Health, demographic change and wellbeing

Personalising health and care: Advancing active and healthy ageing

H2020-PHC-19-2014

Research and Innovation Action

Deliverable 3.2

User state modelling and collaborative platform localisation:

Semantic interpretation of user status and activities based on detailed user models and FriWalk localisation based on multiple-modalities and across platforms

Deliverable due date: 01.02.2017	**Actual submission date: 08.03.2017**
Start date of project: February 1, 2015	**Duration: 42 months**
Lead beneficiary for this deliverable: UNITN	**Revision: 1.0**
Radu L. Vieru (UNITN), Daniele Fontanelli (UNITN), David Macii (UNITN), Payam Nazemzadeh (UNITN), Pashalis Padeleris (FORTH), Stefan Wakolbinger (SIEMENS), Johannes Scheiblauer (SIEMENS), Josef Alois Birchbauer (SIEMENS)	
Internal reviewer: Antonis Argyros (FORTH), Josef Alois Birchbauer (SIEMENS),	

The research leading to these results has received funding from the European Union's H2020 Research and Innovation Programme - Societal Challenge 1 (DG CONNECT/H) under grant agreement n°643644

Dissemination Level		
PU	Public	**x**
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contents

EXECUTIVE SUMMARY 4
CHAPTER 1 5
INTRODUCTION 5
1.1 ROLE OF THIS DELIVERABLE WITHIN WP3 5
1.2 DELIVERABLE OBJECTIVES 5
1.3 DELIVERABLE ORGANIZATION 5
CHAPTER 2 6
USER-CENTRIC SENSING (BIOSIGNALS / BIOMETRICS) 6
2.1 EMOTIONAL VALENCE AND PAIN ESTIMATION 6
2.1.1 PAIN LEVEL ESTIMATION 6
2.1.2 EMOTIONAL VALENCE 7
2.2 NON-CONTACT HR MEASUREMENTS 9
2.3 GAIT ANALYSIS 13
2.3.1 OPTIMIZATION 13
2.3.2 TIP TOE STAND DETECTION 13
2.3.3 INTEGRATION OF NEW CAMERA DEVICES 14
2.3.4 INTRODUCTION OF AN INTERMEDIATE LEVEL REPRESENTATION 16
2.4 *FRIWALK* BODY DETECTOR MODULE 18
2.4.1 EVALUATION OF THE FORTH HUMAN BODY TRACKER (*FHBT*) 18
2.4.1.1 HUMAN BODY MODEL 19
2.4.1.2 DATASET & EVALUATION METRICS 21
2.4.1.3 COMPARATIVE EVALUATION 22
2.4.1.4 *FHBT* LIMITATIONS 25
2.4.1.5 CLINICAL DATASETS 26
2.4.1.6 TASK-SPECIFIC BODY TRACKING 26
2.5 MECHANICAL SENSORS FOR USER DETECTION 28
2.6 CLINICAL DATA EXPLORER 29
CHAPTER 3 31
ACTIVITY ANALYSIS IN THE CLINICAL SCENARIO 31
3.1 SPPB TEST 31
3.1.1 BALANCE TEST 32
3.1.2 GAIT SPEED 33
3.1.3 CHAIR STAND TEST 34
3.1.4 ISOMETRIC/ISOTONIC EXERCISE 35
CHAPTER 4 36
USER STATE MODELING (USM) 36
4.1 INTRODUCTION 36
4.2 INPUTS, INTERFACES: EXTERNAL SPECIFICATIONS 36
4.3 GENERAL STRUCTURE 38
4.4 DETAILED STRUCTURE & INTERNAL SPECIFICATIONS 39
CHAPTER 5 45
COLLABORATIVE LOCALISATION 45
5.1 SENSING TECHNOLOGIES AND ALGORITHM DESCRIPTION 45

5.1.1 Models description 46
5.1.2 Cooperative Localisation Algorithm 49
5.1.3 Results 51
5.2 Optimal QR code placement 54
5.2.1 Problem formulation 55
5.2.1 Simulation results and analysis 57
5.3 Application of Optimal Placement for Collaborative Localisation in a Public Scenario 58
5.3.1 Optimal placement with trajectory observations 59
CHAPTER 6 64
RELATION WITH OTHERS WPS 64
BIBLIOGRAPHY 65

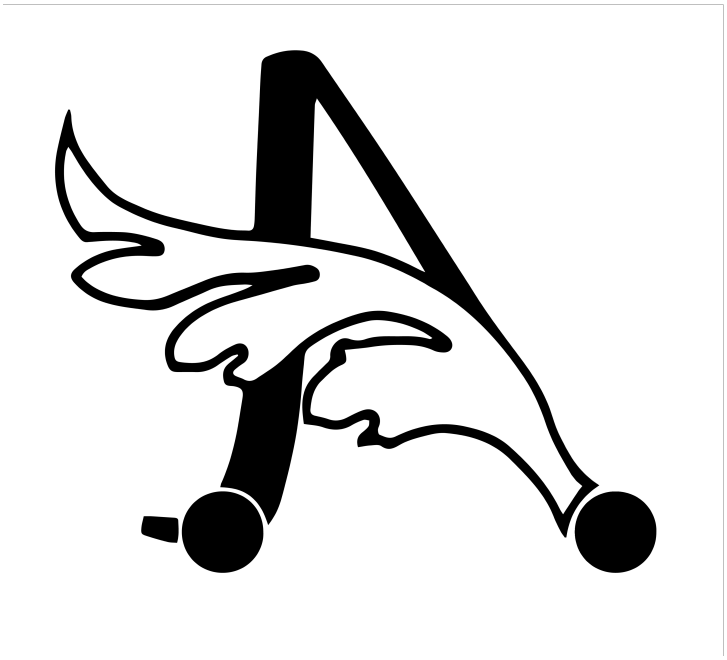
Executive Summary

In this deliverable we present the set of algorithms and sensors currently adopted on the *FriWalk* and developed within Task 3.1 – "Perceiving the user state". This deliverable refers to the following WP3 objectives:

- **Objective 3.1**: "*360 degree user perception and modelling*" - Perception of the user, including location, physiological and emotional state, level of attention, as well as her/his current activities/actions;
- **Objective 3.3**: "*Collaborative sensing*" - Cross-platform sensing and information sharing allows for insights beyond the viewing angle of a single walker.

To address Objective 3.1, we have studied solutions for sensing the biosignals, and define the associated biometrics, by means of the synergistic use of contactless sensors and mechanical sensory data of the *FriWalk*. We then fully exploit the motion capabilities of the ACANTO platform and the user centric sensing to conceive automated therapeutic exercises to be carried out in the clinical scenario. Due to the relatively high amount of information that can be collected from the available sensors, a User State Model is designed and implemented to extract meaningful data and measure the effectiveness of the planned exercise. This is of primarily importance for rating the non-clinical scenario exercises, since usually the therapeutic experts define the clinical scenario exercises with proper target metrics.

Related to Objective 3.3 is the work carried out for collaborative platform localisation, which presents an effective way to share the information among the team of *FriWalks* carrying out a group activity. For localisation, since absolute reference is needed, we propose an experimental study to minimize the number of landmarks to be deployed in a real environment taking into account the accuracy limits of the single platform.



