of inferring emotional and engagement status of the player.

In this Section, we will describe how a multi-modal/multi-sensor gameplay input capturing setup can be deployed for use with the already implemented or planned early prosocial games, providing a thorough mapping of the data acquisition processes to their respective game environments. This will serve as a general guide towards the development of future prosocial games later in the project's lifetime.

5.1 Introduction to the implemented capturing platform

As is apparent from previous Sections in this document, a multi-channel acquisition of observations is supported as part of the work carried out in Task 3.1 of the ProsocialLearn project. Each channel has its own architecture, strictly related to the nature of the observation. The observations themselves are categorized under the audio/visual sensory data (Section 2) and in-game/player background knowledge (Sections 3 and 4) classes.

During the project's first system requirements and architecture report (D2.3), an explanation of the in-game data flow, in which the observation acquisition manager, sensors and processing are mounted on the client-side, was presented. Observations are collected and processed locally on the user's machine and are only authorized to leave the platform according to specifications set by the client regarding data sensitivity. As a reminder, we append this data flow in Figure 33 – ProsocialLearn architecture in game data flow, defined in D2.3, to better emphasize on the positioning of the observation acquisition pipeline in the overall platform architecture. In this respect, observation sensors and I/O devices create input streams to the observation acquisition platform, while a two-way I/O stream with the game client is used to handle direct player input interactions with the game. From the sensors' detailed presentation provided in Section 2.1, it is apparent that any type of sensory hardware used as part of the observations acquisition can also be considered an input device. Synchronization of the game client and any sensory hardware is also required, as sensors and input devices need to collect game-related data and observe player behavior patterns during gameplay, to properly infer prosocial affect signs and trigger adaptation. Furthermore, any device used to directly provide input to the game environment needs to obtain signals during which player activity is analyzed and translated to in-game commands (for example, Kinect gestures or voice commands). During these timeframes, observations still need to be acquired, as they may provide cues towards player engagement (i.e. vigorously shouting out commands as opposed to indifferently repeating keywords).

Seeing how observations are made locally, the activation and synchronization of all available sensors needs to be handled by a generic framework for input device communication. The framework needs to ensure that each sensor is turned on and properly initialized at the precise moment, for acquiring observations with respect to and throughout the duration of the gameplay session only.

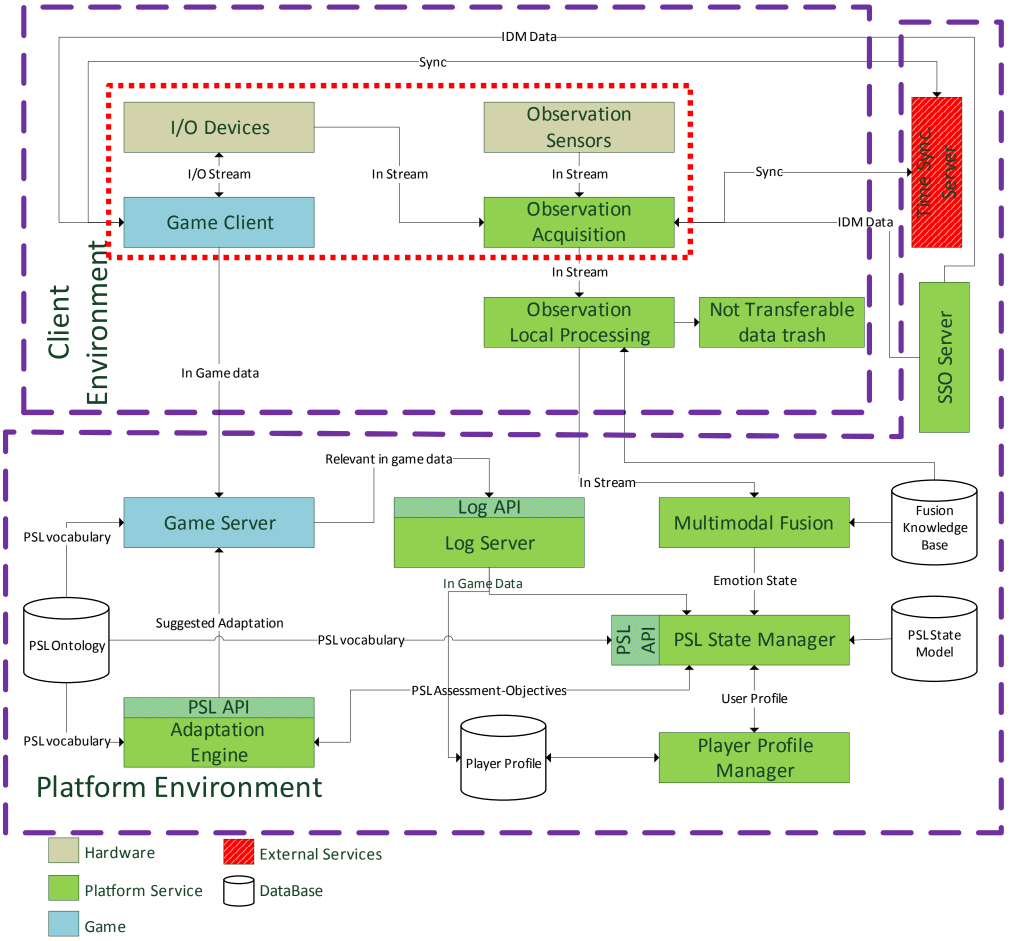


Figure 33 – ProsocialLearn architecture in game data flow, defined in D2.3

The red dotted line area represents the observations acquisition pipeline, i.e. sensor/device communication with the game client.

Simple interactions between the game client and all possible sensory input devices can be carried out, for example, through the use of local sockets. Observation channels need to be open and closed. We envisage different possibilities for doing so; for example, a function in the game calling the start/stop or it can be a function exposed by the platform itself. Another approach would be to support a combination of the two. The logic binding between game and observation channels is performed via the platform SSO.

