Integrating Emotion Recognition Tools for Developing Emotionally Intelligent Agents

Samuel Marcos-Pablos[1]*, Fernando Lobato Alejano[2], Francisco José García-Peñalvo[1]

[1] Department of Computer Science, Universidad de Salamanca, (https://ror.org/02f40zc51), Salamanca (Spain)
[2] Faculty of Informatics, Pontifical University of Salamanca, Salamanca (Spain)

Received 21 April 2022 | Accepted 1 July 2022 | Early Access 19 September 2022

Abstract

Emotionally responsive agents that can simulate emotional intelligence increase the acceptance of users towards them, as the feeling of empathy reduces negative perceptual feedback. This has fostered research on emotional intelligence during last decades, and nowadays numerous cloud and local tools for automatic emotional recognition are available, even for inexperienced users. These tools however usually focus on the recognition of discrete emotions sensed from one communication channel, even though multimodal approaches have been shown to have advantages over unimodal approaches. Therefore, the objective of this paper is to show our approach for multimodal emotion recognition using Kalman filters for the fusion of available discrete emotion recognition tools. The proposed system has been modularly developed based on an evolutionary approach so to be integrated in our digital ecosystems, and new emotional recognition sources can be easily integrated. Obtained results show improvements over unimodal tools when recognizing naturally displayed emotions.

Keywords

Artificial Intelligence, Digital Ecosystems, eHealth, Emotionally Intelligent Agents, Human Computer Interaction.

DOI: 10.9781/ijimai.2022.09.004

I. Introduction

THE growing interest in the field of emotional recognition in the area of human-computer interaction (HCI) has fostered the development in recent years of numerous solutions aimed at making emotional recognition technologies available to inexperienced users. Although these technologies have been successfully employed in many fields such as online sales, trend analysis, or the study of user behavior in social networks, there is still a long way to go before they are fully developed and capable enough to be used in other fields.

The motivation of the present work is on the use of these emotional recognition tools for healthcare, and to incorporate them in our digital health ecosystems [1], [2]. There are many different applications for emotionally intelligent agents in healthcare. They can be used to help patients understand their emotional state, and to fill the information gaps when patients interact with health professionals helping health practitioners to increase the emotional understanding of their patients. Also, the emotional data gathered from the patient can be added to health records helping doctors in diagnosis and understanding of mental problems for example depression, schizophrenia, etc. Another application example are agents able to deliver personalized therapy. As more users employ technology to access health services, these services can be handled by intelligent agents. By endowing these agents with artificial empathy, we can increase the acceptance of users towards these new technologies.

* Corresponding author.
E-mail address: samuelmp@usal.es

However, the fact that many users still show reluctance to artificial intelligence (AI) means that special care must be taken when developing the agent's emotional intelligence, and even more in a delicate field as is human healthcare. The 'control degree' that emotional intelligence has over agent behaviors can make its actions better suited from an empathic interaction point of view but may generate unexpected behaviors leading to user rejection. However, there is a tendency towards the acceptance of AI which should increase as the users get more and more used to technology. Considering social robots as embodied agents, early studies in human-robot interaction in home environments suggested that users do not want a robot companion to be a friend, but to perform the tasks they are intended for. Contrary to these results, new studies suggest that robots able to accentuate their own personality are preferred by users [3].

Based on this tendency, the objective of this work is to integrate different commercially available emotion recognition tools into health services provision. However, current emotion recognition tools are generally unimodal, in the sense that they focus on a single communication channel (e.g., facial expressions, text, voice prosody, skin temperature, etc.) and provide an output which is associated to the activation of the universal emotions (namely: anger, fear, surprise, disgust, joy, and sadness). This approach is suitable for certain tasks that focus on specific affective aspects and not on the whole emotional spectrum (e.g., emotion recognition in an online review). On the other hand, multimodal emotion recognition by combining data collected from various communication channels can be employed for a wider range of applications, and provide surplus information with an increase in accuracy [4]-[7], therefore making it more suitable for health applications.

Please cite this article in press as:
S. Marcos-Pablos, F. Lobato Alejano, F. J. García-Peñalvo. Integrating Emotion Recognition Tools for Developing Emotionally Intelligent Agents, International Journal of Interactive Multimedia and Artificial Intelligence, (2022), http://dx.doi.org/10.9781/ijimai.2022.09.004

This paper presents an approach for multimodal emotion recognition using Kalman filters for the fusion of available discrete emotion recognition tools to be incorporated into emotionally intelligent agents. The proposed system has been modularly developed based on an evolutionary approach so to be integrated in our digital health ecosystems and allowing new emotional recognition sources to be easily integrated. The paper has been divided into five sections. The second section describes the background behind emotional intelligence and emotion parameterization for multimodal emotion fusion. The third section describes the proposed method and an implementation for testing purposes. The fourth section shows and discusses the results of the proposed approach. Finally, last section summarizes the main conclusions of this work.

II. Background

A. Emotionally Intelligent Agents' Main Characteristics

The different scientific disciplines where agents are applied means that different definitions of intelligent agents can be found in the literature. Originally robotics was the primary driver for agent-based research, however current agents include software mimicking or acting on behalf of humans (i.e., software agents) or internet robots (i.e., web-bots) [8]. From a general perspective, three major categories of agents can be distinguished: human agents, hardware agents and software agents [9]. Using the analogy of human agents, intelligent hardware and software agents are defined as capable of generating goals, performing actions, communicating messages, sensing environment, adapting to changing environments, and learning.

Continuing with the human analogy, emotional intelligence is defined as the accurate appraisal and expression of emotion in oneself and in others, the effective regulation of emotion in oneself and others, and the use of feelings to motivate, plan, and achieve day-to-day actions [10]. On the other hand, empathy can be defined as the cognitive ability to infer the thoughts and feelings of others and then developing a similar response to what the other person is feeling [11]. Empathy can be divided into cognitive empathy, or the ability to understand another's mental state; and affective empathy, described as the ability to respond with appropriate emotional reaction to the mental states of another.

From the above definitions, it can be inferred that endowing an agent with emotional intelligence involves sensing the environment, learning from the environment, and generating goals and actions adapted to the environment conditions and state. Therefore, emotional intelligent agents need to be provided with three main abilities: sense, compute, and act. In addition, as emotional intelligent agents are likely to interact with human users, their final goal should consider the environment with the focus on the user and provide a certain degree of empathy during human-computer interaction or collaboration. This means that the agent must be capable of capturing users' emotions (sense), appraisal of captured emotions to regulate its internal state (compute), and finally perform tasks where actions are regulated by the computed "emotional" state (act).

