INSENSION

www.insension.eu

Personalized intelligent platform enabling interaction with digital services to individuals with profound and multiple learning disabilities

TECHNICAL REQUIREMENTS FOR THE NON-SYMBOLIC BEHAVIORAL SIGNALS RECOGNITION COMPONENTS

Deliverable D2.1

Lead authors

Carmen Campomanes-Alvarez[1], B. Rosario Campomanes-Alvarez[1], Maciej Stroiński[2], Marcin Szymański[2], Arkadiusz Radziuk[2], Andrejaana Andova[3], Erik Dovgan[3], Torsten Krämer[4], Meike Engelhardt[4]

[1]CTIC, [2]PSNC, [3]JSI, [4]PHHD

Partners:

PSNC

Pädagogische Hochschule HEIDELBERG University of Education

Jožef Stefan Institute

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780819.

CTIC centro tecnológico

Stowarzyszenie na tak

HARPO

This deliverable has been produced as part of the INSENSION project implemented within the European's Union Horizon 2020 Programme, Industrial Leadership Pillar, Topic ICT-23-2017 Interfaces for Accessibility.

This deliverable is confidential / restricted / **public**.

Nature of this deliverable: **Report** / Prototype / Demonstrator / Other

Consortium partners contributing to this deliverable: CTIC Centro Tecnológico (CTIC), POZNAN SUPERCOMPUTING AND NETWORKING CENTER (PSNC), JOŽEF STEFAN INSTITUTE (JSI) PÄDAGOGISCHE HOCHSCHULE HEIDELBERG (PHHD).

Lead Author(s): Carmen Campomanes-Álvarez (CTIC), B. Rosario Campomanes-Álvarez (CTIC), Maciej Stroiński (PSNC), Marcin Szymański (PSNC), Arkadiusz Radziuk (PSNC), Andrejaana Andova (JSI), Erik Dovgan (JSI), Torsten Krämer (PHHD), Meike Engelhardt (PHHD).

Reviewer(s): Mitja Luštrek (JSI)

Version/revision: 1.11

Delivery date: 31.08.2018

Versioning:

Version	Date	Name, organisation
1.0	10.08.2018	Carmen Campomanes, CTIC
1.01	21.08.2018	Mitja Luštrek, JSI
1.1	29.08.2018	Carmen Campomanes, CTIC
1.11	31.08.2018	Michał Kosiedowski, PSNC

EXECUTIVE SUMMARY

This document presents the technical requirements for the designing and implementing the components for detection of non-symbolic behavior signals of people with profound and multiple learning disabilities. Four components are discussed: for gesture recognition, facial expression recognition, vocalization recognition and physiological parameters monitoring.

Concerning the gesture and facial expression recognition components, the existing methods oriented to people without disabilities have been studied. On the one hand, gesture recognition research to date has focused on movement (the atomic element of motion) and activity (a sequence of movements or static configurations). These systems need a human body model to represent and recognize the motion over the detected human body in images or videos. Normally, these models consist of segments, which represent all the parts of the body. On the other hand, facial expressions are the facial changes in response to a person's internal emotional states, intentions or social communications. From a computer vision point of view, facial expression analysis refers to computer systems that attempt to automatically analyze and recognize facial motions and facial feature changes from images. Facial expression analysis includes both measurement of facial motion and recognition of expression. The general approach to automatic facial expression analysis consists of face acquisition, facial data extraction and representation, and facial expression recognition.

Currently, this kind of systems is based on machine learning algorithms, which need a gesture and facial expression database systematically constructed, in this case, adapted to people with PMLD. Thus, a carefully control environment arises indispensable. To define the requirements for this particular scenario, some control experiments have been performed and test the state-of-the-art techniques in people with PMLD.

To define the requirements of this particular scenario, a control set of experiments have been performed in order to test the behavior of the state-of-the-art methods in people with PMLD. As a result, high requirements of the system are needed in terms of processing memory, as well as proper environment conditions.

With regard to the vocalization recognition module, existing literature was surveyed. A classical Markov Model-based approach was selected for control experiments. The results show the module will likely not present considerable hardware requirements; however, they suggest the accuracy of recognition may strongly depend on recording conditions and the quality of training example annotations.

Regarding physiological parameters monitoring, non-contact devices are considered since they are most comfortable for people with PMLD. The main such method is monitoring of photoplethysmogram (measurement of the amount of blood in tissue) with a camera. This enables the extraction of parameters such as heart rate and its variability, which are indicators of psychological state. We have tested a few such methods with moderate success. They require well-lit images, advisable resolution is 1280 x 720 pixels, and frame rate should be 20–50 frames per second.

This report is structured in five sections. The first one introduces the general aspects of the project's implementation, the second summarizes the state of the art of the four main components and the third one details the methods, materials and requirements for them. The fourth section presents a set of experiments performed to verify each recognition technique. Finally, the conclusions are summarized in the fifth section.

TABLE OF CONTENTS

FIGURES

TABLES

LIST OF ACRONYMS

H2020	-	Horizon 2020 - The Framework Programme for Research and Innovation
EU	-	European Union
ICT	-	Information and Communication Technology
PMLD	-	Profound and Multiple Learning Disabilities
AI	-	Artificial Intelligence
HCI	-	Human Computer Interaction
NN	-	Neural Network
HMM	-	Hidden Markov Model
DTW	-	Dynamic Time Warping
CNN	-	Convolutional Neural Network
AFEA	-	Automatic Facial Expression Analysis
HOG	-	Histogram Oriented Gradient
DCT	-	Discrete Cosine Transformation
PCA	-	Principal Component Analysis
AVEC	-	Audio/Visual Emotion Challenges
LDA	-	Linear Discriminant Analysis
AU	-	Action Unit
FACS	-	Facial Action Coding System
FERA	-	Facial Expression Recognition Analysis
GMM	-	Gaussian Mixture Model
SVM	-	Support Vector Machine
MLP	-	Multi-Layer Perception
STE	-	Sub-band Temporal Envelope
SPD	-	Sub-band Probabilistic Distance
ANN	-	Artificial Neural Network
GTCC	-	Gammatone Cepstral Coefficients
MFCC	-	Mel-Frequency Cepstral Coefficients
PMVDR	-	Perceptual Minimum Variance Distortion less Response
DNN	-	Deep Neural Network
EM	-	Expectation-Maximization

NMF	-	Non-negative Matrix Factorization
BLSTM-RNN	-	Bidirectional Long Short Term Memory-Recurrent Neural Network
LSTM	-	Long Short-Term Memory
KNN	-	k-Nearest Network
HR	-	Heart Rate
HRV	-	Heart Rate Variability
BR	-	Breathing Rate
RGB	-	Red Green Blue
PPG	-	Photoplethysmogram
EDA	-	Electro Dermal Activity
ECG	-	Electrocardiogram
GSR	-	Galvanic Skin Response
ST	-	Skin Temperature
IR	-	InfraRed
Fps	-	Frame per second
ELAN	-	Eudico Linguistic Annotator
TS	-	Time Sampling
ES	-	Event Sampling
BVP	-	Blood Volume Pulse
SDK	-	Software Development Kit
CAFFE	-	Convolutional Architecture for Fast Feature Embedding
CUDA	-	Compute Unified Device Architecture
GPU	-	Graphical Processing Unit
PCM	-	Pulse Code Modulation
AIC	-	Akaike Information Criterion
BIC	-	Bayesian Information Criterion
ToT	-	Test on Train
PXV	-	Pseudo Cross Validates
DSP	-	Digital Signal Processor
PAF	-	Part Affinity Field
ASR	-	Automatic Speech Recognition

MSE - Mean Squared Error

MAE - Mean Absolute Error

1 INTRODUCTION

The main purpose of the INSENSION project is to design and develop an information and communication technology (ICT) platform to help understand and support the needs of people with profound and multiple learning disabilities (PMLD). To do this, it is necessary to process the information collected from them and the world around them and to communicate their needs to others with the use of advanced technologies previously not available to them. Although learning symbolic communication is not generally impossible for people with PMLD, non-symbolic behavior is often used. Consisting of a number of atomic reactions, it varies from gestures and facial expressions to vocalizations and physiological reactions.

Understanding requirements related to the technical construction of such a system demands to carefully consider methods for recognizing these non-symbolic behavioral signals. Therefore, in the course of the work reported in this deliverable we performed a number of activities leading to the definition of initial assumptions on the way the recognition components could be built. These assumptions are based not only on literature review and previous experiences of the project team, but also on the actual verification of the candidate methods with the use of real data collected from the target end users of the INSENSION system. This allowed to identify technical requirements for the signal recognition components on several levels, that is the level of the method to be used as the basis of a given components, the hardware and related requirements, as well as requirements concerning further development of these recognition components.

Recognizing non-symbolic behavioral signals as described in the first paragraph of this section is possible thanks to the use of advanced technologies based on artificial intelligence (AI). Related to this, a great amount of work has been published in recent years. However, recognition of signals in people with PMLD presents special characteristics to take into account that makes this problem into an important challenge. Each individual uses varying methods to communicate their needs, so an individualized system is required. In addition, the technology should work in a relatively noisy environment and be able to recognize the non-symbolic behaviors that occur in natural situations without influencing the normal living of the people with PMLD. Therefore, this scenario demands a specific approach to technological modules responsible for tracking behaviors in such a way that they could seamlessly adapt to the individual end user needs. For that reason, the final goal of this project is to develop specific tools able to recognize communication patterns of people with PMLD through gesture, facial expression, vocalization and physiological parameters recognition.

In this deliverable, a preliminary state-of-the-art study of each recognition component (i.e. gesture, facial expression, vocalization and physiological parameters) has been summarized in Section 2. Based on these studies, the most remarkable methods have been selected – these are explained in Section 3. Then in Section 4, we discuss first experiments of these promising methods performed on data collected from six individuals with PMLD. These experiments were performed in order (1) to understand if building of the planned signal recognition components is possible using adaptations of existing methods, and (2) to draw Finally, we outline conclusions of this work in Section 5, summarizing the foreseen technical requirements.

2 STATE OF THE ART

In this section, the state of the art of the four recognition components is described.

2.1 GESTURE RECOGNITION

Gesture is the use of motions of the limbs or body as a means of expression, to communicate an intention or feeling. The majority of hand gestures produced by speakers are meaningfully connected to speech. These communicative hand movements were defined along a "gesture Kendon's Continuum" as five different kinds of gestures in [1]:

1) Gesticulation: spontaneous movements of the hands and arms that accompany speech.
2) Language-like gestures: gesticulation is integrated into a spoken utterance, replacing a particular spoken word or phrase.
3) Pantomimes: gestures that depict objects or actions, with or without accompanying speech.
4) Emblems: familiar gestures such as "V for victory", "thumbs up", and assorted rude gestures (often culturally specific).
5) Sign languages: Linguistic systems, such as American Sign Language, which are well defined.

Another classification of gestures is depicted in Figure 1 and was defined in [2].

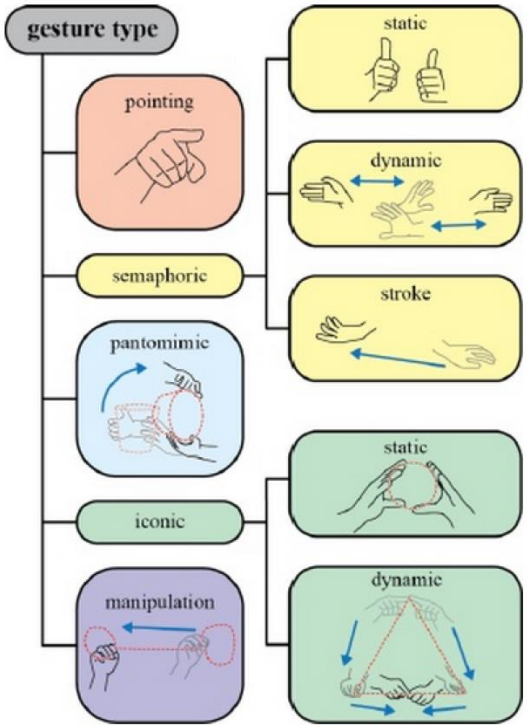

Figure 1. Classification of gestures proposed in *[2]* with examples for each gesture *type*

It is important to note that most of these gestures are not used for communicating by people with PMLD. Nevertheless, they are described as part of a general gesture recognition system, which is the base of this project approach.

Ninety percent of the gestures found in human iterations are unconscious or spontaneous (gesticulation in Kendon's Continuum), which accompanies speech in communicative situations. Despite this importance of spontaneous gesture in normal human-to-human interaction, most research to date in human-computer

interaction (HCI), and most virtual environment technology, focuses on emblems and sign languages, where gestures tend to be less ambiguous, less spontaneous and natural, more learned, and more specific of the culture. The computer science community mostly has attempted to integrate emblematic gestures (e.g. the thumbs up gesture, or putting one's palm out to mean stop), that are employed in the absence of speech, and emotional facial displays (e.g. smiles, frowns, looks of puzzlement) [3].

Some gestures have both static and dynamic elements, where the pose is important in one or more of the gesture phases; this is particularly relevant in sign languages. When gestures are produced continuously, each gesture is affected by the gesture that precedes it, and possibly, by the gesture that follows it. There are several aspects of a gesture, which may be relevant and therefore may need to be represented explicitly in computer vision systems. Four aspects of a gesture were defined in [4], which may be important to its meaning:

(a) Spatial information – where it occurs, locations a gesture refers to;

(b) Pathic information – the path which a gesture takes;

(c) Symbolic information – the sign that a gesture makes;

(d) Affective information – the emotional quality of a gesture.

Because gestures are highly variable, from one person to another and from one example to another with a single person, it is essential to capture the essence of the gesture – its invariant properties – and use this to represent the gesture. Besides the choice of representation itself, a significant issue in building gesture recognition systems is how to create and update the database of known gestures [5]. In general, a system needs to be trained through some kind of learning, there is often a tradeoff between accuracy and generality – the more accuracy is desired, the more unspecific training is required. In addition, systems may be completely trained when in use, or they may adapt over time to the current user.

Static gesture or pose recognition can be accomplished using template matching, geometric feature classification, neural networks (NNs), or other standard pattern recognition techniques to classify pose. Dynamic gesture recognition, however, requires consideration of temporal events. This is typically accomplished by using techniques such as time-compressing templates, dynamic time warping, hidden Markov models (HMMs) and Bayesian networks.

Currently, computer vision systems for recognizing gestures look similar. The components of a gesture recognition system are:

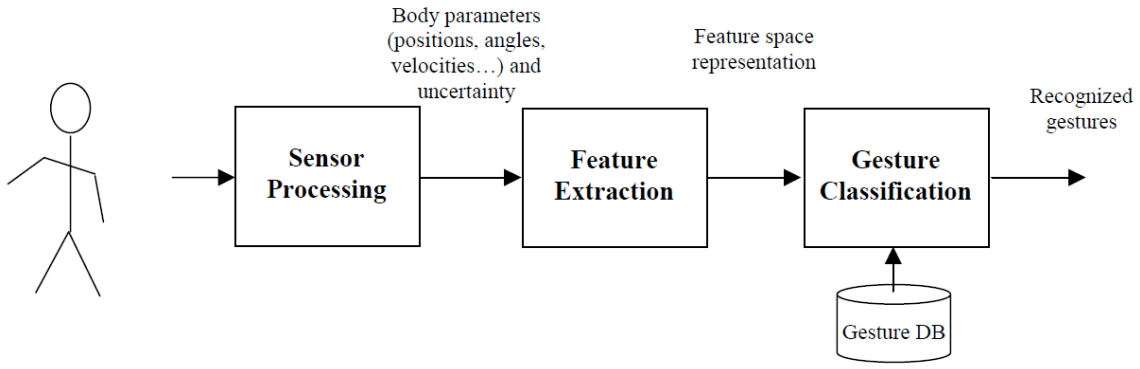

Figure 2. Gesture recognition system stages [6].

Vision-based interfaces use one or more cameras to capture images, at a frame rate of 30Hz or more, and interpret those images to produce visual features that can be used to interpret human activity and recognize gestures. Typically, the camera locations are fixed in the environment, although they may also be mounted on moving platforms or on other people.

Unlike sensors worn on the body, vision approaches to body tracking have to contend with occlusions. From the point of view of a given camera, there are always parts of the user's body that are occluded and therefore not visible – e.g., the backside of the user is not visible when the camera is in front. More significantly, self-occlusion often prevents a full view of the fingers, hands, arms, and body from a single view. Multiple cameras can be used, but this adds correspondence and integration problems. The occlusion problem makes full body tracking difficult, if not impossible, without a strong model of body kinematics and perhaps dynamics.

Vision-based systems for gesture recognition vary along a number of dimensions, most notably [6]:

- Number of cameras – More than one, combining different types, like early (stereo) or late (multi-view).
- Speed and latency – Considering if the system is real-time (i.e., fast enough, with low enough latency, to support interaction) or not.
- Structured environment – Restrictions on the background, the lighting, the speed of movement, etc.
- User requirements – If the user wear anything special (e.g., markers, gloves, long sleeves) or anything disallowed (e.g., glasses, beard, rings), etc.
- Primary features – The kind of low-level features computed (edges, regions, silhouettes, moments, histograms).
- Two- or three-dimensional representation.
- Representation of time – The temporal aspect of gesture represented and used in recognition (e.g., via a state machine, dynamic time warping (DTW), HMMs, time-compressed template).

The gesture recognition approaches can be classified into three major categories: (a) model based, (b) appearance based and (c) motion based. Model based approaches focus on recovering three-dimensional model parameters of articulated body parts. Appearance based approaches use two–dimensional information such as gray scale images or body silhouettes and edges. Motion based approaches attempt to recognize the gesture directly from the motion without any structural information about the physical body. In all these approaches, the temporal properties of the gesture are typically handled using DTW or statistically using HMMs.

Static gesture or pose recognition can be accomplished by a straightforward implementation of using template matching, geometric feature classification, neural networks, or other standard pattern recognition techniques such as parametric Eigen space to classify pose. Dynamic gesture recognition requires consideration of temporal events, typically accomplished by using techniques such as time – compressing templates, DTW, HMMs, and Bayesian networks.

Body gestures recognition includes tracking full or partial body motion (e.g. movement of waist or chest, shoulder shrug etc.), recognizing body gestures (postural shifts, angular distance, upright position with ankles locked etc.), and recognizing human activity. Activity may be defined over a much longer period than what is normally considered a gesture; for example, two people meeting in an open area, stopping to talk and then continuing on their way may be considered a recognizable activity.

A taxonomy of motion understanding was proposed in [7], in terms of:

- Movement – the atomic elements of motion.

- Activity – a sequence of movements or static configurations.
- Action – high-level description of what is happening in context.

Most research to date has focused on the first two levels.

In order to observe or analyze the characteristics of human motion and verify or evaluate the developed algorithm and its application system, a gesture database systematically constructed in carefully controlled environment is indispensable. In [5], a full-body gesture database was collected with 2D video data and 3D motion data of 14 normal gestures, 10 abnormal gestures and 30 command gestures for 20 subjects.

In addition, these systems also need a 3D human body model to represent and recognize the motion over the detected human body. Normally, these models consist of segments, which represent all the parts of the body. In order to establish all the possible movements of the human body, a hierarchical system is created, which indicates the base position of each segment and the 3D translation and rotation difference between adjacent frames at each segment. This model automatically extracts the skeleton to classify seven human actions: standing, walking, running, jumping, falling, lying and sitting [8].