The framework is generic in that, it utilizes the same protocol to communicate with a multitude of sensory hardware and I/O devices described in Section 2.1, regardless of their type, making the process of plugging in or omitting a specific sensor before a session is started a trivial process.

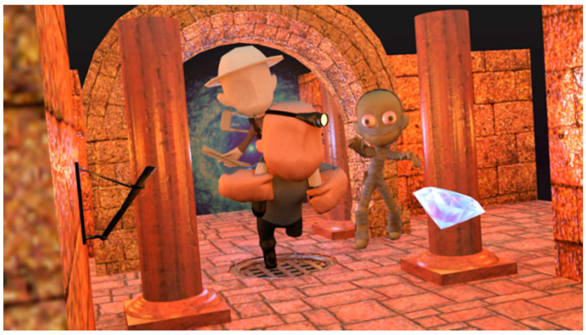
5.2 Mapping to the ProsocialLearn games' environments

Throughout the course of this report, we have elaborated on our observation acquisition pipeline per category (sensory, in-game and static), modality (audio, visual, body and input) and feature level (speech, facial expression, eye gaze, motion, interaction patterns, chat). In this Section we present how these observations can be applied to games developed as part of the ProsocialLearn project, to aid future game developers in choosing how and why to incorporate these observations in their own games. A comprehensive summary that maps observations to intended yielding results with respect to prosocial player affect and engagement is presented in Table 26. As a further aid, we conclude this Section by providing an overview of the initial games developed for small scale experimental studies conducted for analyzing and extracting the features presented in Section 2 and Section 3. Each game is reserved a Table summarizing the entire range of feature modalities supported, as well as the actual observed features themselves. We present these studies in hopes of helping incoming SMEs later in the lifetime of the project by providing a baseline and target outputs on which to build their own gaming worlds.

Category	Modality	Hardware	Type	Level	Target output
Sensory Observations	Audio	Microphone	Speech	Low-level	Player affect
				Two-class	
				Five-class	
	Visual	Camera	Facial Expression	Low-level	Player affect / Engagement
				AU	
				Emotions	Player Affect
			Gaze	Low-level	Player affect / Engagement
				Attention	Engagement
	Body	Kinect/Kinect2 LEAP Motion Controller	Motion	3D Body / Head / Hand	Player affect / Engagement
In-game Logs	Input	Mouse Keyboard	Interaction pattern	Behavioral	Player affect / Engagement
		Keyboard	Chat	Sentiment analysis	Player Affect
Static data	Questionnaire	-	Contextual	Basic	Player affect / Engagement
			Player Profile	Psychometric	Player Affect
				Third-party input	

Table 26 -Mapping overview of observations acquisition to target outputs.

5.2.1 Path of Trust



Path of Trust (PoT) is a two-player, endless running maze game in which players strive to collect treasure while trying to avoid mummies and traps. One player assumes the role of the *Guide*, a character assumed to have explored the maze before and therefore able to navigate through the corridors via a top-down view of the dungeon map. The other player is put in control of the endless running *Muscle* character, which is responsible to navigate the maze, but has no information on the layout, of the maze or the room contents. The duo navigates the maze by having the Muscle carry the Guide on his back. During each game cycle, the characters find themselves in one of the maze's many rooms, which end up in junctions leading up to 3 different directions. Before entering the room, the Guide has already been shown a short glimpse of the contents in each of the adjacent rooms and has to pick a decision for the Muscle character to follow. The Muscle then gets a small time window to decide whether to actually heed the Guide's advice and proceeds to take one of the available routes. Afterwards, the Muscle gets a short time to either collect the treasure or avoid a hazard running inside a small corridor leading up to the next room/junction while the Guide is again shown the contents of the adjacent rooms there, from which point the cycle repeats. Characters are rewarded for collecting treasure, yet the Muscle player is rewarded twice the amount of points for each treasure piece collected. The game offers a way for players to switch characters in the form of a magic portal, and therefore players get the chance to reap equal rewards by planning their character swaps carefully. The game furthermore features a time limit and 5 different endings for players to reach. PoT was designed to help children understand the benefits of *Cooperation* and when to express *Trustworthiness*, as described in D2.2. The Muscle player must learn to trust the Guide to provide guidance away from danger and towards mutual interests while the Guide must learn to trust the Muscle to listen to directions. This trust and cooperation element is a crucial component to progress in the game, as players may be caught by one of the mummies and thus lose hard-earned treasure or fall into a trap and instantly lose the game, meaning none of the players actually succeeds. Figure 34 – Path of Trust gameplay cycle for Guide (top) and Muscle (bottom) players shows screenshots of the game taken during a complete game cycle.

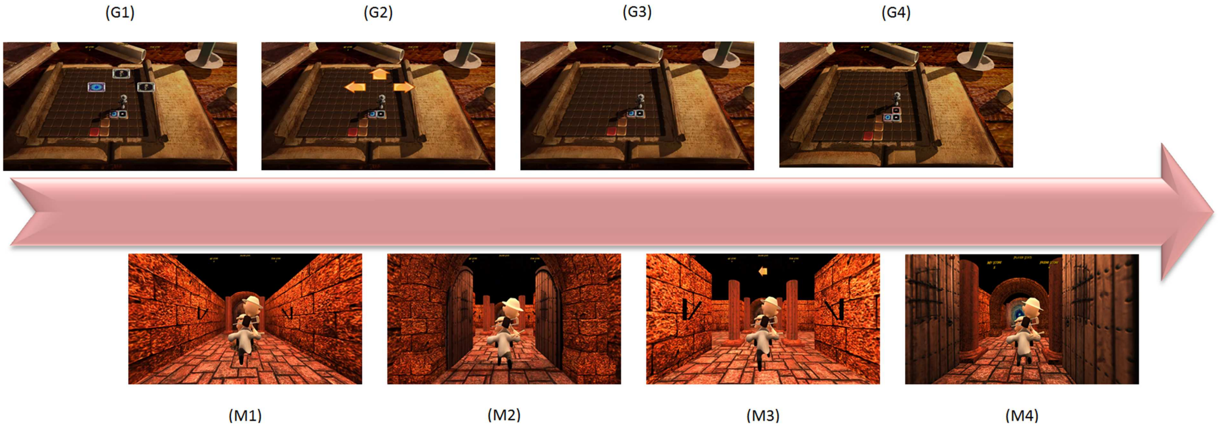


Figure 34 – Path of Trust gameplay cycle for Guide (top) and Muscle (bottom) players

