

Design of a Virtual Assistant to Improve Interaction Between the Audience and the Presenter

S. Cobos-Guzman^{1*}, S. Nuere², L. De Miguel¹, C. König³

¹ Universidad Internacional de la Rioja, Escuela Superior de Ingeniería y Tecnología (ESIT), Logroño, La Rioja (Spain)

² Universidad Politécnica de Madrid, UPM, Escuela Técnica Superior de Ingeniería y Diseño Industrial. Madrid (Spain)

³ Universitat Politècnica de Catalunya, UPC, BarcelonaTech. Barcelona (Spain)

Received 6 August 2020 | Accepted 27 June 2021 | Published 16 August 2021



ABSTRACT

This article presents a novel design of a Virtual Assistant as part of a human-machine interaction system to improve communication between the presenter and the audience that can be used in education or general presentations for improving interaction during the presentations (e.g., auditoriums with 200 people). The main goal of the proposed model is the design of a framework of interaction to increase the level of attention of the public in key aspects of the presentation. In this manner, the collaboration between the presenter and Virtual Assistant could improve the level of learning among the public. The design of the Virtual Assistant relies on non-anthropomorphic forms with 'live' characteristics generating an intuitive and self-explainable interface. A set of intuitive and useful virtual interactions to support the presenter was designed. This design was validated from various types of the public with a psychological study based on a discrete emotions' questionnaire confirming the adequacy of the proposed solution. The human-machine interaction system supporting the Virtual Assistant should automatically recognize the attention level of the audience from audiovisual resources and synchronize the Virtual Assistant with the presentation. The system involves a complex artificial intelligence architecture embracing perception of high-level features from audio and video, knowledge representation, and reasoning for pervasive and affective computing and reinforcement learning to teach the intelligent agent to decide on the best strategy to increase the level of attention of the audience.

KEYWORDS

AI, Human-machine Interaction, Learning, Multimedia, Virtual Assistance.

DOI: 10.9781/ijimai.2021.08.017

I. INTRODUCTION

IN the academic world, the mission of the professor/teacher/lecturer (presenter) is to communicate knowledge in an efficient way [1]. This general objective can be achieved thanks to the experience of the presenter and the own skills acquired during the years of teaching. However, even if the presenter has great strategies for communication, it is difficult to get the attention of students or the public when interest lacks in the presentation's topic [2]. Feedback about the attitude and level of attention of the audience could be useful for the presenter to adapt and personalize the presentation according to the requirements of the audience.

Nowadays, there are several applications of virtual agents [3]-[5] or autonomous robots that can interact with the public using virtual or mechatronic faces [6]. Social interaction between humans and robots is paramount in applications where a high level of collaboration is required such as in hospitals or industry. For this reason, the design of a robot to generate an instantaneous positive reaction in humans is key for fostering human-robotic interaction.

Habitually, roboticists put much effort into the design of a new robot on physical embodiment properties such as locomotion, manipulation, haptic interactions, and face mechatronic interaction. However, for social interaction, technologies such as virtual agents or chatbots which focus on communication abilities rather than physical embodiments have been developed.

Therefore, this study presents a novel design of a virtual agent using artificial intelligence methods for creating a framework for effective collaboration between the public and the presenter. This framework is thought for improving the interaction of presentations in large venues (e.g., auditoriums with 200 people) where direct visual contact of the presenter and the audience is difficult. Thus, the new virtual agent can be used for: virtual interaction, preliminary design for a new future mechatronic system and to provide a hybrid combination of humans and virtual assistants that can help to increase the level of attention during a presentation.

The contribution of this research is the design of a Virtual Assistant with four levels of interaction to increase the audience's attention during a presentation. The four levels of interaction are derived from the intelligent system proposed. The intelligent system must detect different behavioral patterns of the audience such as divert/distract attention, people asleep, positive, or negative participation, and confused participation.

* Corresponding author.

E-mail address: salvador.cobos@unir.net

Furthermore, the study presents a set of psychological analyses of the virtual assistant's graphic design to validate the suitability for its use in the different levels of interaction. The psychological analysis considers the level of education and type of public to recommend the use of the virtual assistant with a given level of interaction. The information about the audience's profile can be used as an input in the system for adapting the mode of interactions.

This work also presents a design of the system's architecture, which is based on 3 different modules as follows: i) the first module is based on video and image recognition in order to identify different patterns (e.g., type of public and level of attention); ii) the second module analyzes audio and will contain natural language processing algorithms to recognize the phrases of the presenter and the feedback information from the public. iii) the third module contains an intelligent agent that can guarantee in real-time the synchronization with the presenter and the presentation; Moreover, the third module will depend on the recognition of patterns from the first and second module (perception modules) to correctly synchronize the collaboration between the presenter and the virtual assistant. For this, the system involves a knowledge representation and reasoning module to attribute a semantic meaning to the high-level perceptions regarding the attention and emotion. Knowledge representation is important to model the context of the current state of the world so that intelligent decisions can be taken. Reinforcement learning is proposed as the solution to teach the intelligent agent to decide on the best interaction with the audience at each moment.

For example, information such as applauds, and the expression of faces captured through image analysis provide feedback about the attitude of the audience in each moment, so that the Virtual Assistant can decide to take certain actions to improve the level of attention, i.e., activate a given interaction mode of the Virtual Assistant.

The remaining part of this article is organized as follows: Section II describes the design of the Virtual Assistant tackling the visual appearance followed by a discussion of the proposed design and the results of the evaluation of the virtual assistant animations by different publics (Section III). Section IV describes the design of the artificial intelligence system. Conclusions and future lines of work are presented in Section V.