Chapter 1
Introduction

1.1 Role of this deliverable within WP3

One of the key ideas of ACANTO is to learn as much as possible about the user of the *FriWalk without* the necessity to have this information actively provided by the user, since we want to ease the burden for our target group and not to pose an additional challenge. In other words this means continuous observation and perception of the user's state. Some of the observations will be relevant only at the time of the measurement, some will be meaningful by aggregation over a longer period of time, some of them indicate physiological conditions with medical relevance (e.g. with respect to therapeutic goals) while others address the motivational level or mood of the person. In any case, the means to gather all this information are the sensors. These sensors are deployed on the *FriWalk*/*FriTab* or alternatively also on the user, while the latter option is considered very carefully since we want to acquire data in an utmost non-obtrusive way whenever possible.

The purpose of user-centric sensing is at least twofold: First the derivation of automated activity analysis in the context of therapeutic exercises to support our clinical scenario, and second the aggregation of the sensory information to build up a semantically meaningful user state model that allows for automatic activity evaluation and measuring the impact of activities on the user's well being.

Since the user is logically "linked" to the *FriWalk* we can implicitly localize the user via platform localisation. However, the platform localisation is different from the perspective that it can use intrinsic (e.g. odometry) as well as extrinsic (e.g. GPS) cues. Furthermore, localisation can also be performed in a relative manner, i.e. with respect to other *FriWalk* units by introducing of a novel concept of collaborative localisation that reflects an aim of ACANTO in a very natural manner: fostering group activities and social contacts amongst older adults.

1.2 Deliverable objectives

The aim of this deliverable can be summarized as follows:

- Identify suitable solutions for **sensing** of **biosignals and biometrics** in the context of *FriWalk*/*FriTab*.
- Exploitation of user-centric sensing for **activity analysis in the clinical scenario** (i.e. automated therapeutic exercises).
- Conception and implementation of a **user state model** (USM) that is derived from the sensory information
- Implementation and evaluation a novel concept for **collaborative** or **synergic localisation** of an ensemble of *FriWalk* units.

1.3 Deliverable organization

The rest of the deliverable is organized as follows: In chapter 2 we present various options on user-centric sensing including non-contact measurements for emotional valence, pain estimation, heart rate, gait pattern and human body analytics. In addition we present a unified graphical user interface for presenting this plurality of information for the purpose of clinical data exploration. In the subsequent chapter 3 the sensory information is used to build automated therapeutic exercises. Chapter 4 presents the conception and implementation of a semantically meaningful users state model (USM) derived from the plurality of sensory and intermediate level information presented in chapter 2. Chapter 5 reports on concepts for localizing the *FriWalk* - and thus implicitly the user - in the environment. In particular chapter 5 elaborates on a novel concept for *collaborative* or *synergic* localisation, on the minimization of landmark placements and on an outlook for the actual application of the collaborative platform localisation in public scenarios.

Chapter 2
User-Centric Sensing (Biosignals / Biometrics)

2.1 Emotional Valence and Pain Estimation

In ACANTO, we are interested in building a user state model as rich as possible, using technologies with a low invasiveness mark. This includes both aspects related to the general emotional state of the user (mid-long term), as well as more short-term physiological condition aspects showing some medical relevance. Of course, when it comes to analysing the emotional state, the face is known to be the index of the mind, meaning that one's mental constitution is exhibited by his facial features. We argue that, while this is a generous statement, some degree of information is indeed encoded into the appearance of the face. We leverage the current state of research in face analysis to develop solutions for estimating the level of pain and the emotional valence/arousal by means of facial analysis. Pain estimation and monitoring is of great value for the clinical scenario, especially while patients are performing rehabilitation exercises. The emotional valence/arousal, on the other side, bring substantial colour in painting the long term user state model.

The proposed solutions are based on OpenFace [69], an open source tool implementing state-of-the-art computer vision and machine learning techniques for facial behaviour analysis. OpenFace includes components for facial landmark detection, head pose tracking and facial action unit detection and intensity estimation. All these components have been validated in extensive experiments, thus forming a solid block to build upon. In particular, we are interested in evaluating the facial action unit intensity and head pose estimators as feature vectors for subsequent face analysis tasks.

2.1.1 Pain Level estimation

Automatic pain recognition has recently become a valuable component in applications from health care, ranging from monitoring patients in intensive care units, rehabilitation scenarios or assessing of chronic lower back pain [7]. Current research on automatic pain detection is based on automatic analysis of facial expressions, since it has been shown that facial cues are very informative for pain detection [10]. Based on this observation, we propose a simple, yet effective solution for automatic pain level estimation using as features facial action unit intensities. We validate our approach on a publicly available dataset [4] containing videos of subjects experiencing pain while being asked to perform physical exercises.

Figure 1: Examples from UNBC-McMaster Pain Shoulder Archive. The first two rows are labelled at their peak pain intensity frame with a PSPI score of 6, whereas in the last row, the peak pain intensity frame is labelled with a PSPI score of 14.

Our system is based on Random Regression Forests [79], widely used for their computational efficiency and their powerful feature selection mechanism. We grow each tree by sequentially splitting non-leaf nodes following a classical information gain maximization principle. Assuming

$V = \{v_1, v_2, ..., v_N\}$ is one N-dimensional feature vector, we define a set of K binary tests $\Theta = \{t_1, t_2, ..., t_K\}$, where $t_k : v_i > \tau$. The values of the index i and threshold τ are randomly generated. Additionally, trees are grown following a bootstrap aggregating strategy, such that to avoid over fitting. The winning binary test in Θ is selected as the one that maximizes the information gain associated to the splitting output. The process continues until the stopping criterion is met (i.e. a predefined minimum amount of samples reach the splitting node), at which point a leaf node is formed. Leaves store the class posterior probabilities computed from the samples that reach them.

We train and validate our system on UNBC-McMaster Shoulder Pain Expression Database [4], a dataset containing 200 video sequences from 25 different subjects showing participant's faces (who were suffering from shoulder pain) while performing a series of active and passive physical tests using their affected and unaffected limbs on two separate occasions (see Figure 1 for some examples). A total of over 48K frames were labelled in terms of action unit presence (FACS) as well as the Prkachin and Solomon Pain Intensity (PSPI) metric [6]. PSPI gives a pain intensity estimate on a scale between 0 (no pain) and 15 (maximum intensity) based on psychometric properties of facial expressions.

The baseline system reported in [4] solves a binary classification problem in which the positive class is considered to be any frame labelled with a PSPI score different than 0. Different feature sets are used for classification, namely SPTS – encoding shape parameters of AAMs fitted onto subjects' faces and CAPP – accounting for appearance parameters of the same models. Experiments are carried out on a leave-one-subject-out fashion and, as performance measure, the Area under the ROC curve (AUC) is reported. For the same problem formulation, we train our random forest classifier containing 100 trees, which we stop growing below 1000 samples per node. The leaves store class probability measures derived from the samples that reach them. As features, we use the 17 AU intensity predictions estimated using [69]. Results are depicted in Table 1. As seen in the table, our RF-based approach generates better results (average AUC of 82%, which corresponds to an average Equal Error Rate - EER of 0.38) than any of the two feature types derived from the AAMs, but it is not able to overcome the fusion of the two.