Many different approaches exist for capturing user's emotions, which include capturing movements [12], physiological parameters [13] or voice [14]. Deep learning is also widely employed for identifying face and body expressions in 2D and 3D [15]. Apart from trying to improve accuracy rates, recent systems focus on a multimodal approach to diminish the effects produced by variation of emotional display between users [15]. However, to date many of the non-invasive emotion sensing proposals (i.e., using cameras or microphones) encounter problems with emotion recognition in uncontrolled environments. For that reason, latest research is focused on capturing user's emotions under high variability conditions (light, noise, occlusions, etc.) [16].

In emotionally intelligent agents, as in human psychology, emotions are recognized as functional in decision-making by influencing motivation and action selection. Thus, emotion appraisal is needed after sensing to regulate the agent's internal state which in turn will determine the following actions to take. Emotions can be seen as a response to a certain stimulus that elicits a tendency towards a certain action, and as complex feedback signals that shape behavior. Therefore, processing emotions should be approached from a dual perspective: motivated action and feedback. Therefore, emotional intelligence in software agents can be seen as how the system processes emotions, focusing on how input is translated through an algorithm to an output and whether it contains a knowledge of past events or history. There are several types of algorithms for this purpose, which include fuzzy models, Markov models, neural networks, probability tables, reinforcement learning and unsupervised machine learning approaches such as K-means, K-medoids or self-organizing maps [3].

Acting is related to the tasks or operations the agent oversees conducting. For emotionally intelligent agents, actions should be adapted to the environment conditions and state. This means that not only agent's actions should be derived from the users' emotional state, but preferably the agent should return the user corresponding emotional feedback. The way such emotional reaction is expressed highly depends on the agent's degree of anthropomorphism, and can be as simple as a text message, color, or sound. However, as the agent anthropomorphism increases (e.g., in virtual avatars or robots), it turns necessary to employ natural communication channels (e.g., voice synthesis, facial animation) to match the agent's behavior with its appearance and avoid falling into the uncanny valley [3].

From the three characteristics described above, this paper focuses on capturing users' emotions (sense) using a multimodal approach based on different emotional sources.

B. Emotion Parametrization for Multimodal Recognition

Recognizing emotions from sensed signals is a complicated task but can be divided into two main steps: feature extraction and classification. During feature extraction, signal processing techniques are employed to produce a set of numerical values from the sensed signal (i.e., audio, video, EMG, and others). These numerical values collect certain features of the signal so that they can be processed by a computer. The extracted features are then processed by a classifier, which is complemented by a scoring function to produce the final emotional estimation. After feature extraction, emotion classification aims to provide with meaning to the observed features.

Multimodal emotion recognition can be seen as the fusion of information from different sources. There are two main approaches to fuse emotional data: feature-level fusion or early fusion, and decision-level fusion or late fusion [4]. Feature-level or early fusion fuses the features extracted from various sources such as visual, text and audio features into a unique feature vector which is sent for analysis. As the correlation between the features is performed at an early stage, feature level fusion can provide better results. However, putting all features in the same format is not an easy task and can induce to errors, as the features obtained from different channels can differ in many aspects. On the other hand, in decision-level or late fusion approaches the features of each input channel are examined and classified independently, and then their results are fused to obtain a decision vector as an output. The advantage is that the fusion of decisions obtained from different tools is easier as these tools usually have the same form of data. Additionally, each input channel can be processed with the most suitable classifier for its particular features.

Whichever approach is taken for computerized emotion recognition (unimodal, feature-level or decision-level fusion), it is necessary to parameterize human emotions, so that recognition can be done using quantitative computational techniques. In general terms, emotion parameterization can be divided into simple and dimensional approaches. Simple discrete models associate a set of detected patterns in the sensed emotional sources to the basic core of 6 universal emotions (namely: anger, fear, surprise, disgust, joy, and sadness). Such patterns do not need to be related with the psychology behind the expression, and unimodal recognition tools and many feature-level techniques use this approach.

On the other hand, dimensional approaches parameterize emotions as a lineal combination of different psychological dimensions. Emotion parametrization has been widely studied in psychology, and the most accepted dimensional models of emotion are the circumplex model, the vector model , and the Positive Activation - Negative Activation (PANA) model [17]. Multimodal decision-level emotion recognition systems employ this dimensional approach.

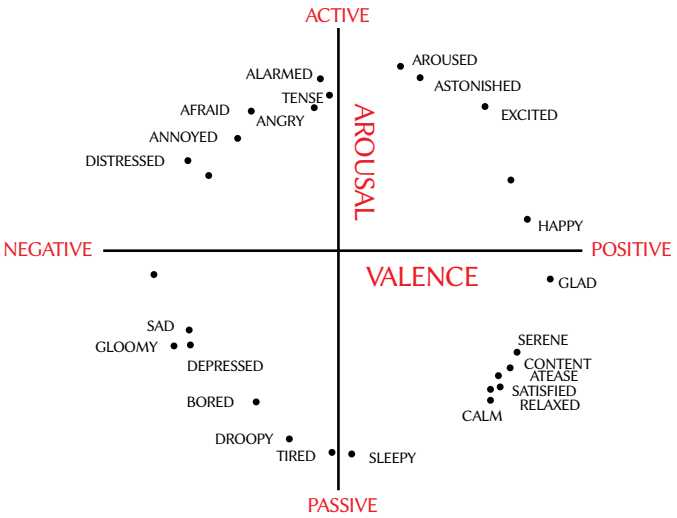

Fig. 1. The circumplex model of emotions.

In the circumplex model of affect, emotions can be categorized by 2 dimensions: valence, from unpleasant (negative) to pleasant (positive); and arousal, from passive (weak emotion) to active (strong emotion). By varying the values of each dimension, emotions can be plotted on two coordinate axes (see Fig. 1). Emotions are distributed in space with dimensions of arousal and valence in a circular pattern centered on medium arousal and neutral valence. In the vector model valence is modeled as binary, so emotions are plotted in a v-shape around the positive valence axis. The PANA model is like a 45-degree rotation of the circumplex model, where the two axes are: Positive Activation (PA), which goes from active, elated, and excited, to drowsy, dull, and sluggish; and Negative Activation (NA), which goes from distressed, fearful, nervous to calm, at rest, and relaxed. Vector models have been found to better describe emotional properties of text, whereas circumplex and PANA models have been identified for describing emotion in words, prosody and facial expressions [17].