Systems that analyze human motion in virtual environments may be quite useful in medical rehabilitation and athletic training. For example, Extreme Motion Studio is an application that provides full body motion detection from a 2D webcam. The tool permits create care plans, insurance requirements, stroke rehabilitations, etc.
Recent studies use deep learning for detecting pose of human bodies in 2D images. In particular, a real time algorithm to efficiently detect the 2D pose of multiple people in images was presented in [9] based on the previous pose detection algorithms proposed in [10]. The approach uses a nonparametric representation to learn to associate body parts with individuals in the image. The architecture encodes global context, allowing a greedy bottom-up parsing step that maintains high accuracy while achieving real time performance, irrespective of the number of people in the image. The architecture is designed to jointly learn part locations and their association via two branches of the same sequential prediction process. Same approach was followed in [11] for detecting hand motion in images. The result of these studies is the OpenPose system, which is able to efficiently detect 18 key points of human bodies, 2x21 key points of hands and 70 key points of faces in 2D videos. In the same line, DeepEyes algorithm [12] is a deep learning methodology for different uses in cases of video recognition, like automotive industry, banks, gaming industry, health, smart home, marketing, security, etc.

Analysis, recognition and synthesis of natural gestures are still an ongoing research. An extensive evaluation of convolutional neural networks (CNNs) on general video classification is provided by [13] using the Sports-1M dataset. They compare different frame fusion methods to a baseline single-frame architecture and conclude that their best fusion strategy only modestly improves the accuracy of the baseline. In [14], Neverova et al. present an extended overview of their winning solution for the ChaLearn LAP 2014 gesture recognition challenge and achieve a state-of-the-art score on the Montalbano dataset. They propose a multi-modal 'ModDrop' network operating at three temporal scales and use an ensemble method to merge the features at different scales. They also developed a new training strategy, ModDrop, which makes the network's predictions robust to missing or corrupted channels. Temporal ordering models have also been applied in the context of complex activity recognition [15]. They mainly focus on inferring composite activities from predefined, semantically meaningful, basic-level action detectors. A representation for events is presented that encodes statistical information of the atomic action transition probabilities using a HMM.

Achieving accurate, efficient and real-time systems is an extremely complex task. The latest works on gesture recognition can be found in the IEEE Face and Gesture Recognition Conference held every two years.

2.2 FACIAL EXPRESSION RECOGNITION

Facial expressions are the facial changes in response to a person's internal emotional states, intentions or social communications. From a computer vision point of view, facial expression analysis refers to computer systems that attempt to automatically analyze and recognize facial motions and facial feature changes from images.

The accomplishments in the related areas such as psychological studies, human movement analysis, face detection, face tracking, and recognition make the automatic facial expression analysis possible. Automatic facial expression analysis can be applied in many areas such as emotion and paralinguistic communication, clinical psychology, psychiatry, neurology, pain assessment, lie detection, intelligent environments, and multimodal HCI.

Facial expression analysis includes both measurement of facial motion and recognition of expression. The general approach to automatic facial expression analysis (AFEA) consists of three steps: face acquisition, facial data extraction and representation, and facial expression recognition.

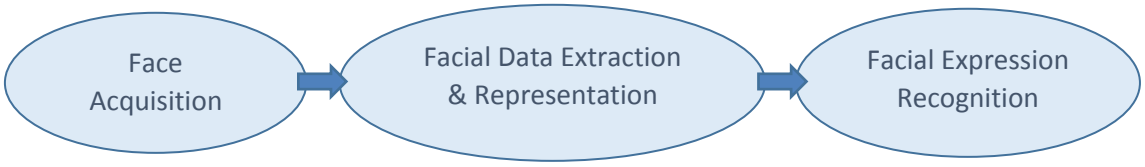

Figure 3. Facial expression recognition stages.

The acquisition of faces and facial features from an arbitrary uncontrived image is a critical precursor to recognition. A robust scheme is needed to detect the face as well as determine its precise placement to extract the relevant data from an input image. This is necessary to properly prepare the image's 2D intensity description of the face for input to a recognition system. This detection scheme must operate flexibly and reliably regardless of lighting conditions, background clutter in the image, multiple faces in the image, as well as variations in face position, scale, pose and expression. The geometrical information about each face in the image that we gather at this stage will be used to apply geometrical transformations that will map the data in the image into an invariant form. By isolating each face, transforming it into a standard frontal mug shot pose and correcting lighting effects, it should be limited the variance in its intensity image description to the true physical shape and texture of the face itself. The boosted cascade with simple features proposed in [16] becomes the most popular and effective design for practical face detection. The simple nature of the features enable fast evaluation and quick early rejection of false positive detections. Meanwhile, the boosted cascade constructs an ensemble of the simple features to achieve accurate face vs. non-face classification. The original Viola-Jones [16] face detector uses the Haar feature, which is fast to evaluate yet discriminative enough for frontal faces. A number of improvements to the Viola-Jones face detector have been proposed in the past decade. Most of them follow the boosted cascade framework with more advanced features.
However, much more reliable solutions exist nowadays based also on a method invented in 2005 called Histogram of Oriented Gradients (HOG) [17].

After that, the next step is to extract and represent the facial changes caused by facial expressions. In facial feature extraction for expression analysis, there are mainly two types of approaches: geometric feature-based methods and appearance-based methods. The geometric facial features present the shape and locations of facial components (including mouth, eyes, brows, nose, etc.). The facial components or facial feature points are extracted to form a feature vector that represents the face geometry. When the single facial features are hardly resolved in detail, implies that the overall geometrical configuration of the face features is sufficient for discriminations. The overall configuration can be described by a vector or

numerical data representing the position and size of the main facial features: eyes and eyebrows, nose and mouth. This information can be supplemented by the shape of the face outline. Face recognition, although difficult, presents a set of interesting constraints that can be exploited in the recovery of facial features. The first important constraint is bilateral symmetry. Another set of constraints derives from the fact that almost every face has two eyes, one nose, and one mouth with a very similar layout. Although this may make the task of face classification more difficult, it can ease the task of feature extraction.

With appearance-based methods, image filters are applied to either the whole-face or the specific regions in a face image to extract a feature vector. These feature extraction methods extract novel features (e.g. holistic features) from the initial representations. They map an input representation onto a lower dimensional space to discover a latent structure from the representation. This transformation can be non-adaptive or adaptive (learnt from training data). The most popular non-adaptive transformation is the discrete cosine transformation (DCT) whereas the most popular adaptive transformation is principal component analysis (PCA). PCA computes a linear transformation that aims at extracting decorrelated features out of possibly correlated features. Under controlled head-pose and imaging conditions, these features capture the statistical structure of expressions efficiently PCA is used by many systems including the winner of the Audio/Visual Emotion Challenges (AVEC) continuous challenge that consisted of evaluating dimensional affect models.
A supervised alternative to the unsupervised PCA is linear discriminant analysis (LDA). LDA uses label information to learn how to discriminate between differently labelled representations, and group similarly labelled representations. LDA can handle more than two classes as it considers only whether two arbitrary samples have the same or different labels. Most affect recognition systems train LDA using multiple classes simultaneously [18].
Alternative training schemes are also proposed. Kyperountas et al. [19] proposed a scheme where multiple LDA models are involved, and each model discriminates between a pair of classes.

Depending on the different facial feature extraction methods, the effects of in-plane head rotation and different scales of the faces can be removed by face normalization before the feature extraction or by feature representation before the step of expression recognition.

Facial expression recognition is the last stage of AFEA systems. The facial changes can be identified as facial action units (AUs) or prototypic emotional expressions. Studies of automatic facial expression recognition have made a significant progress in the last two decades due to the advances in machine learning and computer vision techniques. The current research can be classified in two types: the recognition of the appearance of facial actions and the recognition of the emotions conveyed by the actions.

The first set of systems usually relies on the facial action coding system (FACS) [20]. FACS consists of 44 facial AUs, which are codes that describe certain facial configurations. Thirty AUs are anatomically related to contraction of a specific set of facial muscles and they are shown in Figure 4.

Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
					
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
					
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink
Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
					
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
					
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
					
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

Figure 4. FACS action units (AU)

The anatomic basis of the remaining 14 is unspecified and they are referred to as miscellaneous actions in FACS (Figure 5).

AU	Description
8	Lips toward
19	Tongue show
21	Neck tighten
29	Jaw thrust
30	Jaw sideways
31	Jaw clench
32	Bite lip
33	Blow
34	Puff
35	Cheek suck
36	Tongue bulge
37	Lip wipe
38	Nostril dilate
39	Nostril compress

Figure 5. Miscellaneous action units

Facial expression can be defined as the combination of these AUs. Figure 6 shows some examples.

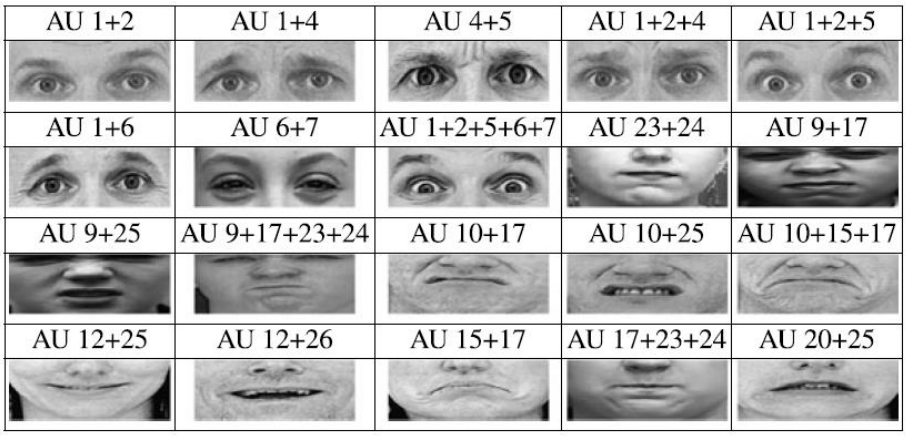

Figure 6. Some examples of combination of FACS action units

The production of a facial action has a temporal evolution, which plays an important role by interpreting emotional displays. The temporal evolution of an expression is typically modeled with four temporal segments: neutral, onset, apex and offset. Neutral is the expressionless phase with no signs of muscular activity. Onset denotes the period during which muscular contraction begins and increases in intensity. Apex is a plateau where the intensity usually reaches a stable level, whereas offset is the phase of muscular action relaxation.

AUs and temporal segments are properly analyzed in psychology and their recognition enables the analysis of sophisticated emotional states such as pain [21] and helps distinguishing between spontaneous and posed behavior [22]. By this way, it is possible to define new different gestures or expressions made by people with PMLD. For that reason, the state of the art of the emotion recognition classifiers is explained below. The systems that recognize emotions consider basic or non-basic emotions. Basic emotions refer to the affect model developed by Ekman et al. [20], who argued that the production and interpretation of certain expressions are hard-wired in our brain and are recognized universally. The emotions conveyed by these expressions are modeled with six classes: joy, sadness, surprise, fear, anger and disgust. Basic emotions are believed to be limited in their ability to represent the broad range of everyday emotions. More recently, researchers considered non-basic emotion recognition using a variety of alternatives for modelling non-basic emotions. One approach is to define a limited set of emotion classes (e.g. relief, contempt). Another approach, which represents a wider range of emotions, is continuous modelling using affect dimensions, like arousal, valence, power and expectation. The above-listed affect models were evaluated in a number of affect recognition competitions. The facial expression recognition analysis (FERA) [23] focus on evaluating posed expressions, near-frontal recordings, or both. This makes it hard to tell how existing expression recognition approaches perform under conditions where faces appear in a wide range of poses (or camera views), displaying ecologically valid expressions. The main obstacle for assessing this is the availability of suitable data, and the challenge proposed here addresses this limitation. The AVEC [24] aimed at comparison of multimedia processing and machine learning methods for automatic audiovisual depression and emotion analysis.

Except from a small number of unsupervised knowledge-driven approaches, all affect recognizers use machine learning techniques. As any machine learning application, the performance of an affect recognition system depends on the quality and quantity of training data as well as the selected machine learning model.

In the case of supervised machine learning systems, the training data should be properly labelled. Labelling data is a challenging and laborious task, particularly for spontaneously displayed expressions and emotions. The annotation of spontaneously displayed emotions is challenging mainly due to the subjective perception of emotions, which is often addressed by using multiple annotators. Spontaneous AUs require frame-by-frame annotation by experts, and unlike posed AUs, where the subjects are instructed to display a

particular (usually single) AU, the annotator has to deal with an unknown facial action which may be a combination of AUs, as occurring in this project.

Concerning the output of this kind of systems, affect recognition approaches give a label of an emotion or facial action. In addition, recent studies provide also the intensity of the displayed emotion or facial action. Meanwhile, for automatic unit recognition, the output can be enhanced significantly by providing the temporal phase of the displayed AU. Besides, several studies recognize combinations of AUs rather than individual AUs in order to render the output more suitable to spontaneous behavior [25].

The main challenges in automatic facial actions recognition are head-pose variations, illumination variations, registration errors, occlusions and identity bias. Spontaneous affective behavior often involves head-pose variations, which need to be modeled before measuring facial expressions. Illumination variations can be problematic even under constant illumination due to head movements. Registration techniques usually yield registration errors, which must be dealt with to ensure the relevance of the representation features. Occlusions may occur due to head or camera movement, or accessories such as scarves or sunglasses. Dealing with identity bias requires the ability to tell identity-related texture and shape cues apart from expression-related cues for subject independent affect recognition. While being resilient to these challenges, the features of a representation shall also enable the detection of subtle expressions [25].

The most recent facial expression recognition systems can be categorized in terms of basic emotion, AU or non-basic emotion recognition systems.

Basic emotion recognition has mostly been analysed on posed data, and systems have been evaluated using the average recognition rate or average area under the curve metrics. Although the recognition of posed basic emotions is considered as a solved problem, it is still used for proof of concept of spatial [26] and spatio-temporal representations [27], [28] as well as novel statistical models [29].

AU recognition has been studied both for posed and spontaneous data. The problem is typically formulated as a detection problem and approached by training a two-class (positive vs. negative) statistical model for each AU. In this setting, results are reported using metrics such as Area Under the Curve, F1-measure or 2AFC score [30], [31]. Two well-studied non-basic emotion recognition problems are dimensional affect recognition and pain recognition. In [32], where affect recognition has been performed in terms of quantised affect dimensions, performance has been measured as average recognition rate on four affect dimensions, whereas [33] and [34] considered continuous affect recognition and evaluated performance using the Pearson's correlation— [34] considered also the recognition of depression and evaluated performance using the mean absolute error and the root mean square error.

2.3 VOCALIZATION RECOGNITION

Vocalization recognition aims at detecting instances of non-linguistic sounds produced vocally by a person under surveillance. Such sounds include laughing, wailing, heavy breathing etc. Similarly to other media of expression (i.e. facial, gestural), not all of these sounds have to correspond to interpretable messages communicated by a particular person with PMLD (however, it is not the task of the vocalization module to decide which of the event types are meaningful).

A classical approach to the recognition problem would be to (1) perform signal parametrization for uniformly spread short-time audio frames, usually by strict mathematical transformations (e.g. Cepstral Coefficients every 10ms for overlapping windows of 25ms), (2) then feed such streams of data into a chosen type of a statistical process-modeling framework (e.g. based on Hidden Markov Models).

A resurgence of neural network models, observed over last 6-7 years in speech and image recognition in particular, is leaving an imprint on vocalization and sound event recognition as well. Some of those most recent papers represent what can be seen as hybrid approaches, ie. fusing connectivist feature extraction with more 'classical' detection frameworks, while some present holistic end-to-end solutions.

The methods are quite diverse. The current project will be attempting to propose a hybrid solution, that will combine a modern machine learning approach with a classical statistical framework.

2.3.1 SIGNAL PROCESSING AND STATIC FEATURES

The following papers provide important background on acoustic signal processing and modelling of short-time audio features:

In [35], the authors investigate the applicability of various subsets of audio features, aiming at automatic detection of non-linguistic sounds from vocalizations. Subsets of features were formed based on ranking the relevance and the individual quality of several audio features. During the audio parameterization process, every input utterance is converted to a single feature vector. Next, a subset of this feature vector is fed to a classification model, which aims at estimation of the unknown sound class.

The development of a gender-independent laugh detector with the aim to enable automatic emotion recognition is described in [36]. Different types of features (spectral, prosodic) for laughter detection were investigated using different classification techniques such as Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs) and Multi Layer Perceptron (MLP), often used in language and speaker recognition. Acoustic measurements showed differences between laughter and speech in mean pitch and in the ratio of the durations of unvoiced to voiced portions, which indicate that these prosodic features are indeed useful for discrimination between laughter and speech.

In [37], the authors propose a classification method of ten types of short-time sound events based on probabilistic distance SVMs. Parametric approach of characterizing sound signals using the distribution of the sub-band temporal envelope (STE), and kernel techniques for the Sub-band Probabilistic Distance (SPD) under the framework of SVM are studied. It is shown that generalized gamma modeling is well devised for sound characterization, and that the probabilistic distance kernel provides a closed form solution to the rapid calculation of divergence distance.

The issue of cough detection using only audio recordings is addressed in [38], with the ultimate goal of quantifying and qualifying the degree of pathology for patients suffering from respiratory diseases, notably mucoviscidosis. A large set of audio features describing various aspects of the audio signal is proposed. These features are assessed in two steps. First, their intrinsic potential and redundancy are evaluated using mutual information-based measures. Secondly, their efficiency is confirmed relying on three classifiers: artificial neural network (ANN), GMM and SVM. The influence of both the feature dimension and the classifier complexity are also investigated.

In [39], gammatone cepstral coefficients (GTCC) are investigated for cough recognition. The accuracy of GTCC comparing with mel-frequency cepstral coefficients (MFCC) is evaluated on a designed cough dataset following a 10 fold cross-validation scheme. Considering the imbalance of that dataset, weighted SVM is applied as the base classifier. The results indicate that GTCC surpass MFCC in modeling cough signals. With combination of GTCC and MFCC, a better performance is achieved. This paper provides a better feature representation prototype in cough recognition.

The work described in [40] (see below) compares two alternative acoustic front-end features: MFCC and perceptual minimum variance distortion less response (PMVDR).

The work presented in [41] describes algorithm development in support of a pilot study of uberculosis patient coughing. Several signal-processing approaches to event detection and classification are presented, with eventual goal of development of a low-cost ambulatory cough analysis system.

In [42] (see below), deep neural networks (DNNs) are applied to model acoustic features in cough detection.

2.3.2 EVENT MODELING AND DETECTION

The following references were found to focus on acoustic event modeling and detection, in particular methods using HMMs.

The proposal in [43] provides an examination of two methods for optimization of HMM configurations for better classification and recognition of nonverbal vocalizations within speech, like filled pauses, laughter, breathing, hesitation, etc. An in-depth analysis of the discussed methods is provided.

The authors in [44] present an HMM-based automatic system for recognition of bird species from audio field recordings. It includes an improved unsupervised modelling of individual bird syllables and duration modelling. The acoustic signal is decomposed into isolated segments, each segment containing a temporal evolution of a detected sinusoidal component. Modeling of bird syllables is performed using HMMs. A set of syllables of bird vocalizations is discovered in an unsupervised manner by employing dynamic time warping and agglomerative hierarchical clustering. A novel iterative maximum likelihood procedure is used to train individual HMMs for syllables of each species. Modelling of the state duration is employed in a post-recognition stage by combining the likelihood of the acoustic and duration modeling. Evaluations demonstrate that the use of the proposed un-supervised iterative HMM training procedure and the duration modelling allow to build a system that recognizes bird species with high accuracy.

Another research [45] proposes the use of HMMs to automatically detect cough sounds from continuous ambulatory recordings. The recognition algorithm follows a keyword-spotting approach, with cough sounds representing the keywords. The results suggest that HMMs can be applied to the detection of cough sounds from ambulatory patients.

