Virtual Reality
https://doi.org/10.1007/s10055-018-0357-0

Exploiting sensing devices availability in AR/VR deployments to foster engagement

Nicholas Vretos[1] · Petros Daras[1] · Stylianos Asteriadis[2] · Enrique Hortal[2] · Esam Ghaleb[2] · Evaggelos Spyrou[3] · Helen C. Leligou[4] · Panagiotis Karkazis[4] · Panagiotis Trakadas[4] · Kostantinos Assimakopoulos[4]

Received: 12 October 2017 / Accepted: 6 July 2018

Abstract

Currently, in all augmented reality (AR) or virtual reality (VR) educational experiences, the evolution of the experience (game, exercise or other) and the assessment of the user's performance are based on her/his (re)actions which are continuously traced/sensed. In this paper, we propose the exploitation of the sensors available in the AR/VR systems to enhance the current AR/VR experiences, taking into account the users' affect state that changes in real time. Adapting the difficulty level of the experience to the users' affect state fosters their engagement which is a crucial issue in educational environments and prevents boredom and anxiety. The users' cues are processed enabling dynamic user profiling. The detection of the affect state based on different sensing inputs, since diverse sensing devices exist in different AR/VR systems, is investigated, and techniques that have been undergone validation using state-of-the-art sensors are presented.

Keywords Affect state detection · Engagement · Interpretation of interaction · Multimodal affect state detection

1 Introduction

Even though the learning goals, the available resources (time, equipment, teachers and teaching facilities) and the contexts may change, learning needs appear throughout our lifetime irrespective of age, gender, nationality, culture and learning intricacies. Education can be distinguished in formal, non-formal and informal, depending on the setting; it may target a homogeneous or heterogeneous learners' group, and it may target large or small learners' groups and diverse settings (e.g., home, school, enterprise premises). The common goal of any educational procedure is to make each and every learner acquire the desired skill or piece of knowledge in the *minimum time*, with the *maximum pleasure/convenience*. To achieve this goal, researchers agree that *engagement* is of key importance.

Although augmented reality (AR) or virtual reality (VR) technologies have been available for several years, it is only recently that they have developed and matured to a level that enables rapid penetration in the consumer space and educational environments. Enhancing and extending the learning experience is at the heart of what VR can offer students, and it is possibly one of the most powerful of all technologies that could help change how we learn forever. Virtual and augmented reality (VR/AR) aspires to contribute in fostering learners' engagement almost in any educational context, realizing Albert Einstein's words "the only source of knowledge is experience." Moreover, it has penetrated the educational sector as a promising tool, helping teachers face one of the biggest timeless issues which is engagement; AR, with its ability to combine both digital and physical worlds, while VR, with its ability to completely immerse users in new environments, bring new dimensions to teaching and learning. Virtual and augmented reality in education can lead to increased knowledge retention exploiting the unique multisensory experience. This technology fits the needs of user groups of all ages, of almost all cultures as it shifts learning from a text-based process to an *experience*-based process and of diverse learning needs. The next question

✉ Panagiotis Trakadas
pkarkazis@isc.tuc.gr

[1] Centre of Research and Technology Hellas, Thermi, Thessaloniki, Greece

[2] University of Maastricht, Maastricht, The Netherlands

[3] National Centre for Scientific Research "Demokritos", Agia Paraskevi, Athens, Greece

[4] Technological Educational Institute of Sterea Ellada, Psahna, Halkida, Greece

Published online: 14 July 2018

that may arise is the range of learning materials. Although not unlimited, available materials support different subject areas from improving creative writing to understanding science and math topics through enhanced visualization and immersion with current systems supporting a certain level of flexibility in learning material creation.

Although VR may enhance the learning experience, the pedagogists insist that achieving engagement directly depends on personal characteristics including not only competence levels but also, and more importantly, *affect state*. According to well-established theories (Csíkszentmihályi 2008), to maximize results, the learner has to remain in the so-called flow state (i.e., feeling neither frustrated nor bored), which depends on the level of challenge/difficulty and on their skills. Flow has been defined by Csíkszentmihályi (Csíkszentmihályi 2008) as "an optimal psychological state that people experience when engaged in an activity that is both appropriately challenging to one's skill level, often resulting in immersion and concentrated focus on a task; this can result in deep learning and high levels of personal and work satisfaction." When the challenge does not match the user's skills, she/he feels either bored (i.e., when dealing with a too easy task for her/his skills) or anxious (i.e., when experiencing a too difficult task). *To keep the learners in the flow state, their affect state has to be constantly and in real time detected, i.e., real-time dynamic profiling has to be implemented. Once they come close to boredom, the challenge has to be intensified, while when they come close to anxiety, the challenge has to be slightly relaxed.*

In this paper, we propose the exploitation of sensing devices which are available in AR/VR installations to detect the affect state of the user and to tailor the experience not only to her/his personal preferences (which is the case today) but most importantly to her/his affect state which changes dynamically. To prove that this approach is realistic, we investigate how this can happen through a variety of sensor types and we describe practical implementation of techniques well-grounded to algorithms available in the literature. The implemented techniques have been evaluated using *publicly available datasets* as well as into *real-life deployments*. Such an integrated approach is expected to boost engagement with AR/VR application and thus, learning efficiency and user satisfaction.

The rest of the paper is organized as follows: In Sect. 2, we present the modifications required for the realization of the proposed approach in a typical AR/VR system architecture. In Sect. 3, we thoroughly investigate how affect detection can be achieved through different sensors/modalities and include relevant results. Then, in Sect. 4, we discuss the benefits, challenges and prospects of the proposed approach. Finally, Sect. 5 concludes the paper.

2 AR/VR system architecture

Adopting a rather simplistic approach, any VR/AR system can be considered to consist of three parts (as also illustrated in Fig. 1): the AR/VR learning material logic (e.g., a game), which is the heart of the system, a component (or components) that deliver/create the "environment" to the user, which is in essence the output of the system to the user, e.g., the speakers and the HoloLens display, and the device(s) that sense the user's activity (inertia, microphone, etc.) which are fed to the heart of the system to trigger a change (the next step) of the experience. The AR/VR logic component can be organized in different ways and may impose diverse computing and storage requirements. Turning our attention to the devices interfacing the user, hardware components for augmented reality typically include sensors and input devices. Modern mobile augmented reality systems use one or more of the following tracking technologies: digital cameras and/or other optical sensors, microphones, accelerometers, GPS, gyroscopes, solid state compasses, RFID and wireless sensors. These technologies offer varying levels of accuracy and precision.