Guide is briefly shown the contents of the adjacent rooms (G1) while the Muscle runs through the corridor (M1). Then the Guide is called upon to pick a direction to suggest to the Muscle (G2), who is in the meantime entering the junction room (M2). The Muscle is displayed the Guide's direction and has to choose whether to follow while arriving at the junction (M3), as the Guide awaits the outcome (G3). The Muscle then enters a corridor connecting to the chosen room and takes action in accordance to the found content (M4), while the Guide progresses towards the next space on the top-down map (G4). The cycle then repeats at {G1, M1}.

Two players connect to a server which propagates game-related information between the two to maintain a balance in the game. The game clients are responsible for rendering the game and deploying the NUI, while the server updates the game state on both clients based on inputs received in the previous game cycle. The game can be played either using a traditional approach (e.g. keyboard) or through a gesture-driven NUI. More specifically, two distinct sensor-driven configurations are set up according to the generic framework for input device communication, described in sub-Section 5.1.

One involves the Kinect sensor while the other utilizes the LEAP Motion controller. Students playing PoT using the Kinect have to perform three distinct hand gestures to interact with the game during input acquisition time windows. The same applies for students playing using the LEAP Motion controller. These gestures are shown in Figure 35. Seeing how students playing the game using the

LEAP Motion configuration get to be seated in front of a standard computer monitor, a camera observation modality can be additionally plugged in for the acquisition of multi-modal observations. PoT therefore makes use of almost every possible visual observation acquisition method, as is shown in Table 27.

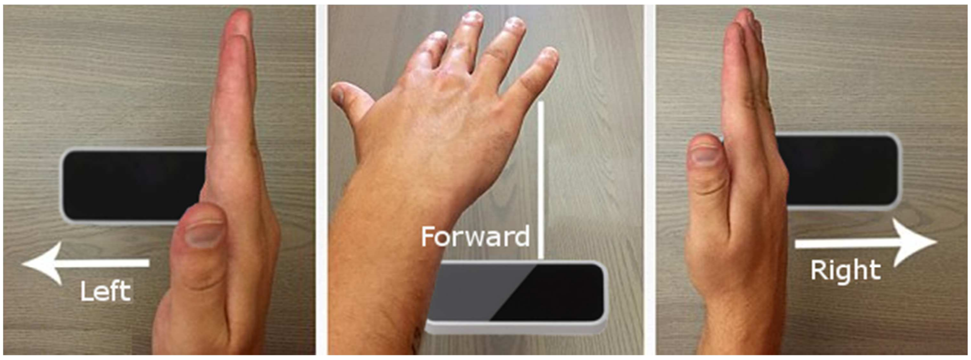


Figure 35 -Hand gestures for controlling player Muscle/Guide actions using LEAP Motion controller.

Sensor	Feature Modality	Feature Level	Actual features
Camera	*Facial Expression Analysis*	*Low-level features*	Eyes, Eyebrows, Mouth.
		AUs	AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU17, AU25, AU26.
		Emotions	Happiness, Sadness, Surprise, Anger, Fear, Neutral.
	Gaze Analysis	*Low-level features*	Gaze Distance, Pupil Diameter, Blinking
		Visual Attention	Attentive/Non-attentive
Microphone	*Speech Analysis*	*Low-level features*	Zero crossing rate, RMS energy, F0, Harmonic noise ratio, MFCC
		Two-class	Idle, Negative
		Five-class	Angry, Empathic, Neutral, Positive, Rest
Kinect	*Body Motion Analysis*	-	3D Body, Head Motion, Hand Motion.
LEAP Motion Controller	*Body Motion Analysis*	-	Hand Motion
Traditional Input Controllers: Keyboard	*Gameplay Data Analysis*	*Behavioral observation data logging*	Individual keystroke indicators

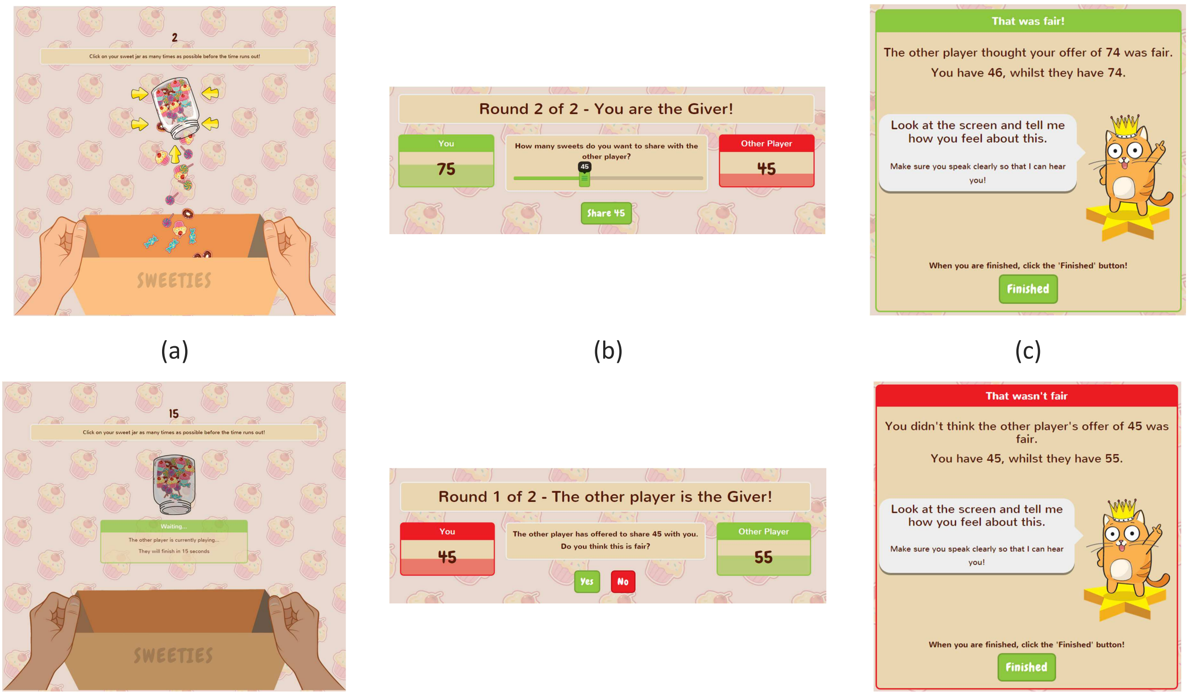
Table 27 -Summary of features mapped onto the PoT game environment.