The following section presents our approach for multimodal decision-level emotion recognition using Kalman filters for the fusion of decisions obtained from different commercially available recognition tools. As these decisions come from different communication channels, the circumplex model is used to parameterize emotions.

III. Methods

A. Acquiring and Parameterizing Existing Emotional Tools

In order to integrate different emotional recognition sources, we have developed a modular architecture which was introduced in [18]. The adopted approach was intended to be seamlessly deployed not only in computational agents, but also in physical agents such as robots. It consists of two main submodules: the facial emotion recognition submodule and the speech emotion recognition submodule, although additional emotion recognition sources can be added. Data coming from different sources is incrementally fused employing Kalman filters, as will be described in the next section. In addition, the developed architecture can be applied to physical agents such as robots, as data is exchanged between the different submodules employing ROS (Robot Operating System) messages.

The emotional recognition system modules have been programmed using both JavaScript and Python. Sound and image capture are performed through JavaScript, along with the different calls to cloud emotional APIs for image recognition and speech-to-emotion transcription. On the other hand, audio splitting and Kalman filtering have been deployed in Python. Python has also been employed to implement some of the message interchange over ROS. In addition, a front-end web page for testing purposes has been developed in html (Fig. 2).

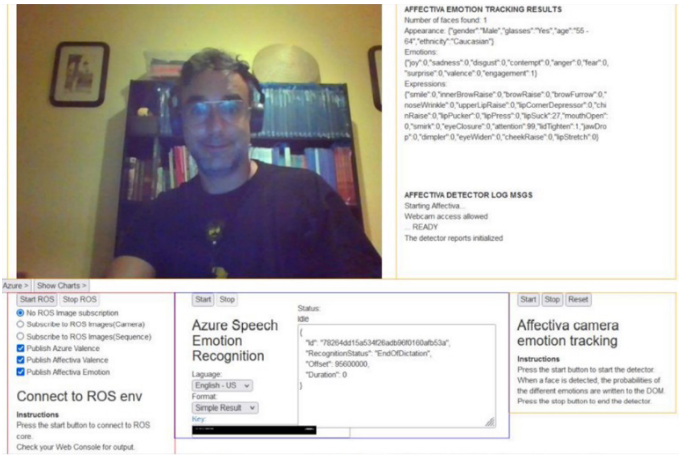

Fig. 2. Developed html interface for testing the proposed architecture.

At present, facial emotion recognition is approached by the combination of two tools: Affectiva's Affdex SDK and Microsoft's Emotional API. A call to these emotion recognition services takes a video or an image as an input and analyzes its emotional content returning a set of values using JSON syntax. After parsing the returned JSON, a value ranging from 0 to 100 is obtained for each of the six universal emotions (anger, disgust, fear, joy, sadness and surprise) [19] plus contempt and neutral. This value indicates the "activation" level for each expression.

Rather than using this value, and to integrate the facial output with other submodules, we have taken a circumplex approach to emotion recognition (See Fig. 1). Although the circumplex model is a continuous model where primary emotions can be expressed at different intensities and can mix with one another to form different emotions, there are studies in the literature that have tried to assess the approximate location of these emotions in the valence/arousal axis [20], [21]. Based on these works, we have converted the returned activation value of each expression into two components ranging from [-1, 1] for both valence and arousal:

$$A_E = 2\alpha_E(\frac{Val_E}{100} - 0.5) \quad (1)$$

$$V_E = 2\beta_E(\frac{Val_E}{100} - 0.5) \quad (2)$$

Where A_E and V_E are the valence and arousal of the expression 'E', whereas Val_E is the returned activation value from the facial recognition tools for that expression. The components (α_E, β_E) have been approximated from the literature as follows: anger (0.8, -0.8); disgust (-0.5, -0.5); fear (0.4, -0.9); joy (0.8, 0.3); sadness (-0.8, -0.2); surprise (0.95, 0.0). As the recognition focuses on the six universal expressions, we have discarded contempt and considered neutral as (0, 0).

A similar approach is followed for speech emotion recognition tools. Namely, current implementation makes use of Vokaturi prosodic-acoustic emotion recognition and Microsoft's Speech-to-Text and Text Analytics. Audio is captured continuously and processed using the SoX – Sound eXchange audio editing software. When a silence in speech is detected, the audio file chunk is sent to the prosodic emotion recognition module. The output of this module is then converted into valence and arousal following a similar approach as the one described before for the facial emotion recognition tools.

On the other hand, emotion recognition in the speech content analysis is divided in two main stages: speech to text, where the audio signal is converted into words, and text sentiment analysis, where text is provided with emotional meaning. After a chunk of raw audio data is captured, it is sent to the cloud service. The cloud service responds with a JSON containing the recognized text, along with other parameters such as the detected language, recognition confidence or the duration of the speech. The second step takes as an input the speech transcript, calls the Microsoft text sentiment analysis and returns a sentiment score which ranges from 0 (which represents a negative sentiment) to 100 (representing a positive sentiment). This sentiment score is translated into a valence value ranging from [-1, 1].

B. Integrating Emotional Sources

To integrate the data computed from the different emotional recognition sources, a sensor fusion approach is taken. Sensor fusion or data fusion is widely used in other fields such as robotics, where data coming from different sensors is merged to increase accuracy and reduce uncertainty. For example, in robot localization data from an inertial navigation system and a global positioning system GPS can be merged using filtering techniques to improve robot's navigation performance. One of the most important features of sensor fusion is that it allows to combine the information from complementary sensors, redundant sensors or even from a single sensor over a period of time. Furthermore, the advantages of using this approach are:

- Redundant information can reduce uncertainty and increase the accuracy with which features are sensed by the system.
- Multiple sensors delivering redundant information increases reliability in case of errors in a data source or when no data is available from a certain source.
- Complementary information from multiple data sources allows the perceived environment to be characterized in a way that would be impossible to perceive using only the information from each data source separately.

To merge the different emotional data sources, we have used the Kalman filter algorithm. This filtering technique is a recursive process that allows to estimate the value of the variables of interest (valence and arousal) based on knowledge of the current and previous observations, together with the description of their noise and errors. Some of the limitations of the Kalman Filter are that it can only be used for linear or linearized processes and measurement systems. However, the nature of the processed captured emotional data in terms of linear combinations of valence and arousal fit this condition. Also, the uncertainty of Kalman filter is restricted to Gaussian distribution, while other filtering techniques such as the particle filter can deal with non-Gaussian noise distribution. In any case, Kalman filtering algorithm tries to converge into correct estimations, even if the Gaussian noise parameters are poorly estimated [22].