Effective, hyper-personalized learning relies on the adaptation of parameters in the learning experience, based upon each learner's affective state. That is, the physical behavior of a learner, expressed by a number of cues in their facial, bodily and vocal expressions, their gaze locality and their physical manipulation of the devices with which they interact, can shed light to their uptake of the process. Once the sensors are already available in such a system and feed the operation of the learning material logic, we propose to use their readings for an additional purpose: to detect the affect state, as shown in Fig. 1. The component named "affect

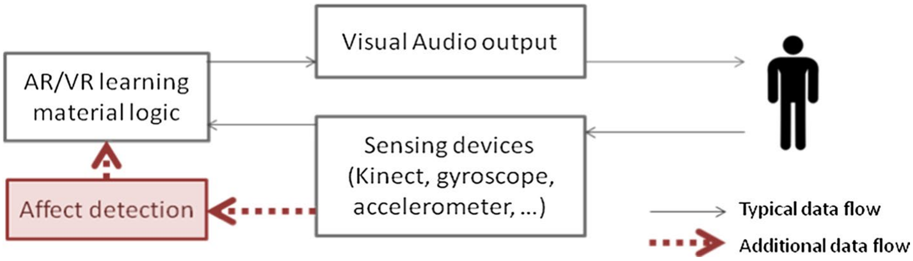
Visual Audio output
AR/VR learning material logic
Sensing devices (Kinect, gyroscope, accelerometer, ...)
Affect detection
Typical data flow
Additional data flow

Fig. 1 Legacy high-level AR/VR system architecture enhanced with affect detection components

detection” receives the readings of the available sensors and estimates the affect state of the learner. Once the affect state is not boredom or anxiety, the logic of the learning material evolves as would evolve without the “affect detection” component implemented (i.e., today). If the affect state is shown to tend to boredom, this is signaled to the logic component and the challenge level is increased. In the case of anxiety detection, the challenge is relaxed, so as to keep the learner in the flow state.

An educational platform that embraces the principle of learning experience adaptation to the learners affect state has been developed under the H2020 MaTHiSiS project (http://mathisis-project.eu/). MaTHiSiS is an educational platform providing every type of learner, in every type of setting, on the device they have at their disposal, with a bespoke, individualized learning experience that is adapted to their personal requirements. With regard to sensory-based affect recognition, MaTHiSiS uses five Sensorial Component modalities: facial expression analysis, gaze detection, skeleton motion analysis, audio analysis and mobile-based inertia sensors analysis. The technical architecture of the platform is illustrated in Fig. 2 and consists mainly of: (a) the MaTHiSiS back-end which implements all the MaTHiSiS algorithms for personalizing the learning experience and for supporting learning in a diversity of environments, (b) the platform agents which are the devices the learner interacts with and can be a laptop, tablet, smartphone, interactive whiteboard, robots, HoloLens or other and (c) the MaTHiSiS front-end gateway which is the user interface through which the teachers, the IT personnel responsible in a school and the platform administrator interact with the platform. Teachers use it to create learning content (e.g., games or exercises) and to define at high level the learning experience suitable for their students (e.g., the subject and the target learning outcomes), IT personnel use it to configure the platform, i.e., define the platform agents (devices) each class will use so that the appropriate learning material reaches the appropriate people, and the administrator uses it to manage the users and their roles.

In each device used by the learners, a different (per device type) “platform agent layer” (PA-layer) instance is implemented comprising of the “Experiencing Service Platform Agent” (ES PA, in Fig. 2), the “Experiencing Service Sensorial Component” (ES SC in Fig. 2) and the “Sensorial Component” (SC in Fig. 2). The latter captures the readings from the sensors available in each device and either processes

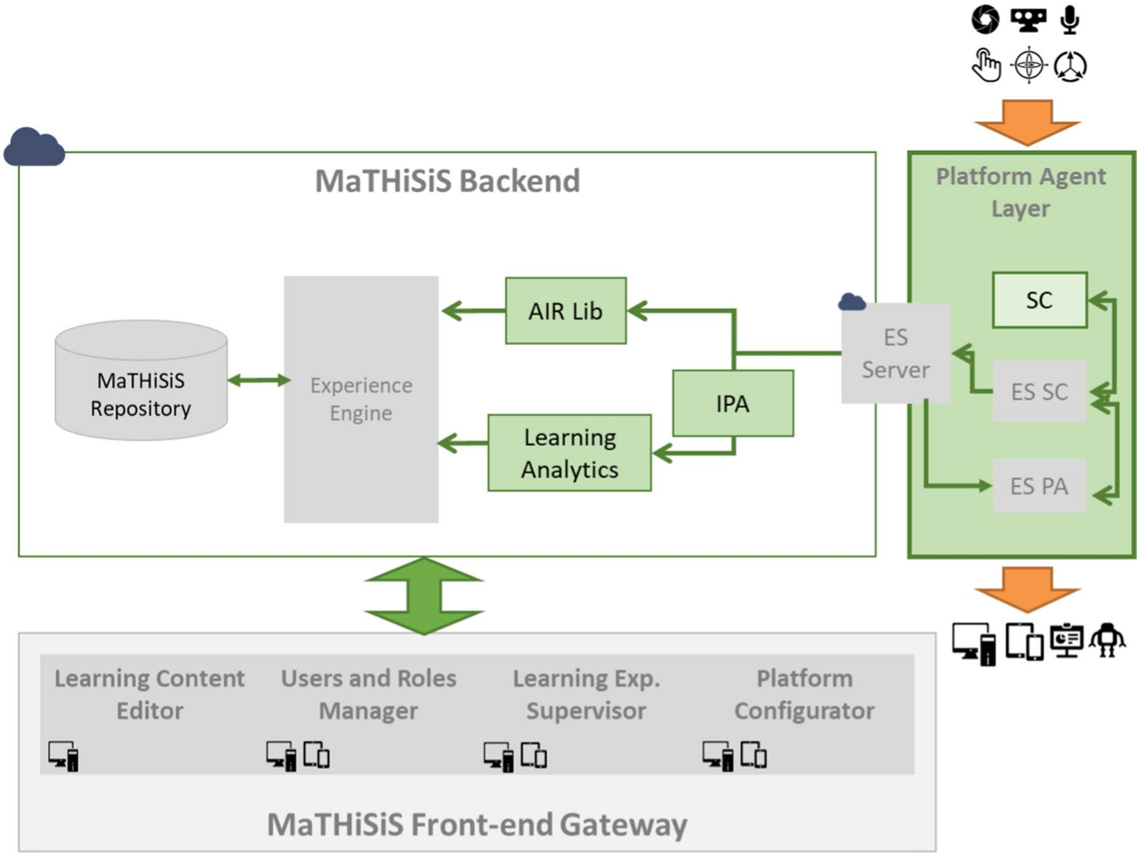

Fig. 2 MaTHiSiS platform architecture

them to derive the user's affect state and delivers it to the ES SC or delivers to the ES SC the raw captured data (which happens in the case of the absence of sufficient processing resources as is the case of the mobile devices). The ES SC forwards the received information to the "Experience Service server" using web socket technology, along with other information that it receives from the ES PA, e.g., regarding the score achieved by the learner. At the back-end, all the information gathered through the platform agents is received by the Affect Information Repository Library (AIR Lib in Fig. 2) component which is responsible for delivering the learner's affect state to the "experience engine." In case, raw data from sensor are received by AIR Lib, a dedicated subcomponent uses them to derive the affect state. This information is then used by the experience engine to define the next learning material and level the user will interact with to ensure the learner is kept at the engagement state. This is then passed to the Platform Agent (through the ES server). The *challenge* then moves to the identification of *affect detection techniques* that lead to results of adequate accuracy with commercially available state-of-the-art sensors.