5.2.2 Kitty King’s Candy Quest

Candy Quest
Click on your sweet jar as many times as possible before the time runs out!
click!
click!
click!
sweeties

Kitty King’s Candy Quest (KKCQ) provides a set of scenarios that present specific moments for pairs of participants to make decisions of *Generosity* and *Fairness* in nature, as described in D2.2. KKCQ is a web-based, two-player game, focused on decision points that deal with prosocial concepts of fairness and generosity. There are four variations of the game, each one contained within the same game package.

A single gameplay cycle is broken down into the following player actions: at the start of the cycle, players complete a short round of collecting candy by clicking on a candy jar. One player is assigned the role of the *Giver*. This player gets all of the candy collected and has to decide how much to share with the other player, who takes on the role of the *Receiver*. The Receiver then decides if the sharing was done in a fair manner. A game consists of several cycles, involving different variants of the above situation with subtle variation that test different generosity and fairness attitudes and responses (i.e. a second variant allows both players to collect candy simultaneously, each player having his own candy jar, while clicking contributes to a shared total). Each of the mini-games takes 1 or 2 minutes to complete. Screenshots of the game flow are provided in Figure 36 – Kitty King’s Candy Quest game flow.



2

Click on your sweet jar as many times as possible before the time runs out!

(a)

Round 2 of 2 - You are the Giver!

You 75

How many sweets do you want to share with the other player?

45

Other Player 45

Share 45

(b)

That was fair!

The other player thought your offer of 74 was fair.

You have 46, whilst they have 74.

Look at the screen and tell me how you feel about this.

Make sure you speak clearly so that I can hear you!

When you are finished, click the 'Finished' button!

Finished

(c)

15

Click on your sweet jar as many times as possible before the time runs out!

Waiting...

The other player is currently playing...

They will finish in 15 seconds

(d)

Round 1 of 2 - The other player is the Giver!

You 45

The other player has offered to share 45 with you.
Do you think this is fair?

Other Player 55

Yes No

(e)

That wasn't fair

You didn't think the other player's offer of 45 was fair.

You have 45, whilst they have 55.

Look at the screen and tell me how you feel about this.

Make sure you speak clearly so that I can hear you!

When you are finished, click the 'Finished' button!

Finished

(f)

Figure 36 – Kitty King’s Candy Quest game flow

Top row displays Giver gameplay, where candy is being collected from a jar (a), then an offer to the Receiver is made (b) and finally a message of the offer’s evaluation is displayed (c). The bottom row shows the gameplay flow from the Receiver point of view. The player waits for the Giver to collect candy (d) and receives the offer, which the player has to decide if they think it’s fair with a Yes/No answer (e). Then, the assessment is displayed to both players (f).

While KKCQ does not take input from any particular observation sensor like in the PoT game, unobtrusive player observation can be collected by plugging in camera and microphone sensors for player behavior monitoring. As shown in Figure 36 – Kitty King's Candy Quest game flow, after each section subjects are asked by the game to respond to their decision, prompted by "Look at the screen and tell me how you feel about this". This allows observations made by the audio channel to analyze speech features, as described in Section 2.2. A complete mapping of observations to the Giver-Receiver game environment is presented in Table 28.

Sensor	Feature Modality	Feature Level	Actual features
Camera	*Facial Expression Analysis*	*Low-level features*	Eyes, Eyebrows, Mouth.
		AUs	AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU17, AU25, AU26.
		Emotions	Happiness, Sadness, Surprise, Anger, Fear, Neutral.
	Gaze Analysis	*Low-level features*	Gaze Distance, Pupil Diameter, Blinking
		Visual Attention	Attentive/Non-attentive
Microphone	*Speech Analysis*	*Low-level features*	Zero crossing rate, RMS energy, F0, Harmonic noise ratio, MFCC
		Two-class	Idle, Negative
		Five-class	Angry, Empathic, Neutral, Positive, Rest

Table 28 -Summary of features mapped onto the *KKCQ* game environment.

6 Conclusions

In this report, a complete summary of observations and input acquisition data was presented, along with general guidelines on applying these observations into future prosocial games and how each modality could map to a game environment.

Three distinct categories of player input data acquisition were defined. These include the sensory observations acquired through audio/visual student behavior monitoring, in-game data logs being constantly recorded during gameplay, and player static data, collected per gaming session and through the player's personal profile, defined for each student. For each of the aforementioned categories, a sufficient set of features were proposed for extraction, and categorized according to the achieved level of player affective abstraction (i.e. low-level observational measurements, emotions, or mid-level representations, such as AUs). All features presented in this report are considered candidates for dynamic fusion algorithm input (Task 3.2). Emotional qualities, being the highest level of features are directly calculated from lower-level features representations, using certified algorithms for person-dependent affect analysis. In this report we have thoroughly presented the state-of-the-art in affect recognition using multi-modal input modalities, and have thus justified our choices of employed algorithms. Where applicable, evaluation schemes were proposed, and experimental results presented, solidifying the choice of components per approach. Where such an evaluation was not possible due to a current lacking in relevant or existing data in the scientific literature, the partners in WP3 of the ProsocialLearn project have proposed their plans on further work to go beyond the requirements of this Task and contribute towards the goals of the project, and the scientific community as well. All of these additional developments being mentioned in this report will be included in future deliverables, later in the project's lifetime.