The research depicted in [42] also proposes using HMMs as a second stage of a two-step cough detection system (with DNNs applied to model acoustic features).

An important requirement that unsupervised approaches based on HMMs must satisfy is the capability of automatically determining a number of distinct states. This can be achieved, for example, by running GMM-based clustering over all audio data and selecting the number of states based on the number of components of the GMM. An unsupervised algorithm for learning a finite mixture model from multivariate data is proposed in [46]. In addition, the presented method, unlike the standard expectation maximization (EM) algorithm, does not require careful initialization of the parameters.

The work performed in [47] describes an effective process for automated detection and classification of frequency-modulated sounds from birds, crickets, and frogs that have a narrow short-time frequency bandwidth. An algorithm is provided for extracting these signals from background noise using a frequency band threshold filter on spectrograms. Feature vectors are introduced and demonstrated to accurately model the resultant bioacoustics signal with HMMs. Additionally, sequences of sounds are successfully modeled with composite HHMs, allowing for a wider range of automated species recognition.

An unsupervised approach [40] is focused on detection of human scream vocalizations from continuous recordings in noisy acoustic environments. The proposed solution is based on signal segmentation, which

employs i.a. Bayesian Information Criteria and mean distances. The performance of proposed system was compared using two alternative acoustic front-end features (i) MFCC and (ii) PMVDR.

In [48], the authors describe an audio-based video surveillance system, which automatically detects anomalous audio events in a public square, such as screams or gunshots, and localizes the position of the acoustic source (so that a video camera can be steered consequently). The system employs two parallel GMM classifiers for discriminating screams from noise and gunshots from noise, respectively. Each classifier is trained using different features, chosen from a set of both conventional and innovative audio features. The location of the acoustic source, which has produced the sound event, is estimated by computing the time difference of arrivals of the signal at a microphone array and using linear-correction least square localization algorithm.

A similar problem and solutions are presented in [49].

The proposal in [50] also deal with audio events detection in noisy environments for a multimedia surveillance application, particularly sounds produced by gunshots. A novel detection approach is proposed, which offers a solution to detect abnormality in continuous audio recordings of public places, with focus on the robustness of the detection against variable and adverse conditions and the reduction of the false rejection rate.

2.3.3 NEURAL NETWORKS

While some overlap with previous subsections may occur, papers that use modern machine learning techniques are grouped here.

In [51], the author introduce a novel tandem approach by integrating likelihood features derived from non-negative matrix factorization (NMF) into Bidirectional Long Short Term Memory-Recurrent Neural Networks (BLSTM-RNNs) in order to dynamically localize non-linguistic events, i.e., laughter, vocal, and non-vocal noise, in highly spontaneous speech. This tandem architecture is compared to a baseline conventional phoneme-HMM-based speech recognizer.

A study on recognition of animal sounds, comparing dynamic and static classification by left-right and cyclic HMMs is presented in [52], RNNs with long short term memory (LSTM), and SVMs, as well as different features commonly found in sound classification and speech recognition.

A NN-based method to automatically identify cough segments is proposed in [53]. Discarding other sounds such a speech or ambient noise, developed for a real-time cough identification technique in continuous cough monitoring systems.

In the aforementioned [42], DNNs are applied to model acoustic features in cough detection. A two-step cough detection system is proposed based on DNNs and HMMs. Different configurations of DNNs are evaluated. Experimental results show that many of the DNN configurations outperform GMM.

In [54], the authors present two different approaches of using DNNs for cough detection. A CNN and a RNN are implemented to address these problems, respectively. The effect of the network size parameters and the impact of long-term signal dependencies in cough classifier performance are also explored. Experimental results showed network architectures outperform traditional methods.

2.3.4 OTHER METHODS

Articles that do not clearly fit any previous category are listed below.

A method to detect oral snoring by extracting the acoustic properties of snoring sounds and using the k-nearest network (KNN) classifier is described in [55].

The research in [56] investigates the impact of narrow-band standard speech coders on the machine-based classification of paralinguistic cues (affective vocalizations) in clinical vocal recordings. In addition, it analyzes the effect of speech low-pass filtering by a set of different cut-off frequencies.

In [57], several pitch and formant measures are used, as well as a multidimensional GMM discriminator to perform classification of utterances as approval, attentional bids, or prohibition. It finds that timbre or cepstral coefficients, as well as changes in pitch, are an important cue for affective messages.

2.4 PHYSIOLOGICAL PARAMETERS MONITORING

A number of studies exist that use physiological parameters to determine the psychological state of an individual. Because our target group are subjects that are not often capable of symbolic communication, determining their psychological state and their needs is important and would help both the subjects and their caregivers. We will attempt this using their physiological parameters such as heart rate (HR), heart rate variability (HRV), breathing rate (BR), etc. The following paragraphs describe the state-of-the-art approaches for all the steps that are required for determining the psychological state from physiological data. These steps are a) the reconstruction of physiological signals from sensor data, b) the calculation of physiological parameters from the physiological signals if this is not performed by the sensor device such as in the case of cameras and microphones, and finally c) the determination of the psychological state from the physiological parameters.

2.4.1 RECONSTRUCTION OF PHYSIOLOGICAL SIGNALS

We will first try to obtain the physiological signals that will be later used as a basis to determine the psychological state of the subjects. Many methods exist that return the physiological signals without any contact with the subject. Since camera, thermal camera and microphone are already used in the project for gesture recognition, facial expression recognition and vocalization recognition, we decided to focus on methods that use these sensors to retrieve the physiological signals. Examples of methods that use data from cameras and microphones are described below. Nevertheless, there also exist approaches that use other noncontact sensors for retrieving physiological signals. For example, heartbeats of the subjects can be detected using a Doppler radar (a specialized device that uses the Doppler effect to measure the velocity of objects at a distance) [58] [59] [60], HR and respiration rate can be measured using Kinect [61], etc. These approaches will not be used since such technologies are not mature enough to justify the usage of these additional sensors.

It should be noted that the non-contact methods for retrieving the physiological signals using cameras and microphones are also new and still in development. Consequently, their results might not be satisfying enough. For that reason, we will also use contact sensors (i.e., wristbands) to collect physiological data, which is a reliable data collection approach. The collected data will be used for the evaluation of noncontact approaches. In addition, if the noncontact approaches will not produce satisfactory results, the contact approaches (including the contact sensors) will also be included in the final project solution. On the other hand, if non-contact approaches will produce good results, the contact approaches and sensors will not be needed.

The following subsections describe the state-of-the-art approaches for collecting physiological data using sensors that are relevant for INSENSION. More precisely, they describe approaches for contact (wristbands) and noncontact -red green blue (RGB) cameras, thermal cameras and microphones- sensors.

2.4.1.1 WRISTBANDS

The most convenient method for obtaining the physiological signals in a non-clinical environment is using a wristband. There are many wristbands for measuring physiological signals, such as Empatica E4, Microsoft Band, Apple Watch, cheap AliExpress wristbands, etc. We will be using Empatica E4, since this is the most reliable wristband. Empatica has a light source to measure the variation in blood volume inside the tissue resulting in the photoplethysmogram (PPG), an infrared thermophile for measuring the skin temperature and an electro dermal activity (EDA) sensor to measure the constantly fluctuating changes in certain electrical properties of the skin.

Although wristbands return reliable results, subjects need to wear them, which might cause increased stress. It is therefore preferable to use non-contact sensors for measuring the physiological signals.

2.4.1.2 RGB CAMERAS

Two main approaches are used in the camera-based methods. The first approach uses the same physiological phenomena as the wristbands do, i.e., it analyzes the variation in blood volume in the skin tissue to retrieve the PPG signal. As a light source, this approach uses the sunlight, which is less intense than the light source of the wristband. Consequently, this approach is very sensitive to different environment conditions. For example, if the environment is not properly lighted it might produce inconsistent results. To detect variations in blood volume, this approach analyzes the changes in RGB intensity of the skin pixels between two sequential video frames. The most basic way of retrieving the PPG signal is by calculating the average of the red, green and blue intensity of all the skin pixels over time [62]. Another approach consists of independent component analysis on the RGB data as color signals and use the most PPG-like resulting signal [63] [64]. Instead of independent component analysis, [65] uses principal component analysis to retrieve the PPG signal. Other approaches do not calculate the average of all skin pixels, but treat the skin pixels independently. For example, [66] tracks the variation of color in each skin pixel independently and afterwards chooses the most PPG-like signal. The method in [67] tracks the vibration of the color space of the skin pixels to retrieve the PPG signal. In [68] the basic color signals are used by a NN to reconstruct various physiological signals. The method in [69] amplifies all the color changes in the facial pixels to follow the blood flow in the head. Although the listed methods seem promising, an independent evaluation on a publicly available dataset showed that they are not precise enough to be used in real-world scenarios [70]. More precisely, this evaluation included three state-of-the-art methods for retrieving PPG from RGB camera video and the results show that there is low correlation between the reconstructed and true PPG.

The second approach analyzes the small head movements that are induced by the pumping of blood into the head [71]. However, it should be noted that these small movements are very subtle and might not be recognized with a low quality camera.

2.4.1.3 THERMAL CAMERAS

Thermal camera recordings are used in many studies for analyzing stress, BR, HR, etc. These studies show that we can see small changes of the temperature each time the heart pumps blood into the major blood vessels. The PPG signal can be retrieved by analyzing these temperature changes [72]. In addition, by analyzing the hot air exhalation of the subjects we can measure their breathing rate [73].

Stress experienced by subjects can be determined by analyzing the temperature of various parts of the face. It has been shown that the nose, the corrugator (the region between the eyebrows) and the forehead respond to emotional and distressing stimuli [74]. Thermal imaging has also been used to classify cognitive workload of the subjects [75] [76].

2.4.1.4 MICROPHONES

Microphones will be used in INSENSION for vocalization recognition. Therefore, it would be convenient if they could be simultaneously used also for retrieving the physiological signals. However, this is a very challenging task and almost no research has been done in this direction. Nevertheless, a method for extracting the heart rate from vowel speech signals has been developed, which analyses the spectrogram of the /i:/ vowel [77]. However, this method cannot measure physiological signals when the subjects do not speak, or if the subjects have issues with symbolic communication as in the case of the INSENSION target group.

2.4.2 CALCULATION OF PHYSIOLOGICAL PARAMETERS

Most of the previously described methods retrieve the PPG signal that can used to calculate various physiological parameters. If the signal is clear enough, we can calculate the HR as well as the HRV by detecting the peaks in the signal [78]. There are also methods that estimate the breathing rate from the PPG signal [79] [80], as well as the blood pressure [81] [82].

2.4.3 DETERMINATION OF THE PSYCHOLOGICAL STATE

The psychological state has been determined based on the physiological parameters in several studies. For example, the PPG signal can be used to determine the subject's stress level [83] [84] [85]. However, these studies were done on stressful situations that were artificially created in laboratory conditions. In real-life, the subjects will probably not have such strong physiological responses to the stressful situations.

Mental engagement was determined from electrocardiogram (ECG), BR, galvanic skin response (GSR), and skin temperature (ST) [86]. To this end, PCA and NNs were applied. The developed system was evaluated with various virtual-reality tasks that were performed by the subjects.

Several studies focused on recognizing human emotions from physiological parameters [87]. During the evaluation, the emotions were typically induced, for example, by showing appropriate movies. Examples of induced emotions were happiness, fear, joy, anger, amusement and frustration, where each study focused only on a specific subset of them. Examples of used physiological parameters are blood volume pressure, GSR, ST, HR, and respiration rate [88]; GSR, ST, skin potential, skin resistance, skin blood flow, and instantaneous respiratory frequency [89]; GSR, HR, RR interval, and ECG [90]; and GSR, ECG, respiration rate and electromyography [91].

3 METHODS, MATERIAL AND THEIR REQUIREMENTS

3.1 COLLECTED MATERIAL

3.1.1 RECORDINGS

3.1.1.1 TECHNICAL ASPECTS OF THE RECORDINGS

To provide materials for further analysis and experiments, dedicated recording sessions for each of the six test persons have been organized in the following facilities:

- "Orzeszek" kindergarten
  - Person A – male, 8 years, characteristics: Struge-Wender syndrome
  - Person F – male, 9 years, characteristics: cerebral palsy, epilepsy
- "Zakątek" school
  - Person B – male, 11 years, characteristics: cerebral palsy, hypotonic form, quadriplegia, erectile pattern in the lower lombs
  - Person C – male, 14 years , characteristics: post-inflammatory hydrocephalus and implanted placental system after purulent meningitis with etiology of e-coli in the non-infant period
- "Kamyk" day care center for adults
  - Person D – female, 41 years, characteristics: significant intellectual disability, able to walk
  - Person E – female, 30 years, characteristics: cerebral palsy, hydrocephalus, epilepsy

The recordings were taken in two phases, in the natural environment of those facilities, during person's daily activities, e.g. meals and individual activities. These recording situations were structured in a specific way to trigger both communication behavior of different content (demand, comment, protest) and different inner states (pleasure, displeasure), which is described in more detail in the deliverable D1.1. During the first phase, a basic setup described below has been used. During the second phase only 4K cameras have been used. The second phase focused on catching situations that were not recorded during the first stage.

Recording setup

The basic recording setup consists of seven cameras and two or three microphones, depending on specific needs, and includes:

- Three 4K cameras: Panasonic HC-x1000.
- Two 3D Full HD cameras: JVC GY-HMZ1ED and Sony HDR-TD-10.
- Two IR cameras: Flir T420 and Flir T620.
- One Asden SGM-2X microphone.
- Handy recorder ZOOM H6 with stereo microphone MSH-6 and shotgun microphone Sennheiser ME67.

The goal of the recordings was to gather video streams from two perspectives, according to Figure 7. In the first perspective, the cameras are placed in front of the recorded person. The second perspective was a lateral one (left or right) in which cameras record the person from the side. Additionally, one camera has been placed in the corner providing a 45-degree angle view on the scene. All cameras have been placed as close to the walls as possible to check whether walls would be good mounting points for the target infrastructure in terms of the distance and the ability to recognize non-symbolic signal.

Each perspective has been recorded by three different types of cameras placed side by side. Such a set consisted of a 4K camera, 3D Full HD camera and an infrared (IR) camera. The aim was to provide different types of materials which may be usable for observation and detection of different behaviors, and to check what kind of recording is most useful in the context of the project's main goal. The idea of using IR and 3D cameras was to check whether they could provide additional useful information, unavailable on a standard RGB camera stream – through a thermographic stream and 3D spatial movement, respectively. All cameras recorded the whole scene with a person with PMLD and their caregiver, as well as their surroundings to observer both the context and interactions.

The directional microphones were usually placed in front of the recorded person (two microphones) and the side of the person, mounted on a camera (one microphone). The microphones have been placed at different heights to check whether the distance from the floor is a factor influencing audio results. The stereo microphone (MSH-6), placed at the height ~1.2 meter above floor has been recorded the general sound background taking into account the direction of the sound source, whereas the Sennheiser ME67 has been placed on the tripod ~ 2 meters above floor. The purpose of that microphone was to record higher quality sounds related to vocalization of the person with PMLD as well as commands and comments of caregivers.

As a result, a set of video streams has been collected for further analysis:

- 3840x2160 pixels resolution, 25 frames per second (fps) high quality video streams.
- 1920x1080 pixels resolution, 25 fps 3D video streams.
- 320x240 pixels resolution, 30 fps infrared video streams.
- 640x480 pixels resolution, 15 fps infrared video stream.

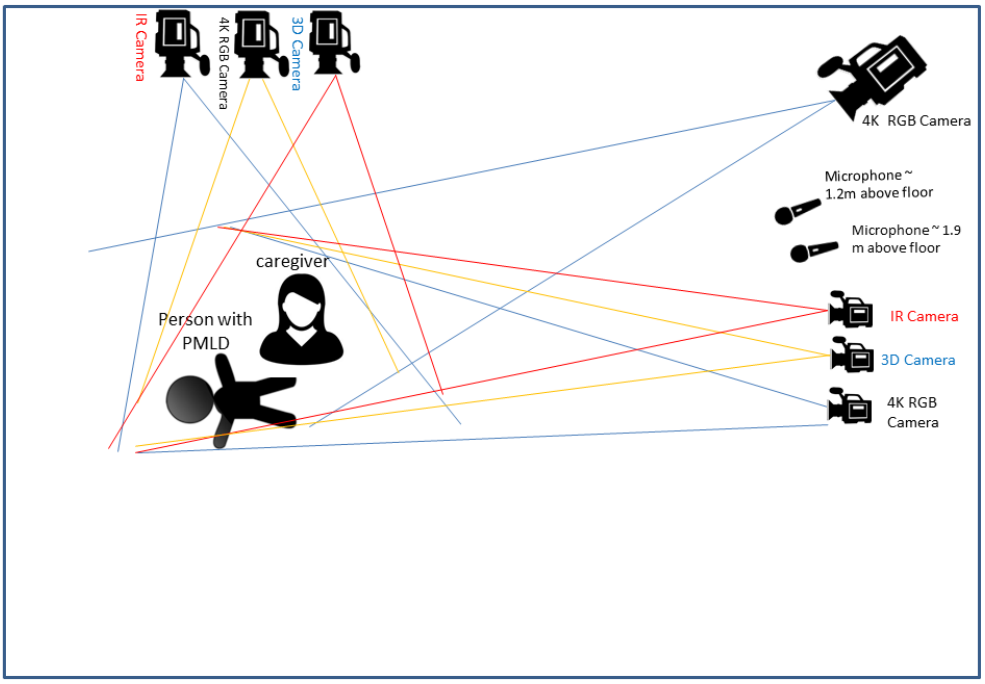

Figure 7. Basic recording setup

Recorded sound streams were:

- Mono, sampling rate 48kHz, depth 24 bits.

- Stereo, sampling rate 48 kHz, depth 24 bits.

Postproduction

Following the above-described sessions, the most promising recordings, from the project point of view, have been chosen for additional processing.

In the first step, the audio from the recordings has been translated to English in the form of subtitles.

In the next step, recordings related to a particular person, but recorded from different perspectives, have been synchronized in time and converted into a mosaic view. The synchronization of those video streams gives the ability to analyze the same scene from different perspectives in exactly the same point in time. In order to simplify the analysis process, final mosaic views have been created. According to specific needs and requirements, the mosaics consisted of two to four streams synchronized and presented together.

In the final step, subtitles have been embedded into the mosaic videos. Recordings created this way (see Figure 8) have been used as an input for an annotation process.

Figure 8. Final mosaic example

3.1.1.2 TECHNICAL ASPECTS OF THE ANNOTATION PROCESS

In the course of data collection, several videos have been recorded, which show the test persons in different communication situations expressing various inner states. These situations were specifically structured in a standardized procedure based on [92] and [93].

In the next step, these recordings needed to be analysed and annotated for two reasons:

(1) To recognize meaningful behaviour signals that express the triggered communication content and
(2) To provide a foundation for training the technological models.

The annotation process was coordinated by a specifically created working group, which consisted of pedagogical and technological specialists of the project. In a first step, the software *ELAN* (EUDICO

Linguistic Annotator)[1] was chosen, an annotation tool for audio and video recordings that uses a tier-based structure. As a free and open source software developed at the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands, and applied in humanities and social sciences research, it provides a procedure in three steps by defining tier types and tiers, selecting time intervals and entering annotations. *ELAN* offers both the use of time sampling TS[2] [94] and event sampling ES[3] [95].

