Computer Vision and Image Understanding 148 (2016) 3–22

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Wize Mirror - a smart, multisensory cardio-metabolic risk monitoring system

Yasmina Andreu[a], Franco Chiarugi[c], Sara Colantonio[b,*], Giorgos Giannakakis[c], Daniela Giorgi[b], Pedro Henriquez[a], Eleni Kazantzaki[c], Dimitris Manousos[c], Kostas Marias[c], Bogdan J. Matuszewski[a], Maria Antonietta Pascali[b], Matthew Pediaditis[c], Giovanni Raccichini[b], Manolis Tsiknakis[c,d]

[a] Robotics and Computer Vision Research Laboratory, School of Computing Engineering and Physical Sciences, University of Central Lancashire, PR1 2HE Preston, UK
[b] Institute of Information Science and Technologies, National Research Council of Italy, Via G. Moruzzi 1, 56124 Pisa, Italy
[c] Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH), N. Plastira 100, Vassilika Vouton, GR-700 13, Heraklion, Crete, Greece
[d] Technological Educational Institute of Crete, Biomedical Informatics and eHealth Laboratory, Estavromenos, GR-71004, Heraklion, Crete, Greece

ARTICLE INFO

Article history:
Received 17 April 2015
Revised 23 March 2016
Accepted 24 March 2016
Available online 12 April 2016

Keywords:
Unobtrusive health monitoring
3D face detection
Tracking and reconstruction
3D morphometric analysis
Psycho-somatic status recognition
Multimodal data integration

ABSTRACT

In the recent years personal health monitoring systems have been gaining popularity, both as a result of the pull from the general population, keen to improve well-being and early detection of possibly serious health conditions and the push from the industry eager to translate the current significant progress in computer vision and machine learning into commercial products. One of such systems is the Wize Mirror, built as a result of the FP7 funded SEMEOTICONS (SEMEiotic Oriented Technology for Individuals CardiOmetabolic risk self-assessmeNt and Self-monitoring) project. The project aims to translate the semeiotic code of the human face into computational descriptors and measures, automatically extracted from videos, multispectral images, and 3D scans of the face. The multisensory platform, being developed as the result of that project, in the form of a smart mirror, looks for signs related to cardio-metabolic risks. The goal is to enable users to self-monitor their well-being status over time and improve their life-style via tailored user guidance. This paper is focused on the description of the part of that system, utilising computer vision and machine learning techniques to perform 3D morphological analysis of the face and recognition of psycho-somatic status both linked with cardio-metabolic risks. The paper describes the concepts, methods and the developed implementations as well as reports on the results obtained on both real and synthetic datasets.

1. Introduction

A healthy lifestyle has become universally recognized as a key factor in disease prevention. Efforts at promoting lifestyle improvements are now considered as a viable and effective way for reducing the incidence of pathologies, such as cardiovascular diseases and metabolic disorders. This coupled with the more active role people aspire to have, so as to shift from passive recipients of care towards actively managing their own health, has opened a new important prevention realm for the assistive technologies. The health related self-monitoring and self-assessment are gaining momentum. Many personal well-being and fitness monitoring tools are available on the market, mainly in the form of wearable devices such as wristbands, smart-watches, eye wear and wearable bio-monitors, as well as dedicated apps on smart-devices such as MyFitnessPal, Endomondo, Argus, Googlefit. It has been shown that these technologies are predominantly embraced by the younger generation (25–34 years old) focused on fitness, and the older users (55–64 years old) mainly interested in improving overall health with the aim of improving the quality of life and the life expectancy. Interestingly, in contrast with the increasing acceptance of wearables, many of consumers stop using the device within six months. In other words, in many cases these tools fail to drive long-term, sustained engagement and as a

* Corresponding author. Tel.: +39 050 6213141.
E-mail addresses: yasmina.andreu@gmail.com (Y. Andreu), sara.colantonio@isti.cnr.it (S. Colantonio).

http://dx.doi.org/10.1016/j.cviu.2016.03.018

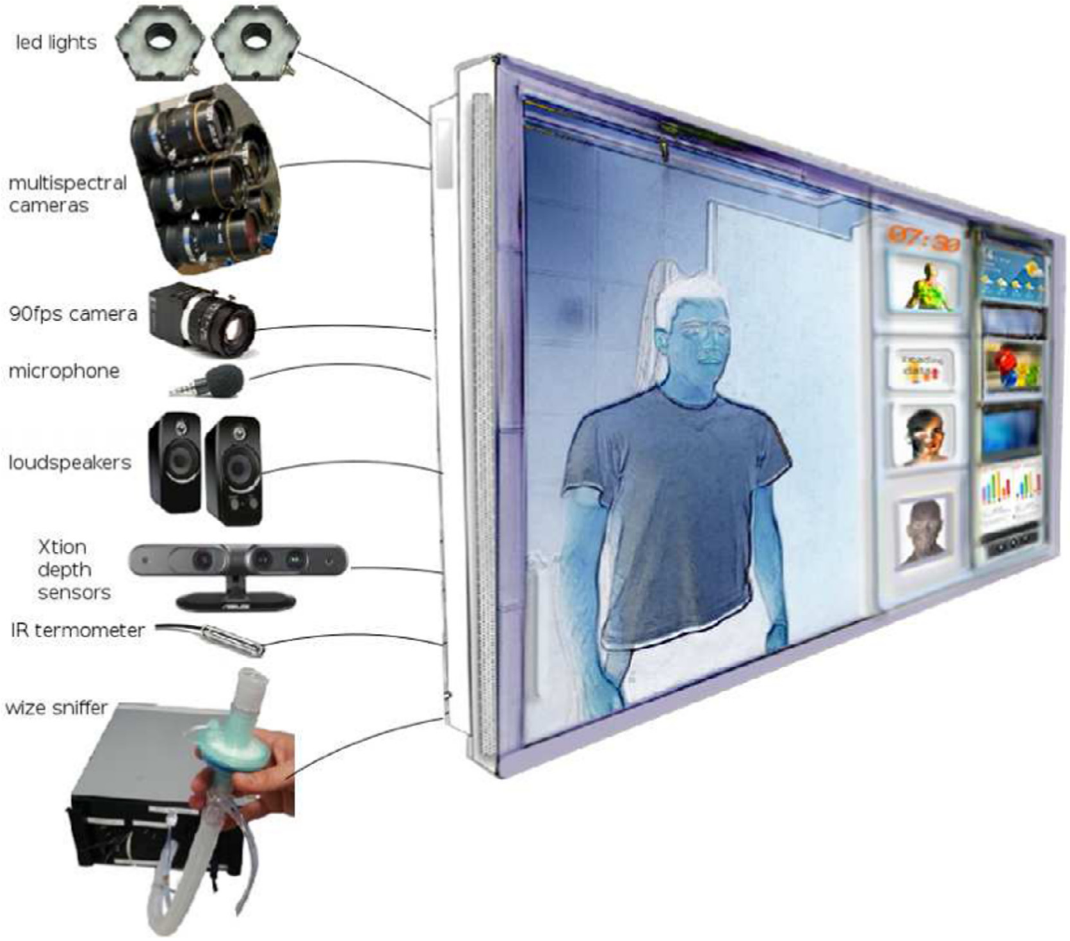

Fig. 1. Illustrative representation of the Wize Mirror. On the right widgets panel with the user graphical interface, optionally this may also include: clock, weather forecast, news, etc. On the left pictorial representation of different devices used for data acquisition.

consequence, they fail to make a long-term impact on their users' health. The authors believe that the key to successful deployment of self-assessment technologies is sustained engagement, based on the promotion of behaviour change towards holistic wellness. Enhancing wellness is an effective way to promote participation and motivate people to change their habits. It is in this context that the European project SEMEOTICONS (SEMEOTICONS, 2013) has been launched. SEMEOTICONS started in November 2013, challenged with the development of a multisensory device in the form of a mirror, called the Wize Mirror, which comfortably fits at home, as a piece of house-ware, but also in pharmacies or fitness centres. By analysing data acquired unobtrusively via a suite of contactless sensors, the Wize Mirror detects on a regular basis physiological changes relevant to cardio-metabolic risk factors. The computation and delivery of a comprehensive Wellness Index enables individuals to estimate and track over time their health status and their cardio-metabolic risk. Finally, the Wize Mirror offers personalized guidance towards the achievement of a correct lifestyle, via tailored coaching messages. The Wize Mirror is designed to meet the two main objectives: stimulating initial adoption and utilization, by providing a positive usage experience; and supporting long-term engagement, by helping people to establish new positive habits. To this end, the main features of the Wize Mirror are: facilitation of daily unobtrusive monitoring; automatic assessment of physiological conditions via advanced integrated sensing and data processing algorithms as well as promotion of sustained behaviour change towards long-term wellness objectives. These functionalities are developed by integrating theories, methods and tools from different disciplines including: computer science, physics, engineering, medicine, psychology, motivation and communication science, social marketing, behavioural theories, and economics.

From the technological perspective the Wize Mirror is a multisensory platform in the form of a smart mirror (see Fig. 1) integrating different sensors, including: 3D optical scanner, multispectral cameras and gas detection sensor, collecting multidimensional data of individuals standing in front of the mirror. These data are processed by dedicated algorithms, which extract a number of biometric, morphometric and colorimetric descriptors, including: AGE- product concentration, cholesterol level, endothelium function, heart rate, heart rate variability, face morphometric parameters as well as indicators of stress, anxiety and fatigue levels. The descriptors are integrated to define a virtual individual model for a wellness index traced over time. The Wize Mirror also offers suggestions and coaching messages, with personalized user guidance, aimed at achieving and then maintaining a healthy life-style.

The guiding principle behind the design of the Wize Mirror has been that it should easily fit into daily-life settings, by maximising non-invasive and unobtrusive interaction with the users. The focus of this paper is on a subset of sensors, methods and processes deployed on the Wize Mirror using medical semeiotics signs. The principle of medical semeiotics considers the face as an

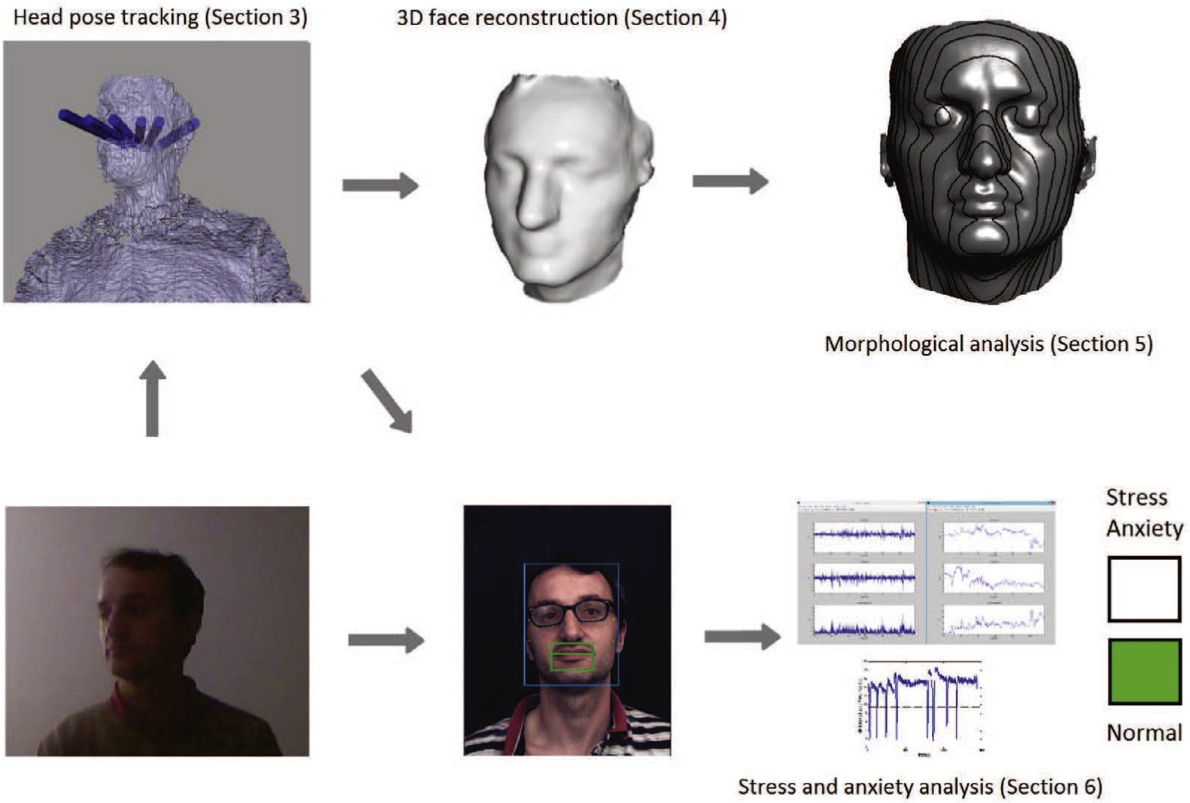
Head pose tracking (Section 3)
3D face reconstruction (Section 4)
Morphological analysis (Section 5)
Stress
Anxiety
Normal
Stress and anxiety analysis (Section 6)

Fig. 2. Processing pipeline of the Wize Mirror methods described in this paper.

important source of information about the health status of individuals, produced by the combination of physical signs and expressive features. Currently, based on their experience, medical doctors acquire the ability of reading and interpreting the complex semeiotic signs of patients' faces. These signs usually suggest how to conduct the medical examination and may contribute to the diagnosis. The paper describes a novel set of techniques developed and implemented to acquire and analyse semiotics signs. More specifically the paper describes the processing pipeline enabling face detection, tracking, partition and 3D reconstruction. Whereas the robust real-time face detection and partition facilitates the described analysis of stress and anxiety forming the psycho-somatic descriptor of the cardio-metabolic risk, the 3D reconstruction provides required information for the estimation of the described overweight and obesity index forming part of the morphometric descriptor of the cardio-metabolic risk. Fig. 2 gives a visual explanation of the processing pipeline. The performance of the proposed techniques is examined in some depth on real and synthetic datasets.

The description of an overall concept of the inexpensive device for self-monitoring and assessment of well-being to promote, improve and maintain a healthy lifestyle is the key novel contribution of this paper. The other technical contributions are linked to different subsystems integrated on the mirror, these include: use of the Kalman filter in conjunction with the random forest for face pose predictions; the processing pipeline integrating face tracking, 3D pose estimation, segmentation, range data scans alignment and fusion for efficient and robust 3D face reconstruction; estimation of body weight and body weight variations using geometric features extracted from the 3D reconstruction of the face; the fusion of motion features from different facial areas for the assessment of the psychological state with focus on stress and anxiety.