Furthermore, in this report, an overview of the current state of early developed prosocial games and mapping of the observations acquisition procedures onto the respective game environments was presented. Towards this end, a generalized framework for unobtrusive capturing of students during gameplay was explored and tied-in with the overall architecture and requirements of the ProsocialLearn platform (D2.3). Early prosocial studies taking place in schools in Greece (WP7) have encouraged us to proceed with this approach. We hope to have set a comprehensive and thorough example on how to build an observations acquisition pipeline, choose algorithms and hardware and draw a mapping of affective/engagement-based outputs within game environments in a way that will allow future game developments being invited to join the project to rapidly prototype new prosocial games and corresponding game adaptation mechanisms (WP4).

7 References

[***Aifanti et al, 2010***] Aifanti, N., Papachristou, C., & Delopoulos, A. (2010, April). The MUG facial expression database. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on* (pp. 1-4). IEEE.

[***APA 2013***] American Psychiatric Association. 2000. Diagnostic and statistical manual of mental disorders (4th edition, text revised). Washington, DC.

[***Apostolakis & Daras, 2013***] Apostolakis, K. C., & Daras, P. (2013, July). RAAT-The reverie avatar authoring tool. In *Digital Signal Processing (DSP), 2013 18th International Conference on* (pp. 1-6). IEEE.

[***Apostolakis & Daras, 2014***] Apostolakis, K. C., & Daras, P. (2014). A framework for implicit human-centered image tagging inspired by attributed affect. *The Visual Computer*, *30*(10), 1093-1106.

[***Argyle and Cook, 1976***] Argyle, M. & Cook, M. (1976). Gaze and mutual gaze, Cambridge University Press.

[***Asteriadis et al, 2009***] Asteriadis, S., Tzouveli, P., Karpouzis, K., & Kollias, S. (2009). Estimation of behavioral user state based on eye gaze and head pose—application in an e-learning environment. *Multimedia Tools and Applications*, *41*(3), 469-493.

[***Bakelman et al, 2013***] Bakelman, N., Monaco, J. V., Cha, S. H., & Tappert, C. C. (2013, August). Keystroke biometric studies on password and numeric keypad input. In*Intelligence and Security Informatics Conference (EISIC), 2013 European* (pp. 204-207). IEEE.

[***Bänziger et al, 2010***] Bänziger, T., & Scherer, K. R. (2010). Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing: A sourcebook*, 271-294.

[***Bänziger et al, 2012***] Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva multimodal expression corpus for experimental research on emotion perception.*Emotion*, *12*(5), 1161.

[***Berger et al, 2011***] Berger, K., Ruhl, K., Schroeder, Y., Bruemmer, C., Scholz, A., & Magnor, M. A. (2011, October). Markerless Motion Capture using multiple Color-Depth Sensors. In *VMV* (pp. 317-324).

[***Baron-Cohen, 1995***] Baron-Cohen S. (1995) *Mindblindness*. MIT, Cambridge.

[***Bettadapura, 2009***] Bettadapura, V. (2009). Face expression recognition and analysis: the state of the art. Emotion, pp. 1–27.

[***Bhatti et al, 2008***] Bhatti, M. W., Wang, Y., & Guan, L. (2008). Language independent recognition of human emotion using artificial neural networks. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, *2*(3), 1-21.

[***Bixler & D'Mello, 2013***] Bixler, R., & D'Mello, S. (2013, March). Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. In*Proceedings of the 2013 international conference on Intelligent user interfaces*(pp. 225-234). ACM.

[***Boone & Cunningham, 1998***] Boone, R. T., & Cunningham, J. G. (1998). Children's decoding of emotion in expressive body movement: the development of cue attunement.*Developmental psychology*, *34*(5), 1007.

[***Bradley et al, 2008***] Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*,*45*(4), 602-607.

[***Busso et al, 2004***] Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., ... & Narayanan, S. (2004, October). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces* (pp. 205-211). ACM.

[***Calvo & Fernández-Martín, 2013***] Calvo, M. G., & Fernández-Martín, A. (2013). Can the eyes reveal a person's emotions? Biasing role of the mouth expression. *Motivation and Emotion*,*37*(1), 202-211.

[***Cambria et al, 2013***] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, (2), 15-21.

[***Camurri et al, 2003***] Camurri, A., Lagerlöf, I., & Volpe, G. (2003). Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques.*International journal of human-computer studies*, *59*(1), 213-225.

[***Chang et al, 2011***] Chang, Y. J., Chen, S. F., & Huang, J. D. (2011). A Kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities.*Research in developmental disabilities*, *32*(6), 2566-2570.

[***Claypool et al, 2001***] Claypool, M., Le, P., Wased, M., & Brown, D. (2001, January). Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces* (pp. 33-40). ACM.

[***Cohn et al, 2004***] Cohn, J. F., Reed, L. I., Moriyama, T., Xiao, J., Schmidt, K., & Ambadar, Z. (2004, May). Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on* (pp. 129-135). IEEE.

[***Cootes et al, 1995***] Cootes, T. F., Taylor, C. J., Cooper, D. H., & Graham, J. (1995). Active shape models-their training and application. *Computer vision and image understanding*, *61*(1), 38-59.

[***Cowie et al, 2010***] Cowie, R., Gunes, H., McKeown, G., Vaclau-Schneider, L., Armstrong, J., & Douglas-Cowie, E. (2010). The emotional and communicative significance of head nods and shakes in a naturalistic database. In *Proc. of LREC Int. Workshop on Emotion* (pp. 42-46).

[***Davidson et al, 1990***] Davidson, R. J., Ekman, P., Saron, C. D., Senulis, J. A., & Friesen, W. V. (1990). Approach-withdrawal and cerebral asymmetry: Emotional expression and brain physiology: I. *Journal of personality and social psychology*, *58*(2), 330.