In addition, an annotation guideline was created in a detailed process of adjustment to clarify which aspects are pedagogically meaningful and at the same time technologically realisable. The result is a guideline based on the paper-based assessment developed earlier in the project in order to increase the compatibility of these two databases (see Appendix A). The guideline is divided into three areas: The area *Communication and Inner States* should be triggered on purpose as mentioned above and is recognisable by means of the area *Behaviours*, which lists a variety of behaviour signals of the test person. The area *Context* needs to be annotated due to its possible impact on both *Behaviour* and *Communication and Inner States*.

These areas are split up into several main categories, subcategories and, finally, nearly 100 tiers with an additional explanation of the type of sampling. Exemplarily, Table 1 shows an extract of the area *Behaviours* and two of its main categories, *Vocalization* and *Facial Expressions*. *Facial Expressions* has all in all three different sub categories of which *Appearance of Eyes* is shown in the example with a further differentiation into several tiers, inter alia, *eye contact* or *eyebrow movement*. These tiers are integrated one-to-one into *ELAN* and should only be annotated by using time sampling when the end-user shows the specific behaviour signals within the intervals of three seconds. In contrast to *Facial Expression*, *Vocalizations* is main category and tier at the same time due to the higher complexity of describing vocalizations with words. For that reason, the technological experts of the working group proposed the use of audio samples to ensure a precise distinction of the different vocalizations. This way, aspects like pitch or volume, which are not assessable objectively can be analysed precisely by means of technology.

Table 1. Extract of the Annotation Guideline

Area	Main Category	Subcategories	Tiers	TS	ES
2. Behaviours	Vocalizations				X
	Facial Expressions	2.1.1. Appearance of Eyes	- eye contact - widened eyes - closed eyes - sleepy eyes - “smiling” eyes - winking - tears - eyebrow movement - frown	X	

Before the actual process of annotating was started, relevant sequences needed to be chosen, in which the test persons show all kind of communication contents and inner states mentioned above in a clear, meaningful and representative way. In addition to a briefing and training session for all annotators, regular meetings, constant exchanges, reciprocal checks and feedback both within the annotators and in

[1] Nijen Twilhaar, J., & van den Bogaerde, B. (2016). Concise Lexicon for Sign Linguistics. Amsterdam: John Benjamins Publishing Company

[2] TS means to analyze the recording in preset regular intervals (e.g., every five seconds or every hour), in which the shown behavior is annotated.

[3] Using ES, only the focused behaviour is annotated when it occurs, with full length without defining intervals.

combination with the technological partners provided a quality of the time-consuming annotation work as high as possible.

3.1.2 PHYSIOLOGICAL PARAMETERS

Besides video and audio data, we will also collect physiological data with wearable, i.e., contact sensors. However, it should be noted that the main goal would be the retrieval of the same set of physiological data from noncontact sensors, i.e., cameras. The contact approaches with wearable sensors will be included in the final project solution only if the noncontact approaches will not produce satisfactory results.

Based on our experience, we selected the Empatica E4 wristband for the collection of physiological parameters, which is currently the most reliable physiological-data collecting wristband.

Empatica E4 has the following sensors:

- PPG sensor that measures blood volume pulse, from which heart rate and heart rate variability can be derived.
- 3-axis accelerometer that captures motion-based activity.
- EDA/GRS sensor that measures the constantly fluctuating changes in certain electrical properties of the skin.
- Infrared thermopile that reads peripheral skin temperature.

It also has the event mark button that enables to tag the events, i.e., annotate the collected data.

Empatica enables to retrieve the following data:

- EDA/GSR.
- Blood volume pulse (BVP).
- Acceleration.
- HR.
- Temperature.

The collected physiological data are managed as follows:

- They can be stored in the wristband's internal memory that allows to record for up to 60 hours with five s. synchronization resolution.
- They can be transferred from the wristband to the Empatica cloud platform via USB.
- The data on the Empatica cloud can be downloaded in the CSV format.
- The data can be transferred in real time from the wristband to the connected device via Bluetooth. In this case, the data are automatically uploaded to the Empatica cloud platform from the connected device.
- E4 software development kit (SDK) enables to develop interfaces for transferring the data from the wristband to the Android/ iOS user applications.

3.2 GESTURE AND FACIAL EXPRESSION RECOGNITION SYSTEMS

3.2.1 PRE-PROCESSING OF THE MATERIALS

The material required for developing the recognition systems are the recordings with their correspondent annotations. The original material needs to be previously processed in order to generate the dataset for training and testing the computer vision algorithms.

The pre-processing step consists on extracting the slots of the video in which the person with PMLD appears doing a certain gesture or facial expression. After the annotation process for each recording, a file with tiers and the correspondent time intervals annotated is generated. The format of the file is shown in Figure 9.

annotation.eaf

```
<?xml version="1.0" encoding="UTF-8"?>
<ANNOTATION_DOCUMENT AUTHOR="" DATE="2018-08-01T14:28:14+01:00" FORMAT="3.0" VERSION="3.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="http://www.mpi.nl/tools/elan/EAFv3.0.xsd">
    <HEADER MEDIA_FILE="" TIME_UNITS="milliseconds">
        <MEDIA_DESCRIPTOR MEDIA_URL="file:///Z:/DATOS/Videos INSENSION/Mikolaj/m_video_1_subs.mp4" MIME_TYPE="video/mp4" RELATIVE_MEDIA_URL=
        "../Mikolaj/m_video_1_subs.mp4"/>
        <PROPERTY NAME="URN">urn:nl-mpi-tools-elan-eaf:df9dacb8-8534-418f-ae51-1d03dea5c689</PROPERTY>
        <PROPERTY NAME="lastUsedAnnotationId">1</PROPERTY>
    </HEADER>
    <TIME_ORDER>
        <TIME_SLOT TIME_SLOT_ID="ts1" TIME_VALUE="700"/>
        <TIME_SLOT TIME_SLOT_ID="ts2" TIME_VALUE="1560"/>
    </TIME_ORDER>
    <TIER LINGUISTIC_TYPE_REF="default-lt" TIER_ID="default">
        <ANNOTATION>
            <ALIGNABLE_ANNOTATION ANNOTATION_ID="a1" TIME_SLOT_REF1="ts1" TIME_SLOT_REF2="ts2">
                <ANNOTATION_VALUE></ANNOTATION_VALUE>
            </ALIGNABLE_ANNOTATION>
        </ANNOTATION>
    </TIER>
    <LINGUISTIC_TYPE GRAPHIC_REFERENCES="false" LINGUISTIC_TYPE_ID="default-lt" TIME_ALIGNABLE="true"/>
    <CONSTRAINT DESCRIPTION="Time subdivision of parent annotation's time interval, no time gaps allowed within this interval" STEREOTYPE=
    "Time_Subdivision"/>
    <CONSTRAINT DESCRIPTION="Symbolic subdivision of a parent annotation. Annotations refering to the same parent are ordered" STEREOTYPE=
    "Symbolic_Subdivision"/>
    <CONSTRAINT DESCRIPTION="1-1 association with a parent annotation" STEREOTYPE="Symbolic_Association"/>
    <CONSTRAINT DESCRIPTION="Time alignable annotations within the parent annotation's time interval, gaps are allowed" STEREOTYPE="Included_In"/>
</ANNOTATION_DOCUMENT>
```

Figure 9. Format of the annotation file

A simple tool was developed in order to automatically extract the correct slot of video for each annotated facial expression or gesture. These classified slots of videos were used to train and test the recognition algorithms.

3.2.2 METHODS

The goal of the recognition system is to detect gesture and facial expressions using computer vision techniques by means of the OpenPose library [96].

OpenPose is a library for real-time multi-person key point detection and multi-threading written in C++. OpenPose represents the first real-time system to jointly detect human body and hand key points on single images. In addition, the library's computational performance on body key point estimation is invariant to the number of detected people in the image.

The main functionality of the library is detailed as follows:

- Multi-person 15 or 18-keypoint body pose estimation and rendering. Running time invariant to number of people on the image.
- Multi-person 2x21-keypoint hand estimation and rendering. Note: In this initial version, running time linearly depends on the number of people on the image.
- Flexible and easy-to-configure multi-threading module.
- Image, video, and webcam reader.
- Able to save and load the results in various formats (JSON, XML, PNG, JPG, etc.).
- Small display and user interface for simple result visualization.
- Functionality is wrapped into a simple-to-use OpenPose Wrapper class.

3.2.2.1 GESTURE AND FACIAL RECOGNITION SYSTEM ARCHITECTURE

Gesture, posture and facial expression recognition are built based on OpenPose gesture and facial libraries. OpenPose uses other third party libraries in order to build the methods for training, testing and then,

running the recognition models. Some third party libraries are used to perform the whole recognition system [96]:

- Boost: it provides free peer-reviewed portable C++ source libraries. Boost libraries are intended to be widely useful and usable across a broad spectrum of applications. They are comprised by a set of libraries for the C++ programming language that provide support for tasks and structures such as linear algebra, regular expressions and unit testing [97].

- Convolutional Architecture for Fast Feature Embedding (CAFFE) is a deep learning framework made with expression, speed and modularity. It is developed by Berkeley AI Research and by community contributors. The main features are expressive architecture, extensible code and speed [98].

- Compute Unified Device Architecture (CUDA) is a parallel computing platform and programming model developed by NVIDIA for general computing on graphical processing units (GPUs). With CUDA, developers are able to dramatically speed up computing applications by harnessing the power of GPUs [99].

- The NVIDIA CUDA®, cuDNN is a GPU-accelerated library of primitives for deep neural networks. cuDNN provides highly tuned implementations for standard routines such as forward and backward convolution, pooling, normalization, and activation layers. cuDNN is part of the NVIDIA Deep Learning SDK [100].

- Open source Computer Vision (OpenCV) library [101] is the leading open source library for computer vision, image processing and machine learning, and now features GPU acceleration for real-time operation.

- gflags is a C++ library that implements command line flags processing. It includes built-in support in standard types such as string and the ability to define flags in the source file in which they are used [102].

- Protocol buffers (protobuf) are Google's language-neutral, platform neutral, extensible mechanism for serializing structure data in a faster and simpler way than other technologies such as XML [103].

OpenPose architecture consists of a two-branch multi-stage CNN. CNNs are very similar to ordinary NNs. They are composed of neurons that have learnable weights and biases. Each neuron receives some inputs. The whole network expresses a single differentiable score function: from the row image pixels on one end to class scores at the other. They also have a loss function on the last (fully connected) layer. Therefore, a CNN is comprised of one or more convolutional layers and then followed by one or more fully connected layers as in a standard multilayer NN. The architecture of a CNN is designed to take advantage of the 2D structure of an input image (or another 2D input). This is achieved with local connections and tied weights followed by some form of pooling which results in translations invariant features. Another benefit of CNNs is that they are easier to train and have fewer parameters than fully connected networks with the same number of hidden units [104].

Regarding the OpenPose (Figure 10), each stage in the first branch predicts a confidence map (i.e. a probability density function on the new image, assigning each pixel of the new image a probability) and each stage in the second branch predicts part affinity fields (PAFs). After each stage, the prediction form the two branches along with the image features are concatenated for the next stage [9].In more detail, the detection is composed by three stages:

- Stage 0: The first 10 layers of the CNN are used to create feature maps for the input image.
- Stage 1: A 2-branch multi-stage CNN is used where the first branch predicts a set of 2D confidence maps of body part locations. The second branch predicts a set of 2D vector fields (L) of part affinities, which encode the degree of association between parts.
- Stage 2: The confidence and affinity maps are parsed by greedy inference to produce the 2D key points for all people in the image.

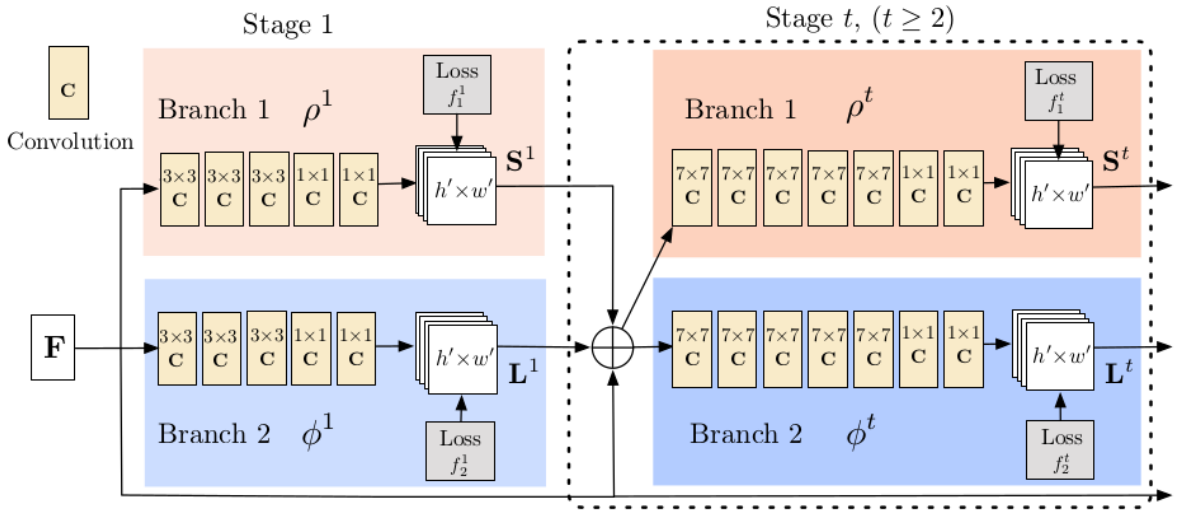

Figure 10. Architecture of the two-branch multi-stage CNN. Each stage in the first branch predicts confidence maps St. Each stage in the second branch predicts PAFs Lt

3.2.2.2 GESTURE RECOGNITION SYSTEM

OpenPose-pose key points are used to classify different gestures like hand on head or raising leg (Figure 11). In Figure 11, which corresponds to the frontal view of a person body, OpenPose detects 18 key points. In some other cases, the library can detect a different number of key points depending on the quality of the image and the human pose.

The recognition application builds a sample vector by computing the distances between a set of reference key points or only one key point to all other pose key points. The distance metric allows users to uniquely classify different human postures.

Using these sample vectors, a machine learning algorithm is training in order to learn how to classify different human gestures or postures.

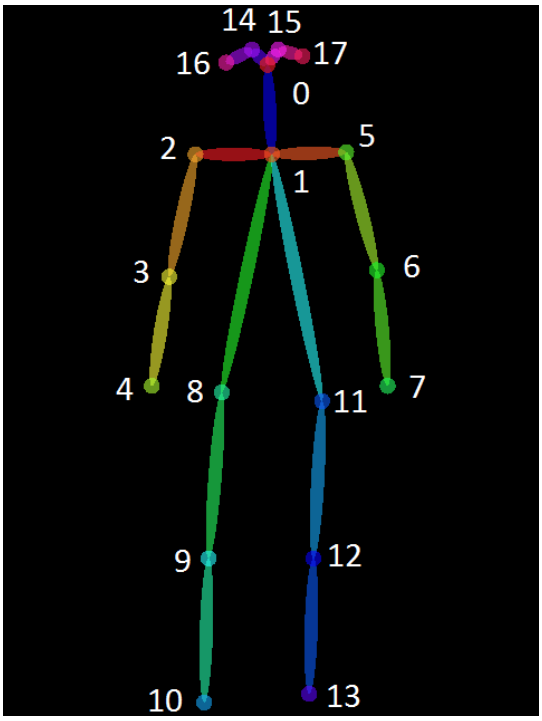

Figure 11. OpenPose pose keypoints

3.2.2.3 FACIAL EXPRESSION RECOGNITION SYSTEM

OpenPose-face key points, illustrated in Figure 12, are used to classify different human face expressions. As in the previous case, the recognition system constructs a sample vector using both diverse distance measures [105] from a reference key point or a set of reference key points (e.g. key point 30, which refers to the tip of nose) to all other face key points.

In order to reduce the high dimensionality of the vector space and to obtain a better classification, key points from point 0 to 16, (chin and jaw) could be ignored. However, as explained in the gesture recognition case, the number of key points can be variable depending on the quality of the image and the human pose.

Using these sample vectors, a machine learning algorithm is training in order to learn how to classify different facial expressions.

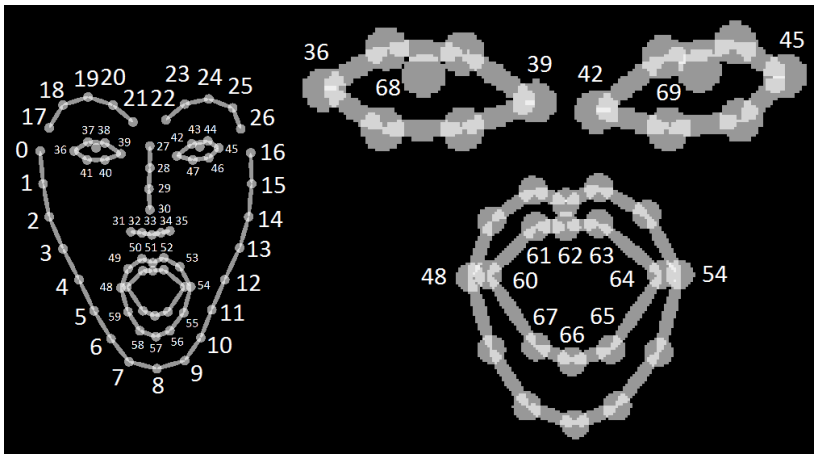

Figure 12. OpenPose face keypoints

3.2.3 REQUIREMENTS

In order to obtain a good performance of the system, a wide variety of images and video sequences with their corresponding annotation should be needed. These instances should have enough quality in terms of clarity (i.e., avoid occlusions of the important parts to detect), lightning conditions, distance to the subject and resolution.

In addition, requirements for the default configuration are outlined as follows:

- NVIDIA GPU version: NVIDIA graphics card with at least 1.6 GB.
- At least 2.5 GB of free RAM memory (assuming that cuDNN is installed).
- AMD GPU version: Vega series graphics card. At least 2 GB of free RAM memory.
- CPU version: Around 8GB of free RAM memory.
- Highly recommended: a CPU with at least eight cores.

3.3 VOCALIZATION RECOGNITION SYSTEM

3.3.1 PRE-PROCESSING OF THE MATERIALS

Vocalization recognition tools, in their current form, expect mono pulse code modulation (PCM) wave file sampled at 16 kHz. For the time being, the extraction of audio layer from video files is being accomplished with the help of third-party tools, e.g. *ffmpeg*.

3.3.2 METHODS

In short, the algorithms used for vocalization recognition are based on Hidden Markov Models (HMMs). First, signal processing (parametrization) modules allow choosing either Mel-scale filterbank, MFCC, or their gammatone counterparts as the main static vector, plus a number of optional features. The static vector can be extended by its frame-domain derivatives of 1st and 2nd order. We are planning to implement the extraction of additional static features from the signal, based on literature (currently only maximum autocorrelation coefficient is used).

In this solution, every unique vocalization type is stored as a list of distinct states of the event and the state transition matrix. Each state is assumed to correspond to one stationary segment of audio observations (a segment that is expected to appear within a modeled event); the stationary signal in a given state is thus represented by its Gaussian Mixture Model (GMM), that describe distributions of parameters mentioned before (e.g.MFCC). It is also possible to extend the models by explicit state duration distributions, which formally makes them semi-Markov rather than HMM.