The remainder of the paper is organised as follows. Section 2 reports on the state of the art of the methods involved in the cardio-metabolic and psycho-somatic analyses performed in this work. Section 3 and 4 provide details about the techniques employed for face pose estimation, tracking, face segmentation and 3D reconstruction. Section 5 shows how the morphological analysis of the face is carried out for assessing cardio-metabolic risks and Section 6 describes the estimation of the face signs used for recognising different psycho-somatic states. Finally, Section 7 summarises the main conclusions of the work presented.

2. State of the art

2.1. Face tracking and 3D reconstruction

Typically, 3D face reconstruction methods integrate raw data from different sensors (colour and/or depth) into a point cloud to produce a 3D face representation. In order to avoid points which belong to the background or the other body parts, 3D head pose estimation and tracking is needed to select the relevant information for processing. By using the pose estimation a face segmentation can be effectively preformed reducing errors in the reconstruction phase. Additionally, the pose estimation and tracking provide information, needed for face normalization and partition, facilitating operations of other subsystems on the mirror, including stress and anxiety analysis as well as multi-spectral measurements.

Head pose estimation and 3D tracking play an important role in the automatic face analysis as an essential pre-processing step. There has been a plethora of methods proposed in literature to solve this problem (Murphy-Chutorian and Trivedi, 2009). Although it is possible to estimate the 3D head pose using only 2D images (Raytchev et al., 2004), the robustness and accuracy of these methods may not be suitable for many practical applications. On the other hand the pose estimation based on 3D data (Smeets et al., 2013) could be very robust and accurate, but the 3D data acquisition is costly and computationally intensive. Indeed, the 3D face reconstruction is more challenging than the head pose estimation. The recent advances in range data (2.5 D) sensing technologies and analysis seem to facilitate a suitable compromise between cost, performance and system complexity. The range/depth sensors are getting cheaper, more reliable and widely used. For that reason, the range data is becoming the modality of choice for solving different detection and estimation problems. For example, there are approaches which use the range data in combination with 2D image data. The method explained in Cai et al. (2010) relies on a regularized maximum likelihood deformable model. The work described in Seeman et al. (2004) is a neural network based system which runs at 10 fps. The approach introduced in Bleiweiss and Werman (2010) is model-based and it can maintain real-time performance. The method in Newcombe et al. (2011) is based on the active appearance model (AAM) and a depth-based constraint. It provides real-time tracking of human faces in 2D and 3D. The mentioned method introduced a new constraint into AAM that uses depth data from sensors like Kinect. To initialise the AAM fitting in each frame, an optical feature tracking is used to provide a location close to the target to improve the convergence. The 3D location accuracy is improved by introducing a depth fitting energy function, which is formulated in a similar way to the iterative closest point algorithm (ICP) (Besl and McKay, 1992). Moreover, the colour-based face segmentation is replaced with the depth-based face segmentation and an L2-regularization term. Using solely depth data, there are methods such as the one described in Fanelli et al. (2011), enabling a real time 3D head pose estimation using consumer depth cameras. That method uses a random regression forest to estimate the pose. The forest regression can be combined with the Kalman filter as described in Mou and Wang (2012), where the Kalman filter is used to refine the noisy regression result. Another approach, which is based on particle swarm optimization is described in Padeleris et al. (2012). Real-time performance is achieved by the methods introduced in Malassiotis and Strintzis (2005) and Choi et al. (2014). The approach in Malassiotis and Strintzis (2005) uses global features and exploitation of prior knowledge along with feature localization and tracking techniques. In the work reported in Choi et al. (2014) a 3D face model is generated from a single frontal image. Then uniformly distributed random points are extracted and tracked in 2D. Given the correspondences, the 3D head pose is estimated using a RANSAC-PnP process. For the low-cost depth cameras, one of the most widely used methods is described in Newcombe et al. (2011). That system is able to accurately map complex and arbitrary indoor scenes in variable lighting conditions. All the input depth data is fused into a global surface model in real-time. The sensor pose is estimated by tracking the global model using a coarse-to-fine iterative closest point (ICP) algorithm, and the data fusion is performed by means of a truncated signed distance function (TSDF). Due to the relatively good results obtained by the low-cost depth cameras, they have become a popular choice for face reconstruction, where a great variety of methods can be found, such as ones described in Hernandez et al. (2015); Huang et al. (2013); Macedo et al. (2013); Zollhofer et al. (2011). The authors of Macedo et al. (2013) presented an extension of the algorithm from Newcombe et al. (2011) to perform a real-time face tracking and modelling. They proposed changing two steps of the original algorithm, preprocessing and tracking. In the pre-processing stage a face detection algorithm is used to segment the face from the rest of the image. For the tracking, they included an algorithm to solve occlusions and real-time head pose estimation to give a new initial guess to the ICP algorithm when it fails. Marching cubes is another well-known technique for reconstruction and modelling. The system developed in Huang et al. (2013) can automatically detect the face region and track the head pose while incrementally integrating the new data in a model. ICP is used for tracking the head pose, then, a volumetric integration method is used to fuse all the data. Afterwards, a ray casting algorithm extracts the final vertices of the model and marching cubes algorithm is used to generate the polygonal mesh of the reconstructed face model. The method in Hernandez et al. (2015) produces face models from a freely moving user without relying on any prior face model. The face is represented in cylindrical coordinates in order to perform filtering operations. The reconstruction is initialized with a depth image, and then the subsequent point clouds are registered to the reference one using ICP. Temporal and spatial smoothing are applied to the updated model. Most of the methods rely on ICP rigid registration algorithm, however, the approach in Zollhofer et al. (2011) introduces the advantages of using a robust non-rigid registration and a deformable model.

2.2. Morphological analysis of cardio-metabolic risk

Back in 1942, D'Arcy Wentworth Thompson expressed the importance of investigating biological form in a fully quantitative manner (Thompson, 1942):

The study of form may be descriptive merely, or it may become analytical. We begin by describing the shape of an object in the simple words of common speech: we end by defining it in the precise language of mathematics; and the one method tends to follow the other in strict scientific order and historical continuity.

We may say that D'Arcy Thompson's vision has come true: in the last century, morphometrics came of age, as the discipline dealing with the quantitative study of form Reyment (1996). In medicine, information about body size and shape has been used traditionally by physicians to assess health or nutritional status and guide treatment, and many efforts have been put to recognize the facial gestalt of some dismorphic syndrome (Hammond, 2007). However, most of the studies correlating anthropometric measurements with cardio-metabolic risk deal with the body rather than the face.

Simple parameters such as the waist circumference and the abdominal sagittal diameter are known to correlate well with the body fat and have been used as predictors for metabolic disorders and cardiovascular risk (Li et al., 2007). The anthropometric measurements collected by a 3D scanner were recently correlated with metabolic parameters in validation studies (Lin et al., 2004; Wang et al., 2006; Wells et al., 2008). A relevant drawback is that these tools are not standardized: parameters strongly depend on the acquisition device and on the subject pose, and do not provide a complete characterization of the body. Interesting results are presented by Velardo and Dugelay (2010): a model for the weight estimation is retrieved via multiple regression on a set of anthropometric features exploiting a large medical dataset for the model training, and validating the method both in ideal and real conditions; here the set of geometric body measurements used is extracted from the 2D body silhouette. More recently Giachetti and colleagues in Giachetti et al. (2015) presented a pipeline for the automatic extraction of health-related geometrical parameters from heterogeneous body scans. Their aim was the computation of parameters independent of the precise location of anatomical landmarks. The parameters computed included total body mesh volume and area, trunk volume and area, maximal and minimal trunk width, maximal trunk section radius and area, eccentricity of an ellipse approximating the body. They correlated the parameters with body fat values estimated with a DXA (Dual-energy X-rays Absorptiometry) scanner, and found that several values were highly correlated with total body less head (TBLH) fat and trunk fat. Moreover, in Velardo et al. (2012) an automatic vision-based system was proposed for estimating the subjects' absolute weight from a frontal 3D view of the user, acquired through a low cost depth sensor.

Potential applications include extreme environments and circumstances in which a standard scale cannot work or cannot be used: in the space for monitoring astronauts' weight, or in the hospitals, for medical emergencies.

Concerning faces, there is no consensus in the literature about which are the facial morphological correlates of body fat. A study reported the relationship between facial adiposity and Visceral Obesity (VO) and suggested that facial characteristics, such as cheek fat, are indicators of insulin resistance (Sierra-Johnson and Johnson, 2004). An increase in some facial dimensions was observed in a study about the face morphology of obese adolescents (Ferrario et al., 2004): the authors observed that the face of obese adolescents was wider transversally, deeper sagittally and shorter vertically than matched controls. Djordjevic et al. (2013) reports an analysis of the facial morphology of a large population of adolescents under the influence of confounding variables. Though the statistical univariate analysis showed that four principal face components (face height, asymmetry of the nasal tip and columella base, asymmetry of the nasal bridge, depth of the upper eyelids) correlated with insulin levels, the regression coefficients were weak and no significance persisted in the multivariate analysis.

The authors in Lee et al. (2012) proposed a prediction method of normal and overweight females based on BMI using geometrical facial features only. The features, measured on 2D images, include Euclidean distances, angles and face areas defined by selected soft-tissue landmarks. The study was completed in Lee and Kim (2014) by investigating the association of visceral obesity with facial characteristics, so as to determine the best predictor of normal waist and visceral obesity among these characteristics. Cross-sectional data were obtained from over 11 thousand adult Korean men and women aged between 18 and 80 years. The study in Lee and Kim (2014) was the starting point of our research. We started by reproducing and evaluating the measurements on 3D data, then we added measures specifically defined on the 3D surface.

2.3. Stress and anxiety analysis

Psychologists and biomedical scientists have studied stress intensively for over 60 years, and the concept of stress has been the subject of scientific debate ever since its first use in physiological and biomedical research (Selye, 1950). Stress was originally defined as the non-specific response of the body to any unpleasant stimulus. Later, the concept was refined by distinguishing between the terms 'stressor' and 'stress response': a stressor is a stimulus that threatens homoeostasis and the stress response is the reaction of the organism aimed to regain homeostasis (Koolhaas et al., 2010). Stressful events cause dynamic changes in the human body. They can be observed by changes in the body's response signals, involuntary caused by the autonomic nervous system. Stress has a severe impact on the immune and cardiovascular systems if it is sufficiently powerful to overcome defence mechanisms, (Sharma and Gedeon, 2012). Nevertheless, the stress response evolved to help individuals survive, so that a lack of a sufficient stress response can often result in an inability to cope with a stressor (Romero, 2004). When a person is under stress, an increased amount of stress hormones are released, accompanied by changes in heart rate, blood pressure, pupil diameter, breathing pattern, galvanic skin response, emotion, voice intonation and body pose. Common techniques for detecting stress include the analysis of physiological signals such as the electroencephalograph, blood volume pulse, heart rate variability, galvanic skin response and electromyograph (Sharma and Gedeon, 2012). The manifestation of stress through visible facial expressions enables non-invasive techniques for detecting and analysis (Sharma and Gedeon, 2012). Facial muscle movements, such as head and mouth movements, have been used to determine stress. Eye gaze spatial distribution, saccadic eye movements, pupil dilation, blink rates, eyebrow movements and mouth deformation are features able to show stress presence (Sharma and Gedeon, 2012). In addition, jaw clenching, grinding teeth, trembling of lips, and blushing are also signs of stress (The American Institute of Stress, 2015c).

Anxiety is a very common psychosomatic state, felt as an unpleasant mood characterized by thoughts of worry or fear (Harrigan and O'Conell, 1996; Shin and Liberzon, 1996). A person experiencing anxiety has thoughts that are actively assessing a certain situation, sometimes even automatically and outside of conscious attention, and developing predictions of how well they will cope based on past experiences. People with anxiety disorders may also have recurring intrusive thoughts or concerns that may lead to avoiding certain situations out of worry. They may also have physical symptoms such as sweating, trembling, dizziness or a rapid heartbeat (Anxiety, 2015a). Anxiety has been shown to inhibit social relationships, to impede cognition, learning and performance, to contribute to psycho-physiological disorders and is the primary symptom of a variety of disorders. Indeed, dysfunctional levels of anticipation appear to manifest in a number of anxiety disorders including specific phobia, generalized anxiety disorder, social anxiety disorder and panic disorder (Harrigan and O'Conell, 1996). Individuals with elevated anxiety are more likely to have a wide array of medical conditions than those without anxiety, including cardiovascular, autoimmune, and neuro-degenerative diseases, and are at greater risk of early mortality (Niles et al., 2015). Anxiety and depressive disorders are linked to a higher cardio-metabolic risk and a higher incidence of acute cardiovascular events (Sardinha and Nardi, 2012). Given the impact and the frequency with which anxiety occurs, it is critical to investigate its manifestations, particularly those which may reveal anxiety indirectly through non-verbal indices, such as facial movements. Research in non-verbal manifestation of anxiety is not very common (Chiarugi et al., 2014). Ekman and Friesen (1971) reported that, when a negative effect is experienced, it is often masked by another effect that the individual considers more appropriate. Anxiety is a composite effect with a strong connection to fear and therefore, when someone is anxious, we expect to identify facial movements related to fear such as raised eyebrows, stretched lips horizontally, raised and tensed upper eyelid which widens the eye, lip bite, lip wipe and increased eye movement (Ekman and Friesen, 1971). Other anxiety specific manifestations are shared with stress, such as increased eye blink rate (Harrigan and O'Conell, 1996) and shortened breath (Anxiety, 2015b). Increased blinking is associated with increased sympathetic nervous system activity that increases involuntary responses when people are emotionally aroused (Harrigan and O'Conell, 1996), as a result, during anxiety the overall activity of facial muscles increases (Gunes and Piccardi, 2007).