Classifier	Linear SVM [4]			RF(our implementation)
Features	SPTS	CAPP	SPTS + CAPP	AU Reg Scores
AUC (%)	76,9	80,9	83,9	82,0

Table 1: Pain intensity estimation results on UNBC-McMaster Database

2.1.2 Emotional Valence

In psychological terms, the emotional valence is a measure of the intrinsic attractiveness or averseness towards an event, object or situation. Emotional arousal, on the other side, is a measure for the condition of sensory alertness, mobility and readiness to respond to a stimulus. Amongst the channels that encodes emotional information, the face provides an important mean for communicating both affect valence and arousal.

We propose to solve the problem of estimating the emotional state using facial cues, by exploiting existing work in facial expression recognition and by noticing there has been a significant body of research [8][9] focusing on linking the space of discrete emotional states with the continuous AV space. It all started with a study performed by J.A. Russel [8] who discovered that emotion-related words coming from five different languages share pan cultural properties, in the sense they are occupying similar positions in the arousal-valence circular space. This means that concepts such as happiness, fear, anger, sadness, etc., they all have a well-defined place on the AV circumplex model (see Figure 2).

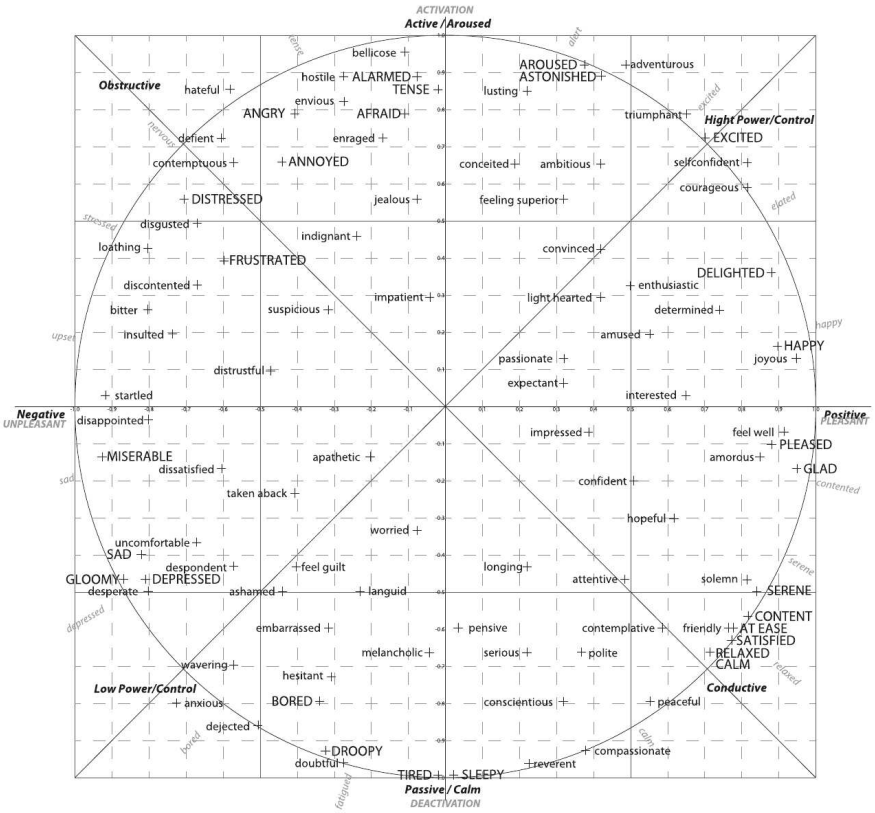


Figure 2: The two-dimensional circumplex space model. Image taken from [5] by combining data from [9] and [8]

Based on the above studies and motivated by the lack of existing face datasets annotated with AV labels, we propose to learn a model for predicting discrete emotional states and, subsequently, map the discrete predictions onto the circumplex space model. More specifically, we train a RF classifier, as previously explained, using as features the action unit intensity estimates from [69] on data coming from two well known publicly available facial expression datasets: CK+ [3] and BU3DFE [11]. Both these datasets contain face images labelled with the 6 universal facial expressions plus the neutral one. After filtering out the contempt class from CK+ and selecting the last two intensity facial expressions from BU3DFE, we ended up having a little over 1700 samples for training/evaluation. We trained RFs containing 100 tree, following a subject-independent 10-fold cross-validation, for which results are reported in Table 2 (average normalized recognition rate) and Figure 3 (confusion matrix).

	Mean normalized recognition rate
Training set	87.63%
Test set	75.56%

Table 2: Average normalized recognition rate for facial expression recognition on both (joint) CK+ and BU3DFE

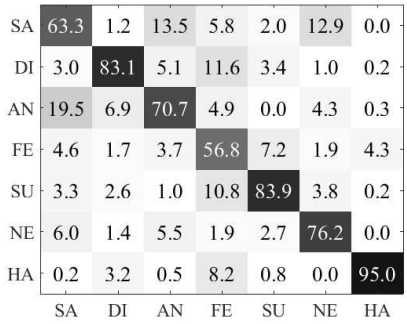


Figure 3: Confusion matrix obtained for facial expression recognition on both CK+ and BU3DFE

As expected, happiness and surprise are the most difficult to confuse expressions, result in line

with the majority of work in face analysis. The mapping values between categorical expression labels and corresponding continuous AV values are reported in Table 3As a post-processing stage, we apply temporal filtering to smooth the final predictions and integrate unstable behaviour.

	Sadness (SA)	Disgust (DI)	Anger (AN)	Fear (FE)	Surprise (SU)	Neutral (NE)	Happiness (HA)
Arousal	-0.40	0,49	0.79	0.79	0.02	0,00	0.17
Valence	-0.81	-0.68	-0.41	-0.12	-0.92	0.00	0.90

Table 3: Mapping values between discrete facial expression labels and continuous AV values following the circumplex space model [5]

2.2 Non-contact HR Measurements

One of the key aspects when designing the components to build up the user state model is non-obtrusiveness. Concerning the heart-related measurements, the COTS devices are generously well represented on the commercial market, ranging from uncomfortable but more accurate belts (e.g. Zephyr's Bioharness 3[1]) to more comfortable but slightly less accurate wristbands. In the latter case, the market has been recently flooded with photoplethysmography (PPG) devices[2] measuring the light absorption properties of the skin, modulated by blood flow. One of those devices is Fitbit's "Charge HR" which is used in ACANTO. Studies confirmed the correlation between the Fitbit-measured heart rate and the actual heart rate [81][82]. For resting and modest exercise – the scenarios we will mostly face in ACANTO – even stronger correlations were shown in the publications.

Our comparison on heart rate of the "Fitbit Charge HR" and the "Hexoskin smart garment" – with the latter treated as ground truth – showed positive results as well: In a "modest exercise scenario" the average difference between the two devices was 5.2 beats per minute, which is an acceptable deviation for the user state model. See Figure 4 for the heart rate time series.