The Kalman filter represents the system state by using two equations, where variables can be matrixes:

$$x_k = Ax_{k-1} + Bu_k + w_{k-1} \quad (3)$$

$$z_k = Hx_k + v_k \quad (4)$$

Equation (3) indicates that the data values x_k are a linear combination of its previous value, a control signal u_k and a process noise w_k-1. In our case, there is no control signal so the second term can be discarded. Equation (4) indicates that any measurement value is a linear combination of the data value and the measurement noise. While entities A, B and H are in general form matrices, in our case they can be modelled as numeric and constant as in many other signal processing problems. As described in the literature, they can be considered to be equal to 1 for simple processes [22]. Noise values w_k- and v_k are considered as Gaussian, and an approximation to their mean and standard deviation had been initially obtained from the literature on the selected emotional tools [23] – [25].

Once the system had been modelled, we have made use of a python implementation of the Kalman filter. The filtering process estimates the output at a particular state from measured data by following two steps. Firstly, the state of the system is predicted:

$$\hat{x}_{\bar{k}} = A\hat{x}_{k-1} + Bu_k \quad (5)$$

$$P_{\bar{k}} = AP_{k-1}A^T + Q \quad (6)$$

Secondly, the collected observations are incorporated once they have been corrected:

$$K_k = P_{\bar{k}}H^T(HP_{\bar{k}}H^T + R)^{-1} \quad (7)$$

$$\hat{x}_k = \hat{x}_{\bar{k}} + K_k(z_k - \widehat{Hx}_{\bar{k}}) \quad (8)$$

$$P_k = (1 - K_kH)\,P_{\bar{k}} \quad (9)$$

Where X_k is the current estimation, K_k the Kalman gain, z_k the measured value and Q is the covariance of the process noise. The process starts from an initial estimation of cero for x_0 and an error covariance matrix P_k estimated from the literature. During iteration, a prediction step is first performed, based on difference between consecutive emotion measures to compute (5) and (6). Then, as new measurements arrive (7), (8) and (9) are used to compute the estimated valence and arousal values and update the error. These estimated values are the system output at a particular state and are also used as input for the next prediction step.

IV. Results and Discussion

The system has been tested on two types of pre-recorded databases: posed and natural, as there is an important distinction between spontaneous and deliberately displayed emotions. Apart from being initiated in two different parts of the brain, but also elicited facial expressions do not look identical. Spontaneous facial expressions are characterized by synchronized, smooth, symmetrical facial muscle movements whilst posed expressions are subject to volitional real-time control and tend to be less smooth, with more variable dynamics [26].

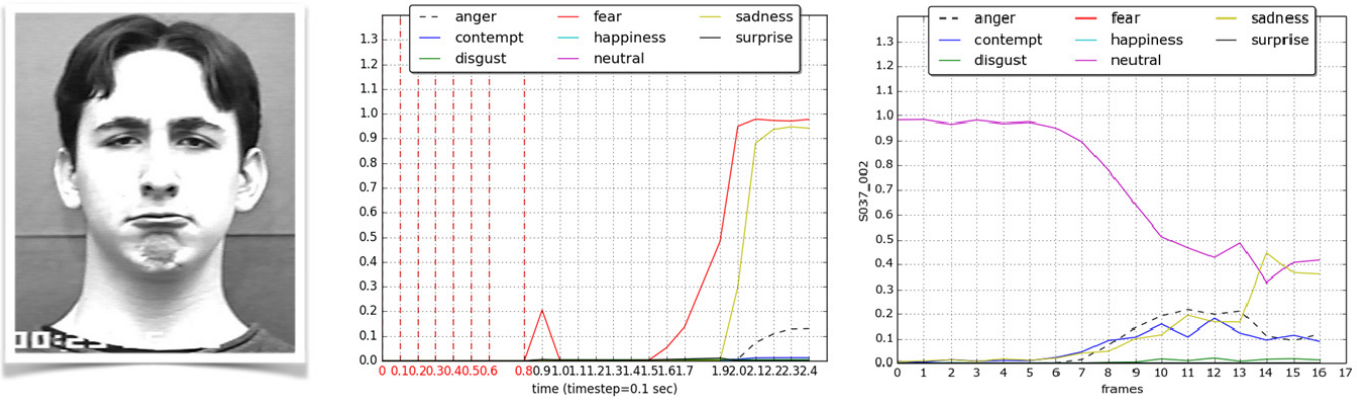

Fig. 3. Recognition results for Affectiva (center graph) and Microsoft (right graph) face emotion recognition APIs, tested on person 37, sequence 2 of the Cohn-Kanade database where the expression is labelled as sad. Affectiva was tested as live sequence whereas Microsoft was tested frame by frame. Red vertical lines indicate lost and repeated calls to Affectiva's service until the service is ready.

The selected posed databases are the Cohn-Kanade expression database [27] and the Ryerson Emotion database [28]. The Cohn-Kanade database consists in approximately 500 frontal camera image sequences from 100 subjects. Accompanying meta-data include annotation of FACS action units (AUs) (i.e., micro-expressions) [19]. However, image sequences have no sound and are only labelled in terms of action units but not emotional expressions. It is therefore necessary to translate the AU labelling for each sequence into its corresponding emotional expression. For our experiments, we have made use of the emotional labelling developed by Buenaposada et al. [29] who selected a subset of 333 sequences of 92 people from the Cohn-Kanade database and labelled them with their corresponding emotional expression. In the Ryerson Emotion database, video samples are collected from eight subjects, speaking six languages (English, Mandarin, Urdu, Punjabi, Persian, and Italian). The Ryerson Emotion database contains 720 audiovisual emotional expression samples, where subjects were provided with a list of emotional sentences and were directed to express their emotions as naturally as possible by recalling the emotional happening, which they had experienced in their lives. A frontal camera was used to record the samples in a quiet and bright environment, with a simple background.

As for natural emotional expressions, the Belfast Induced Natural Emotion Database was used. It contains recordings of mild to moderate emotionally colored responses to a series of laboratory-based emotion induction tasks [30]. The recordings are accompanied by information on self-report of emotion and intensity, continuous trace-style ratings of valence and intensity, the sex of the participant, the sex of the experimenter, and the active or passive nature of the induction task. An excerpt of the results obtained for the three databases can be found at (https://bit.ly/3OArZnp).