3 Affect state detection based on diverse sensing devices

The sensorial component on the Platform Agents and the AIR Lib component in the back-end are the basis of the recognition of the learners' affect states. Their goal is to gather (physical) behavioral cues of the learner and apply machine learning techniques in order to interpret them into comprehensive affective cues that tell the story of the learner's uptake of the learning objective(s). Its role is to provide for the learner's affective state. In the MaTHiSiS system outlined above, a variety of algorithms for affect detection has been implemented and tested per modality, as it will be described in the forthcoming sections. All adopted algorithms utilize machine learning techniques. Thus, appropriate training of the algorithms needs to take place prior to the normal operation of the system. At the following, we shall refer to students without disabilities as "mainstream" students.

3.1 Facial expression analysis

The facial expressions are often considered as the strongest indicative communication tool of human emotions. They may expose people's feelings and mood state, from simple spontaneous emotions like happiness and disgust to time-dependent affective expressions states like anxiety, boredom and engagement during a current task and/or a situation. This allows the person's interaction counterpart to understand their affective state and adjust its behavior according to the person's underlying feelings. Facial images are one of the data cues that will be captured through the Sensorial Component by means of different types of cameras across devices.

For the extraction of facial expressions, a graph-based method (Kim et al. 2013) has been adopted. More specifically, the face is represented as a graph, which is formed by points extracted from specific areas. The variation of muscle movements on the face during the expression of different emotions leads to different positions of points on the image and may generate different graphs. The input of the algorithm is an image. Then, facial landmarks are detected using the Supervised Descent Method (Xiong and De la Torre 2013). For instance, such landmarks may be the nose, the eyes, the brows, the mouth, etc. These points are tracked, so that the movement of the facial muscles is followed over time. Assuming that all landmarks are connected, they may be considered as a graph. We then make the hypothesis that the density of the graph differs in each facial expression. More specifically, we use spectral graph analysis, through which a feature vector is extracted. This vector depicts areas of density in the graph by using the graph's Laplacian matrix and solving the eigen decomposition problem for the eigenvectors corresponding to the first and second greatest eigenvalues which capture information regarding different density areas of the initial graph. Such areas in the specific problem are those of eyes, mouth and nose.

More specifically, the Laplacian matrix L of a graph G is defined as $L = D - A$, with D denoting the degree matrix and A the adjacency matrix of G. $A(i, j)$ is computed as: $A(i,j) = 1 - e^{\frac{(-|x_i - x_j|)}{d}}$, where |•| denotes the Euclidean distance, x_i, x_j any two given landmark points and d a constant depicting the variance of the overall distance between the facial landmarks. In order to normalize between different image scales and sizes (i.e., for recognition "in the wild"), the symmetric Laplacian matrix is adopted as it is considered to be a more robust option: $L^{\mathrm{sym}} = D^{-1/2} L D^{-1/2}$. Then, its eigen decomposition follows: $L^{\mathrm{sym}} v_i = \lambda_i v_i$.

For the classification, support vector machines (SVM) are used. The initial evaluation of the algorithm is performed using images from the well-known public available Cohn–Kanade (CK) database (Kanade et al. 2010), leading to very satisfying results. Although this dataset involves expressions of the six basic Ekmanian emotions, namely Anger, Disgust, Fear, Happiness, Sadness and Surprise, a correlation of the aforementioned emotions with affective states was retrieved in Russell's Core Affect Framework (Baker et al. 2010). A direct mapping of the spontaneous emotions to affect states conveys this correlation. Using this mapping, Sadness corresponds to Boredom, Happiness to Engagement and Surprise, Anger, Fear to Frustration. The performance of this algorithm using Cohn–Kanade dataset to predict affective stated reached a classification score that rounds up close to 100% accuracy. Results per emotion are depicted in Table 1.

3.2 Gaze estimation

Gaze estimation refers to an emerging computer vision research topic that is defined as the process of determining the eye's point of regard, usually with respect to a specific plane such as a computer screen but also in the more general case of the eye's orientation or a person's "look at" direction. Recent gaze estimation methods aspire to estimate a person's gaze accurately, invariant of the head pose, the lighting conditions and the eye's appearance, using low-cost commodity hardware and simple setups. Gaze-based methods can be classified into three general categories: (a) shape (Feature)-based ones, (b) Appearance-based and (c) hybrid ones, combining elements from the two previously mentioned ones. In another recent work, deep networks are employed to estimate a user's gaze direction (see Zhang et al. 2015). In this work, head poses and eye images are the input to a convolutional neural network (CNN). The output is the gaze direction in the camera coordinate system. Data collection was performed by the RGB cameras (webcams) of laptops, while the evaluation was based on the public MPIIGaze dataset (Zhang et al. 2017) containing images by 15 users. Data were recorded by users applying the data collection software on their laptops in situ without any special guidelines given about the head pose, the time of execution or the illumination conditions. The MPIIGaze dataset consists of 213,659 images with large variation in both illumination and head pose.