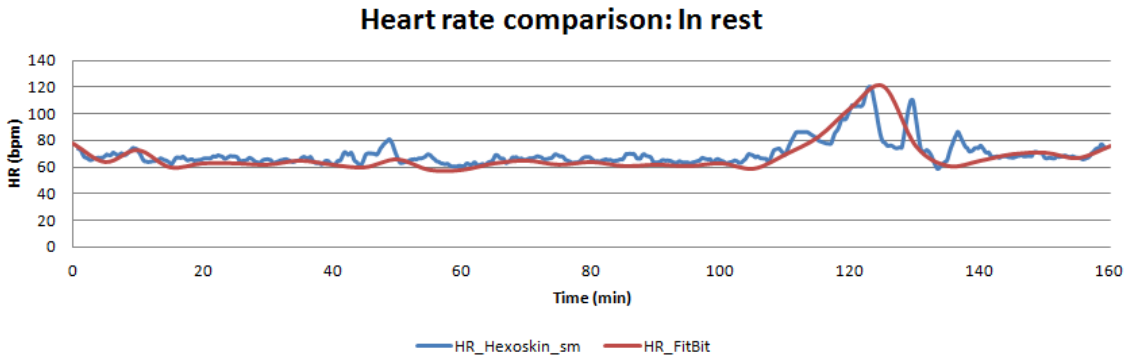


Figure 4: Comparison of the heart rate measurements of the "Hexoskin smart garment" (blue, treated as ground truth) and the "Fitbit Charge HR"(red) during a "modest exercise scenario", within which both the resting heart rate (between minutes 0 and 100) and the heart rate during modest exercise (between minutes 100 and 160) was measured

However, recent studies [22][19] have shown that, while practically invisible to the human observer, skin colour changes due to the blood flow can be captured on face videos and used to measure the heart rate (HR). Motivated by these findings, we set up to investigate, as a research

[1] https://www.zephyranywhere.com/

[2] Nowadays, wristband-like devices, such as smart watches, equipped with PPG modules have become ubiquitous. Many companies (e.g. Samsung, Apple, Empatica, Fitbit) have now such products available at affordable prices, most of the times accompanied by smartphone apps for data storage and processing.

direction, the possibility of estimating the heart rate from face videos.

Ever since the seminal work of [22], computer vision community has focused on improving this non-invasive HR estimation technique and, despite the considerable progress registered in the last years, there are still open issues. One particularly important problem is the lack of current approaches to operate in naturalistic conditions, dealing with variations caused by facial expressions, head movement and changing illumination conditions, such as in ACANTO. To address these challenges, we present a framework for remote HR estimation from visual data, able to output the HR measurement while simultaneously selecting the most reliable face regions. This strategy allows us to discard noisy features, due to spontaneous target's movements and facial expressions and thus, cast more reliable estimates on HR.

Estimating the heart rate from face videos involves capturing subtle changes (affecting colour [25] and motion [13] induced by the internal functioning of the heart. This has become possible thanks to significant improvements in face tracking and alignment techniques [12][21][17][26]. Previous work has successfully addressed HR estimation from videos only in laboratory controlled conditions, constraining the subjects from facial expressions and mimics [19][25][13]. Another limitation of previous approaches is the ability of providing short term estimates. Indeed [19][18][20] provide only average HR measurements over a long time interval, failing to capture short-term phenomena such as sudden changes due to specific medical conditions or emotional context, frequently met amongst the elderly population.

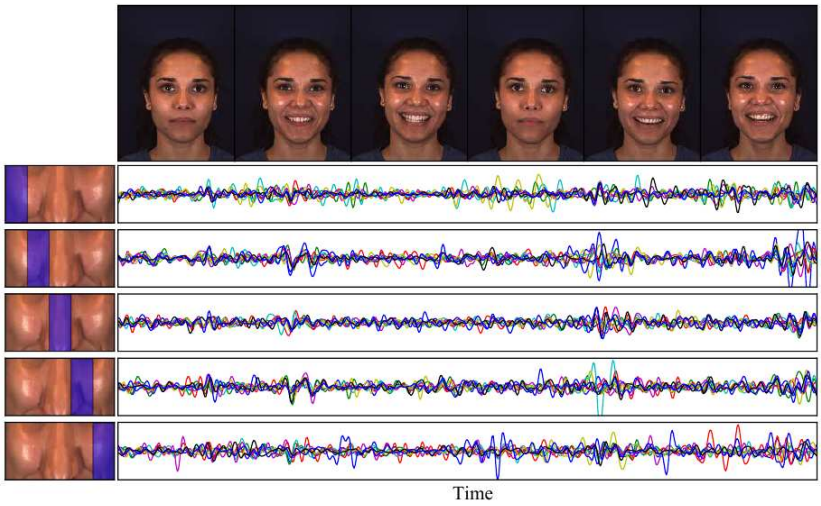


Figure 5: Given a video sequence, automatic HR estimation from facial features is challenging due to target motion and facial expressions. Facial features extracted over time in different parts of the face (purple rectangles) show different temporal dynamics and are subject to noise, as they are heavily affected by movements and illumination changes. We propose a novel approach to simultaneously estimate the HR signal and select the reliable face regions at each time for robust HR prediction.

In order to tackle the aforementioned limitations we introduce SAMC, a novel self-adaptive matrix completion approach for HR estimation from face videos and provide an extensive evaluation on two datasets: the MAHNOB-HCI, previously used for HR recognition research [18], and a spontaneous dataset with heart rate data and RGB videos (named MMSE-HR), which is a subset of the larger multimodal spontaneous emotion corpus (MMSE) [27] specifically targeted to challenge HR estimation methods. Inspired by previous methods, we track the face in a given video sequence, so to follow rigid head movements [18], and extract chrominance features [15] to compensate for illumination variations. Importantly, most previous approaches preselect a face region of interest (ROI) that is kept constant through the entire HR estimation. However, the region containing useful features for HR estimation is a priori different for every frame since major appearance changes are spatially and temporally localized (Figure 5). Therefore, we propose a principled data-driven approach to automatically detect the face parts useful for HR measurement, that is to estimate the time-varying mask of useful observations, selecting at each

frame the relevant face regions from the chrominance features themselves.

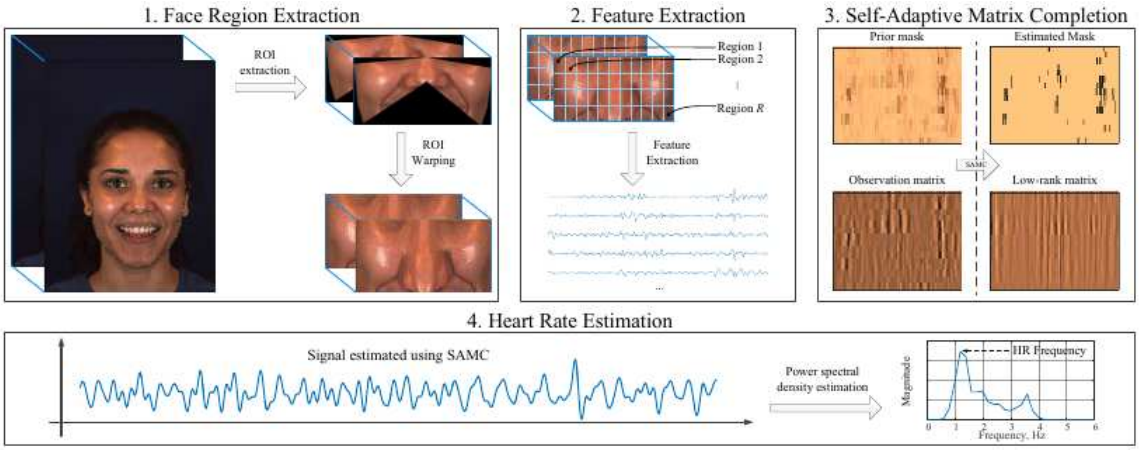


Figure 6: Overview of the proposed approach for HR estimation. During the first phase, we automatically detect a set of facial key points and use them to define a ROI. This region is then warped to a rectangular area and divided into a grid. For each small sub-region, chrominance features are computed (Phase 2). We then apply SAMC on the matrix of all feature observations to recover a smooth signal, while selecting from which sub-regions the signal is recovered (Phase 3). Welch's method [24] is used to estimate the power spectral density and thus the HR frequency (Phase 4).