Facial emotion recognition APIs have been tested independently on the Cohn Kanade database to assess their performance. Overall, accuracy of both Affectiva's SDK and Microsoft's Emotional API match what is described in the literature [23] – [25]. However, Affectiva's precision (the share of correctly predicted images out of all images predicted as one category) is lower than Azure. That is, it tends to output an emotion even if not sure of what the correct emotion is. On the other hand, it has higher sensitivity (the share of correctly predicted images out of all images truly in the respective category) as Azure outputs a neutral value when unsure. Fig. 3 shows an example of this behavior. The image on the left shows the final frame (out of 16) of the sequence corresponding to person 37, sequence 2 of the Cohn-Kanade database. This sequence is composed of 16 frames that go from neutral pose to expression apex and was labelled as sad. Middle graph shows the Affectiva output, where fear and sadness are obtained as output with values close to 1. Right graph shows Microsoft output, where sad expression is only dominant during the last frames of the sequence, near the expression apex. In fact, in this case it is not very clear whether the person in the picture is sad or angry, but Affectiva still gives two high results for sadness and fear. However, there are some expressions, such as anger, that are better recognized by Affectiva and not by Microsoft. Therefore, merging these two tools can help reducing errors.

We also tested if the emotion recognition was dependent of the frame rate and if the recognition considered previous frames, as it was not clear in the case of Affectiva. In the case of Microsoft, emotion recognition is done frame by frame as stated in their documentation. As Cohn-Kanade sequences were recorded at 30 frames per second, we called the recognition services at different frame rates. Fig. 3 shows the Affectiva service being called at 10 fps. The red vertical lines correspond to lost service calls as Affectiva takes almost 1 second to initialize. Obtained results indicate that even though Affectiva accepts video as input, the recognition is little dependent of previous outputs and frame rate.

The problem with using facial recognition APIs that analyze emotions frame by frame is that there may be recognition peaks in emotion sequences even though the averaged resulting expression is correctly recognized. If analyzing micro-expressions (e.g., rising eyebrows), AU timing is related to the amount of time needed for each associated muscle to contract. On the other hand, emotional expression timing is context dependent, but it is unlikely in a real scenario to have an expression of emotion that only lasts a fraction of a second. This can lead to errors when recognizing real emotions and if the result is combined with other recognition sources. Fig. 4 shows recognition results on subject 3 sequence ha5 of the Ryerson emotion database, labelled as happiness. Left graph shows Microsoft APi output, whilst right graph shows the Kalman filter results combining Microsoft and Affectiva outputs. As it can be seen, the Kalman filter output smooths the results and removes recognition peaks.

Overall face emotion recognition tools work well on posed expressions and obtain great results on both Cohn Kanade and Ryerson databases, classifying above 80% of the sequences in the correct category. Results match those found in the literature [31], including the pattern of classifying fearful expressions as surprise or sadness, along with Microsoft's tendency to misclassify anger sadness and fear as neutral (see Fig. 3). In addition, tools that have been trained on posed and deliberate expressions fail to generalize to the complexity of expressive behavior found in real-world settings [32].

Fig. 5 (left graph) shows the results of applying just face emotion recognition tools to the analysis of natural expressions. It can be

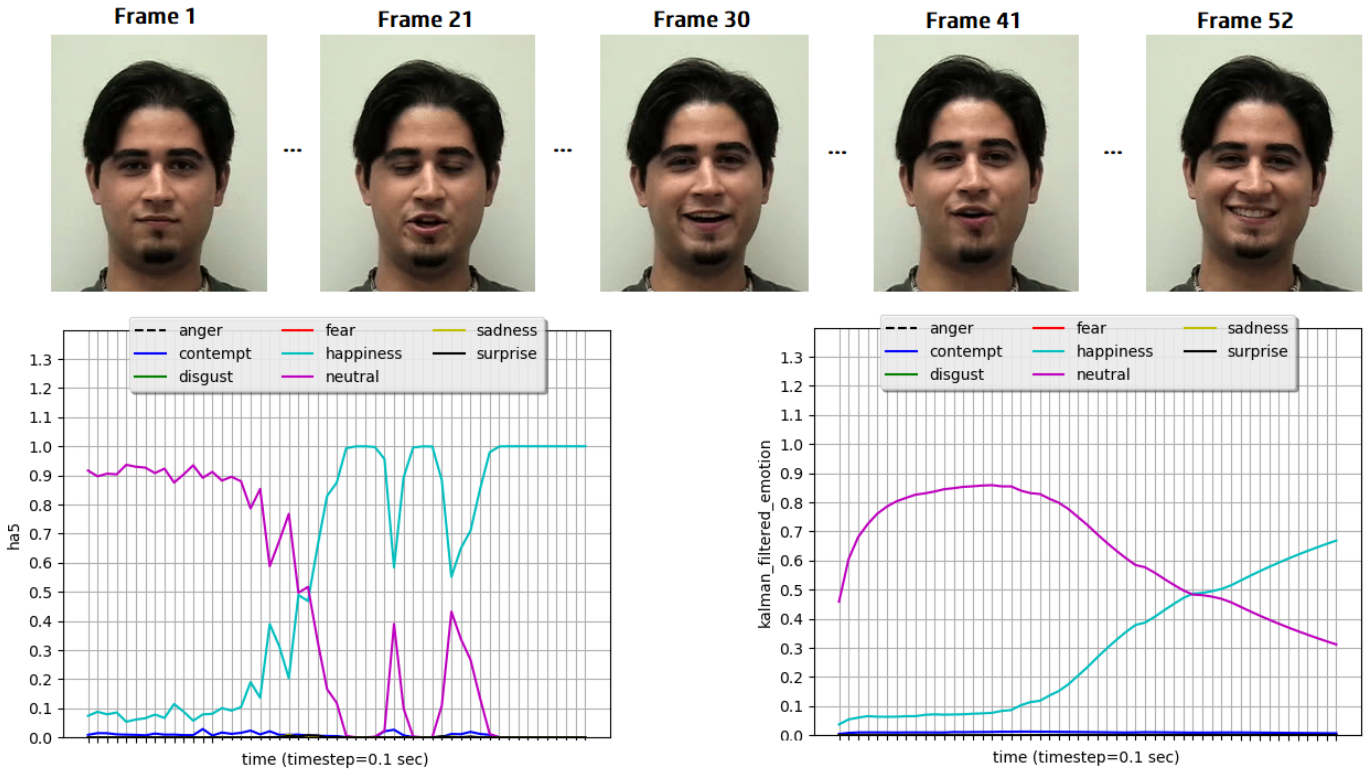

Fig. 4. Recognition results on subject 3 sequence ha5 of the Ryerson emotion database. Left graph shows Microsoft APi output, whilst right graph shous the Kalman filter results of combining Microsoft and Affectiva outputs.

observed that both tools fail to clearly output an emotion, and only in some of the initial and central frames happiness is recognized over neutral. However, images correspond to sequence 79d of the Belfast Induced Natural Emotion Database, where the task which elicited the emotion is encoded as active and social and the targeted emotion is frustration. The annotation of the sequence has negative values for valence and arousal on average, being lower in the case of arousal.