Table 1 Experimental results of facial analysis

Emotion	Accuracy (%)
Anger	100.00
Disgust	86.37
Fear	60.00
Happiness	100.00
Sadness	75.00
Surprise	100.00

We have experimented with a regression CNN-based gaze estimation algorithm; the concept of this algorithm is similar to the MPIIGaze method (Zhang et al. 2015). Initially, face frontalization was a relatively costly operation processing-wise compared to the other steps of the pipeline (face detection, landmark detection and head pose estimation). Therefore, a perspective warping step was opted for, instead of the face frontalization; this was faster processing-wise and also achieved higher accuracy after training. The processing pipeline is depicted in Fig. 3. For face detection, Li et al.'s SURF cascade method (Leifman 2016) is employed, while for the landmark detection, a cascade of boosted regression forests (Rikert and Jones 1998) regresses the positions of the facial landmarks in around 12 ms. For head pose estimation, by utilizing the 2D detected landmarks, correspondences

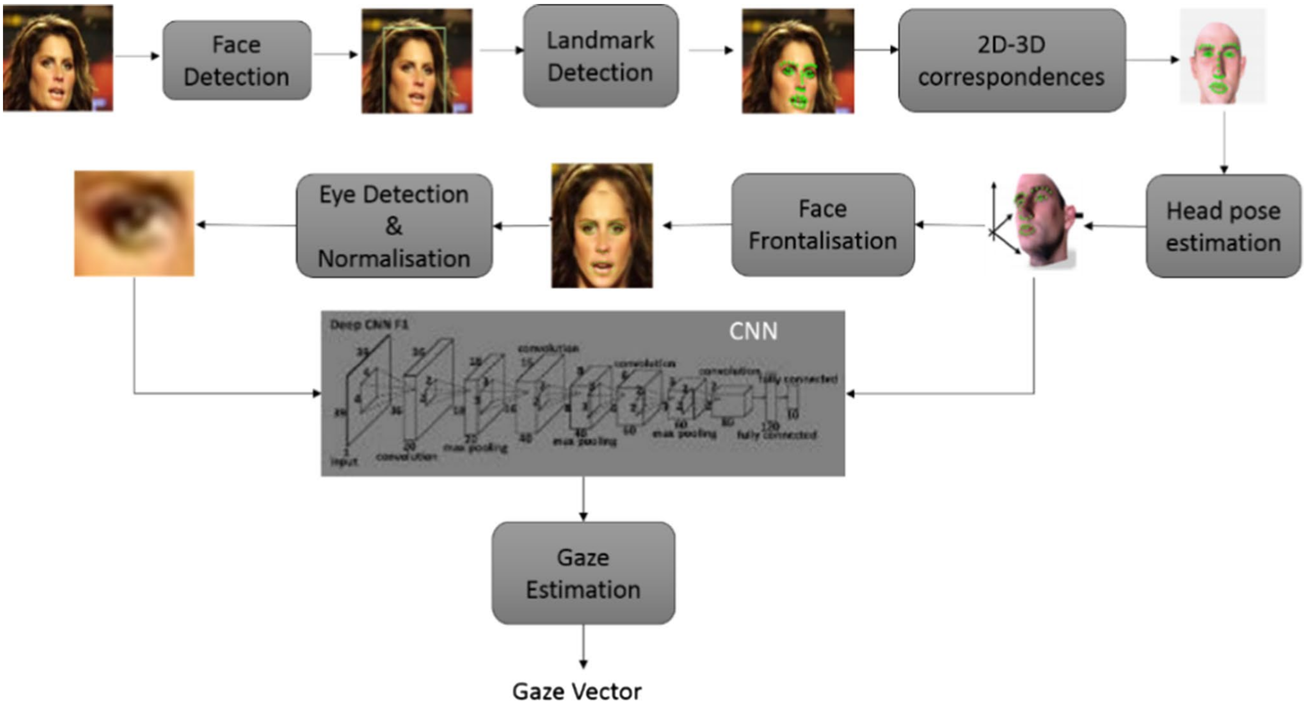

Fig. 3 Overall gaze estimation pipeline

with pre-annotated 3D positions may be now established. These are annotated on a generic 3D mean facial shape head model and are used to estimate the user's head pose by fitting the 3D model data to the 2D image correspondences via nonlinear optimization. The result is the head's pose (rotation and translation) with respect to the coordinate system defined by the camera. For data normalization, the generic 3D mean facial shape is rotated and translated according to the head pose extracted previously. Then, the 3D eye position of the left and right eye is estimated, and a vector from each eye looking at the target is predicted. Moreover, similar to Zhang et al. (2015), the normalization is done by scaling and rotating the camera-captured image so that the eye image is centered at the midpoint of the eye corners from a fixed distance *d* and so that the horizontal axes of the head coordinate system and the camera's coordinate system are aligned. The head pose is then parameterized as a 2D angle, and the eye images are contrast-enhanced. Regarding gaze estimation, the vector of each eye along with the respective image are fed into a pre-trained deep CNN which then regresses a feature vector representing the user's gaze direction.

After training the network and achieving similar accuracy to that of Nottingham Trent University (2017) by using a subset (15%, randomly chosen) of the dataset as a test set, it was also tested on another publicly available dataset, EYEDIAP (Mora et al. 2014). The EYEDIAP dataset contains low-resolution (VGA) video capturing 94 video sequences of 16 participants looking at three different targets (discrete and continuous markers displayed on a screen, and floating physical targets) under both static and free head motion, while it includes two different illumination conditions on some participant recordings. Moreover, along with the database, the authors include a framework of calculating performance measures, with respect to each visual target (discrete, continuous or floating).

The gaze estimation network was tested on multiple sequences from EYEDIAP and achieved a mean angular error on the VGA sequences of 10.46° with high head mobility, and 9.55° with no head mobility. On the contrary, on the challenging non-frontal viewpoint and the HD camera the mean angular error achieved was 18.22° with high mobility and 17.43° with none. In comparison, the method in Zhang et al. (2015) that outperforms other state-of-the-art methods achieves an accuracy of 13.9° on the MPII dataset and 10.5° on the EYEDIAP dataset. The results for mobile head pose are illustrated in Fig. 4, where in the left-hand side figure comparison with the results presented in Nottingham Trent University (2017) for the MPIIGaze dataset is included in green.

Apart from the CNN regression-based approach, a Conditional Local Neural Fields (CLNF) approach OpenFace (Baltrušaitis et al. 2016) was also investigated. After testing with the same datasets, the OpenFace achieved a mean angular error on VGA sequences of 11.17° with mobile head pose activity, therefore underperforming on the particular conditions over the previously described methodology for the particular setting (VGA images). On the HD sequences, the mean angular error achieved was 15° with mobile head pose activity, therefore outperforming the previous method in this setting (HD images). The lower performance in VGA images can be attributed to their low resolution and thus, the lower quality of the eye images that prevent the accurate localization of the eye landmarks. Therefore, it can be deducted that the initially implemented method is more appropriate for VGA images (e.g., a standard non-HD web camera or a NAO robot's camera) while the OpenFace method would be better served in HD resolutions (e.g., Kinect).