II. DESIGN OF THE VIRTUAL ASSISTANT

A. Design of the Appearance

During the last years, many virtual assistants have been designed with different shapes and voices, both in the academy and in the industry. Some virtual agents as Eliza [7] have feminine aspects, trying to evoke human characteristics. But trying to replicate human-like is not as easy and is not all about just benefits as there are so many consequences related to psychology, specifically with how we perceive and interact within the perception.

There are many studies regarding the concept of "uncanny valley". This concept explores the idea that there is something strange in the level of anthropomorphism of something that is not alive, in a biological way, but looks like. Professor Masahiro Mori [8] introduced this concept into robotics, after Sigmund Freud's article 'Das Unheimliche' [9]. The original term comes from Ernst Jentsch [10]. The creation of intelligent agents that work efficiently and effectively 'would be impossible to solve without understanding and using mechanisms of social-emotional cognition' [11]. It is important to keep in mind that the way how we perceive will incise directly in our interaction. In this context, we have tried to avoid this kind of perception though creating a virtual interaction.

A study compared three different animated objects during a cognitive task and their impact on the behavior and performance of primary school children [12]. The results revealed that animated objects were well-accepted as interacting partners and had a positive impact on their emotional state. Also, it shows that 3D objects (animal and robot) with more "live" characteristics elicited more positive behaviors, and the animal-like object decreases attention. Other studies from this field of research encourage managers to take care of the appearance and behavior of robots and promote collaboration between managers and researchers to define the limit of anthropomorphism [13]. Considering all this information, we have developed several designs for the robotic appearance and behavior of a Virtual Assistant aimed to improve the attention of the audience during a presentation.

B. Personification

In the XXI century, the resource of personification is one of the most used rhetorical figures in the field of animation and character design, but historically it has been used in the field of literary-fabulous creation and 2D animation (cartoons), since its inception. This resource has enormous potential to evoke empathy and closeness with the personified object, but at the same time, in the words of Radoslav (1996), the personification can "constitute an educational instrument, capable of producing profound positive effects" [14].

First, an analysis is made of the objects that may be suitable for the personification of the assistant. The training context in which the project is registered is online, therefore, suitable hardware elements are chosen for this purpose. Of all those available, the webcam is chosen for two compelling reasons. The first is the element that visually connects the participants in a virtual face-to-face educational context. Second, because it is considered that its nature is ideal to find formal and conceptual aspects that support the feeling of life that you want to give it.

C. Form

Trying to follow the advice of Andrew Jimenez, from Pixar Animation Studios [15], about the importance of not getting complicated when creating a character, since the important thing is that the idea in your head is raised simply and naturally [16], of all the types of webcams on the market, inspiration is chosen that model whose form is simple and at the same time versatile for its reorientation as "living being".

A model with a round central body is chosen, with the objective cantered, side light pilot on and base in one piece. From there, they begin to make sketches to choose which of their attributes are potentially suitable to support their mobility and characterization when presenting emotionally to the assistant.

At first, the objective is considered as a mouth adding the eyes and eyebrows. It is also necessary to add some element that supports the sensation of life and its expressive load, for which the power cable of the webcam is used as an arm that helps to emphasize the expressiveness of the assistant and his attitude in each of the poses. This aspect has special relevance as it already happened with other references from the world of animation, consulted for this design, such as, for example, the character in the movie *Monsters, INC.* [16], Mike Wazowski who was originally designed as a spherical body with two legs, without arms. But its creators realized that the arms would give him that feeling of reality and more suitable mobility to interact with his co-stars, resulting in closer and less strange to the viewer.

But the facial result is overloaded, losing the feeling of a webcam in which the most important thing is the objective. Thus, it is decided to simplify the model, turning the objective of the inspiration model into the eye of the assistant. Thus, the base of the webcam is modelled on the different attitudes that the assistant must simulate the movement of the feet.

D. Color

“No color is meaningless” [17], and for this reason, the colors that will be part of the assistant have been specifically chosen.

The wizard is white, with shades of blue. The lights and shadows in their movements will give rise to shades of gray. White color favors attention to the object, freeing itself from all colors. It is a bright color and favors the contrast in combination with blue color. Associated as Eva Heller [17] indicates to what is empty and light, it will highlight the expressions of the assistant.

For its part, the assistant’s eye is blue on a white background, which according to Eva Heller is associated with intellectual qualities. Its typical combination is blue and white, both being related to intelligence, science, and concentration. Likewise, and according to the survey carried out by this author, the blue color is the most appreciated with 45%. Another fact to consider is the symbolism associated with the combination of different colors, and specifically blue, white, and gray, in this case, produced by the shadows produced by white, is associated with intelligence. Fig. 1 shows the preliminary sketches of the virtual assistant’s concept.

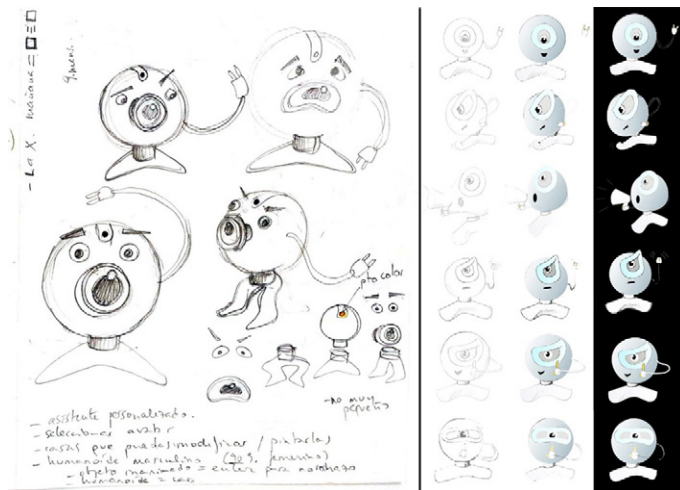


Fig. 1. Preliminary sketches of the virtual assistant’s conceptualization.