3.3.2.1 ESTIMATION OF MODELS

The training procedure consists of two phases:

- Unsupervised audio frame clustering, using the GMM method: the number of distinct states of the model is also determined at this stage – the information criterion: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) or mixture or both is calculated for this purpose; the clustering is currently top-down, ie. as long as the chosen criterion increases, the largest component (distribution) is split into two, and all observations are reassigned across new set of distributions; the phase is finished by transferring every component of the final GMM to the HMM model as the initial distribution of a single-mixture hidden state;

- Reestimation until convergence, using Expectation-Maximization (EM) method: the observation distributions held in Markov states, the state transition matrix and optionally state duration distributions are iteratively readjusted to better model the training examples; classical Baum-Welch is run in case of standard HMM models, more complex duration distribution-aware algorithms are implemented for the semi-Markov case.

3.3.2.2 DETECTION OF EVENTS

The Markov model-based detection framework assumes that an event matching a given model can start and end on any time across a recording, that there can be zero, one or many events of this kind, but the events corresponding to the same model cannot overlap (however, concurrent models, i.e. different vocalization types belonging to the same person, are currently treated independently, and thus their detections can overlap).

A Token-Passing algorithm is implemented (its complexity is dependent on whether state duration distributions are explicitly modeled). A *token* represents an event hypothesis that is *partially* matched to some range of input observations, and 'currently' occupies one of the states. As mentioned before, a new event can potentially start at any frame. Each token thus remembers its supposed start time, as well as its accumulated cost from this point: acoustics (GMM emission probability), transition/duration distributions and constant insertion penalties all contribute to the total cost; in fact, those factors are stored separately in a token data (the reason is that the pruning and the final scoring, both explained below, may use different sets of weights).

With each observation, ie. short-time audio frame, tokens are passed from state to state according to the transition matrix (whenever a state has more than one possible successor, tokens currently occupying this state will be 'cloned' before being propagated). Whenever two tokens meet at the same time the same state (in explicit duration case: also having the same *current state entry time*), the dearer one is discarded. To speed up computation, heuristic pruning is implemented, that puts out least promising candidates within a set of tokens belonging to the same model. Necessarily, both operations discussed here are based on average cost per frame from the supposed *event* start, as opposed to partial cost from the start of the sentence (the latter is used in automatic speech recognition (ASR); note the importance of well-tuned insertion penalties for handling of very short candidates in "average cost" case).

Whenever a token reaches a final state of the model, it undergoes a final scoring: a weighted sum of acoustic and transition/duration average per frame cost (plus a term penalizing very short candidates) must be lower than a given threshold. Currently, the threshold value has to be tuned manually (it can be done separately for each model). Finally, from overlapping candidates, only the most likely are preserved. The implementation of this last rule of the algorithm is not yet fully on-line, meaning that the results for a provided recording are presented all at once.

3.3.3 REQUIREMENTS

The procedures currently assume:

- Audio format of mono PCM 16-bit, 16 kHz;
- The training examples of signals of interest to be precisely annotated by experts (the training tool actually requires an input wave file to be precisely cropped to a signal range that contains a training example, and stored one example per file, but this is only a temporary technical limitation).

We expect that the accuracy of recognition will depend primarily on:

- Number of examples for each vocalization model;

- The quality of recordings (possible methods for voice tracking from multi-channel audio are considered, but not implemented).
- The accuracy of training material annotation;
- The success of extra parametrization features.

The detection procedures are efficient enough to run in real-time on a contemporary consumer-class PC for a realistic number of unique signals (event types) that are concurrently recognized, unless a very computationally expensive (likely neural network-based) analysis is introduced in the future. The procedures are currently written in C++ and compiled under Microsoft Windows®, however there are no fundamental obstacles to compile these to run under other operating systems.

3.4 PHYSIOLOGICAL PARAMETERS MONITORING SYSTEM

3.4.1 PRE-PROCESSING OF THE MATERIALS

All of the camera-based methods use the face of the subject as the “field of interest” in their research. For that reason, using the OpenCV library, we detected the face of the subject. Since the other parts of the frame rather than the face of the subject are not needed, we deleted them and only kept the frame of the face of the subject.

The skin-based approach uses the information skin pixels contain about the blood flow of an individual to retrieve the PPG signal. Because we only want the skin pixels, it is necessary to implement a method that discriminates well skin and non-skin pixels. Here we describe two methods. The first method is fast, simple and works well on our dataset. This method however might not be robust for all skin types. Therefore, the second model uses machine learning to learn itself.

In the first method, we transform the RGB color space into the YCbCr color space. This color space is commonly used for video compression and does therefore contain less redundant information. Using some reasonable thresholds on the values in the color space, we classify which pixels belong to the skin and which do not.

The second method applies one-class support vector machines to classify the skin pixels. The SVM model is obtained using unsupervised clustering on the first three frames, while from the forth frame on starts classifying the skin pixels. As features, we use the relations between the values of the RGB color space as well as the YCbCr color space. This method of skin classification returns less precise results, but works on all skin types.

Both of the skin classification methods classify some of the non-skin pixels as skin. To avoid false positives, we select only the pixels that are most likely to be skin. To do so, we calculate the mean value of all the skin pixels returned by the classifier and remove some proportion of the pixels whose position in the YCbCr color space is not near the mean value.

3.4.2 METHODS

We will focus on obtaining the PPG signal from camera recordings. This PPG signal is often referred in the related work as rPPG and therefore, we will also use this term from now on.

There are two approaches for obtaining the rPPG signal from a camera recording – color based approach and motion based approach. We will describe the methods we used for both of the approaches.

3.4.2.1 COLOR-BASED APPROACH

Since retrieving the rPPG signal from video recording appears to be quite difficult, we tried seven different methods. Five of the methods are inspired by the related work, while we developed the other two ourselves. Some of the steps between different methods overlap with each other.

Color-based method 1

By sequencing the averaged value of the red, green and blue intensity of all the skin pixels, the method creates three different traces. Because all of these traces contain information about the blood flow, we normalize them and transform them with independent component analysis (FastICA algorithm [106]) [63] [64]. This method still returns three signals, so based on the power spectrum of the frequencies included in the signal, we choose the one with most frequencies in the range [0.6 Hz, 4 Hz] as rPPG.

Color-based method 2

Another method also uses the three normalized traces as the previous method, but instead of applying the independent component analysis, assumes that the three traces are dependent on each other and therefore uses principal component analysis to extract only one trace, which represents the rPPG signal [65].

Color-based method 3

As well as the first and the second method, the third method also uses the mean of the red, green and blue intensity all the skin pixels [62]. To retrieve the rPPG signal, this method makes a linear combination from the red green and blue traces, resulting with two other traces X and Y calculated as follows:

$$X = 3R - 2G$$

$$Y = 1.5R + G - 1.5B.$$

The X and Y traces are afterwards filtered and combined to retrieve the rPPG signal.

This method was already implemented in the BOB library[4] and therefore, we use this library to retrieve the rPPG signal.

Color-based method 4

The forth method is an upgrade from the third method. While the third method retrieves the rPPG signal by averaging the RGB values of all skin pixels from the face, this method uses all skin pixels as independent components that retrieve pulse traces. The pulse traces are afterwards combined into a single rPPG signal [66].

Since face region pixels might disappear from the frame (e.g., the subject has moved his head from the left side to the right), tracking them for a long-time period might not be efficient. Therefore, using the Farneback tracking algorithm [107], we are tracking each set of pixels detected from the first frame only for three seconds. This tracking technique returns the optical flow for each pixel in the image and not only for the corners. In order to prevent big tracking error, we use forward-backward flow tracking [108] to filter all pixels whose backtracked coordinate distance is larger than one.

[4] https://pypi.org/project/bob.rppg.base/

In order to produce the pulse traces, we calculate the difference of the RGB values between two sequential pixels and normalize them. Afterwards, we apply the steps from the third method for each skin pixel from the head that was not filtered out, and get a 3-second interval rPPG signal. Using principal component analysis, we combine all pulse traces into a single rPPG signal. The 3-second interval rPPG signals are merged together into one the resulting rPPG signal.

Color-based method 5

This method uses all skin pixels from one frame to define the space this frame is represented in [67]. By tracking the changes in this space, we retrieve the rPPG signal. To accomplish this, a covariance matrix is computed to represent the space. The blood flow into the skin causes changes in this covariance matrix for each frame. By calculating the eigenvectors and the eigenvalues of the covariance matrix, we get a representation of the color space for the skin pixels. The rotation between two eigenvectors of sequential frames presents the changes of the color space. This rotation is also related to different relative PPG contributors and therefore, by concatenating the rotation between the first opposing to the second and the third eigenvector, rPPG-like traces are retrieved. The eigenvalues are also influenced by the pulsatile blood and thus are used to scale the changes in the two rPPG-like signals. The final rPPG signal is produced as a combination of the two rPPG-like signals. This method is also implemented as part of the BOB library[4], which we use to retrieve the rPPG signal.

Color-based method 6

Because the results of the previous four methods were not satisfying enough, we decided to try to filter their output rPPG signal using deep learning. To accomplish this, we used a LSTM model as well as an auto-encoders model. As input to the model, we used the rPPG signal retrieved by one of the previous algorithms. As target, we set the original PPG signal measured by a finger PPG sensor. The window in the LSTM model was set to five seconds, while the network architecture consisted of two LSTM layers. The auto encoders consisted of two encoding and two decoding feed-forward layers.

Color-based method 7

Inspired by the sixth method, we decided to put some of the skin pixels as input into the LSTM model and set as target the original PPG signal.

3.4.2.2 MOVEMENT-BASED APPROACH

While other methods use the skin as a basis for analyzing the blood flow of the subject, another method is using the head oscillations [71]. In this method, we track the flow of the top features using the Lucas-Kanade flow-tracking algorithm [109]. We select only the vertical flow vectors and because the heart beats with frequency in the interval of [0.6 Hz, 4 Hz], we filter the frequencies that do not belong to this interval. Afterwards we apply principal component analysis and similarly as in the previous state of the art methods, we select the most rPPG-like signal as a result.

3.4.2.3 AGGREGATION OF RECONSTRUCTED SIGNALS

The PPGs reconstructed using presented methods will be aggregated using a meta machine learning model. To this end, we will evaluate several meta learners to determine the most appropriate one for our data. Additionally, we will focus on meta learners that are capable of estimating the quality of underlying models in order to increase the quality of the reconstructed PPG.

3.4.2.4 CALCULATION OF PHYSIOLOGICAL PARAMETERS

The physiological parameters will be estimated by detecting the peaks in the PPG signal, which correspond to heartbeats. This way, heart rate and heart-rate variability can be derived. In addition, features will be computed from the morphology of the segmented PPG cycles, with which we will attempt to estimate respiratory rate and blood pressure.

3.4.2.5 DETERMINING THE PSYCHOLOGICAL STATE

Based on the physiological parameters, the psychological state of the subjects can be determined. For this purpose, we will adapt state-of-the-art methods from this field to the specific requirements of our target group. The main goal is to distinguish between three psychological states (displeasure, neutral, pleasure); however, we will also attempt to determine other states. It should be noted that this psychological state is only a partial estimation, which will be fed into a global estimator that will also consider data from gesture, facial and vocalization recognition systems.

3.4.3 REQUIREMENTS

Because all of our methods use video recordings to measure the physiological parameters, the quality of the camera will be of crucial importance. A frame rate of at least 20 fps is enough to measure the heart rate of a subject. Heart-rate variability can also be measured at 20 fps, but it is more precise with a higher frame rate. To extract physiological measures such as blood pressure and breathing rate, more subtle information contained in the PPG signal is needed and consequently higher frame rates. We have not hard data on how high, but around 50 fps seems reasonable.

Since we are analyzing small movements and changes in the color of the face, the resolution of the images is very important. Results from related work show that resolution of 1280x720 pixels can be used to extract rPPG [71].

Most of the evaluated methods do not have any specific hardware or software requirements. However, deep learning methods typically require nVidia GPUs. Since these methods seem the most promising, we would require GPUs to greatly accelerate learning and evaluation of the developed models.

Considering that most of the methods use the effect of photo plethysmography, the subject will have to be in a well-lit environment.

4 EXPERIMENTS AND RESULTS

4.1 GESTURE RECOGNITION SYSTEM

4.1.1 CLASSIFIER FOR THE GESTURE RECOGNITION SYSTEM

Following the method explained in Section 3.2.2.2, a set of OpenPose-face key points were used to classify different human gestures.

For this initial experiment, four gestures were selected: “hand_on_hand”, “foot_on_foot”, “raising_right_arm” and “raising_left_arm”. A sample vector codified each gesture by calculating the L2 Euclidean distance measure [105] from a reference key point to the rest of points that represent the body. In this case, the reference key point was associated to the neck, value equal to zero (Figure 11).

Hence, each gesture was represented by a vector of L2 distances. It is important to note that each gesture can be described by “n” vectors of distances. The latter occurs because there could exist different gesture poses and orientations for the same expression, so the distances between key points could be varied for the same gesture.

4.1.1.1 MATERIALS AND SETUP

In this initial experiment, three video files were used. These files contained diverse behavior scenarios of two subjects. Each video-file was divided into frames, which contained the most relevant information in terms of gestures.

In total, a batch of 149 samples, which were associated to a sequence of five vectors that corresponds to the frames of 18 dimensions (i.e. 18 keypoints per frame), was considered. The batch input shape of the layer was then (149, 5, 18), and the “input_shape”, not including the samples dimension, was (5, 18).

Once the input dataset was built, one part of the data was selected for the training phase and the other part was used for the test stage in a proportion of 67:33, respectively.

A Time Distributed Feed Forward (dense) NN was used to train the recognition system using the set of training gestures (Figure 13).

```
Build Keras Timedistributed-LSTM Model...
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
time_distributed_1 (TimeDist (None, 5, 36)             1332
_________________________________________________________________
dropout_1 (Dropout)          (None, 5, 36)             0
_________________________________________________________________
time_distributed_2 (TimeDist (None, 5, 72)             2664
_________________________________________________________________
dropout_2 (Dropout)          (None, 5, 72)             0
_________________________________________________________________
time_distributed_3 (TimeDist (None, 5, 36)             2628
_________________________________________________________________
dropout_3 (Dropout)          (None, 5, 36)             0
_________________________________________________________________
time_distributed_4 (TimeDist (None, 5, 18)             666
_________________________________________________________________
dropout_4 (Dropout)          (None, 5, 18)             0
_________________________________________________________________
time_distributed_5 (TimeDist (None, 5, 9)              171
_________________________________________________________________
dropout_5 (Dropout)          (None, 5, 9)              0
_________________________________________________________________
lstm_1 (LSTM)                (None, 9)                 684
_________________________________________________________________
dense_6 (Dense)              (None, 5)                 50
=================================================================
Total params: 8,195
Trainable params: 8,195
Non-trainable params: 0
```

Figure 13. Results obtained during the training phase of the NN algorithm for gesture recognition

4.1.2 RESULTS OF THE GESTURE RECOGNITION SYSTEM

After training the algorithm, a validation of the model was performed with the previously selected test dataset.

Figure 14 shows the total performance of the model during the test stage. On the other hand, Figure 15 depicts the resulting confusion matrix, i.e. the false positive and false negative rate, as well as the true positive and negative results.

It can be seen that the system, using the test dataset, wrongly predicts only three cases for “hand_on_head” when the true expression would be “raising_right_arm”. The confusion matrix determines that the model predicts “raising_right_arm” and the true expression would be “foot_on_foot” and finally the algorithm suggests, in one case, “foot_on_foot” when it would be “hand_on_head”. The obtained error test rate (the rate of wrongly classified instances by using the test dataset) was 0.10.

Therefore, using this preliminary training and test dataset, an accuracy equal to 0.90 was achieved for the whole system, i.e., the gesture recognition model for a set of four body gestures. This means that if the system receives a video file as input, it will be able to identify which gesture corresponds to the body expression shown in the recording with an accuracy of 0.90.

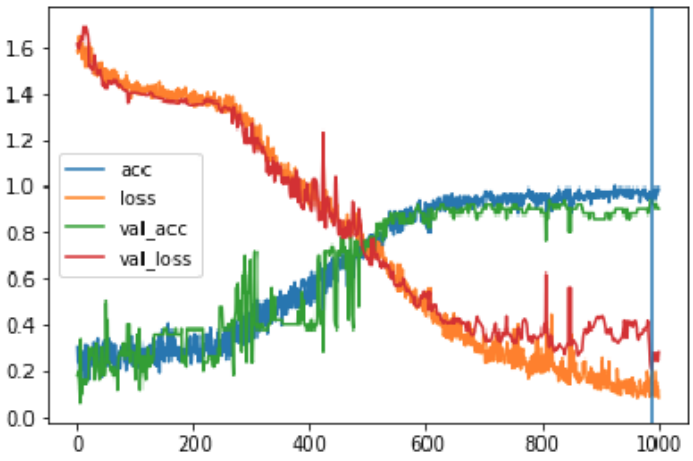

Figure 14. Detail of the performance results during the classifier training and test of the gesture recognition system

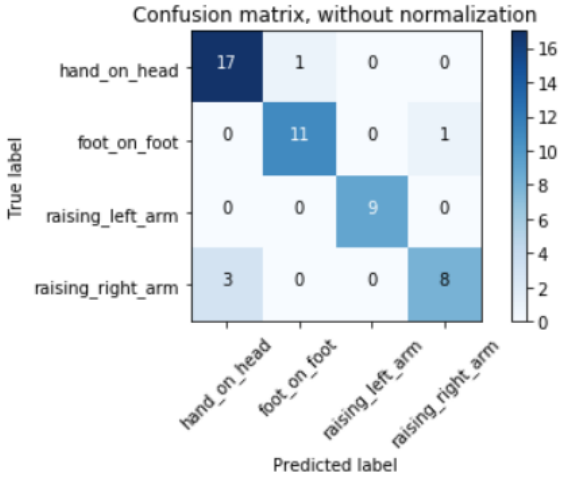

Figure 15. Confusion matrix obtained by the model for the gesture recognition system

After building the final gesture recognition model, the experimentation was carried out using the pre-defined set of video files of PMLD patients. Some results from the experiments are shown below (Figure 16 to Figure 19).