3.3 Inertia sensors

Recently there have been many studies that support the potential usage of recognizing users' emotional states through various inertia sensors such as accelerometer and gyroscope (Kim et al. 2013). Inspired by the research performed by Coutrix (2012), an affect recognition system which exploits the expression through 3D gesture using aforementioned sensors can be implemented. As

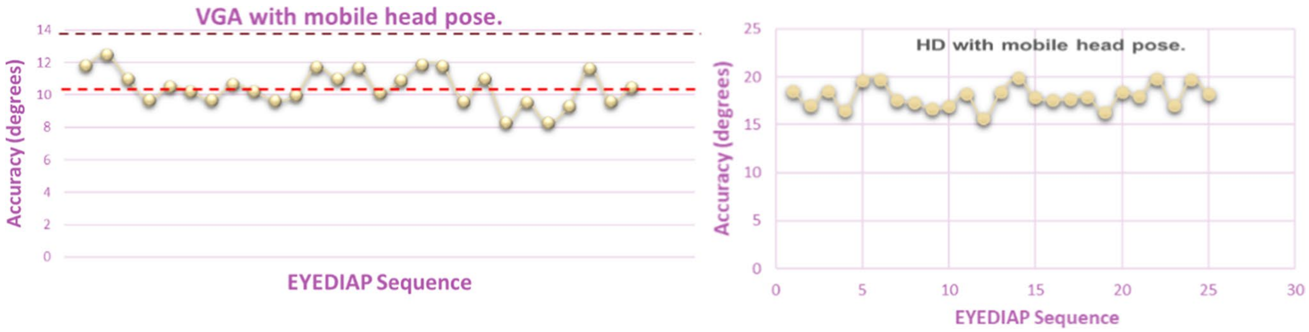

Fig. 4 Results for mobile head pose

demonstrated in Coutrix (2012), 3D descriptors contribute to emotion expression while interacting and using mobile devices. In this work, a high number of significant correlations were detected in 3D motion descriptors of gestures and the arousal dimension. This study can be expanded, in order to infer affective states which are commonly experienced during the learning process, focusing our effort in the recognition of engagement, boredom and frustration. This 3D and continuous space can be accurately mapped to affective states from the Theory of Flow. The features extracted are analyzed in order to detect common patterns which can allow the system to infer the affect state of the user. Ideally, these features will help in the identification of erratic movements or unexpected behaviors such as the lack of motion or interactions with the devices. This information could denote frustration or boredom, respectively.

Due to the absence of public datasets, we decided to gather data to investigate the relationship between users' movements while using smartphones and their affective states. The data were recorded on an android phone using a K6DS3TR gyroscope and accelerometer by STMicroelectronics. For this purpose, a mobile game application was developed. During the game play, sensorial data of accelerometer and gyroscope were recorded. The two sensors use the right-handed coordinate system, where x-axis increases as user moves toward the screen, y-axis increases as the user moves to the top of the screen and the z-axis is perpendicular to the screen. These x, y and z coordinates are recorded with sampling rate of 50 Hz. Data were gathered in sessions for multiple subjects in an adjusted environment such as a classroom or at home. Diversity of the subjects was considered such that they have different profiles in terms of age, and gender. In total, there were 24 subjects, 18 were males and 6 were females. Their ages ranged between 18 and 53 with a mean of 25.6 and 7.6 standard deviation.

Subjects were asked to use the developed mobile application in their usual way when they are bored, engaged or frustrated to gather sensorial data aligned with the Theory of Flow. Prior to the game play and data collection, they were given instructions to sit down on a chair, to avoid putting their elbows or arms on flat surface, and to use the device either with one or two hands. Each subject played the game eight times, and the affective state was recorded/sampled every 20 s, for each affective state. The resulted total number of sessions (samples) per subject was 24. Consequently, the final number of samples was 576. It is important to note that each sample was labeled by the intended affective state which the subject simulated while using the mobile application.

Following the step of data collection, feature extraction and analysis are applied on the raw gesture logs from the 3D sensorial outputs of accelerometers, and gyroscopes. Features were computed on the 3D vector in both time and frequency domains. In the time domain, features were computed for the acceleration and jerk, in the x, y and z directions separately, and on the 3D vector length of the three directions. Acceleration and Jerk are the third and fourth derivative of displacement with respect to time. For each orientation and 3D vector length, the maximum, minimum, mean, median, variance and amplitude of these descriptors have been calculated, for both absolute and signed values. These features were computed on both the raw acceleration and high-pass filtered acceleration values. In the frequency domain, a 1D Fourier transform was applied to perform the spectrum analysis. Then, the following set of features was computed: the gap (G) that maximizes the difference between the most and least important frequencies in the spectrum of acceleration signal, the number (N) of important frequencies, as well as the most important (M) frequency. This set of features was obtained for the x, y and z projections of raw acceleration signal and for the high-pass and low-pass filtered signals of acceleration.

Data collection, modeling and feature extraction methods were tested to study how they are correlated to users' movements and annotations. Prior to the supervised classification, each sample's features were standardized to have zero mean and unit variance. For the classification task, support vector machines (SVM) implementation provided by scikit-learn (Pedregosa et al. 2011) was used. The evaluation is validated using threefold cross-validation. Table 2 provides the confusion matrix of the three affective states. In this matrix, each cell i, j reflects the percentage of cases where sentiment i was detected by the algorithm while sentiment j was experienced by the learner as marked by the teacher. For example, when boredom is detected by the algorithm, 64% of cases the sentiment was actually "boredom" while 31.8% was engagement (i.e., the algorithm confused the actual sentiment with boredom) and 4.2% of times the algorithm confused frustration with boredom. The results show high classification rates of the three states which indicate the correlation between the 3D features while using of a mobile phone and users' affective states. In addition, the Cohen's kappa value and the average precision are 50.4 and 68%, respectively. It also supports the usability of our approach to study this correspondence. Furthermore, the high detection of frustration is due to the fact that subjects were more expressive when it comes to events such as losing game or

Table 2 Confusion matrix in terms of classification rate for the three affective states for the inertia sensors

	Boredom (%)	Engagement (%)	Frustration (%)
Boredom	64	31.8	4.2
Engagement	30.4	59.2	10.4
Frustration	9.1	15.2	75.7

facing higher level of challenge. When it comes to boredom or engagement, expressivity through device's movement was reduced substantially. Most subjects paid less attention to the mobile device when they were bored, and when they were engaged, they tended to hold upright the device with both hands and more concentration.

3.4 Skeleton motion analysis

Human action recognition became a necessity for applications in surveillance, human–robot interaction, robot perception, etc. A variety of methods has been introduced, each one applying a different feature extraction and classification methodology. In the latter decades, handmade features have dominated in the field of action recognition. Those features were carefully selected to enable the interpretation of the performing action. However, more recently, based on the deep learning approaches, automatically generated features were introduced (Du et al. 2015). Through a deep learning framework and a large corpus of data samples, deep learning methods have provided us with very promising results, overcoming the handmade methodologies. After a thorough review of the available literature, we proposed a novel method called Speed Relation Preserving Slow Feature Analysis (srpSFA) in Kim et al. (2013).