Above results show the convenience of adding other emotional sources to the recognition process, especially for the detection of spontaneous emotions. To incorporate emotion recognition from speech, captured audio is split in chunks using the SoX – Sound eXchange audio editing software. Audio file chunks can be used for prosody and text sentiment analysis. Different software has been tested for prosody analysis (Vokaturi, Beyond Verbal and openSmile). The best obtained results have been for Vokaturi, although its precision was found below 40% for both Belfast and Ryerson databases. However, as it also has very low sensitivity, prosody can be incorporated in the system provided that neutral recognition is discarded, and the error is considered during the filtering process.

On the other hand, emotion recognition from text using Microsoft Text Analytics provides improved results. As a test, we have extracted 200 reviews of New York City hotels from TripAdvisor. For each review we have saved the title, the content, and the score of the overall experience. The returned sentimental score by the text recognition tool is then compared with the number of stars assigned by the reviews' authors. The comparison is considered successful if the score is less than 40% and the assigned stars were 1 or 2 or if the score is more than 60% and the corresponding stars were 4 or 5. The reviews with 3 stars (which is supposed to be a neutral score) were not taken under account since it is unlikely that those reviews were effectively 100% neutral. Using these rules, the obtained accuracy is above 85% for positive reviews and of 89% for negative reviews.

Given this good performance, the analysis of emotions in text was therefore used to correct the deviations of the other sources of emotional recognition. That way, right graph on Fig. 5 shows the overall result of the system for recognition, where face recognition data output is translated in terms of valence and arousal and corrected as the speech recognition results are obtained.

The results show how the overall recognized emotion goes from the positive valence and arousal values obtained from the face recognition tools to negative values and close to depressed. Emotion of this sequence is self-catalogued in the Belfast database as frustration, which is associated to low levels of positive arousal and high levels of negative valence (see Fig. 1). However, we consider that obtained results are quite close to the real emotion displayed in the sequence as the person in the video speaks with a lot of laziness. In any case, obtained results are way closer to the real self-reported emotion than those obtained by the facial recognition tools.

It should be noted, however, that although obtained results for natural sequences are greatly improved, the system still fails to recognize some of the analyzed sequences from the Belfast database. These errors may be associated to two limitations of the current approach: on the one hand, by applying Kalman filtering the valence bias of the estimates obtained from the facial recognition tools can be corrected thanks to the accuracy and precision of the speech content emotion recognition measures. On the other hand, the arousal bias is only corrected by the prosodic emotion analysis, which is found to have low accuracy and precision, thus making it difficult to model. In addition, inter-subjects' variability for displaying emotions may affect the modelled error bias, as it has been built from averaged values obtained from the literature. To overcome these limitations, thanks to the modularity of the proposed architecture, additional emotional recognition sources and tools can be added in future implementations. Also, alternative approaches to the circumplex model for modelling emotions can be explored [33], as there are emotions such as fear