3. Face 3D pose estimation and tracking

The proposed approach is based on processing single depth data frame at a time, using a random forest model for face detection and face/head pose regression (Fanelli et al., 2011) and then applying the Kalman filter tracking (Henriquez et al., 2014) to the results from random forest pose regression. As result the random noise of the pose estimates are reduced leading to smoother pose trajectories. Finally, a personalised mask alignment is performed to further improve accuracy of the face pose estimates. The multi-level iterative closest point algorithm registration (Quan et al., 2010) method is applied for face alignment. The personalised mask construction process is explained in Section 4. The proposed face tracking has been designed to track the face pose in real-time within a depth image sequence from the depth sensor. The implemented approach relies on algorithms which are not computation-

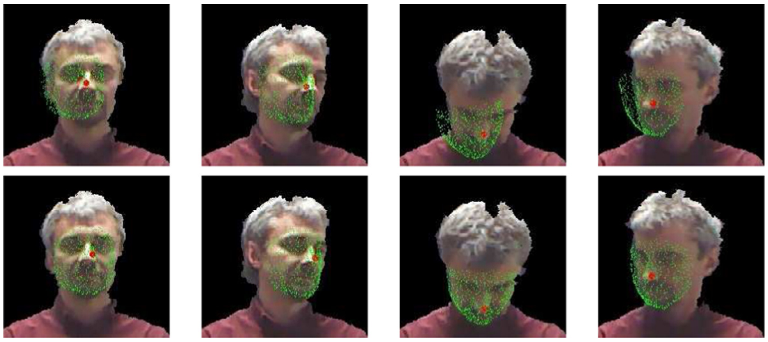

Fig. 3. Comparison between detected and tracked faces obtained with the random forest and Kalman filter (first row), and the refined head pose using the ICP algorithm (second row). The depth images have been coloured in order to facilitate the visualization, the colour information is not used in the process.

ally expensive. The high computational complexity is only required in the training phase, but this phase is performed off-line. Therefore, a face pose can be estimated in each video frame in real-time using a single core processor (2GHz). The face pose tracking results are subsequently used for 3D face reconstruction, described in Section 4, which in turn is used in the face morphological analysis for cardio-metabolic risk assessment (see Section 5). The face pose is also used to perform face partition required as a preprocessing step for the stress and anxiety analysis described in Section 6.

3.1. Face pose estimation

In the first stage of the face tracking process, the face pose is estimated using the approach described in Fanelli et al. (2011). A discriminative random regression forest is used to classify depth image patches between two different classes (face or no face) and perform a regression in the continuous spaces of position and orientation. The trees in the forest are trained to maximise two different measures (classification and regression). The data used for training are depth images captured with the Kinect sensor. Each one is labelled with the 3D face pose (x, y, z, pitch, yaw, roll). The optimisation function consists of two main parts as it is shown in Eq. 1, the class uncertainty U_C and the regression entropy U_R. There are also other parameters such as the depth of the node d, and a λ parameter to balance the importance of classification and regression depending on the depth of the tree node.

$$\underset{k}{\operatorname{argmax}}(U_C + (1.0 - e^{-\frac{d}{\lambda}})U_R). \tag{1}$$

Once the training has been done, the resulting forest can be used for classification and regression of the face pose from a depth image. This process consists of extracting several patches from the image and passing them through the forest. At the nodes, each patch is tested with the sub-patch combination generated in the training stage and continues to the left or right depending on the test result. The test function (Eq. 2) includes F_1 and F_2 sub-patches size, integral images of these sub-patches ($I(q)$) and the threshold (τ).

$$|F_1|^{-1}\sum_{q\in F_1} I(q) - |F_2|^{-1}\sum_{q\in F_2} I(q) \geq \tau. \tag{2}$$

When a patch arrives at a node, the sub-patches are extracted and their integrals are calculated. Depending on the result, the patch is sent left or right. When the sample arrives at a leaf, it produces one vote encoded by the information stored in that leaf. The leaf could be a face leaf or a non-face leaf. After all the patches have passed through all the trees, all the votes are processed by a bottom-up clustering to remove outliers. All the votes inside the distance of the average head diameter are grouped together. Then 10 mean shift iterations are executed in order to localise the centroid of the clusters. Afterwards, if the number of votes exceeds the threshold, a face is considered as detected. The pose result is obtained from the mean of the values stored in the leaves whose votes were selected.

3.2. Face pose tracking

The pose parameters, as estimated by the algorithm described in the previous section, are often noisy when they are applied to individual images in a video sequence. This is due to the detection was performed without imposing any temporal constraints. To reduce the random error in the pose estimation and to avoid some missed detections, a tracking method is used for processing of video sequences. This method is explained in detail in Henriquez et al. (2014). The method uses the Kalman filter to perform head pose tracking, by filtering the measurements provided by the face detector. Additionally, it can detect outliers and handle the missing measurements and introduces adaptive covariance estimation, which is useful, for example, when the average head movement speed varies. The noise covariance is updated based on the variance estimates of the most recent measurements using a sliding window.

3.3. Face alignment based on 3D data

This section describes a technique developed for alignment of a personalised 3D mask to the depth data using the iterative closest point (ICP) registration algorithm (Quan et al., 2010). Such mask alignment is used in order to further increase pose estimation accuracy. The personalised mask is built for each user, utilising the 3D reconstruction algorithm described in Section 4. When the face is detected in the 3D space, the personalised mask is translated and rotated using the pose parameters calculated in the tracking stage. The rotation matrix is defined by the three Euler angles, and the translation vector containing the coordinates of the head centre (*x, y, z*). All the points belonging to the mask are transformed by using a rigid transformation model. After applying the transformation estimated by the tracker, the mask can fit the input data or being slightly misaligned (see Fig. 3) due to the error in the face pose estimation. To tackle this problem, the location and orientation are refined by applying a rigid registration process between the personalised mask and the input depth data using the correspondence search.

For 3D face alignment the real-time processing was achieved as result of a relatively small number of corresponding points used. With the face pose estimation, the 3D model is initialized close to a correct matching position. Additionally, the random sampling is used in the multi-resolution registration scheme, reducing even more the number of correspondences to be estimated. Random sampling improves also convergence due to reduced correlation bias between points used at the different resolution levels (see

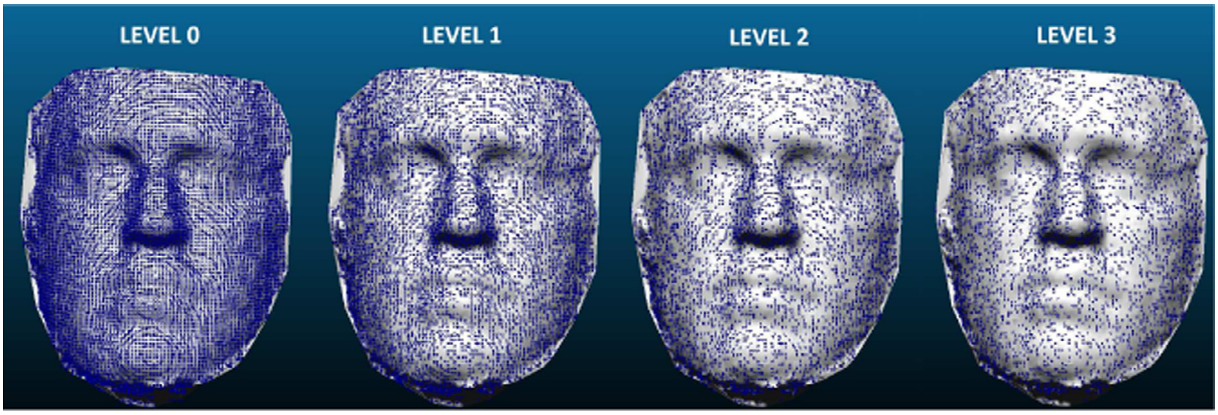

Fig. 4. Example of sub-sampling for four different levels to perform the multi-resolution registration.

Table 1
Sensitivity (True positive rate, TPR) and accuracy (Positive predictive value, PPV) experiments. First row contains the different thresholds used to consider a detection true positive. If the distance between the detected nose position and the ground truth is smaller than the threshold, it is a true positive, otherwise it is considered a false negative. A total of 607 images were processed. RF represents the method described in Fanelli et al. (2011), whereas WM represents the proposed method.

		5	10	15	20	25
TPR	RF	7	44	76	90	93
	WM	**28**	**72**	**88**	**95**	**97**
PPV	RF	7	46	80	**95**	**98**
	WM	**30**	**75**	**89**	**95**	97

Fig. 4). Furthermore, in order to keep the real-time processing constraint at a high frame rate, only four iterations of the ICP are executed as the results showed to be suitable for the post-processing by other functionalities of the system.

3.4. Experimental results

As already explained (see Fig. 2), in the processing pipeline described in this paper, the face pose estimation is used to facilitate the 3D face reconstruction (explained in the next section) and the face detection for the stress and anxiety analysis (introduced in Section 6). To maximise the data spatial resolution the Wize Mirror camera acquiring images for the stress and anxiety analysis (S&A camera) is equipped with a narrow view lens. It is therefore essential to accurately detect the face position in front of the mirror so the acquisition from that camera could be suitably triggered. To evaluate the effectiveness of the proposed solution for that purpose a set of experiments was carried out. They consisted of applying the method from Fanelli et al. (2011) and the proposed method to detect faces in three different sequences, with 607 image frames in total. Each of those frames is labelled with the nose position, therefore the sensitivity (true positive rate) and the precision (positive predictive value) of the methods were estimated depending on the distance between the ground truth and the estimated nose position. As in all the sequences there is a face present in each frame, the true and false positives are defined by a threshold. Five different thresholds were used in the experiments. When the detected nose position is further than the threshold from the ground truth, it is considered as a false positive (i.e., the face may not be fully included in the field of view of the S&A camera). If the distance is smaller than the threshold, the result is counted as a true positive. When the face is not detected, it is considered a false negative. The Table 1 shows the results corresponding to the true positive rate (TPR) and positive predictive value (PPV) for the both tested methods. It can be observed that the proposed method is the one with bigger TPR for all the thresholds. This, in part, is because the proposed method on average has smaller number of false negatives. In terms of the PPV, it can be seen in Table 1 that the results provided by the proposed method improved the detection in most of the distance thresholds (5–25). Additionally, a qualitative comparison can be made by looking at the results showed in Fig. 3. It can be observed, that the orientation results provided by RF method (Fanelli et al., 2011) (shown in the top row) are not as good as for the proposed method (shown in the bottom row). This is despite the fact, that for the images shown in that figure the RF had obtained similar results to the proposed method in the nose distance experiments.

4. 3D face reconstruction

The 3D reconstruction process is based on calculating the different positions of the sensor and merging the 3D data from captured frames to reconstruct the scene (Newcombe et al., 2011). The sensor pose is calculated by tracking the depth data relative to a global model using the iterative closest point algorithm. Afterwards, a truncated surface distance function is applied to merge the new data with the reconstructed model. Finally the surface is predicted using a ray casting algorithm. In order to extract from the depth data only the information representing the face, the range data segmentation is needed. This step eliminates background objects, body parts or hair from the reconstruction process. Without the segmentation the reconstruction can be noisy or/and heavily distorted. The proposed method introduces a modification to the technique proposed in Newcombe et al. (2011) and extended in Macedo et al. (2013). The additional processing step applies a face segmentation method using an average face model in order to obtain the region of interest for the reconstruction and to invert the face movement to the equivalent sensor movement. Two segmentation stages are applied, the first one is based only on depth information and the second is using an average 3D face model/mask.

4.1. Face segmentation

Normally, a face detection technique is used to localise the face centre and select the region of interest for the reconstruction (Macedo et al., 2013). However, a depth segmentation method can be as well an easy and fast way to remove from the image those background and body parts which can produce deformations in the face reconstruction. The typical objects removed as part of this process include: neck, shoulders or objects in the background. The proposed depth segmentation is a variation of the technique

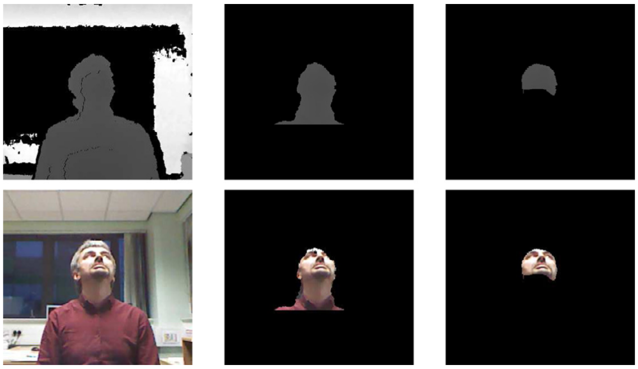

Fig. 5. Comparison between the two different segmentation stages used for preprocessing the input depth data for subsequent 3D reconstruction. Input depth images (left), depth segmentation (centre) and model segmentation (right). The colour images are only used to visualise the segmentation results. The colour information is not used in any of the segmentation methods.

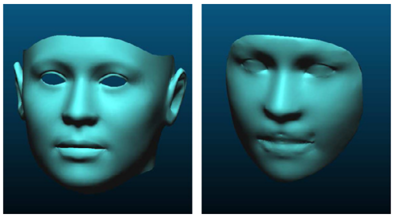

Fig. 6. Average face models used for face segmentation, generated using FaceGen software. Model used for the reconstruction needed in morphological analysis of cardio-metabolic risks (left). Model utilised to build the personalised mask for tracking (right).

proposed in Zollhofer et al. (2011), where using face landmarks as seeds, the rest of the points belonging to the face are found with a flood fill algorithm. In each recursion a four neighbourhood of a current face point are checked in order to evaluate if the depth values change by more than 5 mm. If the change is smaller, the point is added to the segmented face. In the proposed modification of the method the seed is initialised in the 2D projection of the detected 3D head centre, which offers similar results without the need for detecting more facial features. Some examples of face segmentations are shown in Fig. 5.

It can be observed that in most cases the neck and chest patches are included in the segmentation. This can be a problem for the reconstruction process as these extra patches are unreliably included in some frames, producing distortions in the 3D face reconstructions. Additionally this depth segmentation method strongly depends on the posture of the user as it implicitly assumes that the head is always at least 5 mm nearer to the sensor than the rest of the neck or the upper body. As it was explained above, the depth based segmentation method can fail if the threshold to differentiate the face from the neck is not well chosen. The optimal value of this threshold is subject specific and therefore difficult to select. To overcome this problem, a model segmentation approach has been proposed. Based on the face pose estimation, a 3D model is transformed to match the input depth data. The matched model defines the points which are subsequently used for the 3D face reconstruction. Two different average models have been used for this purpose (see Fig. 6). One of them includes the ears and is used for the 3D reconstruction which is the input for morphological analysis of cardio-metabolic risk. The personalised face mask for tracking is built using the model without ears.

When the face is detected in the 3D space, using the method explained in previous sections, the model is translated and rotated using the estimated pose parameters as it is performed for the face tracking. Then, all the points belonging to the model are transformed by using the estimated rigid transformation model and the ICP algorithm. Afterwards, all the points belonging to the model are projected to a depth image using the camera calibration parameters building a depth sparse segmentation. In order to generate a dense and continuous area instead of a set of points, mathematical morphology is applied to the image (dilation and erosion), followed by a contour detection and a flood fill algorithm to remove holes. This technique provides more robust face segmentation for different subjects and varying postures (see Fig. 5).

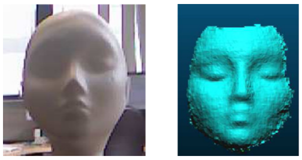

Fig. 7. Plastic head model used for the reconstruction experiments (left), reconstructed model using the proposed method (right).