III. DESIGN OF THE VIRTUAL BEHAVIOR

According to [18], the emotional representation is composed of emotional characteristics, attributes, and attitudes, among others. In this sense, the assistant has an outgoing and close personality, capable of expressing negative or positive attitudes based on the different animations according to the objective to be achieved with each interaction with those attending virtual face-to-face sessions.

Regarding the level of attention in the audience, we consider four states: normal conditions; keep silence; low level of distraction, and high level of distraction. Based on this categorization, we try to reinforce the desired behaviors of the audience: High attention and interaction are rewarded with positive feedback and situations of distraction should be penalized through negative feedback, in this case, “anger face” (Fig. 2 (d)). Also, regarding the condition of interactions within the audiences, we present 4 states: people semi-sleeping; people sleeping, time for questions, and confusing questions. The four designs are clearly differentiating these states and represent a human-like reaction to each one (Fig. 3). About the condition of the responses of the presentation, we distinguish 2 states: positive feedback and positive participation. The virtual design reflects each state using common symbols (Fig. 4).

In the following, we explain how the virtual assistant controls the level of attention of the audience. The first mode is activated when the algorithm detects that people are talking or there are divert of attention. Fig. 2 (a) and (b) show the transition when the intelligent system detects when people are talking in the room, indicating that is important to keep silent in the room. This interaction is useful for the presenter to control the level of attention. Fig. 2 (c) and (d) show interactions when the algorithm detects different levels of distraction.

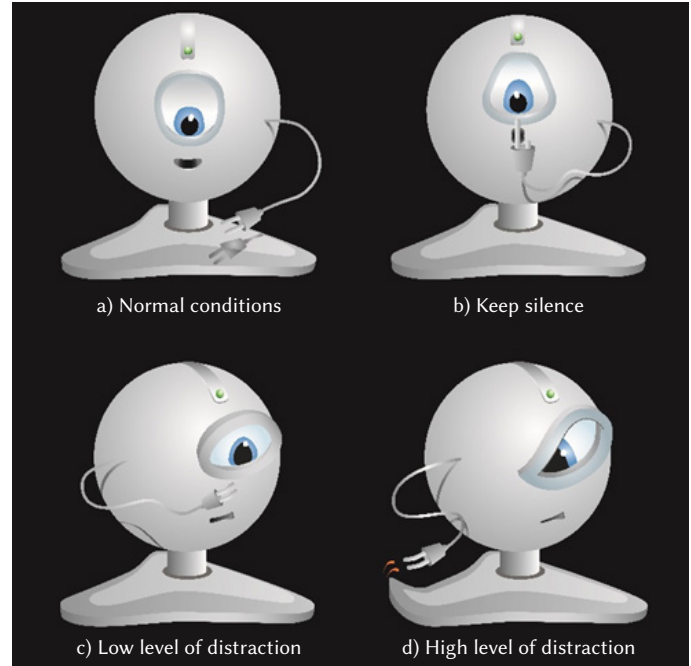


Fig. 2. Modes of interaction; (a) normal conditions; (b) keep silence; (c) low level of distraction and (d) high level of distraction.

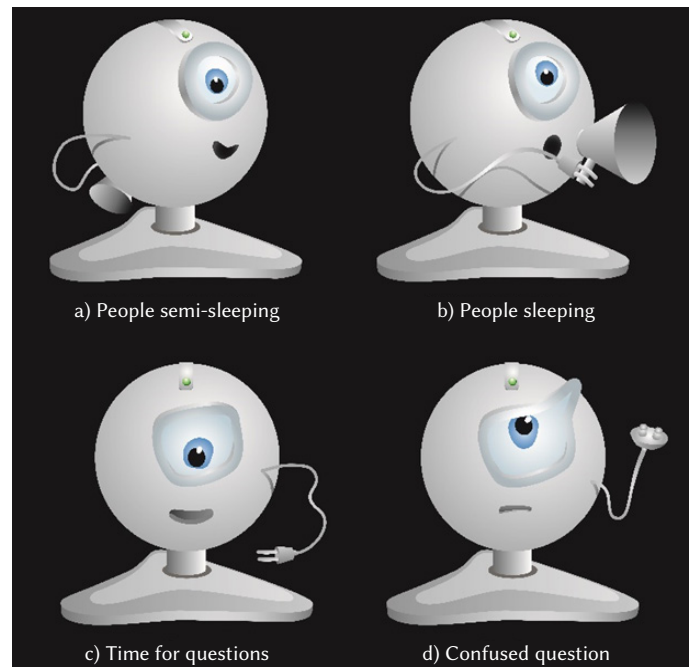


Fig. 3. Modes of interaction; (a) people semi-sleeping; (b) people sleeping (c) time for questions and (d) confused question.

The second mode is activated when the algorithm detects that people are sleeping in the room. Fig. 3 (a) and (b) show the transition when with the help of the presenter the level of attention can be increased.

Moreover, Fig. 3 (c) and (d) show the transitions when people are interacting or asking questions and the question is confusing. Finally, the last mode of interaction is when the system recognizes a positive interaction or comments as is shown in Fig. 4 (a) and (b).