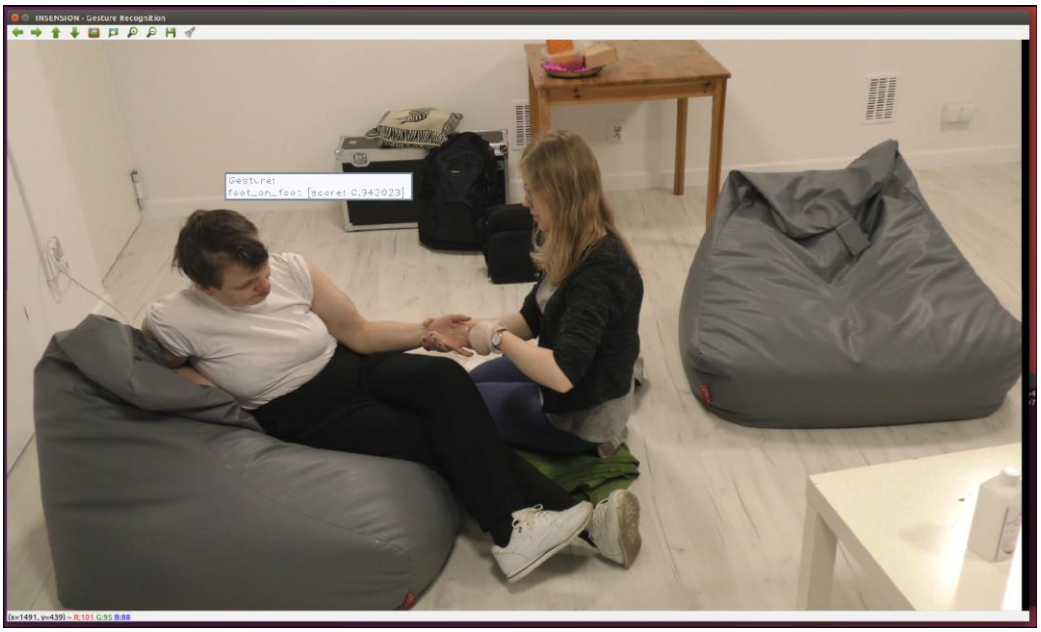

Figure 16. Gesture Recognition System result "foot_on_foot", accuracy = 0.94

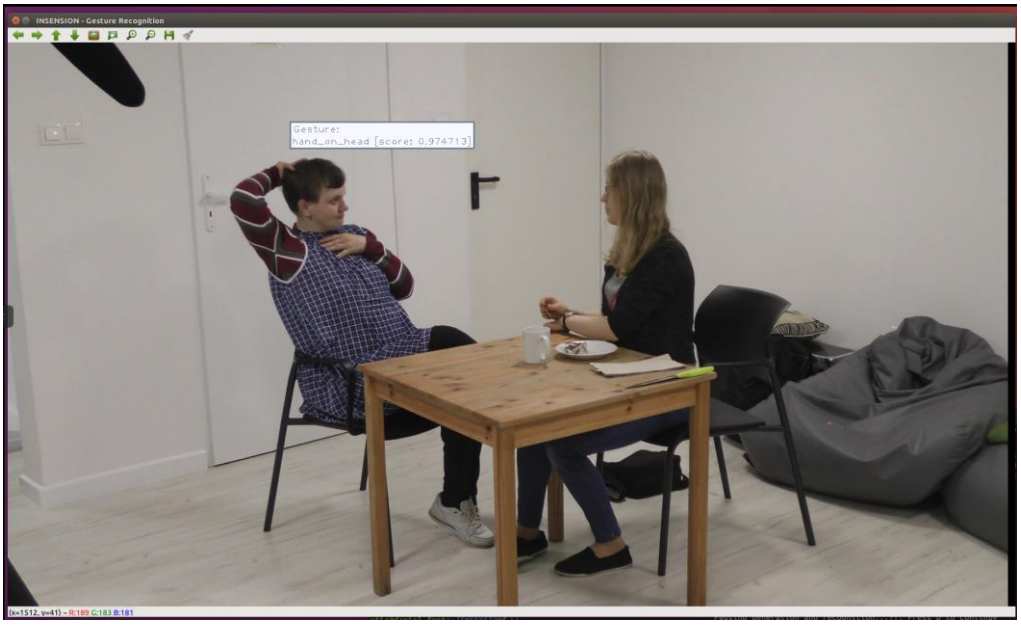

Figure 17. Gesture Recognition System result "hand_on_head", accuracy = 0.97

Figure 18. Gesture Recognition System result "raising_left_arm", accuracy = 0.97

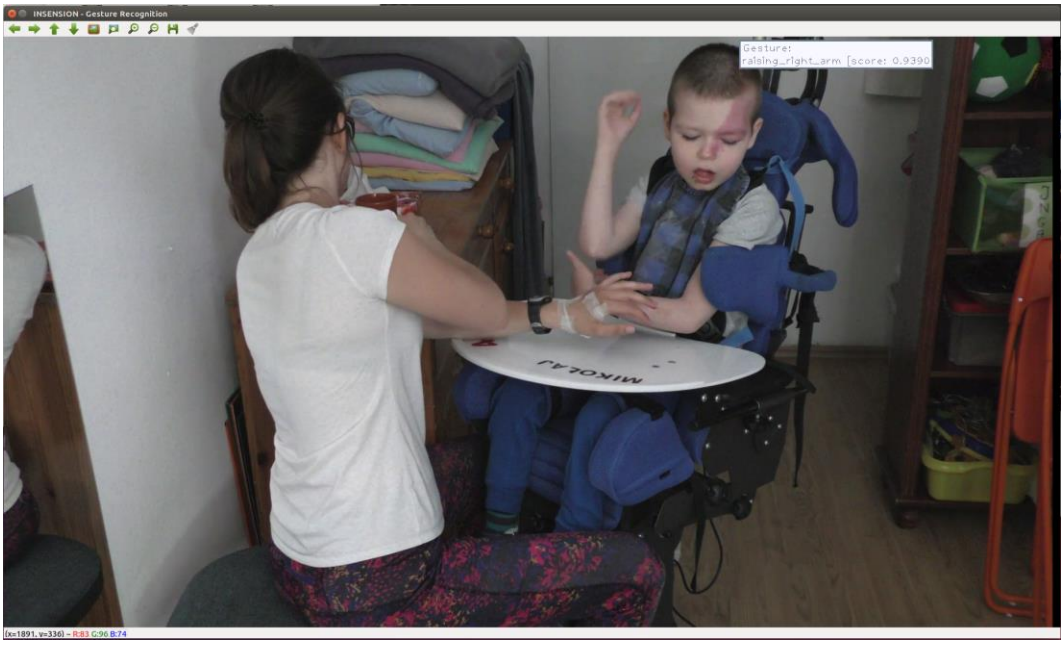

Figure 19. Gesture Recognition System result "raising_right_arm", accuracy = 0.93

4.2 FACIAL EXPRESSION RECOGNITION SYSTEM

4.2.1 CLASSIFIER FOR THE FACIAL EXPRESSION RECOGNITION SYSTEM

In this experiment, four expressions were selected: "closed eyes", "frown", "mouth_open" and "corners_mouth_up". For each expression, a sample vector was built by means of computing the L2 distance measure [105] from a reference key point (key point 30, which refers to the tip of nose) to the rest of face key points. Therefore, a vector of distances determines a particular facial expression. As in the previous case of the gesture recognition system, "n" vectors of distances described each expression,

because there could exist different head poses and orientations for the same expression, so the distances between key points could be varied for the same facial expression.

4.2.1.1 MATERIALS AND SETUP

For these preliminary experiments, several video files were used. These files contained several behavior scenarios of three patients across six recordings. Each video file was divided into frames, which contained the most relevant information in terms of facial expressions.

In total, a batch of 256 samples, where each sample was a sequence of five vectors of 53 dimensions, was considered for this initial experimentation. The batch input shape of the layer was then (256, 5, 53), and the "input_shape", not including the samples dimension, was (5, 53).

Once the input dataset was built, one part was selected for the training phase and the other part was used for the test stage in a proportion of 67:33, respectively.

A Time Distributed Feed Forward (dense) NN was used to train the recognition system for the set of selected expressions (Figure 20).

```
Build Keras Timedistributed-LSTM Model...
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
time_distributed_1 (TimeDist (None, 5, 96)             9312
_________________________________________________________________
dropout_1 (Dropout)          (None, 5, 96)             0
_________________________________________________________________
time_distributed_2 (TimeDist (None, 5, 192)            18624
_________________________________________________________________
dropout_2 (Dropout)          (None, 5, 192)            0
_________________________________________________________________
time_distributed_3 (TimeDist (None, 5, 96)             18528
_________________________________________________________________
dropout_3 (Dropout)          (None, 5, 96)             0
_________________________________________________________________
time_distributed_4 (TimeDist (None, 5, 48)             4656
_________________________________________________________________
dropout_4 (Dropout)          (None, 5, 48)             0
_________________________________________________________________
time_distributed_5 (TimeDist (None, 5, 24)             1176
_________________________________________________________________
dropout_5 (Dropout)          (None, 5, 24)             0
_________________________________________________________________
lstm_1 (LSTM)                (None, 24)                4704
_________________________________________________________________
dense_6 (Dense)              (None, 5)                 125
=================================================================
Total params: 57,125
Trainable params: 57,125
Non-trainable params: 0
```

Figure 20. Results obtained during the training phase of the NN algorithm for the facial expression recognition model

4.2.2 RESULTS OF THE FACIAL EXPRESSION RECOGNITION SYSTEM

After training the algorithm, a validation of the model was performed with the previously selected test dataset.

Figure 21 depicts the total performance of the model during the test stage. Figure 22 shows the confusion matrix obtained, i.e. the false positive and false negative rate, as well as the true positive and negative results.

It can be seen that the system, using the test dataset, wrongly predicts only two cases for "closed_eyes" when the true expression would be "frown". The confusion matrix determines that the model predicts "mouth_open" and the true expression would be "corners_mouth_up" and finally the algorithm suggests, in one case, "corners_mouth_up" when it would be "closed_eyes". The obtained error test rate was 0.0824.

Therefore, using this preliminary training and test dataset, an accuracy equal to 0.9176 was achieved for the whole system, i.e., the facial recognition model for a set of four facial expressions. This means that if the system receives a video file as input, it will be able to identify which facial expression corresponds to the face shown in the recording with a percentage of accuracy of 91.76.

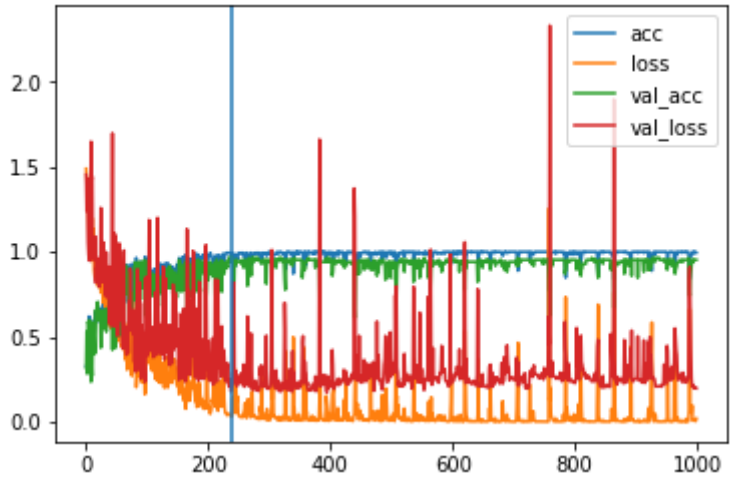

Figure 21. Detail of the performance results during the classifier training and test for the facial expression recognition system

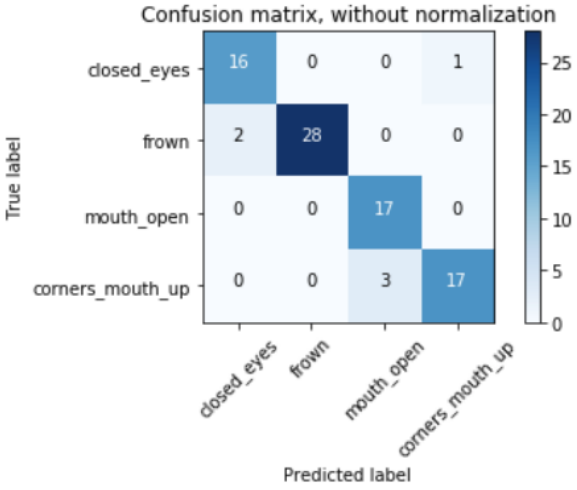

Figure 22. Confusion matrix obtained by the model for the facial expression recognition system

After building the final facial expression recognition model, the experimentation was carried out using the pre-defined set of video files of PMLD patients. Some results from the experiments are shown below (Figure 23 to Figure 25).

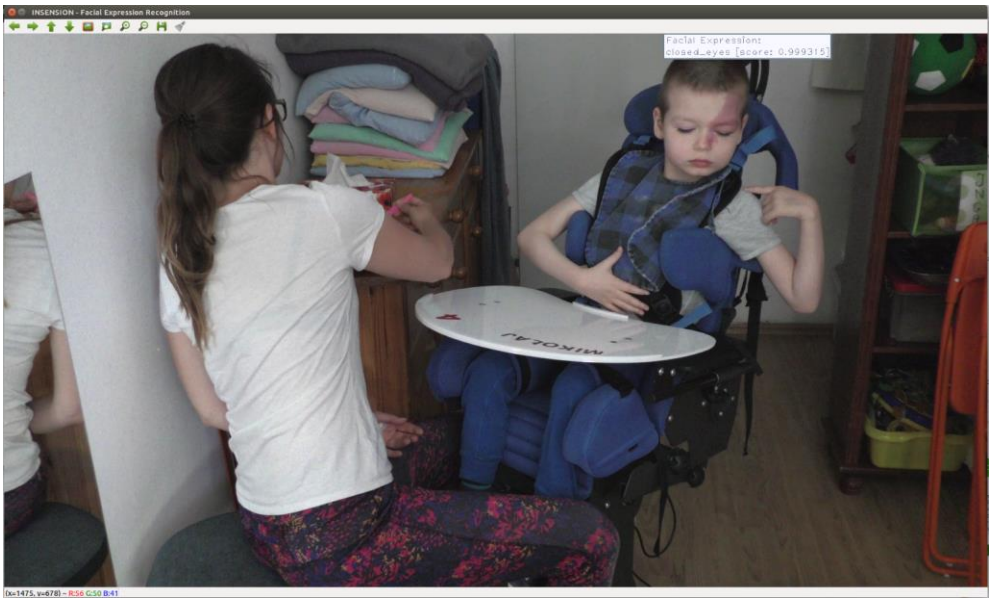

Figure 23. Facial Expression Recognition System result "closed_eyes", accuracy = 0.99

Figure 24. Facial Expression Recognition System result "mouth_open", accuracy = 0.91

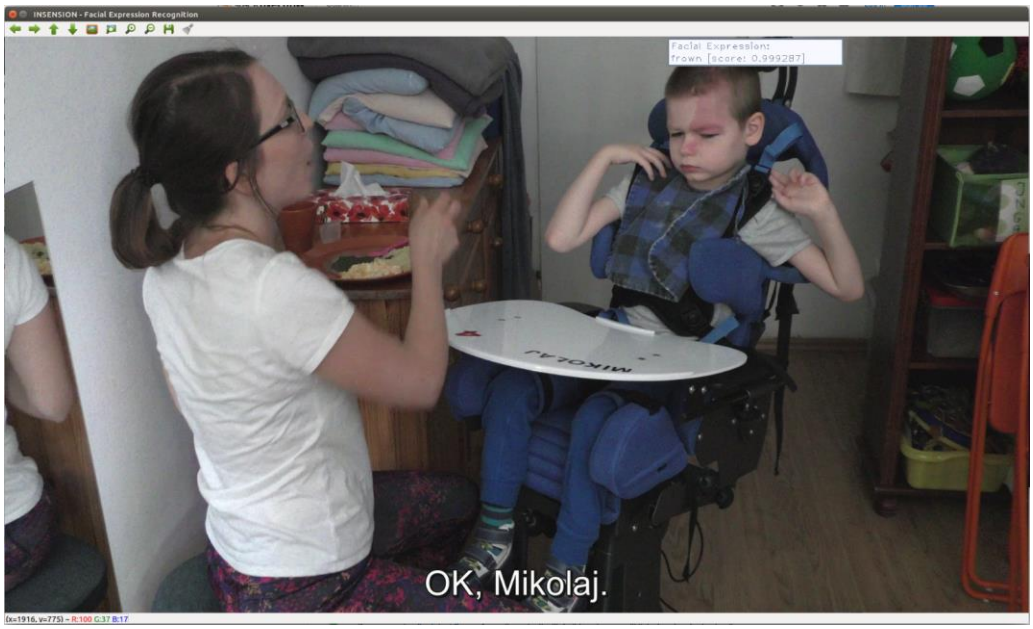

Figure 25. Facial Expression Recognition System result "frown", accuracy= 0.99

4.3 VOCALIZATION RECOGNITION SYSTEM

4.3.1 MATERIALS

For preliminary experiments, vocalizations found in ''orzeszek_03_2018_sounds'' were used. The session contained annotated recordings of two caretakers, across a total of three recordings: Person A demands, Person A protest and Person F demands. The following ''message models'' were built, corresponding to unique vocalizations (labels) that were found in the recordings:

Table 2. Detailed material used in the preliminary experiments for Vocalization Recognition System

Label	No. examples
Person_A.hums	7
Person_A.laughs	3
Person_A.many	1
Person_A.squeals	1
Person_A.wails	9
Person_F.snores	7
Person_F.o-oh	3

4.3.2 SETUP

Two separate experimental setups were assumed:

- Test-on-train setup (ToT), where each model was build based on all its known examples
- Pseudo-cross-validated (pXV) setup, in which for every label except Person_A.many and Person_A.squeals three separate models were build, each excluding about 1/3 of training examples.

Models were always tested on all recording featuring a given person. The fact that in both versions training examples were present in the testing material was enforced by a low number of examples selected for preliminary experiments, especially in the fact that there was no vocalization *label* that would appear

across more than one recording. However, the converse statement is not completely true in case of cross-validated setup, where not all testing examples were used for training. In addition, each particular vocalization (i.e. *realization*) was represented by numerically different values depending on having been calculated from training or testing material (this is because of how signal parametrization works, specifically: when using short-time analysis, the features are calculated with a uniform window-shift constant, e.g. every 10 ms, therefore the exact two parametrizations of a given instance were numerically different if a training example's offset related to the testing file was not a multiple of the window-shift constant; it needs to be stressed, though, that despite being numerically different, they still represent the same 'physical' example).

4.3.3 RESULTS

The detecting program was run at a very high sensitivity (low detection threshold), resulting in a high number of false alarms. The output was then subject to post-processing, which consisted of calculating selected classification measures at different detection thresholds. The final threshold values were selected to minimize the total number of false rejections and false alarms (it is currently assumed that each vocalization model can have a distinct threshold value, but a 'global' threshold was also separately determined).

A number of various digital signal processor (DSP) parametrization and modeling configurations were tested. The best results were obtained for static-only (no delta/acceleration coefficients) generalized MFCC features with ''warp factor'' set to 165 with additional maximum autocorrelation coefficient; the detector was based on HMMs, without explicit duration modeling.

Table 3. Obtained results in the preliminary experiments for the Vocalization Recognition System

Set	Recall	Precision	F1
Person_A.hums (pXV)	19.0	100	32.0
Person_A.laughs (pXV)	22.2	100	36.7
Person_A.wails (pXV)	88.9	75.0	81.4
Person_Fr.snores (pXV)	19.0	100	32.0
Person_F.o-oh (pXV)	100	100	100
pseudo-Cross-Val combined	49.4	74.1	59.3
pCV comb'd (glob.thresh.)	39.1	17.8	24.5
Person_A.hums (ToT)	100	100	100
Person_A.laughs (ToT)	100	100	100
Person_A.many (ToT)	100	100	100
Person_A.squeals (ToT)	100	100	100
Person_A.wails (ToT)	100	100	100
Person_F.snores (ToT)	28.6	100	44.4
Person_F.o-oh (ToT)	100	100	100
Train-on-Test combined	76.5	63.4	69.3
ToT comb'd (glob.thresh.)	73.5	16.9	27.5

4.4 PHYSIOLOGICAL PARAMETERS MONITORING SYSTEM

In this section, we present the results of those methods from Section 0 that we have already implemented.

4.4.1 MATERIALS AND SETUP

To measure the physiological parameters of the subjects in a non-contact way, we need the PPG signal measured with a contact device as the ground-truth. This is commonly done with a professional fingertip PPG sensor, which captures the precise waveform. By comparing our results with the ground-truth PPG signal returned from such a fingertip device, we evaluate how good our results are. Since the materials collected in the project so far do not have such ground truth, we used the COHFACE dataset[5] for the preliminary evaluation of our methods. This dataset consists of 160 videos from 40 different subjects. Each of the subjects is recorded four times at different lighting conditions. The faces of the subjects in this dataset cover approximately only 180x180 pixels, while in the related work these dimensions rise to 1280x720. However, this dataset is the best we encountered so far.

Pre-processing consisted of skin detection only. As explained in Section 3.4.1, we use two different methods to classify which of the pixels belong to the skin. The first one applies thresholds to classify the skin pixels. Figure 26 shows an example of the mask returned by this method.