4.2. Sensor pose estimation

This stage of the process is based on the sensor pose estimation proposed in Newcombe et al. (2011). Originally, that reconstruction method was designed to reconstruct static scenes of rigid objects by moving the sensor and capturing data from different points of view. The sensor pose is calculated by tracking the depth data relative to a global model using the iterative closest point algorithm. The reconstruction requirements for the studied scenario are slightly different, as the sensor is in a fixed position and the person is moving. Some modifications in the above explained method were introduced in order to use it for face reconstruction. The person motion is reversed to estimate the relative motion of the sensor with the head being in a virtual fixed position. The depth image is processed with the segmentation method explained in the previous section, and only the face region is used as input for the reconstruction method described in Newcombe et al. (2011). Hence, when the only information available in the depth data is the user's moving face, the system calculates the equivalent sensor motion with the user's face being still. After segmenting the face, this subsystem tracks the current sensor frame by aligning a surface measurement against the model prediction by minimising the cost function given in Eq. 3. T_k is the new sensor's pose, V_k is the vertex map of the new depth data in the sensor reference frame, $\hat{V}_{k-1}(\hat{u})$ is the predicted vertex map and $\hat{N}_{k-1}$ is the predicted normal map of the model in the global reference frame. The correspondence $u \to \hat{u}$ between vertices is estimated as part of the optimisation process (see Newcombe et al., 2011 for more details).

$$E(T_k) = \sum_{u} \| (T_k V_k(u) - \hat{V}_{k-1}(\hat{u}))^T \hat{N}_{k-1}(\hat{u}) \|_2 \quad (3)$$

4.3. Surface reconstruction

The surface reconstruction is performed by means of a volumetric truncated signed distance function (TSDF) (see Newcombe et al., 2011). After the sensor pose is estimated for a given depth frame, that frame is fused into one single 3D reconstruction containing data from previous depth frames. This global TSDF contains the fusion of the registered depth frames. The reconstructed volume is formed by the weighted average of all individual TSDFs computed for each depth map. This global fusion can be interpreted as denoising, with the global TSDF obtained from multiple noisy TSDF measurements, see Eq. 4 where F_{R_k} are the truncated signed distance values, W_{R_k} the corresponding weights and F the signed dis-

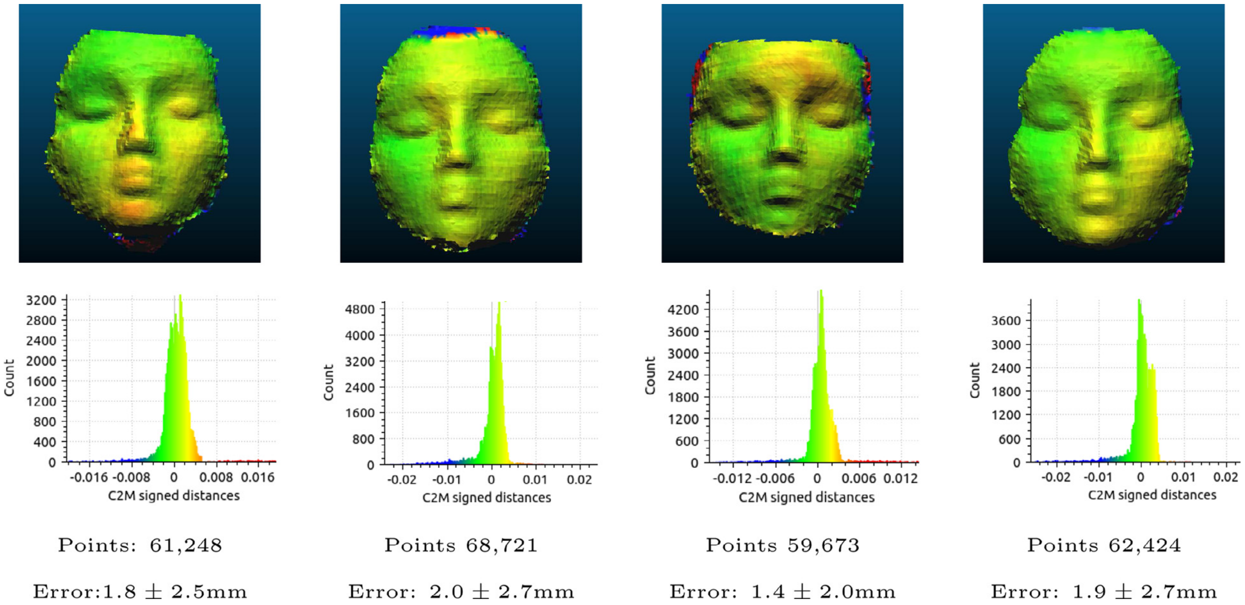

Fig. 8. Comparison between 3D reconstructions obtained using the proposed method. The images on the top represent the signed distance between the two reconstructions: the current reconstruction and a reference reconstruction. The histograms (on the bottom) are calculated with the number of points belonging to the reconstructed face (63,000 points on average) and clustered depending on their signed distance (in meters) to the reference scan. The average error for all the experiments is 1.7 ± 2.4 mm.

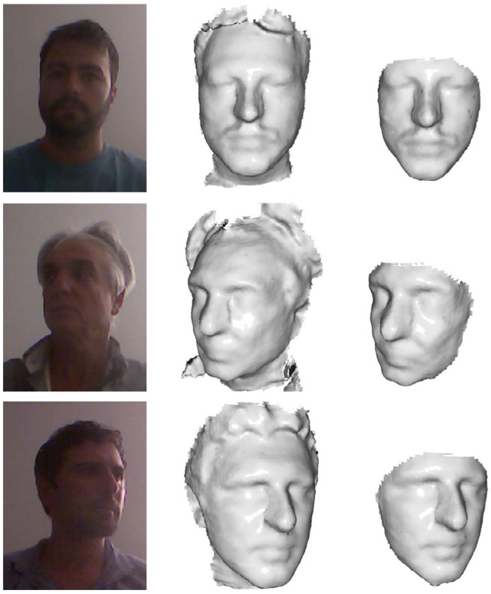

Fig. 9. 3D geometric reconstruction results. RGB image (left), 3D reconstruction for morphological analysis of cardio-metabolic risk (centre), and personalised mask for face tracking (right).

tance function.

$$\min_{F\in\mathcal{F}} \sum_{k} \|F_{R_k} W_{R_k} - F\|_2 \tag{4}$$

After all the input depth maps have been fused to the global model, the reconstruction is complete and a ray casting algorithm is applied in order to estimate the final surface (Newcombe et al., 2011). A sample of the reconstruction results is shown in Fig. 9. Where the middle column shows reconstructions obtained using the depth segmentation technique, and the right column contains the reconstructed faces using the model/mask based segmentation method. It can be seen that the use of model segmentation provides a cleaner face reconstruction which can be used for face tracking and also for morphological analysis.

4.4. Experimental results

The 3D face reconstruction method has been validated through different experiments, using a plastic head model and real faces. Fig. 7 shows, on the left, an image of the plastic head model used in the experiment, and on the right the corresponding reconstructed model using the proposed technique.

The morphological analysis which is subsequently performed on the 3D reconstructions is based on comparing different reconstructions from the same person obtained at different dates. Therefore, it is important that the 3D scanner provides consistent and repeatable results and does not add random error in the reconstructions which may lead to errors in the analysis. To check the stability of the 3D reconstruction obtained using the proposed method, the reconstructions of the plastic head model were repeated multiple times with differently acquired range data. In the first experiment, four different reconstructions were compared to randomly selected reconstruction treated as the reference reconstruction. The plastic head model was scanned five times, from slightly different positions and inclinations in front of the sensor. The reconstruction process requires rotation of the user's face, in this experiment the plastic head model was rotated manually. As it can be seen in Fig. 8, the average error is only 1.7 mm, which indicates that the scanner provides repeatable reconstructions from the same surface independently from the small changes in the position or orientation. This is an important result as it shows that the random reconstruction error, which is difficult to correct, is small.

Another experiment was performed using real faces. The users rotated their heads in front of the sensor and the depth data was captured. The face was tracked and segmented in each frame, and the resulting segmented data was used for reconstruction. The results in Fig. 9 show that the proposed model based segmentation is able to get rid of the hair, neck and shoulder regions (right column in the figure), which otherwise could introduce noise in the subsequent uses of the 3D reconstructed models, for instance if the reconstructed personalised face mask is used for tracking.

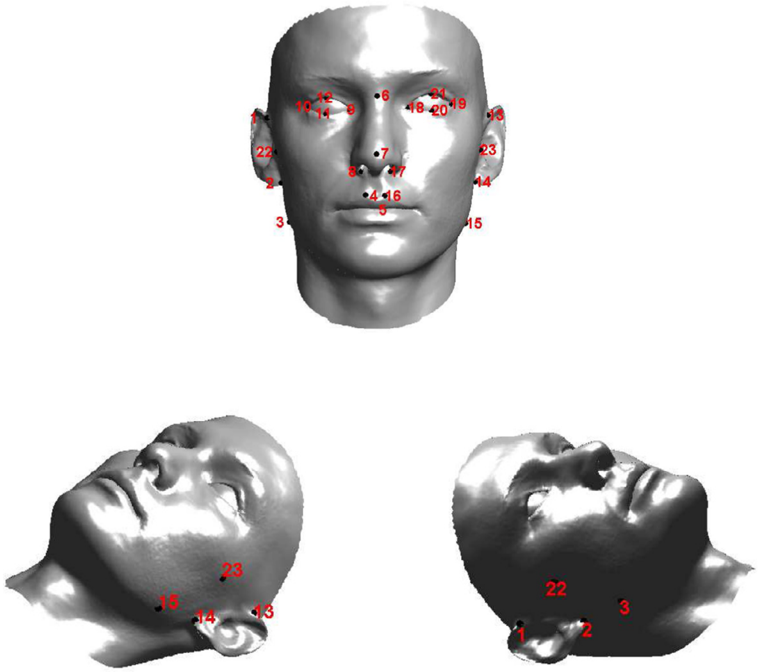

Fig. 10. The 23 landmarks used to analyse faces from the morphological viewpoint.

5. Morphological analysis of cardio-metabolic risk

Our goal is the quantification of patterns in face shape variation due to weight gain. Indeed, according to the semeiotic model of the face for cardio-metabolic risk developed in SEMEOTICONS, the face signs include signs of overweight and obesity. The signs must be computed on a 3D face model reconstructed from range data acquired by a 3D scanner, as described in the previous sections.

Though several authors studied the application of anthropometric analysis to classify normal weight, overweight, and obese individuals, most of the methods in the literature are based on measurements taken on the body of subjects, rather than on their face, as foreseen in SEMEOTICONS' Wize Mirror. Moreover, most of the techniques considering faces are based on measures computed on 2D images rather than on 3D models. Finally, though it is well known that the face is involved in the process of fat accumulation, there is no consensus in the literature about which are the facial morphological correlates of body fat. All these issues make our task a challenging one.

5.1. Landmark-based measurements

The starting point of our research was the study in Lee and Kim (2014), whose authors computed a set of simple linear and planar measurements on 2D face images and evaluated the statistical correlation of each measurement with waist circumference (and hence visceral fat) on a set of 11,347 adult Korean men and women aged between 18 and 80. The measurements included Euclidean distances between the 23 anthropometric landmarks (cf. Fig. 10), and areas of polygons enclosed by the landmarks. Table 2 lists the measurements which were found to have strong correlation with waist circumference (p-value less than 0.005).

Table 2
List of linear and planar measurements which were found to correlate with waist circumference in Lee and Kim (2014). d_E stands for Euclidean distance, d_H horizontal (Euclidean) distance, d_V vertical (Euclidean) distance, and $A(p_1, \dots, p_n)$ is the area of the polygon formed by points $p_1, \dots, p_n$. Fig. 10 explains both the position and the label of the landmarks.

FEATURE	DESCRIPTION
$f1$	$d_H(8, 17)$
$f2$	$d_V(5, 7)$
$f3$	$d_E(3, 15)$
$f4$	$d_E(1, 13)$
$f5$	$d_E(2, 14)$
$f6$	$d_H(22, 23)$
$f7$	$d_E(22, 23)$
$f8$	$A(1, 13, 23, 14, 2, 22)$
$f9$	$A(2, 14, 15, 3)$
$f10$	$f6/f3$
$f11$	$f6/d_V(6, 5)$
$f12$	$f3/d_V(6, 5)$

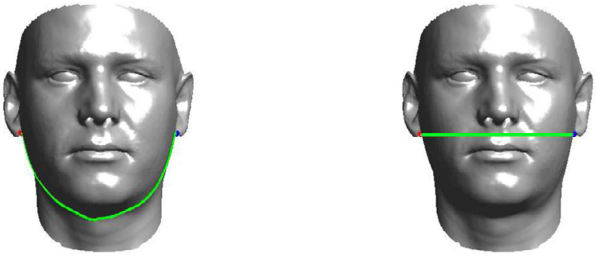

Fig. 11. Geodesic (left) and Euclidean (right) distance between two landmarks.

We implemented the measurements in Table 2 on 3D face data. Moreover, thanks to the availability of complete 3D data rather than 2D images only, we computed additional measures based on geodesic distances between selected anthropometric landmarks. Briefly speaking, geodesic distances measure the shortest path between two points along the surface, that is, the path one would follow if bounded to walk on the surface of the object (Biasotti et al., 2014). Therefore, geodesic distances capture information which is substantially different from their Euclidean counterpart. This can be appreciated in the example in Fig. 11, where the geodesic distance (left) between the two landmarks measures the length of the path passing below the chin, whereas the Euclidean

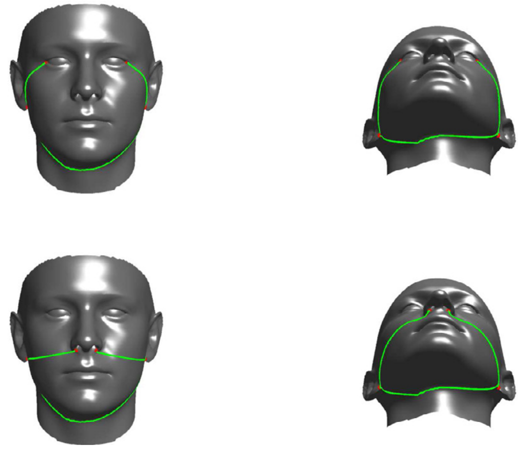

Fig. 12. Two views of each curve passing through four landmarks, on a 3D face model. In the first (resp. second) row is visualized the geodesic path *a* (resp. *b*).

Table 3
The two geodesic-based features computing the length of paths in Fig. 12.

FEATURE	DESCRIPTION
$Lgeod_a$	Length of the geodesic *a*
$Lgeod_b$	Length of the geodesic *b*

distance (right) measures the horizontal distance between the points.