The *Speed Relation Preserving Slow Feature Analysis* (srpSFA) approach works as follows: We define the loss function we want to minimize, and we continue with the appropriate constraints that need to be imposed. Similar to the standard SFA, in our approach the optimal parameters matrix $\boldsymbol{W} \in R^{I \times J}$ needs to be computed, through which the new representations are $\boldsymbol{y}_n^{(t)} = \boldsymbol{W}^T \phi_n^{(t)}$. In order to fulfill the preservation of speed in the feature space, we want to minimize the objective function (loss function):

$$\sum_{ij} \mathbb{E}_t[\left(\dot{y}_i^{(t)} - \dot{y}_j^{(t)}\right)^2 \dot{\Gamma}_{ij}^{(t)}$$

under standard SFA-oriented constrains, i.e., zero mean, unit variance and uncorrelated features for the new mapped node representations. The weight factor $\dot{\Gamma}_{ij}^{(t)}$ penalizes the distance between the new representations $y_i^{(t)}$ and $y_j^{(t)}$. The greater the weight factor, the greater the penalty the new mappings to be "close." A great value of $\dot{\Gamma}_{ij}^{(t)}$ denotes that the speed vectors $\dot{\phi}_i^{(t)}$ and $\dot{\phi}_j^{(t)}$ of the skeleton nodes ϕ_i and ϕ_j in the input space are "close." Thus, minimizing the loss function above, it is ensured that for two speed vectors $\dot{\phi}_i^{(t)}$ and $\dot{\phi}_j^{(t)}$ that are "close" in the input space, their corresponding mappings $\boldsymbol{y}_i^{(t)}$ and $\boldsymbol{y}_j^{(t)}$ will also be "close" in the feature space. Provided the new skeleton node representation of the *n*-th node $\boldsymbol{y}_n^{(t)} = \boldsymbol{W}^T \phi_n^{(t)}$, the matrix notation of the objective function along with the constrains is given by:

$$\min_{\boldsymbol{W}} \operatorname{trace}\left(\boldsymbol{W}^T \dot{\boldsymbol{\Phi}} \operatorname{diag}\left(\dot{\boldsymbol{L}}\right) \dot{\boldsymbol{\Phi}}^T \mathbf{W}\right)$$

$$\boldsymbol{W \Phi \Phi^T W^T = I}$$

To evaluate the algorithm, we used the MSR-Action3D dataset (Li et al. 2010): A Kinect-like depth sensor was used to obtain the recorded skeletons. It consists of 20 different recorded actions performed by 10 different subjects/actors. In addition, each subject repeated each recorded action two or three times. Namely, the actions are *high arm wave*, *horizontal arm wave*, *hammer*, *hand catch*, *forward punch*, *high throw*, *draw x*, draw *tick*, *draw circle*, *hand clap*, *two hand wave*, *side boxing*, *bend*, *forward kick, side kick*, *jogging*, *tennis swing*, *tennis serve*, *golf swing*, *pick up \ throw*. For each skeleton, the 3D joint locations through time were provided. In addition, the connections of the nodes that define the recorded skeleton were also given. Each recording was done in 15 fps. Finally, because of the similarity of the actions, this dataset is considered difficult. In the validation, the method described in Li et al. (2010) was adopted. The whole dataset was split into three subsets namely, the AS1, AS2 and AS3. The data samples that correspond to the subjects with odd identification were used for training, i.e., 1, 3, 5, 7, 9, while the ones that correspond to the subjects with even identification, i.e., 2, 4, 6, 8, 10, were used for testing.

The results presented in Table 3 show that the srpSFA method provides the best performance for difficult (according to the literature) datasets, such as the MSR-Action 3D publicly available dataset, compared to the state-of-the-art reported accuracies. These emotive actions are mapped to the three affective states (engagement, boredom, frustration), and therefore, the accuracy of detected actions in turn yield accurate results in affect recognition. The accuracy exceeds 90% for all tested datasets.

3.5 Speech

It is well-known that human communication in everyday life is mainly carried out by vocalized speech. Speech is used to

Table 3 Experimental results in MSR-Action 3D

Method	AS1	AS2	AS3	Ave.
CVPRW (Li et al. 2010)	72.9	71.9	79.2	74.7
JRTIP (Chen et al. 2016)	96.2	83.2	92.0	90.47
CVPR (Vemulapalli et al. 2014)	95.29	83.87	95.50	94.49
CVPR (Du et al. 2015)	93.33	94.64	95.50	94.49
ICCV (Wang et al. 2015)	–	–	–	96.9
srpSFA	97.83	91.96	99.05	96.28

transfer meaning between individuals. However, apart from meaning, speech also carries emotions that may be related to the speaker's affect state. Even though such emotions may be more easily recognized through visual channels, as discussed in Sect. 3.1, in many practical applications speech may be the single available modality for the recognition of emotions. An example closely related to the goals of the presented system is human–computer interaction through voice–user interfaces. Among various speech-related applications such as automatic speech recognition (ASR), and speaker identification, emotion recognition from speech appears to be the most challenging one. Of course, the task of emotion recognition from speech is a significantly difficult one. Even human experts (e.g., psychologists) often fail to recognize emotions without visual information from the subject. As expected, the task is more difficult for non-experts (e.g., in the context of our work, tutors and/or pedagogists).

Information carried by a typical speech signal may be divided into two distinct types (Anagnostopoulos et al. 2015): (a) the explicit (or linguistic) information, comprising of articulated patterns produced by the speaker; and (b) the implicit (or paralinguistic information, concerning the variation in pronunciation of the aforementioned linguistic patterns). Linguistic information of speech may be qualitatively described, while paralinguistic may be quantitatively measured. To this goal, several spectral features and also features such as the peak and the intensity may be used. A speech segment may be then classified to one or several emotions by using one or a fusion of both types of information. Many emotion recognition approaches from speech are assisted by ASR. The main disadvantage of these is that they are not able to provide language-independent models. Another crucial disadvantage is that there exists a plethora of different sentences, speakers, speaking styles and rates (El Ayadi et al. 2011). Thus, the majority of approaches that aim to be language independent tend to rely on paralinguistic speech information. Nevertheless, even in this case, such information may be significantly diverse, depending on cultural particularities. Additionally, a speaker's potential chronic emotional state may suppress the expressiveness of several emotions. Still, relying solely on paralinguistic information is probably the most appealing approach, when dealing with speakers' emotion recognition.

Within our system, initially each audio signal is transformed to a sequence of feature vectors. Features are extracted on a short-term basis and from 20-ms windows, and afterward, the final feature vectors are formed by concatenating the mean and variance values of the features over a mid-term window of 1 s. The short-term process can be conducted either using overlapping (frame step is shorter than the frame length) or non-overlapping (frame step is equal to the frame length) framing. The extracted features are summarized in Table 4. The aforementioned concatenation of the mean and the variance values results to a feature vector of dimension equal to 68.