The processing pipeline of our approach can be divided into 4 phases (as seen in Figure 6): phase 1 is devoted to processing face images to extract face regions, that are subsequently used in phase 2 to compute chrominance features. Phase 3 consists in the joint estimation of the underlying low-rank feature matrix and the mask using SAMC. Finally, phase 4 computes the heart rate from the signal estimate provided by SAMC.

Assuming a video sequence containing a human face, we use Intraface[3] to localize and track 66 facial landmarks. Many approaches have been employed for face frontalisation [23][16]. However, in order to preserve the underlying blood flow signal, in the current study we define the facial region of interest (see Figure 6 - Phase 1), from which the HR will be estimated. The potential ROI is then warped to a rectangle using a piece-wise linear warping procedure, before dividing the potential ROI into a grid containing R regions. For every region, we compute the average pixel chrominance value, where for every pixel, the chrominance signal is defined as $C = X_f - \alpha Y_f$. The term α is a proportionality factor $\alpha = \sigma(X_f)/\sigma(Y_f)$, where the numerator and the denominator denote the standard deviation of X_f and Y_f respectively. X_f and Y_f are band pass filtered versions of X and Y, defined as: $X = 3R_n - 2G_n$ and $Y = 1.5R_n + G_n - 1.5B_n$. Finally, R_n, G_n and B_n are normalized values of the individual colour channels. Using chrominance features for estimating HR is challenging due to at least two reasons: (i) the features associated to different facial regions are not fully synchronized and (ii) face movement and facial expressions introduce many perturbations (local in space but large in intensity) affecting the features' values. To overcome these problems, we introduce a matrix completion technique embedding a self-adaptation strategy, reducing the synchronization discrepancy by grouping the the rows of the observation matrix that are (near)
linearly dependent and, at the same time, accounting for noisy parts of the data in driving the learning algorithm.

To formalize the problem, we assume R regions from which chrominance features are computed along T video frames. This gives us an observation matrix $C \in \mathbb{R}^{R \times T}$. Solving the matrix completion problem means estimating the low-rank matrix that best approximates C. Minimizing the rank is traditionally replaced with the nuclear norm (due to the NP hard nature of the original problem), thus resulting in:

[3] http://www.humansensing.cs.cmu.edu/intraface

$$\min_{\mathbf{E}} \nu \|\mathbf{E}\|_* + \|\mathbf{E}-\mathbf{C}\|_F^2 \tag{2.1}$$

where ν is a regularization parameter. To Eq. (1) we add a temporal smoothing term (by means of a Laplacian matrix (L)), enforcing the estimated chrominance features to be within the heart rate frequency range, and a binary mask $\mathbf{M} \in \{0,1\}^{R\times T}$ accounting for the unreliable parts of C. Taking into account that $\mathbf{M}$ needs to be learnt, the final optimization problem is of the form:

$$\min_{\mathbf{E},\mathbf{M}} \nu \|\mathbf{E}\|_* + \|\mathbf{M}\circ(\mathbf{E}-\mathbf{C})\|_F^2 + \gamma \mathrm{Tr}(\mathbf{E}\mathbf{L}\mathbf{E}^T) - \beta\|\mathbf{M}\|_1 + \mu\left\|\mathbf{M}-\widetilde{\mathbf{M}}\right\|_F^2 \tag{2.2}$$

Eq. 2.2 is solved using the alternating direction method of multipliers (ADMM [14]), which the low-rank matrix $\mathbf{E}$ and the mask $\mathbf{M}$. Using Welch's power spectral density estimation method [24], from $\mathbf{E}$ we estimate HR.

We validate our approach on two public datasets, MAHNOB-HCI [18] and MMSE-HR [27], for which we report the mean M_e and standard deviation SD_e of the difference between the predicted value and the reference HR, $H_e(i) = H_p(i) - H_{gt}(i)$, over all the video clips. Additionally, we compute the $RMSE$, the mean error rate percentage $M_{eRate} = \sum_{i=1}^{N} |H_e(i)| / H_{gt}(i)$ and the Pearson correlation between $H_p = \{H_p(1), H_p(2), ..., H_p(N)\}$ and $H_{gt} = \{H_{gt}(1), H_{gt}(2), ..., H_{gt}(N)\}$, where N is the number of video clips.

Method	$M_e(SD_e)$	RMSE	M_{eRate}	ρ
Poh, et al., 2010	-8.95 (24.3)	25.9	25.0%	0.08
Poh, et al., 2011	2.04 (13.5)	13.6	13.2%	0.36
Balakrishnan, et al., 2013	-14.4 (15.2)	21.0	20.7%	0.11
Li, et al., 2014	-3.30 (6.88)	7.62	6.87%	0.81
De Haan, et al., 2013	4.62 (6.50)	6.52	6.39%	0.82
SAMC	**3.19 (5.81)**	**6.23**	**5.93%**	**0.83**

Table 4: Average HR prediction: comparison among different methods on MAHNOB-HCI dataset (best performance in bold)

Performance of MAHNOB-HCI is given in Table 4, while HR estimation of MAHNOB-HCI represents a challenging task for early methods, the more recent approaches, [18] and [15], achieve high accuracy. Moreover, our approach outperforms competing methods by a small margin. This can be explained by the fact that MAHNOB-HCI does not contain many sequences with subject's movements and facial expression changes, while SAMC has been designed to explicitly cope with the spatially localized and intense noise they generate.

Method	$M_e(SD_e)$	RMSE	M_{eRate}	ρ
Li, et al., 2014	11.56 (20.02)	19.95	14.64%	0.38
De Haan, et al., 2013	9.41 (14.08)	13.97	12.22%	0.55
SAMC	**7.61 (12.24)**	**11.37**	**10.84%**	**0.71**

Table 5: Average HR prediction: comparison among different methods on MMSE-HR (best performance in bold)

To demonstrate the advantages of our method, we perform similar experiments on the more challenging sequences of the MMSE-HR dataset. Here, we only compare our method against the best-performing approaches from Table 4. Table 5 reports the results of our evaluation. On this more difficult dataset, due to its capacity to select the most reliable chrominance features and ignore the noisy ones, the proposed SAMC achieves significantly higher accuracy than the state-of-the-art.

We find the results obtained by our proposed SAMC to be promising enough for further

evaluation in ACANTO, keeping in mind the benefits of having a completely contact-free heart rate measurement unit and, at the same time, leveraging the hardware setup already decided for both clinical, as well as the daily usage scenarios, which includes a dedicated RGB camera for face analysis. We are also considering alternative commercial solutions (in the form of the wearable PPG wristbands), should this approach prove itself unreliable in practice.