Our idea was to look for geodesic paths able to account for weight variations. We experimented with paths passing through different sets of way-points, and found two sets of way-points generating informative paths (Fig. 12). With the notation used by Farkas notation (Farkas, 1994), the landmarks which define the two geodesic paths *a* and *b* are:

- geodesic path *a: exocanthion* (eye) left - *subaurale* (ear) left - *subaurale* right - *exocanthion* right ;
- geodesic path *b: alare* (nose) left - *subaurale* left - *subaurale* right - *alare* right.

It cannot be assumed that a geodesic path joining landmarks 2 and 14 always goes through the same surface for any real face, e.g. through the neck. Thus, in the real setting a proper constraint should be used in order to ensure the geodesic path passing through the desired surface, e.g. adding a specific extra way-point in the neck region. For the specific set of experiments reported in this paper, it has been visually verified that both the geodesic paths *a* and *b* pass through the desired region of the face.

We computed the lengths of each path, and used them as features to quantify facial changes due to weight gain, as summarized in Table 3.

5.2. Landmark-independent features

A drawback of the measurements above is that they rely on the accurate identification of anatomical landmarks on the 3D face mesh. As suggested in Giachetti et al. (2015), whereas in the case of manual anthropometric measurements landmarks are identified by expert anthropometrists by observation and palpation, automatically locating landmarks with optimal accuracy on 3D acquired data could be difficult. This holds especially for poorly geometrically characterized landmarks, or landmarks located near regions subject to occlusions, for example due to the presence of hair. Since small errors in detecting the landmarks on real data could affect badly the feature computation, we decided to develop a technique based on shape features independent of the precise, optimal location of anatomical landmarks. We defined a set of planar curves given by the intersection of a face mesh with *p* parallel planes perpendicular to the *z*-axis (Fig. 13). We experimented with $p = 10$. Slicing an object and evaluating sections is a classical idea in geometry, which finds many different applications (including 3D printing technology). Among the many properties which can be computed on planar curves (e.g. curvature), we experimented with average and maximum lengths, which are easily computed from scanned data and robust to noise.

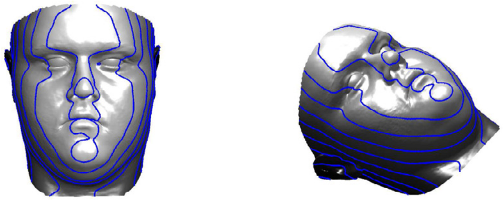

Fig. 13. Sections given by the intersection of the 3D face mesh with equally-spaced planes perpendicular to the *z*-axis.

Table 4
List and description of sectional features.

FEATURE	DESCRIPTION
meanLZ	Average length of the sections
meanAZ	Average area of the polygons enclosed by the sections
maxLZ	Maximum length of the sections
maxAZ	Maximum area of the polygons enclosed by the sections

5.3. Experimental results

Since our essential objective is the description of morphological change over time on a subject, we must check whether our techniques enable us to discover a trend in a longitudinal study. To this end, we generated a dataset of synthetic 3D faces simulating weight changes using a parametric deformable model, namely the Basel Face Model (Paysan et al., 2009). The Basel Face Model provides specific parameters to be tuned for simulating fattening. Moreover, data are labelled with different sets of anatomical landmarks (Farkas and MPEG4-FDP feature point coordinates and indices). These characteristics make the Basel Face Model a natural and effective choice for producing synthetic data to help assessing the techniques we developed.

Twenty-five faces were randomly generated as seeds, and each face was morphed to simulate the process of gaining weight, with 10 equally spaced intervals. This gave a dataset of 250 faces, divided into 10 groups ordered according to increasing fatness. Fig. 14 shows a sequence of fattening faces of the same individual.

In the following we evaluate the features introduced above, with respect to the inter-cluster separability and with respect to the history of an individual. Separability deals with the capability of each feature in classifying a sample by weight, among the whole dataset. The other criterion refers to the ability of reading correctly the weight variations in an individual's history.

5.3.1. Analysis of separability

A first analysis serves to check whether the features listed in Table 2, 3, and 4 are able to separate the faces of people in the 10 groups corresponding to different fatness levels. This can be qualitatively and quantitatively measured by evaluating the inter-cluster separability and intra-cluster homogeneity of the 10 clusters in the embedding space given by the features. Fig. 16 shows the scatter plots for the subjects belonging to three groups of fatness: level 1, in red, level 5, in green, and level 10, in blue) in the embedding

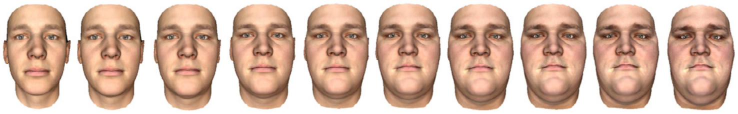

Fig. 14. A sequence of faces generated from the same seed, increasing weight in ten stages.

Table 5
List of all the features, compared each other with respect to the cluster separability. The five best (bold) performing features are: *f*3, *f*6, *f*7, *Lgeod*$_a$, *Lgeod*$_b$.

FEATURE	Cluster separability	Ranking
*f*1	69.11	15
*f*2	198.70	18
***f*3**	36.73	1
*f*4	42.50	8
*f*5	41.21	7
***f*6**	38.54	2
***f*7**	38.58	3
*f*8	53.52	12
*f*9	49.07	11
*f*10	119.90	17
*f*11	44.05	9
*f*12	40.42	6
***Lgeod*$_a$**	39.70	5
***Lgeod*$_b$**	39.19	4
meanLZ	47.92	10
meanAZ	60.45	13
maxLZ	63.44	14
maxAZ	75.06	16

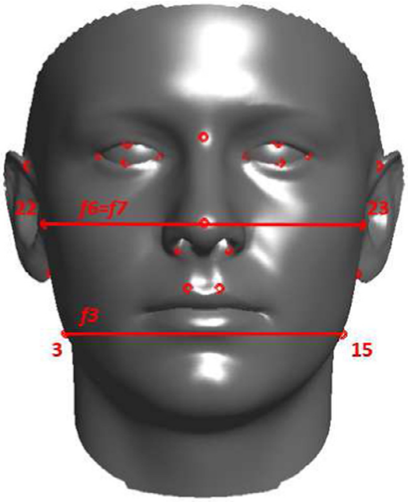

Fig. 15. A visualization of the features *f*3, *f*6, and *f*7. Note: due to the symmetry of the face model used, *f*6 and *f*7 are equal.

space given by the features *f*1 and *f*2, *f*3 and *f*6, *f*11 and *f*12, *Lgeod*$_a$ and *Lgeod*$_b$, *meanLZ* and *meanAZ*. For each feature *f*, the separability can be quantitatively measured by evaluating the *total separation* between clusters. Define μ_i as the centre of the $i - th$ cluster, $i = 1, \ldots, 10$, with 10 the number of fatness levels in our dataset. The total separation is defined as Haldiki et al. (2001)

$$sep = \frac{D_{max}}{D_{min}} \sum_{i=1}^{10} \left(\sum_{j=1}^{10} ||\mu_i - \mu_j|| \right)^{-1}$$

with D_{max} (resp. D_{min}) the maximum (resp. minimum) distance between cluster centers. Table 5 summarizes the results: the best performing features are the lengths of the geodesic paths (showed in Fig. 12), and *f*3, *f*6, *f*7 (in Fig. 15).

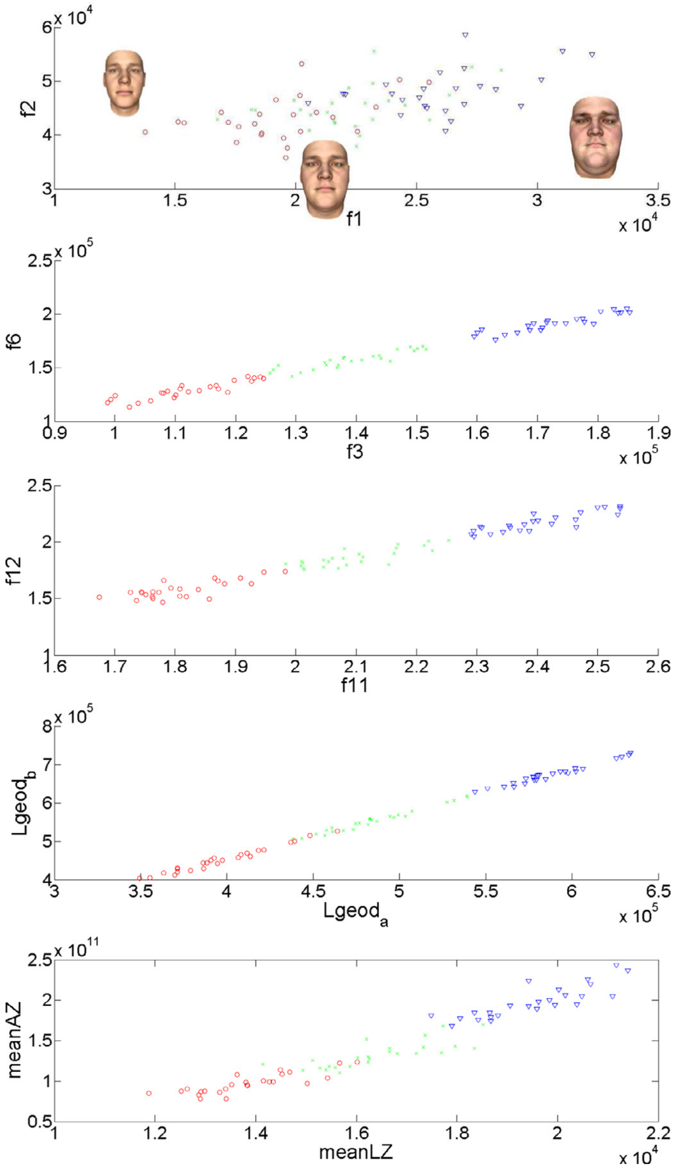

Fig. 16. Scatter plots for the subjects belonging to three groups of fatness: level 1, in red, level 5, in green, and level 10, in blue) in the embedding space given by the features *f*1 and *f*2, *f*3 and *f*6, *f*11 and *f*12, *Lgeod*$_a$ and *Lgeod*$_b$, *meanLZ* and *meanAZ*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

From both a qualitative and quantitative analysis it can be observed that not all the features listed in Lee and Kim (2014) as correlated with waist circumference provide a good separation among people with different fatness levels. Moreover, the length of geodesic paths on the 3D surface provides a comparable or better clustering than the features in Lee and Kim (2014). More notable is the performance of sectional features: though extremely simple to compute and completely independent of the pre-computation of anatomical landmarks, especially *meanLZ* seems to be able to identify facial characteristics correlated with the amount of fat. The

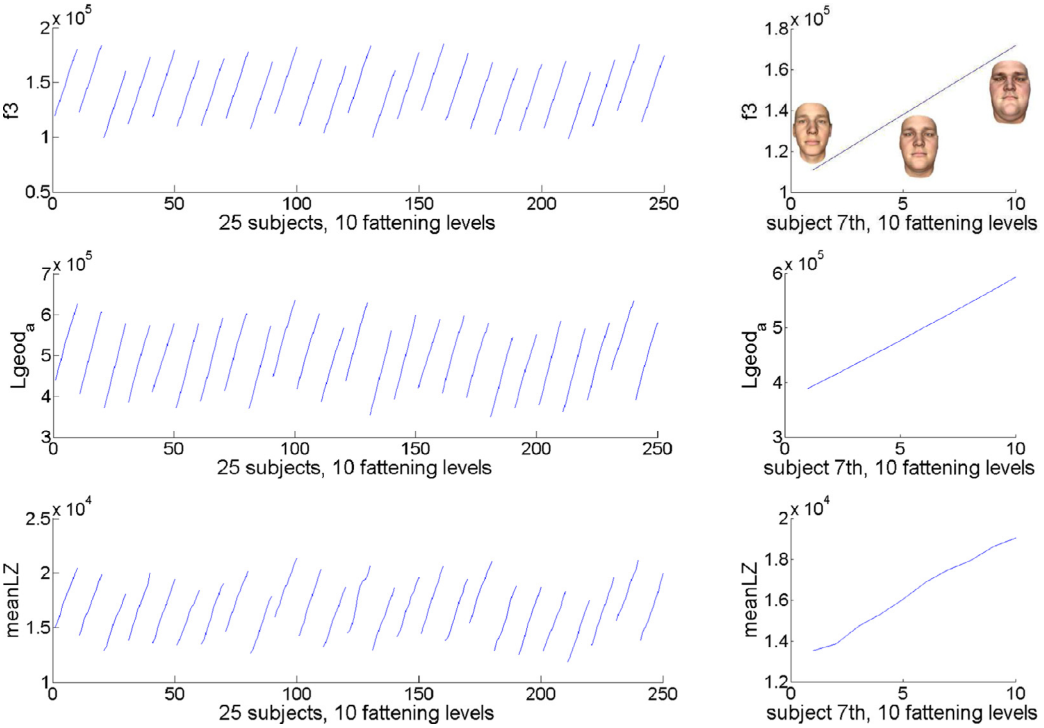

Fig. 17. Graphs of a selection of the features ($f3, Lgeod_a, meanLZ$), computed on the whole dataset; with a zoom on the 7th seed.

performance of sectional features will be further commented in the next section about the monitoring of individual face changes.

5.3.2. Tracking individual changes

Besides evaluating the capability of separating people in different groups, we must also check whether our features enable us to detect morphological changes over time on a subject. In other words, we must check if our features are able to discover a trend in a longitudinal study, by tracking the facial morphological changes on a single individual gaining weight. This is the usage scenario in which the Wize Mirror will operate. A way to do this is visualizing the behaviour of the linear and planar measures on each of the 25 seeds in the dataset along the simulated weight gain. In other words, each individual has a trajectory graph which is made of ten consecutive points. For a given trajectory, we can analyse four attributes, namely *location* (the starting and ending points); *orientation* (the direction of the vector between the endpoints); *size* (the magnitude of the vector between the endpoints); and *shape*. In our context, the location depends on the specific, initial traits of each individual. The orientation is crucial: a consistent orientation would indicate that our technique is able to detect and encode the process of getting weight. The size is a measure of the difference in shape between the thinnest and the fattest morphing of the individual. The shape indicates how the features change along the morphing process.

Fig. 17, first column, shows the trend of the features $f3, Lgeod_a, meanLZ$, computed on the whole dataset; for each plot, the 25 lines represent the 25 seeds and the behaviour of the feature while simulating weight gain on that seed.