For classification of feature vectors to emotions, the well-known and widely used support vector machines (SVMs) are used. SVMs are well-known supervised learning models, and during recent years, they have been extensively used in both classification and regression problems.

All features have been implemented in Python, while for classification, the scikit-learn library (Coutrix 2012) has been used.

For training and evaluation of this technique, we have used three public and well-known datasets: (a) EMOVO (Costantini 2014), containing speech in Italian and for disgust, fear, anger, joy, surprise and sadness; (b) SAVEE (Haq and Jackson 2009), containing speech in English for the same emotions; and (c) EMO-DB (Burkhardt 2005) containing speech in German for anger, boredom, disgust, fear, happiness, sadness and neutral. Note that all emotions have been performed by actors in all three datasets. Since in MaTHiSiS it is planned that emotions/states recognized are going to be based on the Theory of Flow (Csikszentmihalyi 1996),

Table 4 Extracted audio features

Index	Name	Description
1	Zero crossing rate	Rate of sign-changes of the frame
2	Energy	Energy sum of squares of the signal values, normalized by frame length
3	Entropy of energy	Entropy of energy entropy of sub-frames' normalized energies. A measure of abrupt changes
4	Spectral centroid	Spectral centroid spectrum's center of gravity
5	Spectral spread	Spectral spread spectrum's second central moment of the spectrum
6	Spectral entropy	Spectral entropy entropy of the normalized spectral energies for a set of sub-frames
7	Spectral flux	Spectral flux squared difference between the normalized magnitudes of the spectra of the two successive frames
8	Spectral roll-off	Spectral roll-off the frequency below which 90% of the magnitude distribution of the spectrum is concentrated
9–21	MFCCs	MFCCs mel frequency cepstral coefficients: a cepstral representation with mel-scaled frequency bands
22–33	Chroma vector	A 12-element representation of the spectral energy in 12 equal-tempered pitch classes of western-type music
34	Chroma deviation	Chroma deviation standard deviation of the 12 chroma coefficients

i.e., boredom, engagement and anxiety, five of the common emotion classes, namely Happiness, Sadness, Anger, Fear and Neutral, were selected from the aforementioned datasets. These were the emotions deemed to be closest to the ones that we would need to extract within MaTHiSiS. A major difficulty resulting from the choice of datasets is the differences between languages, since besides the linguistic differences, there is also big variability in the way each emotion is expressed. For each classification method, six different experiments were carried out where a single dataset was used for training and the remaining two for testing. We calculated the F_1 score which is the harmonic average of the well-known precision (P—the number of correct positive results divided by the number of predicted positive results) and recall (R—the number of positive results divided by all samples that should be predicted as positive) measures. More specifically, it is calculated as $F_1 = 2 \cdot P \cdot R/1(P + R)$. Results are depicted in Table 5 and are indicative of the capabilities of the adopted approach, emphasizing the fact that in many cases training and testing data derive from different (i.e., cross-language) datasets.

Table 5 Experimental results for speech-based affect recognition

	Emovo	Savee	German
Emovo	0.48	0.22	0.49
Savee	0.29	0.57	0.34
German	0.41	0.26	0.64

Values indicate F_1 score. Line denotes training dataset, while column the testing one

4 Benefits, challenges and prospects

In Sect. 3, we described practical implementations of affect detection schemes using commonly available sensors and provided real-life results. An important *technical challenge* faced during the integration of the presented techniques in MaTHiSiS system which uses diverse devices (mobile phones, tablets, robots, laptops) was the fact that the diverse devices (which have sensors installed on them) come with different computing capabilities. This implies that a robot (with a laptop attached to it, as for example, the turtle-bot or even the NAO robot) has completely different processing capabilities from a smartphone and these processing capabilities are mandatory since our aim is affect detection in real time. In the smartphone case, the device may not have adequate processing resources to execute sophisticated affect detection algorithms based on the inertia sensor readings. In such cases, the device carrying the sensors cannot reach a conclusion regarding the affect state but, instead, it sends the sensor readings to the cloud (through a web socket interface) where ample processing power exists. We consider that in any AR/VR solution, the affect detection algorithm can/will be implemented in the same processing resources where the AR/VR application main logic is executed. This relaxes the requirement for heavy processing in the sensing devices themselves.

Furthermore, we have explored the option to combine the results reached from readings from different modalities to improve affect detection accuracy. In Ghaleb et al. (2017), multimodal emotion recognition is attempted leveraging the properties of each modality using different fusion schemes. The method proposed in Ghaleb et al. (2017) outperforms other approaches existing in the literature for challenging datasets and emotion recognition in the wild. Given that in any AR/VR system usually multiple sensing modalities exist, further exploring the combination of techniques may enable either higher accuracy or balancing the complexity of the algorithm (and relevant processing resources) implemented per modality.

AR/VR applications can *benefit* from affect detection since the evolution of the AR/VR experience does not longer depend only on the physical reaction of the user but also of the emotional reaction. The interactions between the user and the system are no longer used solely for evaluating the physical reaction of the user (and thus his performance) but also for evaluating how he feels. This important information enables not only personalization (tailoring to static user preferences/intricacies) but also adaptation of the experience (educational or other) to the real-time conditions of the user. This adaptation adds significant value to AR/VR applications which nowadays are rapidly penetrating the educational sector, as witnessed by the large investments from governments of different countries across the globe [such as the USA, France and China among others (DigiCapital 2017)] through the education and technology ministries. This market is huge with mobile AR market alone anticipated to reach $108 billion by 2021 (DigiCapital 2017). Apart from education, the aforementioned affect state approach can be exploited in other application sectors where AR and VR penetrate. These indicatively include gaming, healthcare, travel and real estate. The feeling/affect of the user can be taken into account and change the actual parameters (e.g. levels) of the game. This way the user can be kept to the desired emotional state which can be engagement or excitement or other depending on the nature of the game. An adolescent enjoys driving to the edge, while an older person enjoys staying in less stressing situations. In healthcare, AR and VR solutions augmented with affect detection capabilities can extensively be used to curate people suffering from dementia and mild cognitive disabilities. In this sector, prompting the user to act is mandatory and tailoring the pace of the experience to his preferences is safeguarding proper operation. For example, if the user is prompted to play cards, he should neither experience anxiety (being stressed to act/select a card), neither bored (being allowed to endlessly wait his action). Affect detection can ensure he is kept in what his carers feel "ideal" affect states. Similar opportunities exist in the travel and real estate sectors, where the user has to be presented with environments that lead to pleasant experience.