[***Doughtym, 2001***] Doughty, M. J. (2001). Consideration of three types of spontaneous eyeblink activity in normal humans: during reading and video display terminal use, in primary gaze, and while in conversation. *Optometry & Vision Science*, *78*(10), 712-725.

[***Ekman & Friesen, 1978***] P. Ekman and W. Friesen, Facial Action Coding System. Palo Alto, CA: Consulting Psychologist, 1978.

[**Ekman et al, 2002**] Ekman, P., Friesen, W.F., Hager, J.C.(2002). *FACS Manual 2002, Investigator's Guide* p 173-174.

[***El-Nasr & Yan, 2006***] El-Nasr, M. S., & Yan, S. (2006, June). Visual attention in 3D video games. In*Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology* (p. 22). ACM.

[***Eyben et al, 2013***] Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013, October). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia* (pp. 835-838). ACM.

[***Felipe et al, 2012***] Felipe, D. A. M., Gutierrez, K. I. N., Quiros, E. C. M., & Vea, L. A. (2012). Towards the Development of Intelligent Agent for Novice C/C++ Programmers through Affective Analysis of Event Logs. In *Proc. Int. MultiConference Eng. Comput. Sci* (Vol. 1).

[***Gao et al, 2012***] Gao, Y., Bianchi-Berthouze, N., & Meng, H. (2012). What does touch tell us about emotions in touchscreen-based gameplay?. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *19*(4), 31.

[***de Gelder, 2009***] de Gelder, B. (2009). Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3475-3484.

[***Grafsgaard et al, 2013***] Grafsgaard, J., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. (2013, July). Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*.

[***Grubišić et al, 2015***] Grubišić, I., Skala Kavanagh, H. A. N. A., & Grazio, S. (2015). Novel approaches in hand rehabilitation. *Periodicum biologorum*, *117*(1), 139-145.

[***Guna et al, 2014***] Guna, J., Jakus, G., Pogačnik, M., Tomažič, S., & Sodnik, J. (2014). An analysis of the precision and reliability of the leap motion sensor and its suitability for static and dynamic tracking. *Sensors*, *14*(2), 3702-3720.

[***Gunes & Piccardi, 2006***] Gunes, H., & Piccardi, M. (2006, August). A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In*Pattern Recognition, 2006. ICPR 2006. 18th International Conference on* (Vol. 1, pp. 1148-1153). IEEE.

[***Gunes & Piccardi, 2009***] Gunes, H., & Piccardi, M. (2009). Automatic temporal segment detection and affect recognition from face and body display. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, *39*(1), 64-84.

[***Gunes et al, 2015***] Gunes, H., Shan, C., Chen, S., & Tian, Y. (2015). Bodily expression for automatic affect recognition. *Emotion Recognition: A Pattern Analysis Approach*, 343-377.

[***Hajimirza et al, 2012***] Hajimirza, S. N., Proulx, M. J., & Izquierdo, E. (2012). Reading users' minds from their eyes: A method for implicit image annotation. *Multimedia, IEEE Transactions on*, *14*(3), 805-815.

[***Heylen, 2006***] Heylen, D. (2006). Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics*, *3*(03), 241-267.

[***Junjie Shan et al, 2014***] Junjie Shan, Akella, S. (2014). 3D human action segmentation and recognition using pose kinetic energy. *Advanced Robotics and its Social Impacts (ARSO), 2014 IEEE Workshop, 69-75*

[***Kaklauskas et al, 2009***] Kaklauskas, A., Krutinis, M., & Seniut, M. (2009, July). Biometric mouse intelligent system for student's emotional and examination process analysis. In*Advanced Learning Technologies, 2009. ICALT 2009. Ninth IEEE International Conference on* (pp. 189-193). IEEE.

[***Kanade et al, 2000***] Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on* (pp. 46-53). IEEE.

[***Karnan et al, 2011***] Karnan, M., Akila, M., & Krishnaraj, N. (2011). Biometric personal authentication using keystroke dynamics: A review. *Applied Soft Computing*,*11*(2), 1565-1573.

[***Kendon, 1990***] Kendon, A. (1990). *Conducting interaction: Patterns of behavior in focused encounters* (Vol. 7). Cambridge University Press Archive.

[***Kessous et al, 2010***] Kessous, L., Castellano, G., & Caridakis, G. (2010). Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, *3*(1-2), 33-48.

[***Khanna & Sasikumar, 2010***] Khanna, P., & Sasikumar, M. (2010). Recognising emotions from keyboard stroke pattern. *International journal of computer applications*, *11*(9), 1-5.

[***Kim et al, 2011***] Kim, H. D., Ganesan, K., Sondhi, P., & Zhai, C. (2011). Comprehensive review of opinion summarization.

[***Kiritchenko et al, 2014***] Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 723-762.

[***Kleinke, 1986***] Kleinke, C. L. (1986). Gaze and eye contact: a research review. *Psychological bulletin*, *100*(1), 78.

[***Kleinsmith & Bianchi-Berthouze, 2007***] Kleinsmith, A., & Bianchi-Berthouze, N. (2007). Recognizing affective dimensions from body posture. In *Affective computing and intelligent interaction* (pp. 48-58). Springer Berlin Heidelberg.

[***Lange et al, 2011***] Lange, B., Chang, C. Y., Suma, E., Newman, B., Rizzo, A. S., & Bolas, M. (2011, August). Development and evaluation of low cost game-based balance rehabilitation tool using the Microsoft Kinect sensor. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*(pp. 1831-1834). IEEE.

[***Lee et al, 2012a***] Lee, H., Choi, Y. S., Lee, S., & Park, I. P. (2012, January). Towards unobtrusive emotion recognition for affective social communication. In*Consumer Communications and Networking Conference (CCNC), 2012 IEEE*(pp. 260-264). IEEE.