Figure 26. Classified skin using the threshold method

The other skin classification method uses machine learning to classify which of the pixels belong to the skin. The results from this method are shown in Figure 27. The result in Figure 26 is better, but the thresholds used are fitted for our dataset and might not work as well on other video recordings.

Figure 27. Classified skin using the machine learning method

[5] https://www.idiap.ch/dataset/cohface - Idiap Research Institute 2010

4.4.2 RESULTS

Here we present the results of the methods for reconstruction of PPG signal. To evaluate them, we compare our output to the ground-truth PPG signal. We calculate Mean Squared Error (MSE) and Mean Absolute Error (MAE) to measure the correlation between the two signals. In addition, we use a peak detection method to estimate the heart rate and we compared the obtained heart rate to heart rate obtained from the ground-truth PPG. The results are presented in Table 4. In addition, the first 10 seconds of each person are shown in Figure 28-Figure 33.

Table 4. Results of the implemented methods

	MAE between signals	MSE between signals	Error between heart rates
Color-based method 1	0.040	0.150	42.000
Color-based method 2	0.040	0.156	42.000
Color-based method 3	0.160	0.040	20.750
Color-based method 5	0.160	0.400	11.730
Color-based method 6	0.040	0.010	8.750
Motion-based method	0.160	0.040	39.000

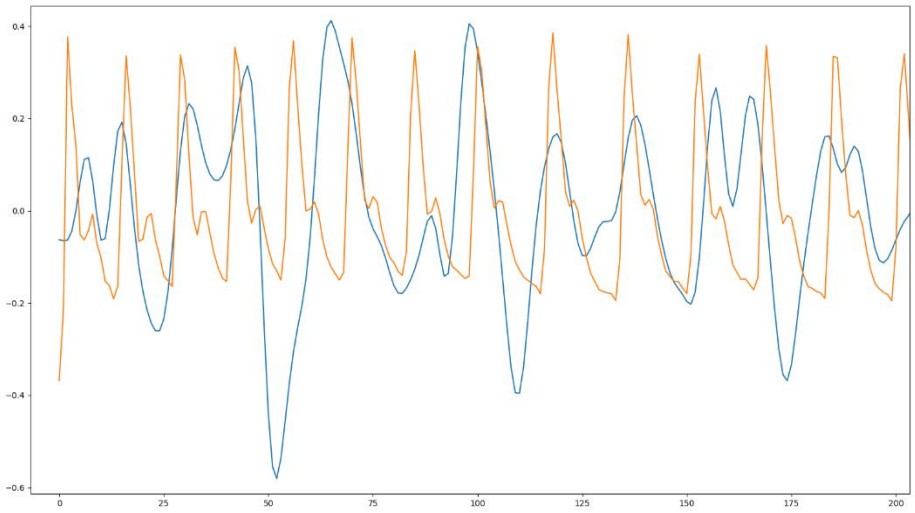

Figure 28. First 10 seconds from color-based method 1. Orange is the ground-truth PPG and blue is the reconstructed PPG. Y axis shows amplitude and x axis shows samples.

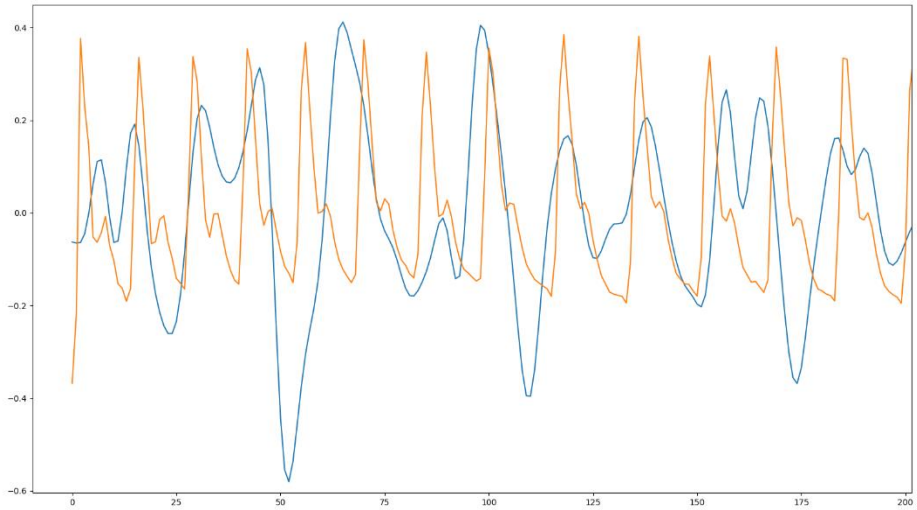

Figure 29. First 10 second of color-based method 2. Orange is the ground-truth PPG and blue is the reconstructed PPG. Y axis shows amplitude and x axis shows samples.

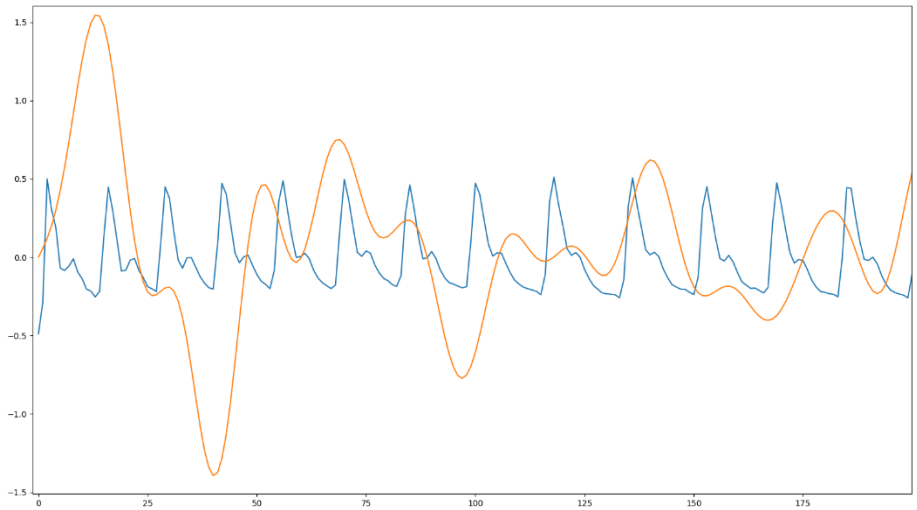

Figure 30. First 10 second of color-based method 3. Blue is the ground-truth PPG and orange is the reconstructed PPG. Y axis shows amplitude and x axis shows samples.

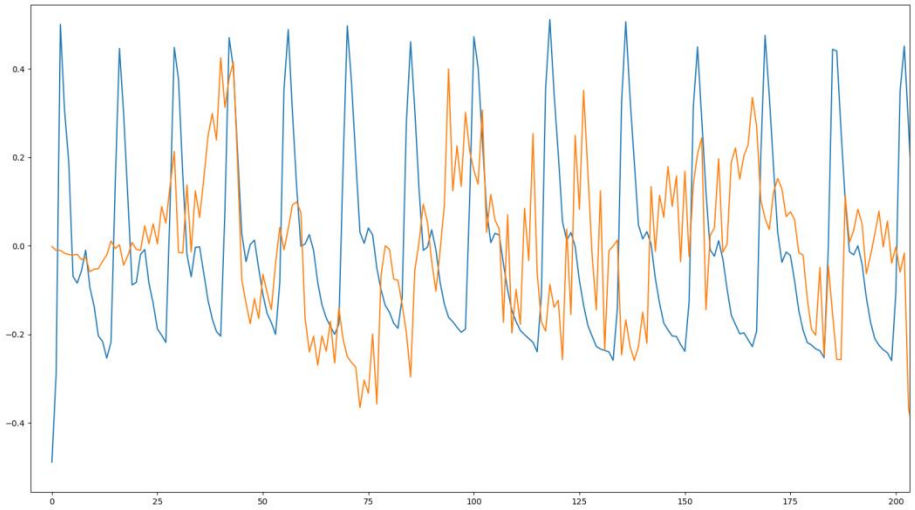

Figure 31. First 10 second of color-based method 5. Blue is the ground-truth PPG and orange is the reconstructed PPG. Y axis shows amplitude and x axis shows samples.

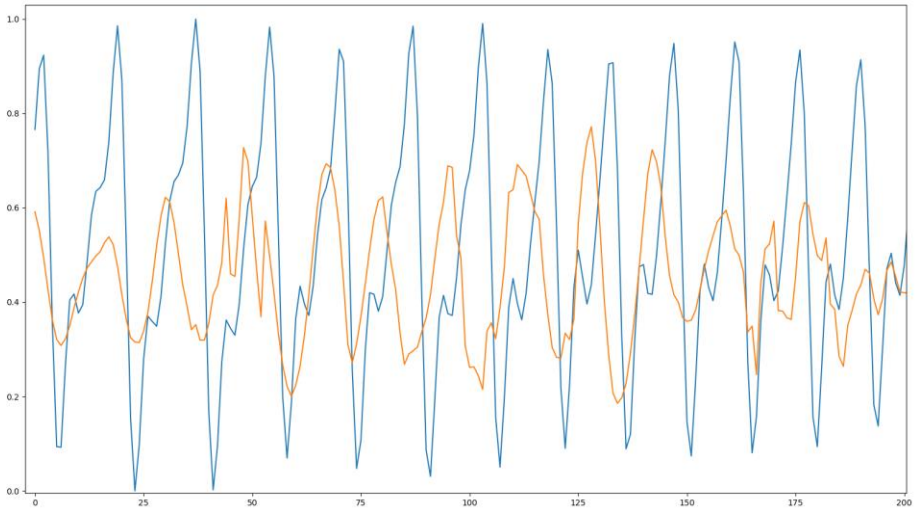

Figure 32. First 10 second of color-based method 6. Blue is the ground-truth PPG and orange is the reconstructed PPG. Y axis shows amplitude and x axis shows samples.

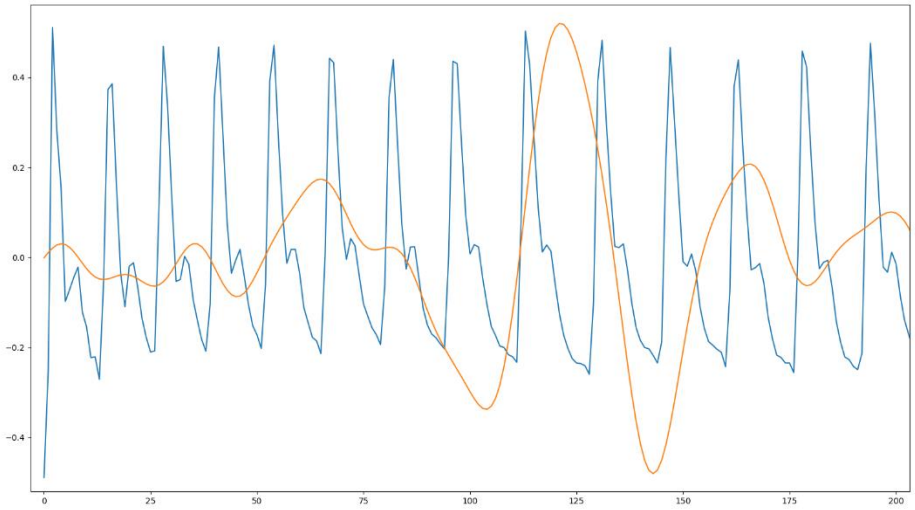

Figure 33. First 10 second of motion-based method. Blue is the ground-truth PPG and orange is the reconstructed PPG. Y axis shows amplitude and x axis shows samples.

These results indicate that the rPPG signal reconstructed from the video recording is noisy and therefore it would be hard to measure any physiological parameters of emotional state with it. The results of color-based method 6 are best in regards to the error between estimated and ground-truth HR, but still not particularly satisfactory. As mentioned in the beginning of this section, higher quality of the recordings could contribute to better results. We will also look for additional methods and improved implementation, as well as combinations of methods.

5 CONCLUSIONS

In this deliverable we presented results of the initial work performed on the design of the components for the recognition on non-symbolic behavioral signals of individuals with PMLD. This work was targeted at understanding the technical requirements concerning building these components, such as consideration of which existing methods of artificial intelligence could be adapted to execute the required functionality, and their practical use as part of the INSENSION system, such as hardware requirements. We summarize conclusions related to these requirements below.

As far as facial expression and body gesture recognition components are concerned, we carried out a set of initial experiments by using the OpenPose technology. The results suggest that this group of techniques is suitable to perform the recognizing task for both: facial expressions and human body gestures. However, a higher dataset is necessary to increase the recognition performance, as well as improving the accuracy of the developed models. Moreover, with regards to the gesture model, a distinction between upper and lower parts of the body should be made in order to enhance the model precision and deal with some complex and particular movements.

For the vocalization recognition system, the results of preliminary experiments suggest, unsurprisingly, a positive correlation between the number of training examples and the accuracy of detection. In addition, those vocalizations that seemed to be easily distinguishable to a human listener (*Person_A.wails*, *Person_F.o-oh*) tended to have higher automatic detection scores, as well. The future work will include primarily extending the parametrization procedures and feature vectors by new features reported across literature. In addition, early training phases, especially unsupervised determination of distinct emitting states, i.e. mixture clustering, may need to be looked into.

The preliminary results of the physiological parameters monitoring are not very good. The reconstruction of the PPG signal from video is nowhere near as mature as the recognition of gestures, facial expressions and vocalizations, so it is not surprising that the problem is proving more difficult. While there are some papers showing good results, there are also published claims that these results are difficult to reproduce. We will definitely try to improve upon these preliminary results in order to reduce the number of devices required for the final INSENSION system. Nevertheless, if needed a strategy to fall back to the contact sensor – a wristband – shall be used.

In summary, a wider experimentation should be performed in the next stages of the project.

With regards to the hardware requirements of the signals recognition subsystem, composed of the four recognition components, Table 5 collects the minimum requirements. It takes into account the requisites of each recognition component.

As it can be observed, the system requirements are relatively high because of the complexity of the problem, especially with relation to the gesture and facial recognition components. Regarding the materials, a large amount of recordings with precised annotations shall be needed for obtaining a good performance of the system. The optimal image resolution varies depending on the recognition component – a minimum can be established as 1280 x 720 pixels that corresponds to the highest identified requirement. The rest of the requisites seems to be compatible among the four recognition components. Preliminary experiments suggest that we may expect that these minimum requirements shall not need redefinition.

Table 5. Technical requirements of the behavioral signals recognition subsystem

System		
NVIDIA graphics card with at least 1.6 GB. At least 2.5 GB of free RAM memory. GPU version: at least 2 GB of free RAM memory. CPU version: around 8GB of free RAM memory. CPU with at least eight cores.		
Materials		
Gesture and facial expression	Vocalization	Physiological parameters
A wide variety of RGB images and video sequences. Minimum size of the person in the image: 1/3 total size of the image. Minimum image resolution of 640x480. Precised annotations.	Audio recordings mono PCM 16-bit, 16 kHz. Precise annotations.	50 fps recordings. 1280x720 pixels RGB recordings.
Environment		
Gesture and facial expression	Vocalization	Physiological parameters
Avoiding occlusions in face and body parts. Good lighting conditions.	Avoiding audibility of over persons with PLMD.	Good lighting conditions.

6 BIBLIOGRAPHY

[1] A. Kendon, “How gestures can become like words.,” Cross-cultural perspectives in nonverbal communication., 1988.

[2] R. Aigner, D. Wigdor, H. Benko, M. Haller, D. Lindlbauer, A. Ion, S. Zhao and J. T. K. V. Koh, "Understanding Mid-Air Hand Gestures: A Study of Human Preferences in Usage of Gesture Types for HCI," Microsoft Research Technicak Report MSR-TR-2012-111, 2012.

[3] H. Gunes, M. Piccardi and T. Jan, “Face and Body Gesture Recognition for a Vision-Based Mutimodal Analyzer,” in VIP'05 Proceedings of the Pan-Sydney area workshop on Visual information proceesing, Sydney, Australia, 2004.

[4] C. Hummels and P. Stappers, “Meaningful gestures for human computer interaction: beyond hand gestures,” in Proc. Third International Conference on Automatic Face and Gesture Recgonition, Nara, Japan, 1998.

[5] B. W. Hwang, S. Kim and S. W. Lee, “A Full-Body Gesture Database for Automatic Gesture Recognition,” in Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR’06), 2006.

[6] M. Turk, “Chapter 9. Gesture Recognition,” in Handbook of Virtual Environments: Design, Implementation and Applications, 2015.

[7] A. Bobick, “Movement, activity, and action: the role of knowledge in the perception of motion.,” in Royal Society Workshop on Knowledge-based Vision in Man and Machine, London, England, 1997.

[8] P. T. Hai and H. H. Kha, “An Efficient Star Skeleton Extraction for Human Action Recognition Using Hidden Markov Models,” in IEEE Sixth International Conference on Communications and Electronics (ICCE), Ha Long, Vietnam, 2016.

[9] Z. Cao, T. Simon, S. Wei and Y. Sheikh, “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” in IEEE Computer Vision and Pattern Recognition (CVPR 2017), 2017.

[10] S. Wei, V. Ramakrishna, T. Kanade and Y. Sheikh, “Convolutional pose machines,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[11] T. Simon, H. Joo, I. Matthews and Y. Sheikh, “Hand Keypoint Detection in Single Images using Multiview Bootstrapping,” in IEEE Computer Vision and Pattern Recognition (CVPR 2017), 2017.

[12] DeepEyes, «The DeepEyes Technology,» [En línea]. Available: https://www.deepeyes.co/.

[13] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar y L. Fei-Fei, «Large-scale video classification with convolutional neural networks,» de Computer Vision and Pattern Recognition, 2014.

[14] N. Neverova, C. Wolf, G. Taylor y F. Nebout, «ModDrop: Adaptive Multimodal Gesture Recognition,» *IEEE Transactions on Pattern Analysis and Machine Intelligence ,* vol. 38, nº 8, pp. 1692-1706, 2016.

[15] C. Sun y R. Nebatia, «Active: Activity concept transitions in video event classification,» de *International Conference on Computer Vision*, 2013.

[16] P. A. Viola y M. J. Jones, «Rapid object detection using a boosted cascade of simple features,» de *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[17] N. Dalal y B. Triggs, «Histograms of oriented gradients for human detection,» de *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005.

[18] S. Liao, W. Fan, A. Chung y D.-Y. Yeung, «Facial Expression Recognition using Advanced Local Binary Patterns, Tsallis Entropies and Global Appearance Features,» de *Proc. IEEE Inf. Conf. Image Process.*, 2006.

[19] M. Kyperountas, A. Tefas y I. Pitas, «Salient Feature and Reliable Classifier Selection for Facial Expression Classification,» *Pattern Recognition,* vol. 43, nº 3, pp. 972-986, 2010.

[20] P. Ekman, W. Friesen and J. Hager, The Facial Action Coding System, London: Weidenfel and Nicolson, 2002.

[21] P. Lucey, J. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett and K. Prkachin, "Automatically detecting pain in video through facial action units," *IEEE Trans. Syst., Man Cybern. B, Cybern.,* vol. 41, no. 3, pp. 664-674, Jun. 2011.

[22] M. F. Valstar, M. Pantic, Z. Ambadar and J. F. Cohn, "Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions," in *Proceedings of the 8th International Conference on Multimodal Interfaces (ICMI '06)*, Banff, Alberta, Canada., 2006.

[23] M. Valstar, B. Jiang, M. Mehu, M. Pantic and K. Scherer, "The first facial expression recognition and analysis challenge," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2011.