A zoom on a single seed (7*th*) is showed in the last column to better appreciate their attributes: the *shape* of each feature is strictly increasing for all, and almost linear; the *orientation* (increasing from left to right) is consistent with fattening. As regards the *size*, we remark that its order of magnitude is 10^5 for $f3$ and $Lgeod_a$, while is 10^4 for *meanLZ*. For $f3$ and $Lgeod_a$, a linear trend is showed, with an average slope (over the 25 seeds) of $6.79 \cdot 10^3$ for $f3$, and $21.27 \cdot 10^3$ for $Lgeod_a$. This means that they are expected to track accurately the evolution of the face morphology while gaining weight, as envisaged in the Wize Mirror usage scenarios.

5.4. Experiments on real data

Our results on a synthetic dataset showed that most of the measurements implemented are able to identify individual weight variation patterns, and to separate thinner from fatter people, to a different extent. Each class of measurements has its pros and cons. Landmark-based measures have the obvious drawback that they require a pre-processing step, which can affect the results on real data. Landmark-independent measures strike a compromise between efficiency and efficacy, according to the Wize Mirror usage scenarios.

The present study on the geometric features able to account for the body weight and body weight change from the 3D facial data is relatively comprehensive but preliminary: a large testing on real face is required to validate all the measurements implemented, then to assess which one is the best performing in the task of monitoring individual weight change. In the next few months, longitudinal validation study will be conducted at three pilot sites on approximately sixty volunteers. This will serve to reinforce findings reported in this paper. In order to verify that the most interesting measurements implemented are feasible to be computed also on real data, a small test has been carried out on ten subjects with the 3D data captured using method described in Section 4. A sample of these results is presented in Table 6, while Fig. 18 shows the scatter plot of $f3$ vs BMI and weight, $LGeod_a$ vs BMI and weight, *meanLZ* vs BMI and weight for all subjects.

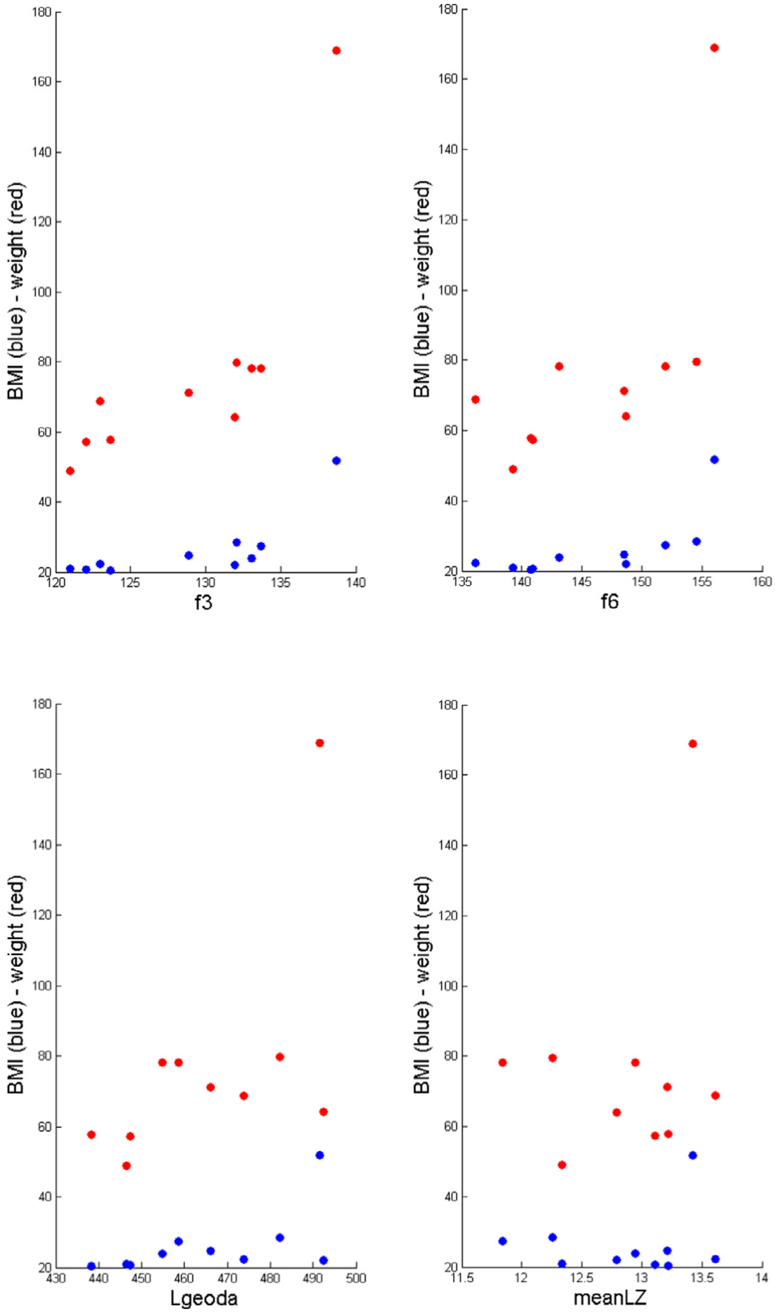

Fig. 18. Selected geometric features: *f*3, *f*6, *Lgeod$_a$*, *meanLZ*, computed on a set of 10 subjects. Results are visualised as scatter plot of each feature vs BMI (blue) and weight (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

6. Stress and anxiety analysis

6.1. Measures for stress and anxiety

As mentioned in Section 2, the facial signs of stress and anxiety are the result of deviating motion patterns of facial musculature. The two main regions that exhibit most of the muscular activity are the eyes and the mouth. The third region is the head itself. In order to cover these major regions in a non-invasive and integrated approach for the detection of stress and anxiety, three methods are applied, each targeting one of the three regions, with the region selection facilitated by the head pose estimation introduced in Section 3.

6.1.1. Eyelid motion

The first method focuses on analysing eyelid related motion, specifically, the blink rate and eyelid opening. It uses active appearance models (AAM) (Cootes et al., 2001), which have been widely applied in facial expression analysis, as well as for facial expression classification (Hamilton, 1959), as they provide a consistent representation of the shape and appearance of the face. AAMs are considered as models containing shape and texture for modelling the human face.

The applied AAM has 68 facial landmarks in total, out of which only 12 are used. The remaining landmarks were not removed since they help in aligning the AAM with the face, especially those on the facial perimeter. Moreover, the usage of a complete (whole face) AAM is useful for extracting additional features such as eyebrow movements, head orientation and lip deformation for future

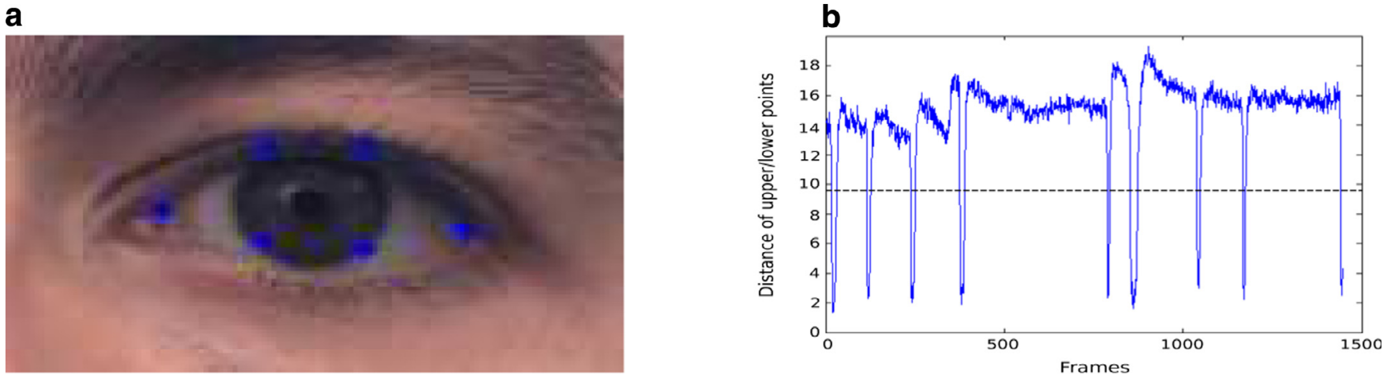

Fig. 19. Eye opening average distance calculation between the two upper and lower eye-lid points (a) and variability of eye average distance and blink threshold (b).

Table 6
Results sample for 4 subjects. For each subject BMI and weight were collected ; and some of the geometric features implemented were computed: *f*2, *f*3, *f*6, *Lgeod$_a$*, *meanLZ*.

FEATURE	Sub 1	Sub 2	Sub 3	Sub 4
BMI	21.7	24.6	28.5	51.8
Weight(kg)	74.2	71.2	79.6	168.8
*f*2	29.50	31.58	34.66	49.31
*f*3	129.57	128.87	132.05	138.70
*f*6	147.61	148.53	154.52	156.04
Lgeod$_a$	472.88	466.18	482.23	491.42
meanLZ	235.3	236.8	232.8	241.1

studies. For extracting the blink rate, the AAM is used in order to segment the eyelid area and to mark out the eyeball perimeter with specific landmarks (six landmark points for each eye). Then, the average distance between the two upper and lower eyelid points as shown in Fig. 19(a) is calculated. Eye blinks can be seen as sharp negative spikes in the extracted signal as shown in Fig. 19(b).

A threshold is established after visual inspection of the data, and an eye blink is detected if the distance remains below that threshold for the next 100 ms. Extreme value analysis is performed on the data, excluding outliers in case a specific subject has motor tics, thus directly affecting the measured eye blinks. Finally, the eye opening is calculated as the mean distance between the points of upper and lower eyelid.

Training and fitting the AAMs

The shape model is built as a parametric set of facial shapes. A facial shape is described as a set of $L \in R^2$ landmarks forming a vector of coordinates $X = [\{x_1, y_1\}, \{x_2, y_2\}, \dots, \{x_L, y_L\}]^T$. Their distribution on the human face is shown in Fig. 20. A common mean model shape is formed by aligning face shapes through Generalized Procrustes Analysis. The alignment of any new estimate leads to the mean shape re-computation and the shapes are aligned again to this mean. This procedure is repeated until the mean shape doesn't change significantly within iterations (cf. Fig. 21). In the next step, Principal Components Analysis (PCA) is employed, projecting data onto an orthonormal subspace in order to reduce data dimensionality. According to this procedure, shapes s are expressed as

$$s = s_0 + \sum p_i s_i \tag{5}$$

where s_0 is the mean model shape and p_i has the model shape parameters.

The appearance model is built as a parametric set of facial textures. A facial texture A of m pixels is represented by a vector of intensities g_i:

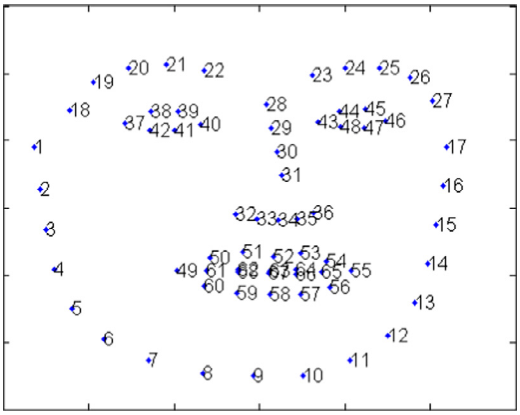

Fig. 20. Spatial distribution of landmarks on human face.

$$A(x) = [g_1 g_2 \dots g_m]^T \forall x \in s_0 \tag{6}$$

As with the shape model, the mean appearance A_0 and the appearance eigen-images A_i are normally computed by applying PCA to a set of shape normalized training images. Each training image is shape normalized by warping the training mesh onto the base mesh s_0 (Matthew and Baker, 2004). After the use of PCA textures A_i can be expressed as

$$A(x) = A_0(x) + \sum \lambda_i A_i(x) \tag{7}$$

where $A_0(x)$ is the mean model appearance and λ_i are the model appearance parameters. It is clear that the model (shape and appearance) depends strongly on the image dataset used for its creation. When the model is created, its fitting to new images I or video sequences turns to be the identification of shape parameters p_i and appearance parameters λ_i that produce the most accurate fit. This non-linear optimization problem pursuits to minimize the objective function

$$\sum_x [I(W(x;\, p)) - A_0(W(x;\, \Delta p))]^2 \forall x \in s_0 \tag{8}$$

where W is a warping function.

6.1.2. Mouth activity

The second method targets motion patterns of the mouth, especially high frequency patterns such as lip twitching, with the aim to provide a quantitative analysis of mouth motion activity. The majority of related work on lip motion analysis deals with automatic lip reading systems that aim to support audio-based speech recognition. In this context, Hojo and Hamada (2009) use space-time interest points, these are extensions of 2D interest point

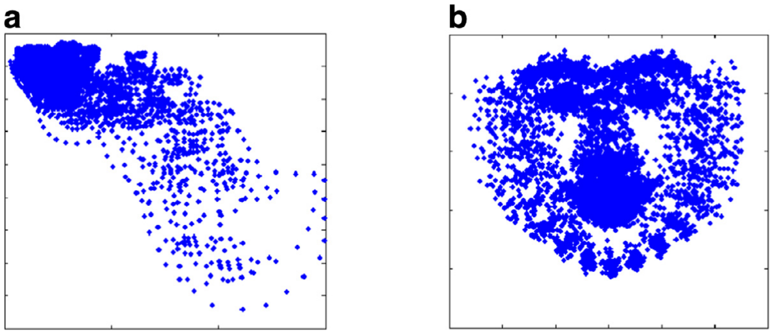

Fig. 21. Landmarks distribution (a); landmarks distribution after GPA alignment (b).

detectors that incorporate temporal information, while Mase and Pentland (1991) use optical flow around the mouth. A further approach for real-time face and lip tracking with facial expression recognition is described by Oliver et al. (2000), who use 2D blob features and a hidden Markov model for their implementation.

The algorithm that was implemented in this work for lip motion analysis uses optical flow, which is a velocity field that transforms one image to the next image in a sequence. It works under two assumptions. The motion must be smooth in relation to the frame rate, and the brightness of moving pixels must be constant. In this work, the velocity vector for each pixel is calculated by using dense optical flow as described by Farneback (2003). The mouth region of interest (ROI) is detected using the mask described in Section 3.3 and split in two horizontal areas for defining the upper and lower lip regions. The upper area has a height of 35% of the total mouth ROI height. The remaining 65% is for the lower lip area, while the width is the same for all ROIs. The maximum velocity is extracted for each of the two ROIs from the computed velocity field, gained by applying optical flow only on the Q channel of the YIQ transformed image, since the lips appear brighter in this channel (Thejaswi and Sengupta, 2008). Finally, for each signal five features are extracted by using a sliding window of 0.5 s in duration and an overlap of 50% over the maximum velocity signal. This short duration reflects the short duration of lip twitches, although a larger duration can be applied for gaining information for long term mouth activity patterns. The five extracted features have been selected among other in order to produce the best results concerning lip twitching detection. These features are:

- The variance of the signal inside the window.
- The skewness of a sample distribution, which is defined as the ratio of the 3rd central moment to the 3/2th power of the 2nd central moment (the variance) of the samples.
- The variance of the time intervals between any two subsequent spikes or transients. This feature is used for estimating the periodicity of the movements based on the observation that rhythmic movements would produce variances close to zero.
- The mean crossing rate, which is the rate of mean crossings along the signal.
- Dominant frequency, which is the frequency with the highest power, derived from the power spectral density, which is calculated with the Discrete Fourier Transform (DFT).