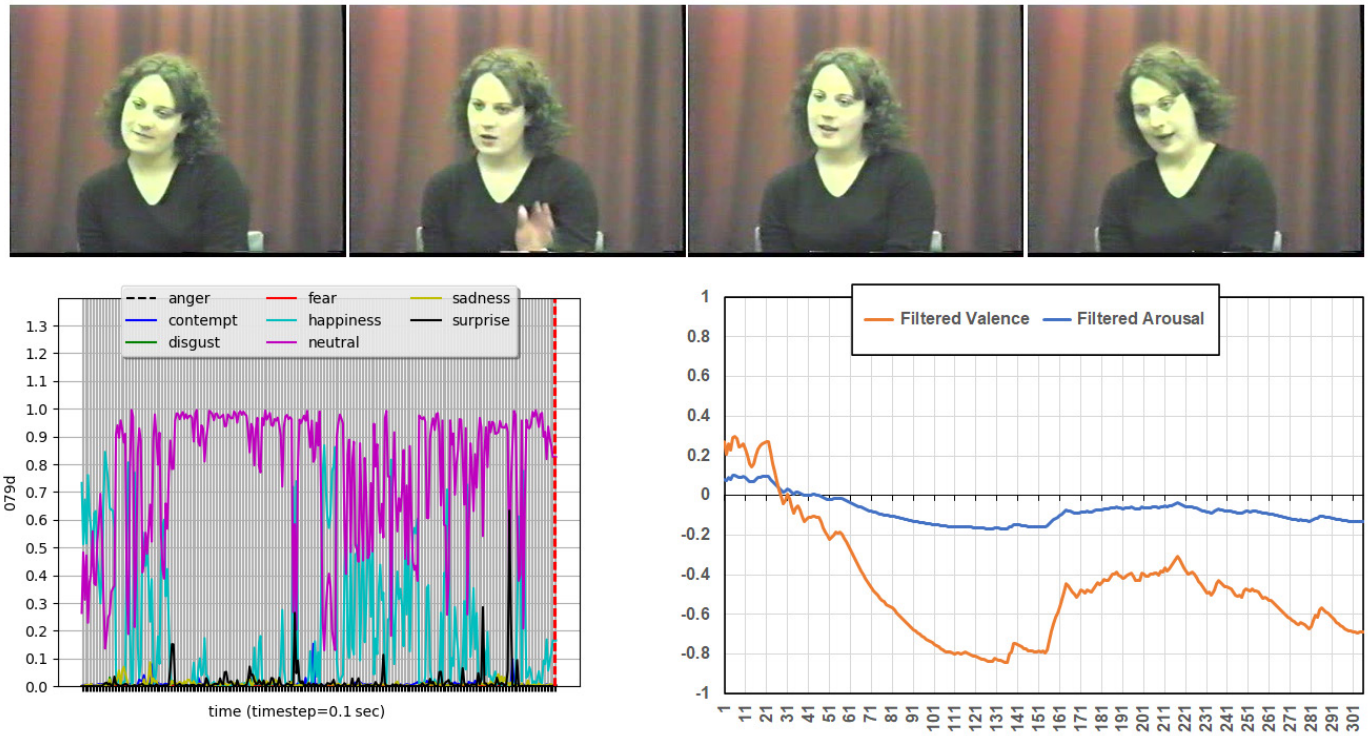

Fig. 5. Emotion recognition results corresponding to sequence 79d of the Belfast Induced Natural Emotion Database, where the task which elicited the emotion is encoded as active and social and the targeted emotion is frustration. The annotation of the sequence has negative values for valence and arousal on average, being lower in the case of arousal. Left graph shows the output obtained from facial emotion recognition tools. Right graph shows the results of the system, where face, voice and text emotion recognition sources are combined using Kalman filters for obtaining the valence and arousal of the displayed expression.

and anger (both are negative and active) that are located in the same quadrant and very close in the 2D space.

V. Conclusion

In this paper we have shown a decision-level fusion of commercially available discrete emotion recognition tools using Kalman filters. The proposed system has been modularly developed based on an evolutionary approach so to be integrated in our digital health ecosystems, and new emotional recognition sources can be easily integrated.

Obtained results show that commercially available tools such as Microsoft or Affectiva face emotion recognition APIs achieve very good recognition rates for posed expressions where no speech is involved, but their accuracy diminishes dramatically when the user communicates naturally. By fusing these tools with other recognition sources such as text analytics or prosody emotion recognition, we obtained great improvements in terms of recognizing emotions in natural databases. Moreover, with the proposed approach the output is expressed in terms of valence and arousal thus providing continuity in the emotional spectrum. This improves its potential for new applications, as allows employing existing recognition tools for more complex tasks as the ones related to health care provision

Future work should focus on adding new tools and emotional source channels to the proposed architecture and to integrate the emotion recognition system in our health ecosystem services.

Acknowledgements

This research was partially funded by the Spanish Government Ministry of Science and Innovation through the AVisSA project grant number (PID2020-118345RB-I00).

References

[1] A. García-Holgado, S. Marcos-Pablos, F.J. García-Peñalvo, "A Model to Define an eHealth Technological Ecosystem for Caregivers", in *New Knowledge in Information Systems and Technologies*, Á. Rocha, H. Adeli, L. P. Reis, and S. Costanzo, Eds. Springer International Publishing, 2019, pp. 422–432, https://doi.org/10.1007/978-3-030-16187-3_41.

[2] S. Marcos-Pablos, A. García-Holgado, F.J. García-Peñalvo, "Modelling the business structure of a digital health ecosystem", in *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2019, pp. 838–846, https://doi.org/10.1145/3362789.3362949.

[3] S. Marcos-Pablos, F.J. García-Peñalvo, "Emotional Intelligence in Robotics: A Scoping Review", In *New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence*, J. F. de Paz Santana, D. H. de la Iglesia, & A. J. López Rivero, Eds. Springer International Publishing, 2022, pp. 66–75, https://doi.org/10.1007/978-3-030-87687-6_7.

[4] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion", *Information Fusion*, 2017, vol. 37, pp. 98–125, https://doi.org/10.1016/j.inffus.2017.02.003.

[5] M.G. Huddar, S.S. Sannakki, and V.S. Rajpurohit, "Attention-based Multimodal Sentiment Analysis and Emotion Detection in Conversation using RNN", International Journal of Interactive Multimedia and Artificial Intelligence, 2021, vol. 6, no. 6, pp. 112-121, http://doi.org/10.9781/ijimai.2020.07.004.

[6] H. Daus and M. Backenstrass, "Feasibility and Acceptability of a Mobile-Based Emotion Recognition Approach for Bipolar Disorder", International Journal of Interactive Multimedia and Artificial Intelligence, 2021, vol. 7, no. 2, pp. 7-14, http://doi.org/10.9781/ijimai.2021.08.015.

[7] M. Magdin, D. Držík, J. Reichel, and S. Koprda, "The Possibilities of Classification of Emotional States Based on User Behavioral Characteristics", International Journal of Interactive Multimedia and Artificial Intelligence, 2020, vol. 6(Regular Issue), no. 2, pp. 97-104, http://doi.org/10.9781/ijimai.2020.11.010.

[8] S. Kirrane, "Intelligent software web agents: A gap analysis", *Journal*

of Web Semantics, 2021, vol. 71, 100659, https://doi.org/10.1016/j.websem.2021.100659.

[9] W. Brenner, R. Zarnekow, and H. Wittig, "Intelligent Software Agents: Foundations and Applications", Springer-Verlag Berlin, 1998, https://doi.org/10.1007/978-3-642-80484-7.

[10] P. Salovey, and J. D. Mayer, "Emotional Intelligence", *Imagination, Cognition and Personality*, 1990, vol. 9, no. 3, pp. 185–211, https://doi.org/10.2190/DUGG-P24E-52WK-6CDG.

[11] W. Ickes, "Empathic Accuracy", *Journal of Personality*, 1993, vol. 61, no. 4, pp. 587–610, https://doi.org/10.1111/j.1467-6494.1993.tb00783.x.

[12] E. van der Kruk, M. M. Reijne, "Accuracy of human motion capture systems for sport applications; state-of-the-art review", European Journal of Sport Science, 2018, vol. 18, no. 6, pp. 806–819, https://doi.org/10.1080/17461391.2018.1463397.

[13] L. Shu, J. Xie, M. Yang, Z. Li, D. Liao, X. Xu, and X. Yang, "A Review of Emotion Recognition Using Physiological Signals", *Sensors*, 2018, vol. 18, no. 7, 2074, https://doi.org/10.3390/s18072074.

[14] M.L. Rohlfing, D.P. Buckley, J. Piraquive, C.E. Stepp, and L.F. Tracy, "Hey Siri: How Effective are Common Voice Recognition Systems at Recognizing Dysphonic Voices?", *Laryngoscope*, 2021, vol. 131, no. 7, pp. 1599-1607, https://doi.org/10.1002/lary.29082.

[15] N. Samadiani, G. Huang, B. Cai, W. Luo, C.H. Chi, Y. Xiang, and J. He, "A Review on Automatic Facial Expression Recognition Systems Assisted by Multimodal Sensor Data", *Sensors*, 2019, vol. 19, no. 8, https://doi.org/10.3390/s19081863.