[***Lee et al, 2012b***] Lee, P. M., Tsui, W. H., & Hsiao, T. C. (2012, January). A low-cost scalable solution for monitoring affective state of students in e-learning environment using mouse and keystroke data. In *Intelligent Tutoring Systems* (pp. 679-680). Springer Berlin Heidelberg.

[***Lin et al, 2012***] Lin, C. C., Chang, C. C., & Liang, D. (2012, March). A new non-intrusive authentication approach for data protection based on mouse dynamics. In*Biometrics and Security Technologies (ISBAST), 2012 International Symposium on* (pp. 9-14). IEEE.

[***Lucey et al, 2010***] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010, June). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on* (pp. 94-101). IEEE.

[***Martin et al, 2006***] Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006, April). The eNTERFACE'05 audio-visual emotion database. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on* (pp. 8-8). IEEE.

[***Massaro, 1998***] Massaro, D. W. (1998). Illusions and issues in bimodal speech perception. In*AVSP'98 International Conference on Auditory-Visual Speech Processing*.

[***Minear & Park, 2004***] Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 630-633.

[***Monaco et al, 2012***] Monaco, J. V., Bakelman, N., Cha, S. H., & Tappert, C. C. (2012, August). Developing a keystroke biometric system for continual authentication of computer users. In *Intelligence and Security Informatics Conference (EISIC), 2012 European* (pp. 210-216). IEEE.

[***Montero et al, 1998***] Montero, J. M., Gutierrez-Arriola, J. M., Palazuelos, S. E., Enriquez, E., Aguilera, S., & Pardo, J. M. (1998, December). Emotional speech synthesis: from speech database to TTS. In *ICSLP* (Vol. 98, pp. 923-926).

[***Nordstrøm et al, 2004***] Nordstrøm, M. M., Larsen, M., Sierakowski, J., & Stegmann, M. B. (2004). *The IMM face database-an annotated dataset of 240 face images*. Technical University of Denmark, DTU Informatics, Building 321.

[***Pang & Lee, 2008***] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis.*Foundations and trends in information retrieval*, *2*(1-2), 1-135.

[***Pantic & Rothkrantz, 2000***] Pantic, M., & Rothkrantz, L. J. (2000). Automatic analysis of facial expressions: The state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *22*(12), 1424-1445.

[***Pantic & Rothkrantz, 2003***] Pantic, M., & Rothkrantz, L. J. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, *91*(9), 1370-1390.

[***Pantic et al, 2005***] Pantic, M., Valstar, M., Rademaker, R., & Maat, L. (2005, July). Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on* (pp. 5-pp). IEEE.

[***Papadopoulos et al, 2014***] Papadopoulos, G. T., Apostolakis, K. C., & Daras, P. (2014). Gaze-based relevance feedback for realizing region-based image retrieval. *Multimedia, IEEE Transactions on*, *16*(2), 440-454.

[***Piana et al, 2013***] Piana, S., Staglianò, A., Camurri, A., & Odone, F. (2013). A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition. In *IDGEI International Workshop*.

[***Pollick et al, 2001***] Pollick, F. E., Paterson, H. M., Bruderlin, A., & Sanford, A. J. (2001). Perceiving affect from arm movement. *Cognition*, *82*(2), B51-B61.

[***Potamianos & Narayanan, 2007***] Potamianos, A., & Narayanan, S. (2007, October). A review of the acoustic and linguistic properties of children's speech. In *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on* (pp. 22-25). IEEE.

[***Potter et al, 2013***] Potter, L. E., Araullo, J., & Carter, L. (2013, November). The leap motion controller: a view on sign language. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration* (pp. 175-178). ACM.

[***Radu et al, 2011***] Radu, P., Sirlantzis, K., Howells, G., Hoque, S., & Deravi, F. (2011, October). A Versatile Iris Segmentation Algorithm. In *BIOSIG* (pp. 137-150).

[***Ringeval et al, 2013***] Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013, April). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on* (pp. 1-8). IEEE.

[***Roether et al, 2008***] Roether, C. L., Omlor, L., & Giese, M. A. (2008). Lateral asymmetry of bodily emotion expression. *Current Biology*, *18*(8), R329-R330.

[***Russell, 2003***] Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, *110*(1), 145.

[***Sadrô et al, 2003***] Sadrô, J., Jarudi, I., & Sinhaô, P. (2003). The role of eyebrows in face recognition. *Perception*, *32*(3), 285-293.

[***Salmeron-Majadas et al, 2014***] Salmeron-Majadas, S., Santos, O. C., & Boticario, J. G. (2014). An evaluation of mouse and keyboard interaction indicators towards non-intrusive and low cost affective modeling in an educational context. *Procedia Computer Science*,*35*, 691-700.

[***Savva et al, 2012***] Savva, N., Scarinzi, A., & Bianchi-Berthouze, N. (2012). Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience. *Computational Intelligence and AI in Games, IEEE Transactions on*, *4*(3), 199-212.

[***Schmidt et al, 2015***] Schmidt, T., Araujo, F. P., Pappa, G. L., & Nascimento, E. R. Real-Time Hand Gesture Recognition Based on Sparse Positional Data.

[***Schowstra & Hoogstraten, 1995***] Schouwstra, S. J., & Hoogstraten, J. (1995). Head position and spinal position as determinants of perceived emotional state. *Perceptual and motor skills*,*81*(2), 673-674.

[***Schuller et al, 2009***] Schuller, B., Steidl, S., & Batliner, A. (2009, September). The INTERSPEECH 2009 emotion challenge. In *INTERSPEECH* (Vol. 2009, pp. 312-315).

[***Schuller et al, 2011***] Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., & Pantic, M. (2011). Avec 2011–the first international audio/visual emotion challenge. In*Affective Computing and Intelligent Interaction* (pp. 415-424). Springer Berlin Heidelberg.

[***Sengupta & Jeng, 2003***] Sengupta, T., & Jeng, O. J. (2003, March). Eye and mouse movements for user interface design. In *Bioengineering Conference, 2003 IEEE 29th Annual, Proceedings of* (pp. 1-2). IEEE.