Finally, the 10 features in total (five for the upper lip ROI and five for the lower lip ROI) are fed into a random forest classifier. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.

6.1.3. Head motion

The head motion algorithm is able to detect and measure movements of a person's head from a 2D video at the actual frame-rate. The algorithm measures the head movements in terms of horizontal and vertical deviations of specific reference points between consecutive frames. In Fig. 22 the flowchart of the algorithm is shown.

As implemented in the Wize Mirror, the algorithm starts with the face detected using the robust face segmentation method explained in Section 4.1. A local ROI has to be selected in absence of, or with very low local movements in order to optimally measure the head motion and to discard movements that are related with facial expressions, such as mouth movements, eye blinks, and other facial expressions. According to Irani et al. (2014) the region between the eyes and mouth is the most appropriate region since it does not contain local movements and has the least possible involvement with facial expressions. After the definition of the ROI, specific reference points (i.e. landmark points) that are located at the four edges of the ROI are selected. Then, a tracker based on optical flow (Lucas and Kanade, 1981) is applied for tracking the landmark point position in each frame. In order to keep only the most stable reference points and discard erratic trajectories, the maximum distance traveled by each point between consecutive frames is calculated and points with a distance exceeding the mode of the distribution are discarded (Balakrishnan et al., 2013). Finally, the reliable reference point trajectories are analysed in order to produce six different time series related to frame by frame movement and speed: the horizontal and vertical scalar components, and the resulting vector (Manousos et al., 2014). From the above time series, the mean, median and standard deviation in both x and y directions and the vector magnitudes of speed and movement have been extracted as representative features.

6.2. Assessment of the performance of each algorithm

6.2.1. Eyelid motion

The algorithm was evaluated using the *Pisa I experiment dataset*. This dataset was acquired during a campaign organised within the framework of the SEMEOTICONS project, where several videos of 23 participating subjects were collected. The videos were collected while participants were: (i) in a neutral state, (ii) while simulating a situation of stress or anxiety, (iii) while performing a stressful task (e.g. Stroop test), and (iv) finally while watching a set of relaxing and stressful images and videos. After the session, the participants were asked to score their stress or anxiety perception. The training of the AAM model was performed using 138 images from the dataset (including all subjects) from a population having both eyes open and eyes closed.

An assessment, performed on 10 videos from five subjects (two videos for each subject) led to an accuracy for the eye blink rate

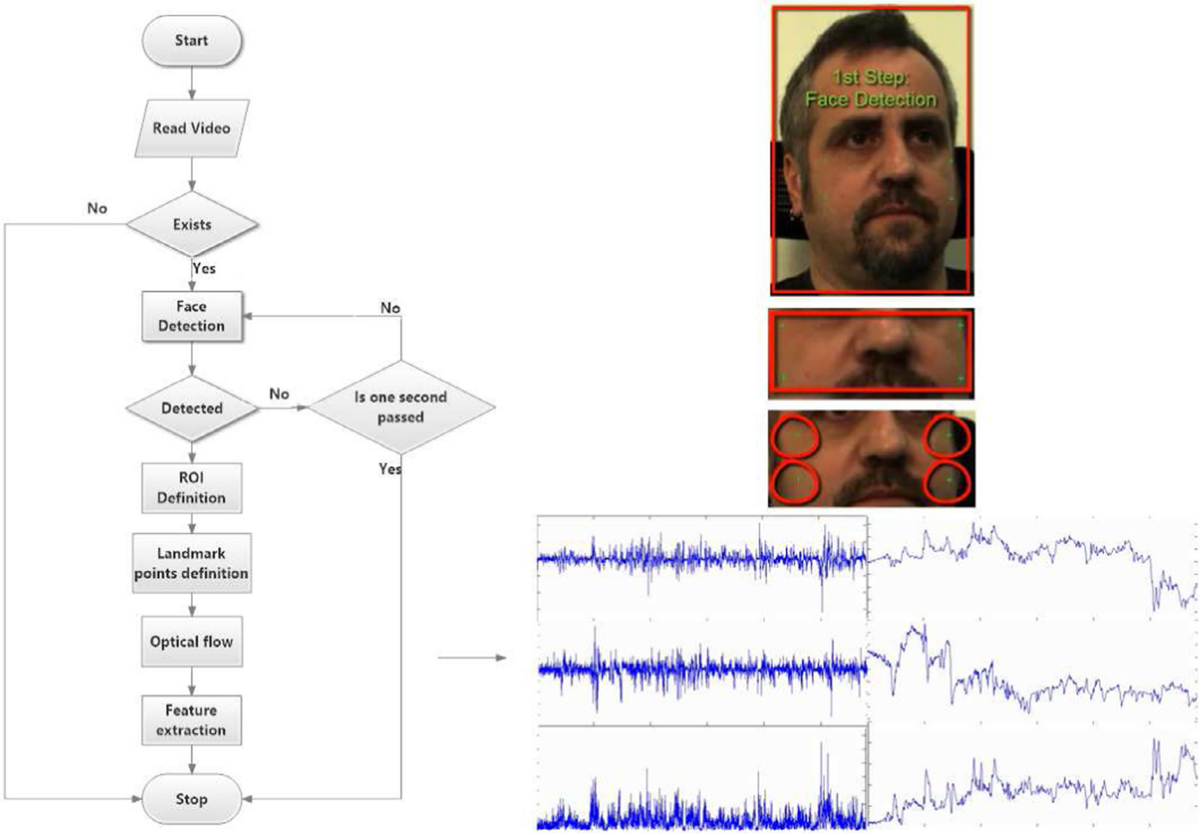

Fig. 22. Flowchart of head motion algorithm.

measurement of about 93.5%. The effectiveness of the eye blink detection algorithm strongly depends on the ability of the AAM to accurately locate and track the eye region landmarks. The most common reasons for errors include very rapid head movements, illumination variations (homogeneous and sufficient illumination is needed), out of plane face pose, eyeglasses and beard.

6.2.2. Mouth activity

The evaluation of the method for measuring mouth activity, especially lip twitching, has been performed on 11 indicative/synthetic video sequences. Since no measured information of the dynamic characteristics of lip twitching could be found, durations and frequencies for a synthetic data set were based on reports for eyelid and muscle myoclonia (Alarcón and Valentn, 2012; Kojovic et al., 2011). The synthetic videos were created using the 3D CAD software DAZ 3D 4.7. Specifically four different video clips showing upper lip twitches were created by editing the "LipTopDown" property of the mouth editor (maximum value: 0.20). In addition to the animated videos, one video showing real lower lip twitching was found on YouTube (only lips visible, otherwise anonymous). The remaining control video sequences included five subjects with no lip twitching from the *Pisa I experiment dataset* and one of a volunteer recorded with a webcam during loud reading. The feature extraction process, as described above gave a total of 1168 instances, 64 representing upper lip twitching, 310 representing lower lip twitching and 794 representing no twitching. The outcome of a stratified 10 fold cross validation showed an overall accuracy of 96.1%. A detailed performance per class is given in Table 7.

Table 7
Detailed performance results of the lip twitching detection algorithm.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
Upper lip twitching	0.734	0.004	0.922	0.734	0.817
Lower lip twitching	0.942	0.007	0.980	0.942	0.961
No twitching	0.987	0.094	0.957	0.987	0.972
Weighted average	0.961	0.066	0.961	0.961	0.961

The results of the classification performance are very satisfactory. A true positive rate of 0.942 and 0.987 for the lower lip twitching and the no twitching classes is a very good result, especially in conjunction with the equally high values for the precision. Regarding the upper lip twitching class, the performance is lower (0.734 for the TP rate). The confusion matrix showed that some of the upper lip twitching instances were falsely classified as no twitching. This might be connected with the fact that all upper lip twitching videos were synthetically produced and compared to real videos for the other two classes. Concluding, the algorithm proves the lip twitching detection possibility.

6.2.3. Head motion

The evaluation was performed in order to determine that the algorithm measures were correct compared to a ground truth and with relative low accuracy errors. For this evaluation a testing setup was developed, where 2D videos were acquired in a predefined scenario. This scenario evaluates the accuracy of motion measurements in comparison to a ground truth. The methodology requires the capturing of videos with specific movements of a person's head in the 2D space, covering specific distances and running at predefined speeds. The testing setup consisted of a flat board with a metric scale in mm printed on the horizontal and vertical axes and a stationary camera positioned at a fixed distance. The motion of the head was simulated by moving a face of a person printed on a second smaller board. The distance of the movements, as well as the speed were measured using the scales and consid-

Table 8
Results of the head motion algorithm, the ground truth and the accuracy.

Y Direction (mm and mm/sec)				
Video	Feature	Ground truth	Measured	Accuracy
YMoveDown1	Speed	14	13.7	98%
	Distance	42	39.9	95%
YMoveDown2	Speed	4	3.4	85%
	Distance	4	3.37	84%
YMoveUp1	Speed	22	21.8	99%
	Distance	45	43.8	97%
YMoveUp2	Speed	29	27.7	95%
	Distance	58	55.3	95%
X Direction (mm and mm/sec)				
Video	Feature	Ground truth	Measured	Accuracy
XMoveLeft1	Speed	26	25	96%
	Distance	127	124.7	98%
XMoveLeft2	Speed	53	42.1	79%
	Distance	128	90	70%
XMoveRight1	Speed	24	22.9	95%
	Distance	115	113.3	98%
XMoveRight2	Speed	52	42.2	81%
(1 missing point)	Distance	127	124	97%

ered to be the ground truth. The duration of the videos was also pre-defined in order to extract the ground truth of speed. Eight different videos were captured with horizontal and vertical movements at various speeds (mm/s) and distances (mm).

The results of the algorithm are reported in Table 8. It is noticeable that the average movement accuracy is about 92% compared to the ground truth, while the average speed accuracy is about 91%. Furthermore, for very small distances (i.e. in YMoveDown2 video) the accuracy of the algorithm is reduced compared to larger distances (since a few pixels may represent a significant percentage error).

6.3. Assessment with respect to detection of stress and anxiety

The three algorithms mentioned above were used in a preliminary study to elaborate a set of facial features for the detection of stress and anxiety as reported in Pediaditis et al. (2015). For this study videos from the *Pisa I experiment dataset* were employed. They were recorded while the subjects were watching three clips, which aimed to elicit the feelings of anxiety, stress and relaxation. For each video the participant was asked to provide a rating for the perceived affect, ranging from 1 to 5, where 1 stands for “Relaxed” and 5 for “Stress or Anxiety”. The latter represented a single class since, according to expert psychologists opinion a correct self-assessment of stress and anxiety cannot be taken for granted. In addition, two psychologists reviewed in a blind manner, and independently the recorded videos. In case of a conflict, a third independent psychologist compared the data of the other two annotators. The selection of the video sequences was performed by accumulating the labels for each class as given from the two experts in conjunction with the subjective rating given by each participant. The selection resulted in 10 videos for the Relaxed class and 12 for the Stress or Anxiety class. In addition to the three algorithms mentioned above, a video-based heart rate estimation method was used (Christinaki et al., 2014) employing blind source separation, as well as a mouth openness detection approach with template matching.

The fusion of all data was performed at the feature level in order to create a single feature vector for the 22 instances (videos). A statistical analysis of the data was initially performed to extract the most prominent features. T-tests and one way ANOVA enabled to identify and eliminate features that did not provide additional information with respect to the dataset. Subsequently, classification experiments were performed using the data mining software Weka v3.7.12, and further feature selection was based on evaluating the worth of a feature by measuring the Pearson’s correlation between the feature itself and the class. After being sorted by their individual evaluation, all features with a rank above 0.25 (32 features) were selected for further classification tests using a multilayer Perceptron artificial neural network (ANN). Leave-one-out cross-validation was chosen as an evaluation method, since it presents the most reliable evaluation for small numbers of instances. Tests involving additional classifiers, such as Naïve Bayes, SVM, Bayes network and Decision tree showed that the ANN returned the best results in terms of balance between the two classes, while the other classifiers showed a tendency to high true positive rates for only one of the two classes. The selected ANN uses sigmoid nodes and backpropagation for training, and the number of hidden layers is calculated based on the count of features plus the number of classes. In order to identify the smallest feature set that classifies both classes the best, given the circumstances, the aforementioned setup was repeated multiple times. Each time the feature with the lowest rank (Pearson’s correlation) was removed, until only one feature was left.

The results showed that with feature sets of 9 and 10 features an overall accuracy of 73% is reached. Some features, such as the eye blink rate or the heart rate, that were expected to play a significant role in the classification process were not employed for the above result. This can be explained by the fact that the study did not take the personal baseline (e.g. heart rate in a relaxed state for each subject individually) into account, which could not be calculated due to the limited number of video clips after selection.

7. Conclusions

This paper describes a part of the work and the results obtained within the framework of the European SEMEOTICONS (2013) project. The aim of the project is to develop a system which monitors the user’s well-being over a period of time and provides suggestions to improve and maintain a healthy lifestyle. The challenge is to create a non-intrusive platform able to acquire multimodal data to detect signs of cardio-metabolic risks. In particular, the techniques presented in this paper make possible to analyse the morphology of the face in 3D and to recognise the psychosomatic status of the person in front of the mirror.

The important aspect of the project is to design and develop an inexpensive system so it could be deployed in a home environment. This prerequisite imposed a set of constraints on the design, in particular the system has to be constructed using affordable sensors. From the output of these sensors, a 3D reconstruction of the face is created and the face is tracked. The proposed face 3D tracking, based on depth data, has shown to be robust, providing good results for face detection accuracy and face spacial position estimation. The tracking is performed in real time, which is a requirement for the subsequent processing of the Wize Mirror multisensory data. This includes analysis of stress and anxiety, both described in the paper, but also multispectral measurements not discussed in this paper. Additionally, the use of the depth sensor has two more advantages: it can be used as a primary sensor for creating 3D face reconstructions, making the mechanical design simpler; and the face pose estimation can be done just once in 3D space with subsequent projections of the estimated 3D pose onto 2D coordinates of the remaining Wize Mirror image sensors.