[16] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments", *Information Fusion*, 2201, vol. 68, pp. 46–53, https://doi.org/10.1016/j.inffus.2020.10.011.

[17] D.C. Rubin and J.M. Talarico, "A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words", *Memory*, 2009, vol. 17, no. 8, pp. 802–808, https://doi.org/10.1080/09658210903130764.

[18] S. Marcos-Pablos, F.J. García-Peñalvo, and A. Vázquez-Ingelmo, "Emotional AI in Healthcare: A pilot architecture proposal to merge emotion recognition tools" in *Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2021, pp. 342–349, https://doi.org/10.1145/3486011.3486472.

[19] P. Ekman and E.L. Rosenberg, "What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)", *Oxford University Press*, 2005, https://doi.org/10.1093/acprof:oso/9780195179644.001.0001.

[20] G. Paltoglou, M. Thelwall, "Seeing Stars of Valence and Arousal in Blog Posts", in *IEEE Transactions on Affective Computing*, 2013, vol. 4, no. 1, pp. 116–123, https://doi.org/10.1109/T-AFFC.2012.36.

[21] M. Olszanowski, G. Pochwatko, K. Kuklinski, M. Scibor-Rylski, P. Lewinski, and R.K. Ohme, "Warsaw set of emotional facial expression pictures: A validation study of facial display photographs" *Frontiers in Psychology*, 2015, vol. 5, https://www.frontiersin.org/article/10.3389/fpsyg.2014.01516.

[22] B. Ristic, S. Arulampalam, and N. Gordon, "Beyond the Kalman Filter: Particle Filters for Tracking Applications", *Artech House*, 2003.

[23] A. Bhattacharjee, T. Pias, M. Ahmad, and A. Rahman,"On the Performance Analysis of APIs Recognizing Emotions from Video Images of Facial Expressions", *17th IEEE International Conference on Machine Learning and Applications*, 2018, pp. 223–230, https://doi.org/10.1109/ICMLA.2018.00040.

[24] A. Mathur, A. Isopoussu, F. Kawsar, R. Smith, N.D. Lane, and N. Berthouze, "On Robustness of Cloud Speech APIs: An Early Characterization", in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 2018, pp. 1409–1413, https://doi.org/10.1145/3267305.3267505.

[25] S.R. Khanal, J. Barroso, N. Lopes, J. Sampaio, and V. Filipe, "Performance analysis of Microsoft's and Google's Emotion Recognition API using pose-invariant faces", in *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion*, 2018, pp. 172–178, https://doi.org/10.1145/3218585.3224223.

[26] W.E. Rinn, "The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions", *Psychological Bulletin*, 1984, vol. 95, no. 1, pp. 52–77.

[27] T. Kanade, J.F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis", in *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 46–53, https://doi.org/10.1109/AFGR.2000.840611.

[28] Ryerson Emotion Database. (n.d.). Retrieved April 23, 2022, from https://www.kaggle.com/datasets/ryersonmultimedialab/ryerson-emotion-database

[29] J.M. Buenaposada, E. Muñoz, and L. Baumela, "Recognising facial expressions in video sequences", *Pattern Analysis and Applications*, 2008, vol. 11, no. 1, pp. 101-116.

[30] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The Belfast Induced Natural Emotion Database", *IEEE Transactions on Affective Computing*, 2012, vol. 3, no. 1, pp. 32–41, https://doi.org/10.1109/T-AFFC.2011.26.

[31] T. Küntzler, T.T.A. Höfling, and G.W. Alpers, "Automatic Facial Expression Recognition in Standardized and Non-standardized Emotional Expressions", *Frontiers in Psychology*, 2021, vol. 12, https://www.frontiersin.org/article/10.3389/fpsyg.2021.627561.

[32] "A New Video Based Emotions Analysis System (VEMOS): An Efficient Solution Compared to iMotions Affectiva Analysis Software". (n.d.). ASTES Journal. Retrieved April 25, 2022, from https://astesj.com/v06/i02/p114/.

[33] Z. Kowalczuk and M. Czubenko, "Computational Approaches to Modeling Artificial Emotion – An Overview of the Proposed Solutions", Frontiers in Robotics and AI, 2016, vol. 3, https://doi.org/10.3389/frobt.2016.00021.

Samuel Marcos-Pablos

He received a Telecommunication Engineer's Degree in 2006, a M.Eng. in robotics in 2009, and a Ph.D. in robotics in 2011 from the University of Valladolid (Spain). He has worked as a researcher at CARTIF's Robotics and Computer Vision Division from 2007 - 2018, where he combined theoretical and field work in the research and development of projects in the area of Social and Service robotics and computer vision. He is currently with the GRIAL research group, and focuses his efforts in the development of ecosystems for the health sector and teaching. Among others, he has authored papers for the journals of Interacting With Computers or Sensors MDPI, as well as conferences such as the IEEE International Conference on Intelligent Robots and Systems and the IEEE International Conference on Robotics and Automation.

Fernando Lobato Alejano

Fernando Lobato Alejano is a Ph.D. in Computer Engineering from the Pontifical University of Salamanca, a Technical Engineer in Computer Systems and has a Degree in Computer Engineering from the Catholic University of Murcia. He also has Master's Degree in teaching, specializing in technology. He is the author of different book chapters and has a multitude of intellectual property registrations, as well as a utility model and a patent in progress. He has been awarded in the cross-border competition for market-oriented prototypes (Prototransfer/Inespo 2018) and currently works as researcher and Professor at the Pontifical University of Salamanca in the areas of Computing Engineering and Management of Technology-Based Companies.

Francisco José García-Peñalvo

He received the degrees in computing from the University of Salamanca and the University of Valladolid, and a Ph.D. from the University of Salamanca (USAL). He is Full Professor of the Computer Science Department at the University of Salamanca. In addition, he is a Distinguished Professor of the School of Humanities and Education of the Tecnológico de Monterrey, Mexico. Since 2006 he is the head of the GRIAL Research Group GRIAL. He is head of the Consolidated Research Unit of the Junta de Castilla y León (UIC 81). He was Vice-dean of Innovation and New Technologies of the Faculty of Sciences of the USAL between 2004 and 2007 and Vice-Chancellor of Technological Innovation of this University between 2007 and 2009. He is currently the Coordinator of the PhD Programme in Education in the Knowledge Society at USAL. He is a member of IEEE (Education Society and Computer Society) and ACM.