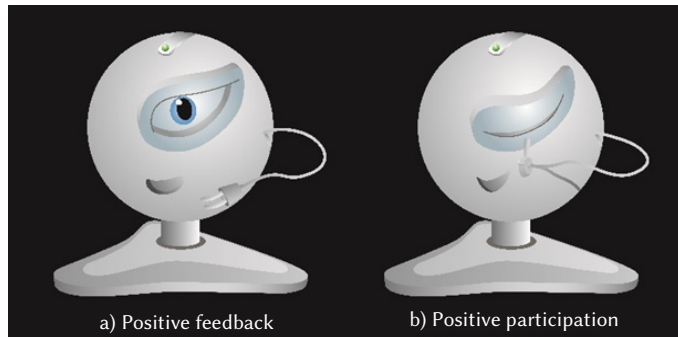


Fig. 4. Modes of interaction; (a) positive feedback; (b) positive participation

A. Discussion of the Virtual Design

Humans use different types of communication in verbal, written, and facial expressions. Face expressions [19] are enormously powerful because they instantaneously communicate emotions and intentions. This characteristic is the key aspect of the design of our artificial virtual assistant. For this reason, we use a smile and a gesture that represents a positive emotional state such as “it is O.K.” represented in Fig. 4 referring to positive interactions or gestures like Fig. 2 (b) that transmit messages such as “keep silence”.

Hence, the design presented in this work has been selected according to the research carried out in section II and III. This analysis helped to encounter the best features for a virtual assistant. Therefore, as a result, our proposed design is a non-anthropomorphic form, as we pretend to reduce the uncanny valley effect.

As a result, the Virtual Assistant’s design incorporates a mixture of robotic and artificial life accessory characteristics. The design includes non-artificial characteristic or “live” characteristics such as the eye, mouth, and the cable-arm to give it an acquainted appearance. With these live characteristics, we try to increase the level of positive behavior in the public. This relies on the fact that humans can recognize patterns of facial expressions nearly instantly as the human learning system has been trained by years for face recognition. Therefore, this natural way of recognition helps to interpret a smile as a positive answer (Fig. 4) and a negative answer as the “anger face” (Fig. 2 (d)).

The color chosen for the principal object (the camera) is white as it represents purity, elegance, and truthfulness. It is also simple and recognizable. But also, it is the counterpart of the color black, and that is why there are some nuances of grey color to emphasize the object. The combination of these colors fosters the contrast of the image, highlighting the color chosen for the blue eye. As Molly E. Holzschlag [20] points out “as white is necessary for contrast and design, it is ideal to mix it with another color that has a stronger and more obvious meaning”. The blue color inspires affection, friendship, confidence, and harmony. According to a survey [17], blue is the color that has more followers (46% men and 44 women) and only a few people do not like it (1% men and 2% women). It is commonly related to positive feelings.

The spherical form is directly in consonance with common traditional web cameras. It represents perfection as it does not have a beginning and an end, as well as it depicts protection and movement [21]. The shape of the figure also inspires stability as it incorporates a foot stand.

These positive and negative interactions are used during the

presentation as input to a reinforcement learning algorithm designed to improve the quality of the presentation. Thus, the proposed architecture can activate the specific modes of interaction as a response to the detection of certain situations employing voice and vision recognition. Therefore, the combination of the different types of interaction with the presenter is a useful tool to support efficient communication of knowledge to the public.

Reinforcement learning has its foundation in psychological principles and practice. This methodology is based on positive reinforcement and punishment (showcased by the virtual assistant). For its implementation, it is important that the intelligent agent of the system can use functions to distinguish the audience’s profile, which is derived from the recognized patterns in the audience through vision and sound processing technologies.

Additionally, the proposed solution represents a creative approach to define the boundaries between biological and artificial life from human expressions for a learning enhancement system avoiding the uncanny valley effect in the design of the virtual assistant.

As we explained in section II, animal representations are biological representations, but they generate negative effects because the audience may get more distracted. In contrast, if the design is too artificial it is possible that people cannot understand its meaning. For this reason, the presented design uses a mixture of features from biological and artificial representations to generate a better interaction with the audience. At the same time, this system will help the presenter to emphasize efficiently the knowledge during the presentation.

B. Evaluation of the Animations

The animations of the virtual design were evaluated using a psychological study based on a ‘discrete emotions questionnaire’ [22]. In this analysis, 34 participants (23 female and 11 male) with an age between 20 and 50 participated in the evaluation of the five animations as follows:

Animation 1: transition from Fig. 2 (c) and 2 (d); animation 2: transition from Fig. 2 (a) and 2 (b); animation 3: transition from Fig. 3 (a) and 3 (b); animation 4: transition from Fig. 4 (a) and 4 (b) and animation 5: transition from Fig. 3(c) and 3(d).