[***Shaker et al, 2011***] Shaker, N., Asteriadis, S., Yannakakis, G. N., & Karpouzis, K. (2011). A game-based corpus for analysing the interplay between game context and player experience. In *Affective Computing and Intelligent Interaction* (pp. 547-556). Springer Berlin Heidelberg.

[***Shotton et al, 2012***] Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., ... & Blake, A. (2013). Efficient human pose estimation from single depth images.*Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *35*(12), 2821-2840.

[***Shultz et al, 2011***] Shultz, S., Klin, A., & Jones, W. (2011). Inhibition of eye blinking reveals subjective perceptions of stimulus salience. *Proceedings of the National Academy of Sciences*, *108*(52), 21270-21275.

[***Soleymani et al, 2012a***] Soleymani, M., Pantic, M., & Pun, T. (2012). Multimodal emotion recognition in response to videos. *Affective Computing, IEEE Transactions on*, *3*(2), 211-223.

[***Soleymani et al, 2012b***] Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *Affective Computing, IEEE Transactions on*, *3*(1), 42-55.

[***Sottilare & Proctor, 2012***] Sottilare, R. A., & Proctor, M. (2012). Passively classifying student mood and performance within intelligent tutors. *Journal of Educational Technology & Society*, *15*(2), 101-114.

[***Sneddon et al, 2012***] Sneddon, I., McRorie, M., McKeown, G., & Hanratty, J. (2012). The belfast induced natural emotion database. *Affective Computing, IEEE Transactions on*, *3*(1), 32-41.

[***Steidl, 2009***] Steidl, S. (2009). *Automatic classification of emotion related user states in spontaneous children's speech* (pp. 1-250). Germany: University of Erlangen-Nuremberg.

[***Stern et al, 1984***] Stern, J. A., Walrath, L. C., & Goldstein, R. (1984). The endogenous eyeblink.*Psychophysiology*, *21*(1), 22-33.

[***Sunstedt et al, 2013***] Sundstedt, V., Bernhard, M., Stavrakis, E., Reinhard, E., & Wimmer, M. (2013). Visual attention and gaze behavior in games: An object-based approach. In *Game analytics* (pp. 543-583). Springer London.

[***Thomaz & Giraldi, 2010***] Thomaz, C. E., & Giraldi, G. A. (2010). A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing*, *28*(6), 902-913.

[***Tian et al, 2001***] Tian, Y. L., Kanade, T., & Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *23*(2), 97-115.

[***Tsai et al, 2012***] Tsai, C. J., Chang, T. Y., Yang, Y. J., Wu, M. S., & Li, Y. C. (2012). An approach for user authentication on non-keyboard devices using mouse click characteristics and statistical-based classification. *International Journal of Innovative Computing, Information and Control*, *8*(11), 7875-7886.

[***Tsoulouhas et al, 2011***] Tsoulouhas, G., Georgiou, D., & Karakos, A. (2011). Detection of learner’s affective state based on mouse movements. *J. Comput*, *3*, 9-18.

[***Tullis & Albert, 2010***] Tullis, T., & Albert, W. (2010). *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufmann.

[***Valstar et al, 2011***] Valstar, M. F., Jiang, B., Mehu, M., Pantic, M., & Scherer, K. (2011, March). The first facial expression recognition and analysis challenge. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* (pp. 921-926). IEEE.

[***Valstar et al, 2012***] Valstar, M. F., Mehu, M., Jiang, B., Pantic, M., & Scherer, K. (2012). Meta-analysis of the first facial expression recognition challenge. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, *42*(4), 966-979.

[***Valstar et al, 2015***] Valstar, M., Girard, J., Almaev, T., McKeown, G., Mehu, M., Yin, L., ... & Cohn, J. (2015). Fera 2015-second facial expression recognition and analysis challenge. *Proc. IEEE ICFG*.

[***Van der Werf et al, 2003***] Van der Werf, F., Brassinga, P., Reits, D., Aramideh, M., & de Visser, B. O. (2003). Eyelid movements: behavioral studies of blinking in humans under different stimulus conditions. *Journal of neurophysiology*, *89*(5), 2784-2796.

[***Villani et al, 2006***] Villani, M., Tappert, C., Ngo, G., Simone, J., Fort, H. S., & Cha, S. H. (2006, June). Keystroke biometric recognition studies on long-text input under ideal and application-oriented conditions. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on* (pp. 39-39). IEEE.

[***Viola & Jones, 2001***] Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* (Vol. 1, pp. I-511). IEEE.

[***Volkova et al, 2014***] Volkova, E., de la Rosa, S., Bülthoff, H. H., & Mohler, B. (2014). The MPI Emotional Body Expressions Database for Narrative Scenarios. *PloS one*,*9*(12), e113647.

[***Vrochidis et al, 2011***] Vrochidis, S., Patras, I., & Kompatsiaris, I. (2011, April). An eye-tracking-based approach to facilitate interactive video search. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval* (p. 43). ACM.

[***Wallbot, 1998***] Wallbott, H. G. (1998). Bodily expression of emotion. *European journal of social psychology*, *28*(6), 879-896.

[***Wei, 2009***] Wei, Y. (2009). Research on facial expression recognition and synthesis. *Master Thesis, Department of Computer Science and Technology, Nanjing*.

[***Williams & Stevens, 1972***] Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*,*52*(4B), 1238-1250.

[***Zhang, 2012***] Zhang, Z. (2012). Microsoft kinect sensor and its effect. *MultiMedia, IEEE*,*19*(2), 4-10.

[***Zhu & Yang, 2002***] Zhu, J., & Yang, J. (2002, May). Subpixel eye gaze tracking. In *Automatic face and gesture recognition, 2002. proceedings. fifth ieee international conference on* (pp. 124-129). IEEE.

[***Zimmermann et al, 2006***] Zimmermann, P., Gomez, P., Danuser, B., & Schär, S. (2006). Extending usability: putting affect into the user-experience. *Proceedings of NordiCHI'06*, 27-32.