[24] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt and M. Pantic, "AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, Mountain View, California, USA, 2017.

[25] S. E., H. Gunes and A. Cavallaro, "Automatic Analysis of Facial Affect: A Survey of Registration, Representation and Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 37, no. 6, pp. 1113-1133, 2015.

[26] K. Sikka, T. Wu, J. Susskind y M. Bartlett, «Exploring bag of words architectures in the facial expression domain,» de *Proc. European Conference on Computer VIsion Workshops Demonstrations*, 2012.

[27] F. Long, T. Wu, J. R. Movellan, M. S. Bartlett y G. Littlewort, «Learning spatiotemporal features by using independent component analysis with application to facial expression recognition,» *Neurocomputing,* vol. 93, nº 0, pp. 126-132, 2012.

[28] P. Yang, Q. Liu y D. N. Metaxas, «Dynamic soft encoders patterns for facial event analysis,» Computer Vision Image Understanding, vol. 115, nº 3, pp. 456-465, 2011.

[29] O. Rudovic, V. Pavlovic y M. Pantic, «Multi-output Laplacian dynamic ordinal regression for facial expression recognition and intensity estimation,» de Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[30] B. Jiang, M. Valstar, B. Martinez y M. Pantic, «Dynamic appearance descriptor approach to facial actions temporal modelling,» IEEE Transactions on System Man Cybernetics. B, Cybernetic, vol. 44, nº 2, pp. 161-174, 2014.

[31] X. Ding, W.-S. Chu, F. De La Torre, J. Cohn y Q. Wang, «Facial Action Unit Event Detection by Cascade of Tasks,» de Proc. IEEE International Conference on Computer Vision, 2013.

[32] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie y M. Pantic, «AVEC 2011-The first international audio/visual emotion challenge,» de Processing International Conference Affective on Computer Intelligence Interaction, 2011.

[33] B. Schuller, M. Valstar, R. Cowie y M. Pantic, «AVEC 2012-The continuous audio/visual emotion challenge,» de Proc. ACM International Conference on Multimodal Interaction, 2012.

[34] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie y M. Pantic, «AVEC 2013-The continuous audio/visual emotion and depression recognition challenge,» de Proc. ACM Intelligence Conference on Multimodal Interfaces, 2013.

[35] T. Theodorou, I. Mporas y N. Fakotakis, «Audio Feature Selection for Recognition of Non-linguistic Vocalization Sounds,» de Artificial Intelligence: Methods and Applications, Cham, 2014.

[36] K. P. Truong y D. A. van Leeuwen, «Automatic discrimination between laughter and speech,» Speech Communication, vol. 49, nº 2, pp. 144-158, 2007.

[37] H. D. Tran y H. Li, «Sound Event Recognition With Probabilistic Distance SVMs,» IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, nº 6, pp. 1556-1568, Aug 2011.

[38] T. Drugman, J. Urbain y T. Dutoit, «Assessment of audio features for automatic cough detection,» de Proceedings of the 19th European Signal Processing Conference, {EUSIPCO} 2011, Barcelona, Spain, August 29 - Sept. 2, 2011, 2011.

[39] J. M. Liu, M. You, G. Z. Li, Z. Wang, X. Xu, Z. Qiu, W. Xie, C. An y S. Chen, «Cough signal recognition with Gammatone Cepstral Coefficients,» de 2013 IEEE China Summit and International Conference on Signal and Information Processing, 2013.

[40] M. K. Nandwana, A. Ziaei y J. H. L. Hansen, «Robust unsupervised detection of human screams in noisy acoustic environments,» de 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[41] B. H. Tracey, G. Comina, S. Larson, M. Bravard, J. W. Lopez y R. H. Gilman, «Cough detection algorithm for monitoring patient recovery from pulmonary tuberculosis,» de 33rd Annual International Conference of the {IEEE} Engineering in Medicine and Biology Society, {EMBC} 2011, Boston, MA, USA, August 30 - Sept. 3, 2011, 2011.

[42] J. M. Liu, M. You, Z. Wang, G. Z. Li, X. Xu y Z. Qiu, «Cough detection using deep neural networks,» de 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2014.

[43] D. Prylipko, B. Schuller y A. Wendemuth, «Fine-tuning HMMS for nonverbal vocalizations in spontaneous speech: A multicorpus perspective,» de 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012.

[44] P. Jancovic, M. Kokuer, M. Zakeri y M. Russell, «Bird species recognition using HMM-based unsupervised modelling of individual syllables with incorporated duration modelling,» de 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.

[45] S. Matos, S. S. Birring, I. D. Pavord y H. Evans, «Detection of cough signals in continuous audio recordings using hidden Markov models,» IEEE Transactions on Biomedical Engineering, vol. 53, nº 6, pp. 1078-1083, June 2006.

[46] M. A. T. Figueiredo y A. K. Jain, «Unsupervised Learning of Finite Mixture Models,» {IEEE} Trans. Pattern Anal. Mach. Intell., vol. 24, nº 3, pp. 381-396, 2002.

[47] T. S. Brandes, «Feature Vector Selection and Use With Hidden Markov Models to Identify Frequency-Modulated Bioacoustic Signals Amidst Noise,» IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, nº 6, pp. 1173-1180, Aug 2008.

[48] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci y A. Sarti, «Scream and gunshot detection and localization for audio-surveillance systems,» de Fourth {IEEE} International Conference on Advanced Video and Signal Based Surveillance, {AVSS} 2007, 5-7 September, 2007, Queen Mary, University of London, London, United Kingdom, 2007.

[49] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci y A. Sarti, «Scream and gunshot detection in noisy environments,» de 2007 15th European Signal Processing Conference, 2007.

[50] C. Clavel, T. Ehrette y G. Richard, «Events Detection for an Audio-Based Surveillance System,» de 2005 IEEE International Conference on Multimedia and Expo, 2005.

[51] F. Weninger, B. Schuller, M. Wollmer y G. Rigoll, «Localization of non-linguistic events in spontaneous speech by Non-Negative Matrix Factorization and Long Short-Term Memory,» de 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011b.

[52] F. Weninger y B. Schuller, «Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations,» de 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011a.

[53] V. Swarnkar, U. R. Abeyratne, Y. Amrulloh, C. Hukins, R. Triasih y A. Setyati, «Neural network based algorithm for automatic identification of cough sounds,» de 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013.

[54] J. Amoh y K. Odame, «Deep Neural Networks for Identifying Cough Sounds,» IEEE Transactions on Biomedical Circuits and Systems, vol. 10, nº 5, pp. 1003-1011, Oct 2016.

[55] T. Mikami, Y. Kojima, M. Yamamoto y M. Furukawa, «Automatic classification of oral/nasal snoring sounds based on the acoustic properties,» de 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012.

[56] S. Fruhholz, E. Marchi y B. Schuller, «The Effect of Narrow-Band Transmission on Recognition of

Paralinguistic Information From Human Vocalizations,» IEEE Access, vol. 4, pp. 6059-6072, 2016.

[57] M. Slaney y G. McRoberts, «Baby Ears: a recognition system for affective vocalizations,» de Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, 1998.

[58] A. Tariq and H. Ghafouri-Shiraz, "Vital signs detection using Doppler radar and continuous wavelet transform," in Proceedings of the 5th European Conference on Antennas and Propagation (EUCAP), 2011, pp. 285-288.

[59] U. Morbiducci, L. Scalise, M. De Melis and M. Grigioni, "Optical vibrocardiography: A novel tool for the optical monitoring of cardiac activity," Annals of Biomedical Engineering, vol. 35, pp. 45-58, 2007.

[60] C. H. Peters, E. D. ten Broeke, P. Andriessen, B. Vermeulen, R. C. Berendsen, P. F. Wijn and S. G. Oei, "Beat-to-beat detection of fetal heart rate: Doppler ultrasound cardiotocography compared to direct ECG cardiotocography in time and frequency domain," Physiological Measurement, vol. 25, p. 585, 2004.

[61] N. Bernacchia, L. Scalise, L. Casacanditella, I. Ercoli, P. Marchionni and E. P. Tomasini, "Non contact measurement of heart and respiration rates based on Kinect," IEEE International Symposium on Medical Measurements and Applications (MeMeA), pp. 1-5, 2014.

[62] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," IEEE Transactions on Biomedical Engineering, vol. 60, pp. 2878-2886, 2013.

[63] M.-Z. Poh, D. J. McDuff and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," Optics Express, vol. 18, pp. 10762-10774, 2010.

[64] M.-Z. Poh, D. J. McDuff and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," IEEE Transactions on Biomedical Engineering, pp. 7-11, 2011.

[65] M. Lewandowska, J. Rumiński, T. Kocejko and J. Nowak, "Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity," in Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS), 2011, pp. 405-410.

[66] W. Wang, S. Stuijk and G. De Haan, "Exploiting spatial redundancy of image sensor for motion robust rPPG," IEEE Transactions on Biomedical Engineering, vol. 62, pp. 415-425, 2015.

[67] W. Wang, S. Stuijk and G. De Haan, "A novel algorithm for remote photoplethysmography: Spatial subspace rotation," IEEE Transactions on Biomedical Engineering, vol. 63, pp. 1974-1984, 2016.

[68] O. R. Patil, Y. Gao, B. Li and Z. Jin, "CamBP: a camera-based, non-contact blood pressure monitor," in Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the ACM International Symposium on Wearable Computers, 2017, pp. 524-529.

[69] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," ACM Transactions on Graphics, vol. 31, no. 4, 2012.

[70] G. Heusch, A. Anjos and S. Marcel, "A reproducible study on remote heart rate measurement," CoRR, vol. abs/1709.00962, 2017.

[71] G. Balakrishnan, F. Durand and J. Guttag, "Detecting pulse from head motions in video," in

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3430-3437.

[72] M. Garbey, N. Sun, A. Merla and I. Pavlidis, "Contact-free measurement of cardiac pulse based on the analysis of thermal imagery," IEEE Transactions on Biomedical Engineering, vol. 54, pp. 1418-1426, 2007.

[73] R. Murthy and I. Pavlidis, "Noncontact measurement of breathing function," IEEE Engineering in Medicine and Biology Magazine, vol. 25, pp. 57-67, 2006.

[74] V. Engert, A. Merla, J. A. Grant, D. Cardone, A. Tusche and T. Singer, "Exploring the use of thermal infrared imaging in human stress research," PloS one, vol. 9, no. 3, p. e90782, 2014.

[75] J. Stemberger, R. S. Allison and T. Schnell, "Thermal imaging as a way to classify cognitive workload," in Proceedings of the Canadian Conference on Computer and Robot Vision (CRV), 2010, pp. 231-238.

[76] B. A. Rajoub and R. Zwiggelaar, "Thermal facial analysis for deception detection," IEEE Transactions on Information Forensics and Security, vol. 9, pp. 1015-1023, 2014.

[77] A. Mesleh, D. Skopin, S. Baglikov and A. Quteishat, "Heart rate extraction from vowel speech signals," Journal of Computer Science and Technology, vol. 27, no. 6, pp. 1243-1251, 2012.

[78] N. Selvaraj, A. Jaryal, J. Santhosh, K. K. Deepak and S. Anand, "Assessment of heart rate variability derived from finger-tip photoplethysmography as compared to electrocardiography," Journal of Medical Engineering and Technology, vol. 32, pp. 479-484, 2008.

[79] K. H. Chon, S. Dash and K. Ju, "Estimation of respiratory rate from photoplethysmogram data using time--frequency spectral estimation," IEEE Transactions on Biomedical Engineering, vol. 56, pp. 2054-2063, 2009.

[80] A. Johansson, "Neural network for photoplethysmographic respiratory rate monitoring," Medical and Biological Engineering and Computing, vol. 41, pp. 242-248, 2003.

[81] X. Teng and Y. Zhang, "Continuous and noninvasive estimation of arterial blood pressure using a photoplethysmographic approach," Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 4, pp. 3153-3156, 2003.

[82] P. H. Jones and W.-M. Wang, "Method of measuring blood pressure with a photoplethysmograph". United States Patent US5269310A, 1993.

[83] F. Mokhayeri, M.-R. Akbarzadeh-T and S. Toosizadeh, "Mental stress detection using physiological signals based on soft computing techniques," in Proceedings of the 18th Iranian Conference of Biomedical Engineering (ICBME), 2011, pp. 232-237.

[84] K.-S. Yoo and W.-H. Lee, "Mental stress assessment based on pulse photoplethysmography," in Proceedings of the 15th International Symposium on Consumer Electronics (ISCE), 2011, pp. 323-326.

[85] F. Bousefsaf, C. Maaoui and A. Pruski, "Remote assessment of physiological parameters by non-contact technologies to quantify and detect mental stress states," in Proceedings of the International Conference on Control, Decision and Information Technologies (CoDIT), 2014, pp. 719-723.

[86] A. Koenig, X. Omlin, L. Zimmerli, M. Sapa, C. Krewer, M. Bolliger, F. Müller and R. Riener, "Psychological state estimation from physiological recordings during robot-assisted gait rehabilitation," Journal of Rehabilitation Research and Development, vol. 48, no. 4, pp. 367-386, 2011.

[87] S. Jerritta, M. Murugappan, R. Nagarajan and K. Wan, "Physiological signals based human emotion recognition: A review," in Proceedings of 7th International Colloquium on Signal Processing and its Applications, 2011, pp. 410-415.

[88] H. Leng, Y. Lin and L. A. Zanzi, "An experimental study on physiological parameters toward driver emotion recognition," in Ergonomics and Health Aspects of Work with Computers, Springer, Berlin, Heidelberg, 2007, pp. 237-246.

[89] C. Collet, E. Vernet-Maury, G. Delhomme and A. Dittmar, "Autonomic nervous system response patterns specificity to basic emotions," Journal of the Autonomic Nervous System, vol. 62, no. 1-2, pp. 45-57, 1997.

[90] R. W. Picard, J. Healey and E. Vyzas, "Toward machine emotional intelligence analysis of affective physiological state," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 10, pp. 1175-1191, 2001.

[91] Z. Guendil, Z. Lachiri, C. Maaoui and A. Pruski, "Emotion recognition from physiological signals using fusion of wavelet based features," in Proceedings of the 7th International Conference on Modelling, Identification and Control, 2015, pp. 1-6.

[92] B. Rotter, G. Kane and B. Gallé, "Nichtsprachliche Kommunikation: Erfassung und Förderung. Geistige Behinderung," vol. 31, pp. 1-26, 1992.

[93] C. Rowland and M. Fried-Oken, "Communication Matrix: A clinical and research assessment tool targeting children with severe communi-cation disorders," Journal of Pediatric Rehabilitation Medicine, vol. 3, pp. 319-329, 2010.

[94] "ELAN," [Online]. Available: https://tla.mpi.nl/tools/tla-tools/elan/ .

[95] N. Twilhaar and B. van den Bogaerde, Concise Lexicon for Sign Linguistics, Amsterdam: John Benjamins Publishing Company , 2016.

[96] "OpenPose," [Online]. Available: http://www.consortium.ri.cmu.edu/projOpenPose.php.

[97] "Boost," [Online]. Available: https://www.boost.org/doc/libs/.

[98] Y. Jia, E. Shelhame, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding.," ArXiv e-prints. Techreport for the Caffe software at http://github.com/BVLC/Caffe/, 2014.

[99] «Cuda,» [En línea]. Available: https://developer.nvidia.com/cuda-zone.

[100] «NVIDIA,» [En línea]. Available: https://developer.nvidia.com/cudnn.

[101] «OpenCV,» [En línea]. Available: https://opencv.org/.

[102] «gflags,» [En línea]. Available: https://github.com/gflags/gflags.

[103] «protobuf,» [En línea]. Available: https://developers.google.com/protocol-buffers/.

[104] A. S. C. Toshev, "Deeppose: human pose estimation via deep neural networks," in IEEE Computer Vision and Pattern Recognition, 2014.

[105] A. Howard, Elementary Linear Algebra, John Wiley & Sons, 2010.

[106] A. Hyvarinen, «Fast ICA for noisy data using Gaussian moments,» de Proceedings of the 1999 IEEE International Symposium on Circuits and Systems, IEEE, 1999, pp. 57-61.

[107] G. Farnebäck, «Two-frame motion estimation based on polynomial expansion,» de Proceedings of the Scandinavian Conference on Image Analysis, 2003, pp. 363-370.

[108] Z. Kalal, K. Mikolajczyk y J. Matas, «Forward-backward error: Automatic detection of tracking failures,» de Proceedings of the 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 2756-2759.

[109] B. D. Lucas y T. Kanade, «An iterative image registration technique with an application to stereo vision,» de Proceedings of the International Joint Conference on Artificial Intelligence, 1981, pp. 674-679.

[110] P. Vos, P. De Cock, K. Petry, W. Van Den Noortgate y B. Maes, «Do you know what I feel? A first step towards a physiological measure of the subjective well-being of persons with profound intellectual and multiple disabilities,» Journal of Applied Research in Intellectual Disabilities, vol. 23, nº 4, pp. 366-378, 2010.

APPENDIX A. SUMMARY OF DATA GENERATED/COLLECTED WITHIN INSENSION PROJECT

Area	Main Category	Subcategories	Tiers	TS[6]	ES[7]
1. Communication and Inner States	Communication		- comment - demand - protest		X
	Inner States		- pleasure - displeasure - neutral		X
2. Behaviours	Vocalizations				X
	Facial Expressions	2.1.1. Appearance of Eyes	- eye contact - widened eyes - closed eyes - sleepy eyes - "smiling" eyes - winking - tears - eyebrow movement - frown	X	
		2.1.2. Movement/ Appearance of Jaw	- drooping - grinding - biting	X	
		2.1.3. Movement of Nose and Mouth	- nose movements - lip movements - tongue movements - tongue out-side - loss of saliva - corners of mouth up - corners of mouth down	X	
	gestures/ movements	2.1.4. Body Posture	- rigid - floppy - jerky - restless - leans to side	X	
		2.1.5. Movement of Head	- rigid - floppy - shaking - nodding - turns head to side - leans to side	X	
		2.1.6. Movement of Left Arm	- rigid - floppy	X	

[6] Time sampling (TS)
[7] Event sampling (ES)

			- jerky - outstretched arm - flexed arm - raising arm - arm close to the body		
		2.1.7. Movement of Right Arm	rigid - floppy - jerky - outstretched arm - flexed arm - raising arm - arm close to the body	X	
		2.1.8. Movement of Left Hand	hand on hand hand on head	X	
		2.1.9. Movement of Right Hand	hand on hand hand on head	X	
		2.1.10. Movement of Left Leg	outstretched leg rubbing flexed leg raising leg	X	
		2.1.11. Movement of Right Leg	outstretched leg rubbing flexed leg raising leg	X	
		2.1.12. Movement of Left Foot	foot on foot	X	
		2.1.13. Movement of Right Foot	foot on foot	X	
		2.1.14. Specific Movements	specific movements		X
3. Context	Persons	3.1.1. Person A	present involved in interaction with test-person involved in interaction with test-person and object		X
		3.1.2. Person B	present involved in interaction with test-person involved in interaction with test-person and object		X
		3.1.3. Person C	present involved in interaction with test-person involved in interaction with test-person and object		X
		3.1.4. Person D	present		X

			involved in interaction with test-person involved in interaction with test-person and object		
	Objects	3.1.5. Object A	test-person acts with object		X
		3.1.6. Object B	test-person acts with object		X
		3.1.7. Object C	test-person acts with object		X
		3.1.8. Object D	test-person acts with object		X
	Background	3.1.9. Sounds	loud (volume) medium (volume) faint (volume) music human voices other noises		X
		3.1.10. Light Conditions	bright dark change in brightness		X