In Sect. 3, we proved that commonly available sensors can deliver accurate affect information. This fact creates new *business prospects* for software/solution developers. The standardization of the way the information of affect state is communicated among devices could have an important impact on the business potential of the relevant solutions. It would allow for the development of affect detection modules (either hardware with appropriate software or software that can be run in already available hardware) which can be flexibly integrated with applications (not only learning) that can exploit affect information. This can further boost AR/VR systems uptake. For example, if there is a web-based learning application that uses affect state information, when a student uses his tablet to exercise at home, the application can receive affect state as detected by the inertia sensors of the tablet, while in a training center, the installation may include a camera or a Kinect device. In the education industry, xAPI (Hansen and Ji 2010) is a format widely adopted to communicate "statements" and could be used to communicate information like "user A is bored." This information can trigger the modification of the difficulty level of the learning material the user interacts with (or any other application-dependent aspect). In case of standardization, a new wave of sophisticated affect detection software modules tailored to different devices or to diverse user groups (e.g., autistic children or elder people with dementia) to offer increased affect detection accuracy can emerge.

5 Conclusions

Augmented and virtual reality systems and applications invade educational environment rapidly capitalizing on their inherent support of enhanced experience. Although AR/VR technologies contribute to the learner engagement, this can be further boosted if the affect state of the learner is taken into account to tailor the experience to the learner's temporal state and to keep her/him in the flow state which has been shown to be the optimal state for learning processes. We proposed to take advantage of the availability of sensors in such systems to detect in real time the affect state and guide/shape the provided experience. We presented a set of practical implementations of techniques that rely on widely available sensors and achieve affect state detection with adequate accuracy for learning environments. Finally, the prospect of the penetration of such techniques in the market is discussed, and innovation opportunities were identified. In our future work, we will use the presented prototype system to evaluate the accuracy of affect detection and the benefits brought in real-life operations in multiple EU countries and diverse user groups including among others children suffering profound multiple learning disabilities, children in mainstream schools and adults in vocational training environments in different countries.

Acknowledgements The work presented in this document has been partially funded through H2020-MaTHiSiS Project. This project has received funding from the European Union's Horizon 2020 Programme (H2020-ICT-2015) under Grant Agreement No. 687772.

References

Anagnostopoulos CN, Iliou T, Giannoukos I (2015) Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artif Intell Rev 43(2):155–177

Baker RSJ, D'Mello SK, Rodrigo MT, Graesser AC (2010) Better to be frustrated than bored: the incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. Int. J. Hum.-Comput. Stud. 68:223–241

Baltrušaitis T, Robinson P, Morency LP (2016) Openface: an open source facial behavior analysis toolkit. 2016 IEEE Winter conference on IEEE applications of computer vision (WACV)

Burkhardt F et al (2005) A database of german emotional speech. In: Proceedings of interspeech, Lissabon

Chen C, Liu K, Kehtarnavaz N (2016) Real-time human action recognition based on depth motion maps. J Real-Time Image Proc 12(1):155–163

Costantini G et al (2014) Emovo corpus: an Italian emotional speech database. In: Chair NCC, Choukri K, Declerck T, Loftsson H, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the ninth international conference on language resources and evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland

Coutrix C et al (2012) Identifying emotions expressed by mobile users through 2D surface and 3D motion gestures. In: Proceedings of the 2012 ACM conference on ubiquitous computing

Csikszentmihalyi M (1996) Flow and the psychology of discovery and invention. Harper Collins, New York

Csíkszentmihályi M (2008) Flow: the psychology of optimal experience. Harper Perennial, New York

DigiCapital (2017) Augmented/Virtual Reality Report Q4. https://www.digi-capital.com/news/2017/01/after-mixed-year-mobile-ar-to-drive-108-billion-vrar-market-by-2021/#.WdyhmTBx3IV. Accessed 10 Oct 2017

Du Y, Wang W, Wang L (2015) Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR)

El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn 44(3):572–587

Ghaleb E, Popa M, Hortal E, Asteriadis S (2017). Multimodal fusion based on information gain for emotion recognition in the wild, intelligent systems conference 2017, 7–8 Sept 2017, London, UK

Hansen DW, Ji Q (2010). In the eye of the beholder: a survey of models for eyes. IEEE Trans Pattern Anal Mach Intell

Haq S, Jackson P (2009) Speaker-dependent audio-visual emotion recognition. In: Proceedings of the international conference on auditory-visual speech processing (AVSP'08), Norwich, UK

Kanade T, Cohn JF, Lucey P, Saragih J, Ambadar Z, Matthews I (2010) The extended Cohn–Kanade Dataset (CK+): a complete expression dataset for action unit and emotion-specified expression, San Francisco, USA

Kim M et al (2013) A touch based affective user interface for smartphone. In: IEEE international conference on consumer electronics (ICCE). IEEE

Leifman G et al (2016) Learning gaze transitions from depth to improve video saliency estimation. arXiv preprint arXiv:1603.03669

Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3d points. In: Proceedings of IEEE international conference in computer vision and pattern recognition workshop (CVPRW)

Mora KAF, Monay F, Odobez J-M (2014) Eyediap: a database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In: Proceedings of the symposium on eye tracking research and applications. ACM

Nottingham Trent University (ed) (2017) Adaptation and Personalization principles based on MaTHiSiS findings. Deliverable for the MaTHiSiS project. http://www.mathisis-project.eu/en/content/d61-adaptation-and-personalization-principles-based-mathisis-findings. Accessed 3/4/2018

Pedregosa F et al (2011) Scikit-learn: machine learning in Python. Journal of Machine Learning Research 12:2825–2830

Rikert TD, Jones MJ (1998) Gaze estimation using morphable models. In: Proceedings of third IEEE international conference on automatic face and gesture recognition. IEEE

Vemulapalli R, Arrate F, Chellappa R (2014) Human action recognition by representing 3d skeletons as points in a lie group. In: Proceedings of the IEEE international conference on computer vision and pattern recognition (CVPR)

Wang L, Zhang J, Zhou L, Tang C, Li W (2015) Beyond covariance: feature representation with nonlinear kernel matrices. In: Proceedings of the IEEE international conference on computer vision (ICCV)

Xiong X, De la Torre F (2013) Supervised descent method and its applications to face alignment. In: 2013 IEEE conference on computer vision and pattern recognition, Portland, pp 532–539

Zhang X, Sugano Y, Fritz M, Bulling A (2015) Appearance-based gaze estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4511–4520

Zhang X, Sugano Y, Fritz M, Bulling A (2017) MPIIGaze: real-world dataset and deep appearance-based gaze estimation. IEEE Trans Pattern Anal Mach Intell