2.3 Gait Analysis

2.3.1 Optimization

While the general proof of concept on "gait analysis on the move" was already given in Deliverable 3.1, Siemens has undertaken significant efforts to improve robustness and performance particularly with respect to the upcoming clinical evaluation, in which the system should work by untrained users and different shoe appearances, shoe materials as well as diverse physiological conditions and behaviours. In the following we list the highlights of our achievements:

- Frame rate has been nearly doubled from 15 to 25 fps (frames per second) on average even though the amount of gait information derived has been increased as presented in chapter2.3.4.

- Outliers in the feet trajectories have been mostly removed resulting in improved consistency of feet pose and stance classification which eliminates the necessity of post processing by raw data smoothing which would reduce the granularity of measurements.

- Quantitative and qualitative tests with various shoes and shoe cover materials have been performed. Since RGBD cameras incompletely acquire some shoes materials, an inpainting algorithm based on [2] has been identified in order to fill holes in the depth map and is currently adapted for our project. While we need to precisely locate the tip of each foot and its angle, the shape of the foot geometry needs to be less precise but is required to be smooth in order to ease the detection process. Inpainting supports this goal by producing smooth depth maps by filling the holes in the data samples of the measured depth map. In addition keeping the frame rate high excludes higher-level geometric shape fitting.

- Estimation of ego motion of the walker based on visual information has been improved and allows for a minimal adapted walker with a gait camera solely without requiring any other instrumentation like wheel odometry. This follows our idea of introducing self-contained OEM like components that walker manufacturers can utilize to upgrade existing walkers to smart walkers. In addition visual ego motion and wheel odometry can complement each other.

- Initialization (i.e. ground plane estimation) has been improved and now can cope with the presence of feet in the FOV reliably during the initialization. This was necessary for usability reasons since its annoying for an elderly person being in need of the support of a walker to step out of the FOV of the walker for initialization of the device.

2.3.2 Tip toe stand detection

Standing on tiptoes for a given amount of repetitions is a relevant training exercise for older adults to strengthen their physiology as pointed out by therapeutic experts in a clinical workshop held in Getafe.

Hence we did implement a classifier for detecting:

- a flat stand with (small instep inclination angle)
- a tiptoe stand (with a high positive instep inclination angle)
- a heel stand (with a high negative instep inclination angle)

Detection works for each foot individually.

To ensure the shoe/foot shape does not influence the detection, a short initialization phase (<3s) is required for each user. At every frame 3D points along the instep are automatically detected (Figure 7 (a)) and a line is fitted through these points for the left and right foot (Figure 7 (b)). During initialization, a sequence of around 50 of the lines' direction vectors is stored, while the user keeps in a steady and flat stand with his toes pointing towards the camera. The median of the sequence provides a robust description of the angular relation between the feet and the ground plane.

Based on this reference, the detection is done by comparing the relative change of the current angle to the initial flat stand angle. In order to prevent detection jitter, the current state is determined as the median over the last 7 frames, which slightly increases response time.

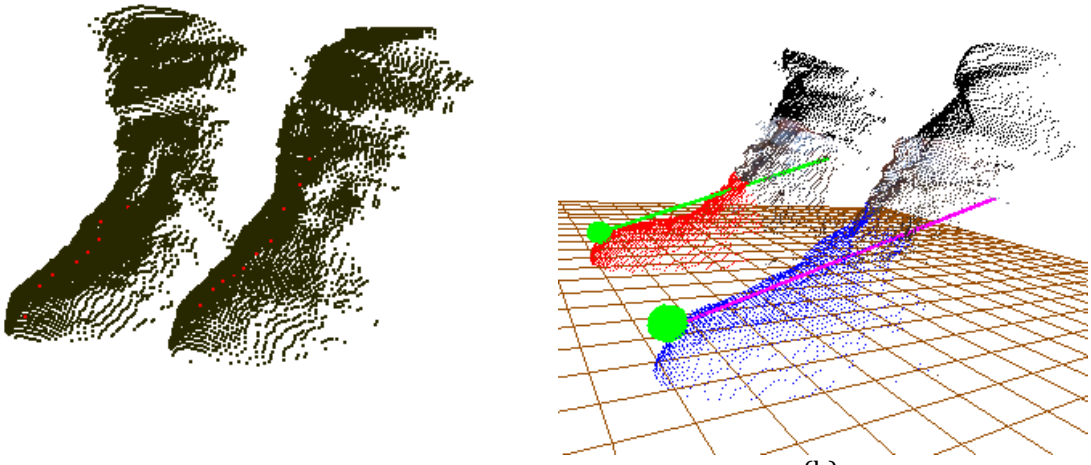


Figure 7: Tip toe detection.
(a) Points along the instep in 3D point cloud. (b) Line fit through the points.

2.3.3 Integration of new Camera Devices

In our previous work we have performed a detailed comparison of various camera devices that produce RGBD information in real-time. For the purpose of gait analytics none of these devices was particularly suitable so we ended up with a compromise and selected the Creative Senz3D. The Senz3D camera device supported our aim of a shorter minimum viewing distance and a high frame rate (i.e. no motion blur) but at the cost of reduced general data fidelity that could only partially be compensated by post processing. In the meanwhile two additional RGB devices have become commercially available (Orbbec Astra S - aka former Fotonic Astra S, Intel Realsense SR300) and both devices support reduced viewing distance while maintaining much higher data fidelity as the Creative Senz3D. In addition Creative Senz3D is at the end of product lifetime, and no longer produced so with respect to a later commercialization a replacement of the sensor did become inevitable. Recently Fotonic has changed the AstraS product housing to a more ruggedized industrial design (Fotonic "P-Series") as compared to the OEM camera module manufacturer Orbbec, while according to the data sheet the internals of the camera module did

stay untouched. In addition a more accurately calibrated device version has been introduced recently (Fotonic "Q-Series"). At least for the public scenario prices of the Fotonic devices now exceed our overall budget constraints.

The following table compares the characteristics of the three RGB-D cameras, whereas Figure 8 gives an indication of data fidelity. Creative Senz 3D shows a very high amount of data noise that needs to be suppressed by data smoothing. Fotonic/Orbbec Astra S as well Intel Realsense SR300 both come with high data fidelity which goes down to a few millimetres error.

	Creative Senz3D	**Orbbec Astra S** (formerly Fotonic Astra S)	**Intel Realsense SR300**
RGB-D measurement principle	TOF	Structured Light	Structured Light
Depth map resolution	320x240 (QVGA)	640x480 (VGA)	640x480 (VGA)
Frame rate	30 fps	30 fps	Up to 60 fps
Range	15 – 99 cm	35 – 280 cm	20 – 120 cm
Depth Sensor FOV (D)	74°	73°	85°
RGB video	1280x720	1280x960	1920x1080
Device Size	108x52x54 mm	165x30x40 mm	110x12x4 mm
Price	$100	$149	$149

Table 6: Comparison of relevant technical data of RGB-D gait cameras.

Creative Senz3D Fotonic/Orbbec Astra S Intel Realsense SR300

Figure 8: Comparison of data fidelity across RGBD cameras.

Astra S and SR300 devices are both suitable for our application while the Intel device due to the smaller minimum viewing distance allows different placement options on the walker. On the other hand the Orbbec Astra S is available as an OEM device without housing which also gives some flexibility in integration. Orbbec also offers a smart camera that bundles the RBD camera

with a processor board in a compact housing. This option needs further investigation and could be attractive for a minimal invasive OEM walker add-on system.