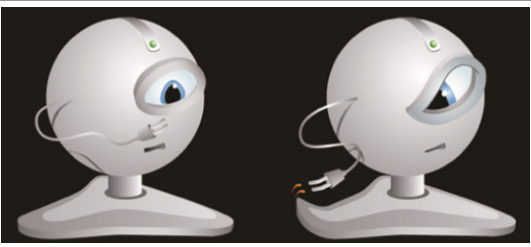
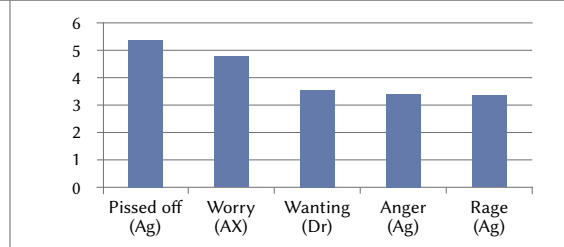
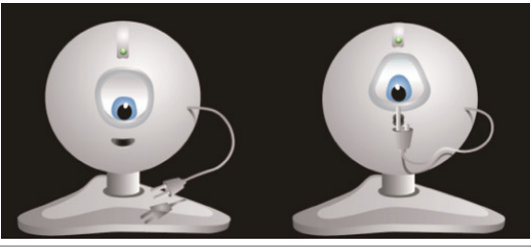
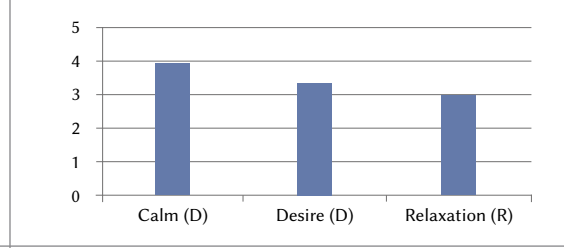
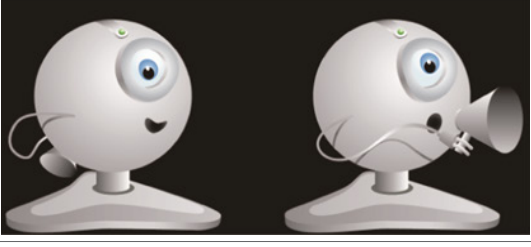
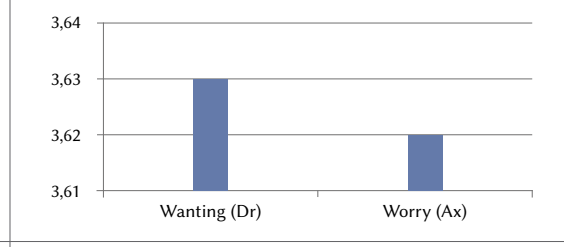


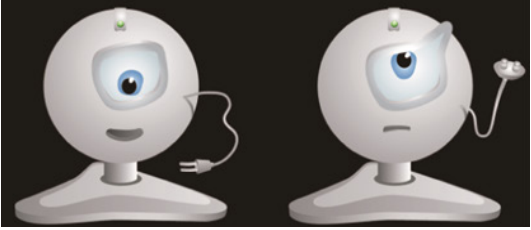
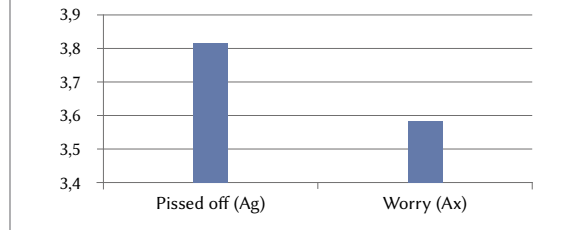
The questionnaire is based on different items that can measure the following emotions: Anger (Ag), Anxiety (Ax), Desire (Dr), Disgust (Dg), Fear (F), Happiness (H), Relaxation (R) and Sadness (S). The emotions were measured on a scale ranging from 1 to 7. The results of the questionnaire are shown in Table I and explained below according to the most relative scores (from 3).

The results of animation 1 generate three emotions of anger as highest values, 1 emotion of anxiety, and 1 emotion of desire. Therefore, this animation achieves the main objective of generating a negative emotion in the public. Animation 2 produced two emotions of relaxation and one of desire for the keep silent animation (Fig. 2 b).

Animation 3 resulted in emotions of desire and one of anxiety for the case ‘people are sleeping’ (Fig. 3 b). Animation 4 yielded three emotions of happiness and two of relaxation. This result is important because the positive feedback animation is evaluated by the public through a combination of positive emotions such as happiness and relaxation. Finally, the animation 5 generates two emotions of Anger and Anxiety.

As conclusion, the results of the survey confirm that the emotional reactions of the people coincide with the emotional effect, which was designed for each animation. These results are important because if any animation does not generate an adequate emotional reaction in the audience, the interaction with the public can be ineffective, resulting in frustration and ignoring of the virtual assistant.

TABLE I. EVALUATION OF THE FIVE ANIMATIONS

<p>Animation 1</p>		
<p>Animation 2</p>		
<p>Animation 3</p>		
<p>Animation 4</p>		
<p>Animation 5</p>		

IV. DESIGN OF THE ARTIFICIAL INTELLIGENCE ARCHITECTURE

The artificial intelligence architecture for the human-machine interaction system comprises three main modules. Two perception modules are responsible for the visual recognition and acoustic analysis (sound analysis and natural language recognition) of patterns from the audience and a third module, responsible for the real-time synchronization between the presenter and the virtual assistant. Fig. 5 describes the organization of the main modules.

A. Perception Modules

The purpose of Module 1 is the recognition of the public’s attitude towards the presentation through the caption of the visual focus of attention. A similar approach has been used recently to analyze the behavior of persons in group meetings [23] using a multi-modal sensor approach. In our approach, the visual patterns are used to detect situations remotely where the public is not putting the appropriate attention. This can be implemented by using face recognition

algorithms, which detect whether the eyes are open or closed [24] or calculate the eye gaze direction [25]. These two variables allow the system to derive information about the level of attention of the audience. The eye gaze direction and head movements have resulted in the key for detecting attention shifts in a previous educational study [26]. Analyzing the eye gaze direction, the system can determine automatically how many people pay attention to the projection of the presentation. Moreover, module 1 must recognize the type of people according to the average age of the audience [27]. Thus, these types of techniques allow the system to determine if people are asleep, divert, or distract attention.