The proposed 3D reconstruction methodology has been shown to have the required properties, including high repeatability. Suitable results have also been obtained for the 3D face morphological analysis. It has been shown that the described features are able to appropriately encode the fat level. Using such features, a regular pattern is produced which can be used to analyse the fattening of the individual. Hence, via 3D shape analysis, it is possible to automatically assess the weight gain which is one of the main factors of cardio-metabolic risk.

Regarding the analysis of stress and anxiety, the proposed algorithms successfully extracted signs of those conditions. Particularly, the signs considered are the eyelid motion, the mouth activity and the head motion. The presented algorithms show the capability of detecting and properly measuring the indicated facial signs with a high accuracy. These signs can then be employed to classify different psycho-somatic states.

The work presented shows that having a lifestyle-compatible device for health self-monitoring and self-assessment is a reality. Moreover, the users will not need to change their habits or interact much with the mirror in order to get a wellness assessment. The process of acquiring data is performed while the users are standing in front of it, possibly as a part of their daily routine. In the current implementation, most of the acquisitions require a very short time, from just a couple of seconds for 3D reconstruction, to one minute for emotion recognition. This non-obstructive characteristic is a key requirement for the successful deployment of a self-assessment system.

Acknowledgments

This work has been supported by the European Community's Seventh Framework Programme (FP7/2013-2016) under the grant agreement number 611516 (SEMEOTICONS). Thanks to the team from The Institute of Clinical Physiology of the National Research Council of Italy, namely Dr. Eng. Giuseppe Coppini, MD Paolo Marraccini and MD Maria-Aurora Morales, for their valuable collaboration in this work.

References

SEMEOTICONS FP7-ICT-2013-10 European project. 2013. URL http://www.semeoticons.eu/
Anxiety. 2015a. URL http://www.apa.org/topics/anxiety/
Anxiety disorders and effective treatment. 2015b. URL http://www.apa.org/helpcenter/anxiety-treatment.aspx
Common signs and symptoms of stress — The American institute of stress. 2015c. URL http://www.stress.org/stress-effects/
Alarcón, G., Valentn, A. (Eds.), 2012, Introduction to Epilepsy. Cambridge University Press, Cambridge, United Kingdom.
Balakrishnan, G., Durand, F., Guttag, J., 2013. Detecting pulse from head motions in video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13), pp. 3430–3437.
Besl, P.J., McKay, N.D., 1992. A method for registration of 3-d shapes. IEEE Trans. Pattern Anal. Mach. Intell. 14 (2), 239–256.
Biasotti, S., Falcidieno, B., Giorgi, D., Spagnuolo, M., 2014. Mathematical tools for shape analysis and description. Synth. Lect. Comput. Graph. Anim. 6 (2), 1–138.
Bleiweiss, A., Werman, M., 2010. Robust head pose estimation by fusion time-of-flight, depth and color. In: IEEE Automatic Face and Gesture Recognition, pp. 116–121.
Cai, Q., Gallup, D., Zhang, C., Zhang, Z., 2010. 3d deformable face tracking with a commodity depth camera. In: European Conference on Computer Vision, pp. 229–242.
Chiarugi, F., Iatraki, G., Christinaki, E., Manousos, D., Giannakakis, G., Pediaditis, M., Pampouchidou, A., Marias, K., Tsiknakis, M.N., 2014. Facial signs and psycho–physical status estimation for well-being assessment. In: 7th IEEE International Conference on Health Informatics (BIOSTEC 2014), Angers, France, pp. 555–562.
Choi, J., Tran, A., Dumortier, Y., Medioni, G., 2014. Real-time 3-d face tracking and modeling framework for mid-res cam. In: IEEE Winter Conference on Applications of Computer Vision, pp. 660–667.
Christinaki, E., Giannakakis, G., Chiarugi, F., Pediaditis, M., Iatraki, G., Manousos, D., Marias, K., Tsiknakis, M., 2014. Comparison of blind source separation algorithms for optical heart rate monitoring. In: Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on 3–5 Nov. 2014, pp. 339–342.
Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 23 (6), 681–685.
Djordjevic, J., Lawlor, D.A., Zhurov, A.L., et al., 2013. A population-based cross-sectional study of the association between facial morphology and cardiometabolic risk factors in adolescence. In: BMJ Open, pp. 1–10.
Ekman, P., Friesen, W.V., 1971. Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. 17 (2), 124–129.
Fanelli, G., Weise, T., Gall, J., Van Gool, L., 2011. Real time head pose estimation from consumer depth cameras. In: Annual Symposium of the German Association for Pattern Recognition, 6835, pp. 101–110.
Farkas, L.G., 1994. Anthropometry of the Head and Face, 2nd ed. Raven Press, New York.
Farneback, G., 2003. Two-frame motion estimation based on polynomial expansion. In: The 13th Scandinavian conference on Image analysis (SCIA'03), Gteborg, Sweden, pp. 363–370.
Ferrario, V., Dellavia, C., Tartaglia, G., Turci, M., Sforza, C., 2004. Soft-tissue facial morphology in obese adolescents: a three-dimensional non invasive assessment. Angle Orthod. 74 (1).
Giachetti, A., Lovato, C., Piscitelli, F., Milanese, C., Zancanaro, C., 2015. Robust automatic measurement of 3d scanned models for human body fat estimation. IEEE J. Biomed. Health Inform. 19 (2), 660–667.
Gunes, H., Piccardi, M., 2007. Bi-modal emotion recognition from expressive face and body gestures. J. Netw. Comput. Appl. 30 (4), 1334–1345.
Haldiki, M., Batistakis, Y., Vazirgiannis, M., 2001. On clustering validation techniques. J. Intell. Inf. Syst. 17 (2–3), 107–145.
Hamilton, M., 1959. The assessment of anxiety-states by rating. Br. J. Med. Psychol. 32 (1), 50–55.
Hammond, P., 2007. The use of 3d face shape modelling in dismorphology. Arch. Dis. Child. 92, 1120–1126.
Harrigan, J.A., O'Conell, D., 1996. How do you look when feeling anxious? facial displays of anxiety. Pers. individ. Differences 21 (2), 205–212.
Henriquez, P., Higuera, O., Matuszewski, B.J., 2014. Head pose tracking for immersive applications. In: IEEE International Conference on Image Processing, pp. 1957–1961.
Hernandez, M., Choi, J., Medioni, G., 2015. Near laser-scan quality 3-d face reconstruction from a low-quality depth stream. Image Vis. Comput. 36, 61–69.
Hojo, H., Hamada, N., 2009. Mouth motion analysis with space-time interest points. In: IEEE Region 10 Conference (TENCON 2009), Singapore, Singapore, pp. 1–6.
Huang, X., Chen, X., Tang, T., Huang, Z., 2013. Marching cubes algorithm for fast 3d modeling human face by incremental data fusion. Math. probl. Eng. 2013, 1–7.
Irani, R., Nasrollahi, K., Moeslund, T.B., 2014. Improved pulse detection from head motions using dct. In: 9th International Conference on Computer Vision Theory and Applications, pp. 118–124.
Kojovic, M., Cordivari, C., Bhatia, K., 2011. Myoclonic disorders: a practical approach for diagnosis and treatment. Ther. adv. neurol. disord. 4 (1), 47–62.
Koolhaas, J., Bartolomucci, A., Buwalda, B., de Boer, S.F., Flgge, G., Korte, S.M., Meerlo, P., Murison, R., Olivier, B., Palanza, P., Richter-Levin, G., Sgoifo, A., Steimer, T., Stiedl, O., van Dijk, G., Whr, M., Fuchs, E., 2010. Stress revisited: A critical evaluation of the stress concept. Neurosci. Biobehav. Rev. 35 (5), 1291–1301.
Lee, B.J., Do, J.H., Kim, J.K., 2012. A classification method of normal and overweight females based on facial features for automated medical applications. J Biomed. Biotechnol.
Lee, B.J., Kim, J.K., 2014. Predicting visceral obesity based on facial characteristics.. BMC Complement. Altern. Med. 14 (248).
Li, C., Ford, E.S., McGuire, L.C., Mokdad, A.H., 2007. Increasing trends in waist circumference and abdominal obesity among u.s. adults. Obesity 15, 216–223.
Lin, J.D., Chiou, W.K., Weng, H.F., Fang, J.T., Liu, T.H., 2004. Application of three--dimensional body scanner: Observation of prevalence of metabolic syndrome. Clin. Nutr. 23 (6), 1313–1323.
Lucas, B.D., Kanade, T., 1981. An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th international joint conference on Artificial intelligence (IJCAI'81), pp. 674–679.
Macedo, M., Apolinario, A., Souza., A., 2013. Kinectfusion for faces: real-time 3d tracking and modeling using a kinect camera for a markerless ar system. SBC J. 3D Inter. Syst. 4 (2), 2–7.
Malassiotis, S., Strintzis, M., 2005. Robust real-time 3d head pose estimation from range data. Pattern Recognit. 38 (8), 1153–1165.
Manousos, D., Iatraki, G., Christinaki, E., Pediaditis, M., Chiarugi, F., Tsiknakis, M., Marias, K., 2014. Contactless detection of facial signs related to stress: A preliminary study. In: EAI 4th International Conference on Wireless Mobile Communication and Healthcare (Mobihealth 2014), Athens, Greece, pp. 335–338.
Mase, K., Pentland, A., 1991. Automatic lipreading by optical-flow analysis. Syst. Comput. Jpn. 22 (6), 796–803.
Matthew, I., Baker, S., 2004. Active appearance models revisited. Int. J. Comput. Vis. 60 (2), 135–164.
Mou, X., Wang, A., 2012. A fast and robust head pose estimation system based on depth data. In: International Conference on Robotics and Biomimetics, pp. 470–475.
Murphy-Chutorian, E., Trivedi, M.M., 2009. Head pose estimation in computer vision: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 31 (4), 607–626.
Newcombe, R., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A., 2011. Kinectfusion: Real-time dense surface mapping and tracking. In: IEEE International Symposium on Mixed and Augmented Reality, pp. 127–136.
Niles, A.N., Dour, H.J., Stanton, A.L., Roy-Byrne, P.P., Stein, M.B., Sullivan, G., Sherbourne, C.D., Rose, R.D., Craske, M.G., 2015. Anxiety and depressive symptoms and medical illness among adults with anxiety disorders. J. Psychosom. Res. 78 (2), 109–115.
Oliver, N., Pentland, A., Brard, F., 2000. Lafter: A real-time face and lips tracker with facial expression recognition. Pattern Recognit. 33 (8), 1369–1382.

Padeleris, P., Zabulis, X., Argyros, A., 2012. Head pose estimation on depth data based on particle swarm optimization. In: Computer Vision and Pattern Recognition Workshop (CVPRW), pp. 42–49.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T., 2009. A 3d face model for pose and illumination invariant face recognition. In: IEEE Proc. of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments, Genova (Italy) - September 2-4, 2009, pp. 296–301.

Pediaditis, M., Giannakakis, G., Chiarugi, F., Manousos, D., Pampouchidou, A., Christinaki, E., Iatraki, G., Kazantzaki, E., Simos, P.G., Marias, K., Tsiknakis, M., 2015. Extraction of facial features as indicators of stress and anxiety. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), Milano, Italy, pp. 3711–3714.

Quan, W., Matuszewski, B., Shark, L.-K., 2010. Improved 3-d facial representation through statistical shape model. In: IEEE International Conference on Image Processing, pp. 2433–2436.

Raytchev, B., Yoda, I., Katsuhiko, R., 2004. Head pose estimation by nonlinear manifold learning. In: IEEE International Conference on Pattern Recognition, pp. 462–466.

Reyment, R.A., 1996. An Idiosyncratic History of Early Morphometrics. In: Marcus, L.F., Corti, M., Loy, A., Naylor, G.J.P., Slice, D.E. (Eds.), Advances in Morphometrics. Springer, US, pp. 15–22.

Romero, L.M., 2004. Physiological stress in ecology: Lessons from biomedical research. Trend. Ecol. Evol. 19 (5), 249–255.

Sardinha, A., Nardi, A.E., 2012. The role of anxiety in metabolic syndrome. Expert Rev. Endocrinol. Metab. 7 (1), 63–71.

Seeman, E., Nickel, K., Stiefelhagen, R., 2004. Head pose estimation using stereo vision for human-robot interaction. In: IEEE Automatic Face and Gesture Recognition, pp. 626–631.

Selye, H., 1950. The Physiology and Pathology of Exposures to Stress. Montreal, Canada: Acta Endocrinologica.

Sharma, N., Gedeon, T., 2012. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. Comput. Methods Programs Biomed. 108 (3), 1287–1301.

Shin, L.M., Liberzon, I., 1996. The neurocircuitry of fear, stress, and anxiety disorders. Neuropsychopharmacology 35 (1), 169–191.

Sierra-Johnson, J., Johnson, B.D., 2004. Facial fat and its relationship to abdominal fat: a marker for insulin resistance? Med. Hypotheses 63, 783–786.

Smeets, D., Keustermans, J., Vandermeulen, D., Suetens, P., 2013. meshsift: Local surface features for 3d face recognition under expression variations and partial data. Comput. Vis. Image Understanding 117 (2), 158–169.

Thejaswi, N.S., Sengupta, S., 2008. Lip localization and viseme recognition from video sequences. In: National Communications Conference (NCC), Mumbai, India.

Thompson, D.W., 1942. On Growth and Form. Cambridge University Press, Cambridge.

Velardo, C., Dugelay, J.-L., 2010. Weight estimation from visual body appearance. In: BTAS 2010, 4th IEEE International Conference on Biometrics: Theory, Applications and Systems, September 27-29, 2010, Washington DC, USA, pp. 1–6.

Velardo, C., Dugelay, J.-L., Paleari, M., Ariano, P., 2012. Building the space scale or how to weight a person with no gravity. In: ESPA 2012, IEEE 1st International Conference on Emerging Signal Processing Applications, January 12-14, 2012, Las Vegas, USA, pp. 67–70. http://dx.doi.org/10.1109/ESPA.2012.6152447.

Wang, J., Gallagher, D., Thornton, J.C., Yu, W., Horlick, M., Pi-Sunyer, F.X., 2006. Validation of a 3-dimensional photonic scanner for the measurement of body volumes, dimensions and percentage body fat.. Am. J. Clin. Nutr. 809–816.

Wells, J.C., Cole, T.J., Bruner, D., Treleaven, P., 2008. Body shape in american and british adults: between-country and inter-ethnic comparisons.. Int. J. Obes. 32 (1), 152–159.

Zollhofer, M., Martinek, M., Greiner, G., Stamminger, M., J., S., 2011. Automatic reconstruction of personalized avatars from 3d face scans. Comput. Anim. Virtual Worlds 22, 195–202.