2.3.4 Introduction of an Intermediate Level Representation

In addition to the raw data stream of foot positions and angles in space a data stream of semantic per-foot information like “foot is visible” or “foot is on ground” as well as higher-level semantic descriptions about stance related poses like “parallel stand” or “angular stand” are introduced. Also some extra specific stands/poses like “tandem”, “tip toe”, “flat” or “on heel” are part of the information and ease the derivation of associated clinical test.

The following table gives an overview on the granularity of information that is made available. The data can be easily serialized, represented as a 32-bit integer value.

Bit	***Flag (description)***	**Bit**	***Flag***
0	*Only one foot visible (l or r)*	15	*Tandem*
1	*Is visible (r. foot)*	16	*Semi-Tandem*
2	*Is visible (l. foot)*	17	*Angular 1*
3	*Is on ground (r. foot)*	18	*Angular 2*
4	*Is on ground (l. foot)*	19	
5	*Flat (r)**		
6	*Flat (l)**		
7	*Tip toe (r)**		
8	*Tip toe (l)**		
9	*On heel (r)**	To	*Not defined yet/ reserved for future use (RFU)*
10	*On heel (l)**		
11	*Parallel*		
12	*Open*		
13	*Closed*		
14	*Side by Side*	31	

*requires user specific initialization

Table 7: Intermediate level semantic representation on gait parameters.

All of these parameters are calculated in real-time at a frame rate of typically 25 fps and are visualized for demo/testing purposed as shown in Figure 9. For the purpose of integration in the clinical activity analysis these quantities are published as ZMQ messages.

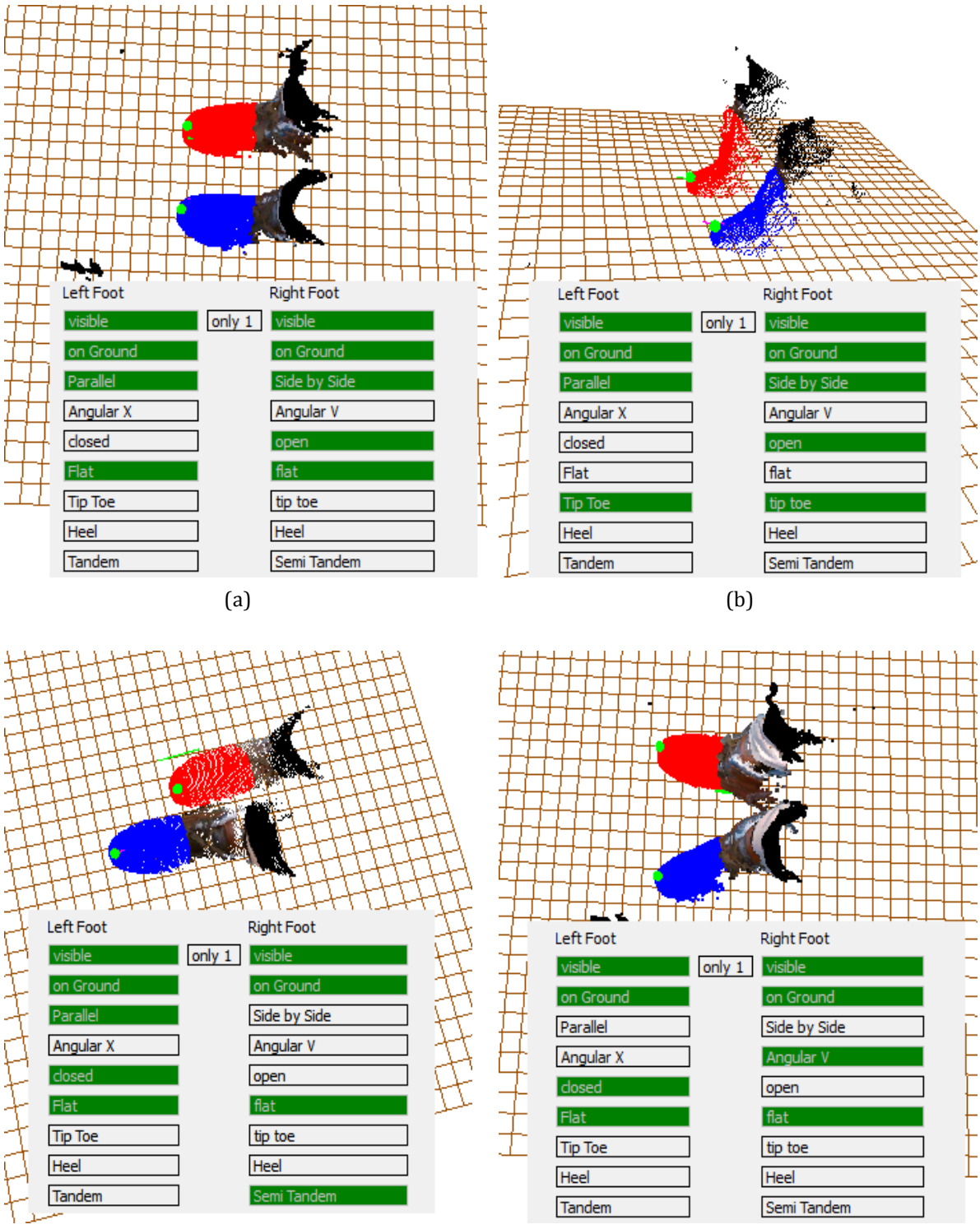


Figure 9: Live-view on intermediate level gait semantics as displayed in the demo-application. (a) Feet in parallel stance. (b) Tip toe stance, a specific stance required to be detected for training exercises (c) Semi tandem stance (as part of the SPPB balance test). (d) Angular V shaped stance.

2.4 *FriWalk* Body Detector Module

The user perception modules on board the *FriWalk* should be able to provide detailed information about the user in real time. Information such as the body posture, stability and gait of the user are valuable for both the clinical functions as well as the accident prevention and guidance capabilities of the device. As part of the ACANTO perception package, FORTH is developing a number of human body detection and tracking modules. The modules provide articulation information from marker-less visual observations obtained by the depth sensor located in the front of the *FriWalk*.

2.4.1 Evaluation of the FORTH Human Body Tracker (*FHBT*)

The baseline body tracking method [44] implemented by FORTH and detailed in deliverable D3.1 can operate with input provided by any of the on board depth sensors, mounted in the front or in the back of the walker. Depending on the sensor used and on the distance of the user to the camera, the method can detect the user and estimate the 3D pose of either the full or the upper part of his/her body (up to 4 meters away). This capability enables the *FriWalk* to monitor the user's upper body while using the device and facilitates the implementation of clinical exercises and "exergames". *FHBT* is a generative, disjoint evidence method that performs tracking by-detection and detection-by-tracking. Body parts are identified independently and then assembled together in a human model[4].

The estimation of the articulated motion of the human body is very important to a number of real world applications, including but not limited to surveillance, gaming, medical rehabilitation, human-robot interaction, smart environments and many others. It is considered to be a challenging problem because of its high dimensionality, the variability of the tracked persons regarding their appearance and sizes, the spatially and temporally extended (self)-occlusions, etc. A number of practical approaches simplify or even avoid these problems by using special hardware that is placed on the environment and/or markers/full body suits [49] worn by the persons to be tracked. However, these are invasive solutions. Unobtrusive, marker-less tracking is definitely preferable since it does not interfere with the environment, the subject and the performed actions.