On the detection of these previous situations, the virtual assistant will be activated with the corresponding interaction model, as described in section III.

Module 2 is used to capture the acoustic feedback from the audience. From a technical point of view, the system must carry out acoustic analysis to recognize certain human activities related to different types of noise. Acoustic scene and event recognition are an

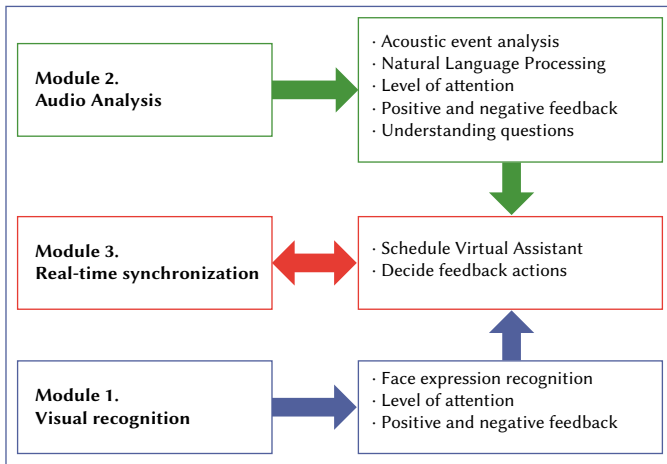


Fig. 5. System architecture of the intelligent system for the virtual assistant.

active field of research, where deep neural network models currently outperform humans in recognizing certain types of acoustic events [28]. In the present system, the focus is on the detection of attention drift primarily based on the level of noise in the room [29]. Another feature of the acoustic module is speech recognition to analyze comments from the audience. Speech recognition relies on complex models for natural language processing [30-32]. Active research during the last decades, especially in the area of neural networks [33] have implemented speech recognition as a helpful human-machine communication technology [34] in virtual assistants. In our system the automatic recognition of speech is designed as an input to the content of the presentation as the algorithm can detect automatically in which slide of the presentation the comment is made. Another functionality of the module is the detection of positive and negative comments and other types of feedback (such as applause). This information will be used for the intelligent agent of module 3 to decide the best mode of the virtual assistant.

The integration of information from several signals (video, audio, or linguistic) is referred to as multimodal feature extraction and fusion providing more reliable and rich information compared to that derived from single-modality signals [35]. Nevertheless, the high dimensionality and the complexity of the input pattern require sophisticated machine learning models for the extraction of high-level features from raw data. Recently, several cloud providers specialized in artificial intelligence technologies, such as Google or IBM, offer services for the automatic extraction of features from video, audio, and natural speech. These services are provided with a scalable and quite affordable cost, so that system developers may integrate such technologies in their software. Other alternatives are open-source software solutions, such as 'Pliers' [36], which provide an extensible framework to automatically implement semantic feature extraction from audiovisual resources.

B. Knowledge Representation

To process the perceptions from the visual recognition or acoustic module, the recognized perceptions must be transformed into treatable information for an intelligent agent by a proper knowledge representation. Semantic networks (or ontologies) provide a formal description of domain knowledge defining the characteristics of entities and the relationship between them. Ontologies use description logics such as OWL as the representational formalism for the domain description providing the ability to apply logical inference on the entities and model a context (state of the world). The ability of reasoning on knowledge representation is an enabler for pervasive (context-aware) computing [37].

Many knowledge bases have been released with the aim of information reuse and sharing. ConceptNet [38], for example, is an important knowledge base describing the semantic of several thousand of concepts. SenticNet [39] is an extension from ConceptNet with the ability to describe sentiment-based annotations. Such representation of human emotions in a computer interpretable format is tackled as affective computing [40]-[41], an interdisciplinary discipline, which aims to enable computer systems to recognize and interpret human emotions. SenticNet considers the categories of admiration, anger, disgust, fear, interest, joy, sadness, and surprise. Recently, OntoSenticNet [42] was released providing a connotative description of emotions to the concepts using polarity values. For the present system, we consider the use of OntoSenticNet, as emotions and attention levels must be attributed to the concepts sensed from the perception module.

The following examples illustrate the conceptualization of three relevant perceptions in the proposed system, namely 'doze_off' (tiredness), 'applause', and 'noise' by OntoSenticNet (Table II). The term 'doze_off' is described as semantically close to the terms 'fall_asleep', 'bed_time', 'become_bored' or 'drowsiness'. The corresponding mood tags are sadness and disgust, and the overall emotive valuation is negative with an intensity of 0.54. The emotional assessment (sentic values) gives a fine-grained evaluation of emotions in different categories. All three analyzed terms show a proper conceptualization regarding the attention and emotional categorization from a human point of view, which confirms the adequacy to use the OntoSenticNet as a knowledge base for emotion characterization in the system.

TABLE II. CONCEPTUAL MAP OF THE TERM 'DOZE-OFF' (1), 'APPLAUSE' (2) AND 'NOISY' (3)

N	Semantic	Mood tags	Sentics	Polarity
1	fall_asleep bed_time become_bored drowsiness	Sadness Disgust	Pleasant (0.48) Attention (-0.08) Sensitive (0.04) Aptitude (0)	Value: Negative Intensity: -0.54
2	watch_play entertainment cheering attend_	Interest Admiration	Pleasant (0.164) Attention (0.24) Sensitive (-0.14) Aptitude (0.28)	Value: Positive Intensity: 0.179
3	Loud uncontrollable obsolete last_clue	Sadness Disgust	Pleasant (-0.44) Attention (-0.15) Sensitive (0) Aptitude (-0.36)	Value: Negative Intensity: -0.21

C. Reasoning Module

Module 3 implements the artificial intelligence agent, who decides on the operating mode of the virtual assistant in real-time. According to [42] "all reinforcement learning agents have explicit goals, can sense aspects of their environments, and can choose actions to influence their environments". In this application, the agent uses the information from the visual and acoustic module as sensing information about the state of the audience. The actions on the audience are taken by deciding a given virtual assistance mode and the goal is to optimize the audience's attention. Thus, when positive answers are detected, the system will activate the positive feedback animation (Fig. 4). However, if a negative answer is detected, the system will activate the animation of "anger face" (Fig. 2 (d)). Moreover, the algorithm will decide which animation activate from negative to positive parameters. Fig. 6 shows a description of the information flow in this human-machine interaction system.

The agent uses a reasoning module to decide on the actions to take in the current state. Reasoning can be implemented with traditional

logic-based symbolism using knowledge representation languages (OWL) or alternatively using machine-learning approaches. Symbolic reasoning applies logic inference to a set of rules and facts (instances in the ontology) deriving higher-level knowledge from the observations, which represent the context of the state. Recent research [43] focuses also on the use of neural networks, capable to project in embeddings the equivalent to the logical reasoning in ontologies. These approaches are interesting, as they overcome some commonplace shortcomings of logic-based reasoning in ontologies, such as sensitivity to noise and missing values.

The ontology to be used in this system must be carefully designed to cover generic descriptions from the behavior of the audience, attitude, level of attention, and questions of the audience about the presentation to model the context. The knowledge regarding emotions of OntoSenticNet is complementary to these representations of entities of other knowledge bases and can be integrated through different APIs in external systems.

Furthermore, the agent should be able to make intelligent decisions. For this, it is necessary to implement a reinforcement learning (RL) system, where the agent learns from its experience of interaction with the environment. In this application, a positive reward is given if the attention of the public increases, while a negative reward is given when the attention decreases.

In RL, an agent observes a state and takes actions, which generates a transaction of the current context to a new state, providing a reward to the agent as feedback of the transaction. The objective of the agent is to learn a strategy that maximizes the reward.

An overview of the literature shows that deep learning models are reported as the state-of-the-art technology for ‘enabling reinforcement learning to scale to problems that were previously intractable’ [44] as these models are capable to process high dimensional input patterns and complex state scenarios. Another interesting work is presented in [45], which gives an overview of recent advances in the integration of natural processing language models in reinforcement applications.

D. Discussion of the Artificial Intelligence System

The modules constituting the artificial intelligence architecture for the proposed human-machine interaction system were described as part of preliminary system design. The implementation of the system requires dedicated work to solve the respective tasks and it is part of the future lines of work.

V. CONCLUSIONS

This work has presented a novel graphic design of a Virtual Assistant with four levels of interaction. An artificial intelligence system has been designed to activate different interaction modes of a Virtual Assistant to improve the communication between the public and the presenter. The proposed system is designed for four levels of attention, such as normal conditions; keep silent; low level of distraction, and a high level of distraction. When one of these levels of attention is recognized by the intelligent agent, a positive or negative interaction of the virtual assistant is induced in order to increase the level of attention of the audience.

Moreover, the presented designs of the assistant rely on non-anthropomorphic forms with “live” characteristics (eye, mouth, and cable-arm). These features help the audience to automatically recognize situations without the need for an explicit explanation of the presenter (e.g., ‘keep silence’). This characteristic makes the system autonomous as the meaning of the interactions is intuitive and easy to understand for humans. An exception is the type of interaction when people are sleeping. In this case, the interaction of the presenter is required.