The methods that use marker-less visual data as their only input fall into three basic categories, the generative, the discriminative and the hybrid ones. Each class of methods has its own characteristics, advantages and disadvantages. Discriminative methods are fast, but rely on a discrete set of training poses whose selection determines the accuracy of the obtained results. Typically, they operate as single frame pose estimation methods, so they do not need to be initialized and they do not drift. The generative approaches provide accurate, physically plausible solutions, typically at a high computational cost. They also require initialization for the first frame, and may suffer from drift and tracking failures that are often irrecoverable. Hybrid methods integrate generative and discriminative components towards combining the advantages of both worlds.

We evaluated three approaches for 3D human pose estimation, one representative of each class. The evaluation has been performed on a dataset annotated with ground truth. The obtained quantitative experimental results help in assessing the relative performance of these methods and in deciding which is preferable in which situation.

The *OpenNI* method [50]: This is a widely employed, purely discriminative method. It is applied on the input of a single RGBD camera.

[4] See also https://youtu.be/ZKlC9PA1IDg

The *HYBRID* method [46]: As a hybrid method[5], it consists of a discriminative and a generative component. The generative, joint evidence component of the method requires input from two extrinsically synchronized RGB-D cameras that is used to reconstruct the 3D volume occupied by the human body. Then, human pose estimation is formulated as an optimization problem that minimizes the discrepancy between the 3D occupancy of hypothesized instances of a human body model and the volume reconstructed from the observations. To track the human pose, solutions for a certain frame are initialized at the vicinity of solutions estimated at the previous frame. However, the solution suggested by the discriminative component (OpenNI) of the method is also considered as a human pose hypothesis to (a) adjust the human model parameters to the tracked individual and (b) safeguard from abrupt human motions as well as from tracking failures.

The *FHBT* method [44]: This is a generative, disjoint evidence method that performs tracking by-detection. Body parts are identified independently and then assembled together in a complete model[6].

Characteristics	**OpenNI**	**HYBRID**	**FHBT**
Method type	Discriminative	HYBRID	Generative
Number of cameras	1	2	1
Auto-initialization	Yes, special pose	Yes, special pose	Yes, any pose
Initialization speed	Slow (>3 sec)	Slow (>3 sec)	Instant (0.03sec)
Auto recovery from failures	Yes	Yes	Yes
Handles various body types	Yes	Yes	Yes
Handles occlusions	No	No	Yes
Moving camera(s)	No	No	Yes
Ensure physical plausibility	No	Yes	No
Mode of operation	Online	Offline	Online
Real time performance	Yes	No	Yes

Table 8: Overview of the evaluated methods with respect to a number of key characteristics and properties

The main characteristics of the evaluated methods are summarized in Table 8. For more details, the reader is referred to the corresponding references. In general, *OpenNI* is a flexible and fast method. As suggested by the extensive evaluation performed in this paper, its accuracy is moderate. One of its main drawbacks is its long initialization time. The *HYBRID* method inherits the long initialization time from *OpenNI*. Moreover, it relies on a more complex setup of two extrinsically calibrated RGBD sensors and its computational requirements are quite high, i.e., near-real-time performance can only be achieved with an elaborate GPU-based implementation on a high-end computer featuring a state of the art graphics card. The experimental results demonstrate that in return, the *HYBRID* method outperforms the two others in terms of accuracy. The *FHBT* method is less accurate than *HYBRID* but on par with *OpenNI*. This fact, together with other advantages of the method (see Table 8), make it an attractive solution to a number of applications that require knowledge of the human body pose.

2.4.1.1 Human body model

The employed articulated model of the human consists of a main body, two legs, two arms and the head (Figure 10). The kinematics of each arm is modelled using six parameters encoding angles. Two parameters determine the shoulder position with respect to the torso, three

[5] See also https://youtu.be/n5irgHVuFwc

[6] See also https://youtu.be/ZKlC9PA1IDg

parameters the upper arm with respect to the shoulder and one parameter the elbow with respect to the upper arm. Six parameters are also used for a leg, three for the root, one for the knee and two for the ankle. Two parameters are used for the head, and three parameters for the articulation between the torso and the hip. The global position of the body is represented using a fixed point on the hip. The global orientation is parameterised using Euler angles. The above parameterisation encodes 35 degrees of freedom (DOFs) human model with each DOF represented by a single parameter.

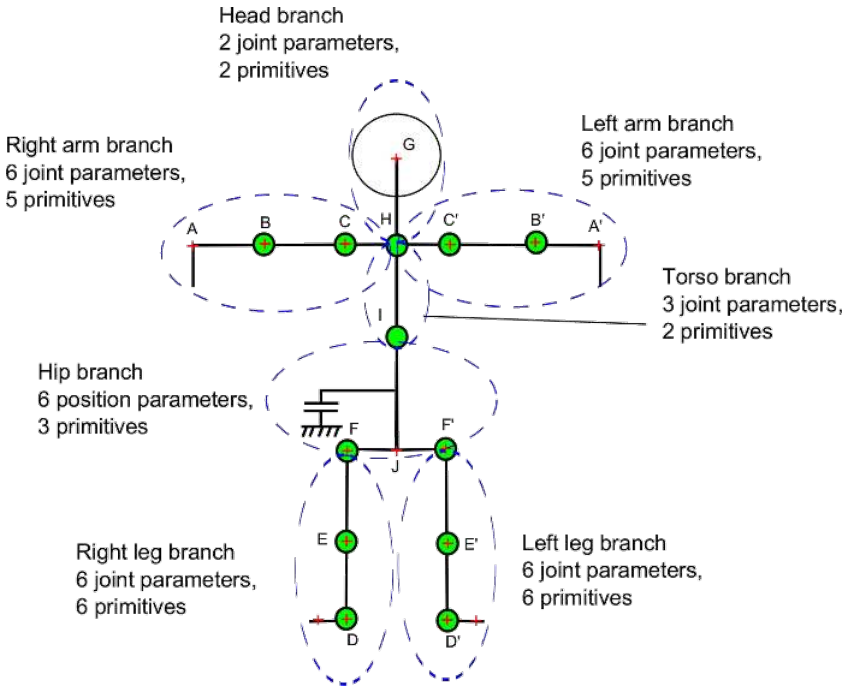


Figure 10 The employed human body model. Model points with a red "+" denote joints whose 3D position is taken into account in defining the tracking error.

On top of the 35 mobilities of this model, 9 parameters control the lengths of certain human body parts. These are the upper body length (UBL), the lower body length (LBL), the shoulders neck distance (SND), the head neck distance (HND), the legs hip distance (LHD), the back arm length (BAL), the forearm length (FAL), the back leg length (BLL) and the front leg length (FLL). Table 9 presents ground truth values for these parameters for the subjects of the employed dataset. The parenthesis next to a parameter name refers to the corresponding body segment(s) in Figure 10.

It has to be noted that individual methods employ their own, internal models for 3D human pose estimation. The model described above and illustrated in Figure 10 is used for the evaluation of the performance of the benchmarked methods, only. Thus, the relation of the above model to the ones used internally by each method has been established and used to bring all results to the same reference frame so as to enable their direct comparison.