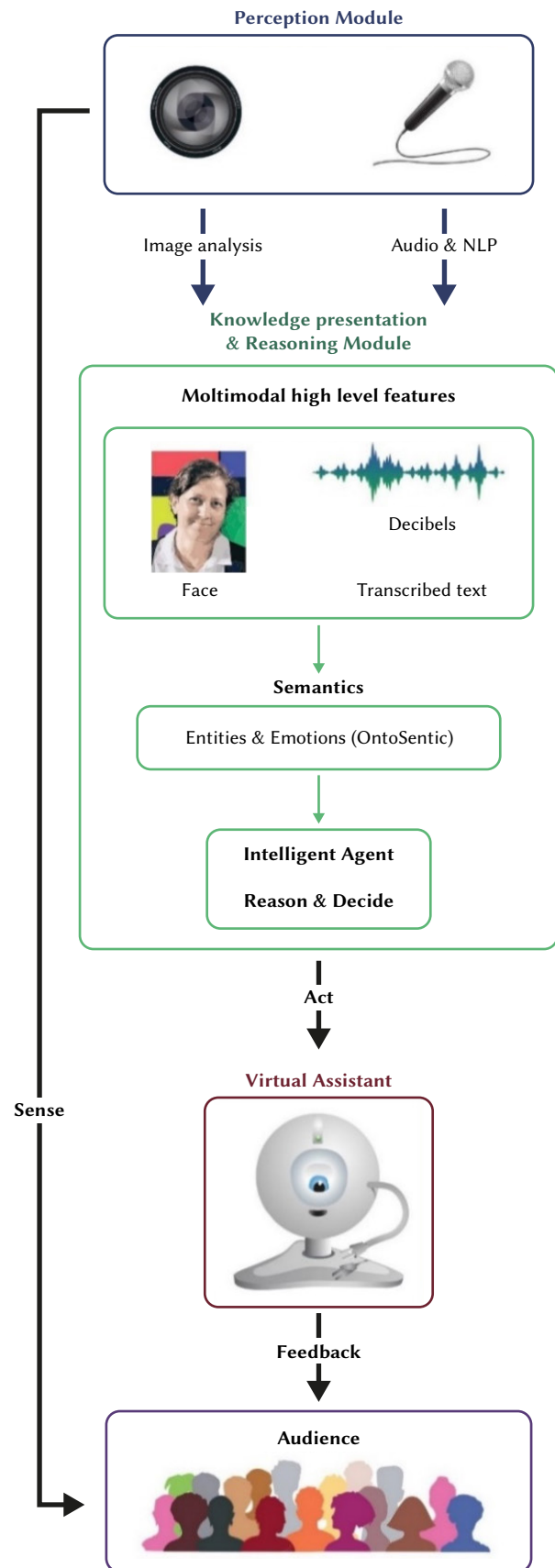


Fig. 6. Description of the human-machine interaction of the system.

Future lines of work will focus on the measurement of the impact of the proposed system on the quality of presentations. The objective is to measure the level of attention of the public comparing conventional ways of presentations and the incorporation of the virtual assistant.

A preliminary study of the intelligent architecture necessary to implement the Virtual Assistant in a human-machine interaction system has shown the need to use several modules. Perceptions of the environment are captured from an audio and visual recognition module extracting high-level features representing behavior or attitudes.

Ontology-based knowledge representation is proposed as a framework for the intelligent agent responsible for the activation of the virtual assistant. Reinforcement learning is recommended for deciding on the best strategy of the virtual assistant to achieve a high level of attention in the audience. The implementation of a proof of concept of the described framework is considered as a future line of work.

ACKNOWLEDGMENT

We would like to acknowledge Gema Fernández-Blanco Martín for her contribution on the psychological analysis of the survey results.

REFERENCES

- [1] B. Alters and C. Nelson, "Perspective: teaching evolution in higher education", *Evolution*, vol. 56, no. 10, p. 1891, 2002, doi: 10.1554/0014-3820(2002)056[1891:pteihe]2.0.co;2.
- [2] P.R. Pintrich, A. Zusho "Student Motivation and Self-Regulated Learning in the College Classroom," in: *Smart J.C., Tierney W.G. (eds) Higher Education: Handbook of Theory and Research. Higher Education: Handbook of Theory and Research, vol 17*. Springer, Dordrecht, 2002, pp. 55-128, doi:10.1007/978-94-010-0245-5_2.
- [3] A. Nijholt, "Towards the Automatic Generation of Virtual Presenter Agents", *Informing Science: The International Journal of an Emerging Transdiscipline*, vol. 9, pp. 097-110, 2006, doi: 10.28945/474.
- [4] R. Looije, M. Neerinx and F. Cnossen, "Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors", *International Journal of Human-Computer Studies*, vol. 68, no. 6, pp. 386-397, 2010, doi: 10.1016/j.ijhcs.2009.08.007.
- [5] C. Bartneck, J. Reichenbach, and A. van Breemen, "In Your Face, Robot! The Influence of a Character's Embodiment on How Users Perceive Its Emotional Expressions", in *Proceedings of the Design and Emotion*, Ankara, Turkey, 2004, pp. 32-51.
- [6] W. Burgard et al., "Experiences with an interactive museum tour-guide robot", *Artificial Intelligence*, vol. 114, no. 1-2, pp. 3-55, 1999, doi: 10.1016/s0004-3702(99)00070-3.
- [7] DeixiLabs, Accessed: Feb. 12, 2020. [Online]. Available: <http://www.deixilabs.com/eliza.html>
- [8] M. Mori, K. MacDorman and N. Kageki, "The Uncanny Valley [From the Field]", *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 98-100, 2012, doi: 10.1109/mra.2012.2192811.
- [9] S. Freud, The Uncanny (1919). Accessed: Jun. 28, 2021. [Online]. Available: <https://web.mit.edu/allanmc/www/freud1.pdf>
- [10] E. Jentsch, "On the psychology of the uncanny (1906)", *Angelaki*, vol. 2, no. 1, pp. 7-16, 1997, doi: 10.1080/09697259708571910.
- [11] A. Chubarov and D. Azarnov, "Modeling Behavior of Virtual Actors: A Limited Turing Test for Social-Emotional Intelligence", in: *Samsonovich A., Klimov V. (eds) Biologically Inspired Cognitive Architectures (BICA) for Young Scientists. BICA 2017. Advances in Intelligent Systems and Computing, vol 636*. Springer, Cham, 2018, doi:10.1007/978-3-319-63940-6_5.
- [12] V. André et al., "Ethorobotics applied to human behaviour: can animated objects influence children's behaviour in cognitive tasks?", *Animal Behaviour*, vol. 96, pp. 69-77, 2014, doi: 10.1016/j.anbehav.2014.07.020.
- [13] S. Kim, B. Schmitt and N. Thalmann, "Eliza in the uncanny valley: anthropomorphizing consumer robots increases their perceived warmth but decreases liking", *Marketing Letters*, vol. 30, no. 1, pp. 1-12, 2019, doi: 10.1007/s11002-019-09485-9.
- [14] K. Sullivan, G. Schumer and K. Alexander, "Ideas for the animated short: finding and building stories" Focal Press, USA, 2008, pp. 64-67.
- [15] K. I. Radoslav, "Televisión, dibujos animados y literatura para niños", *Aisthesis*, 29, pp 33-49, 1996.
- [16] Pixar Animation Studios, Accessed: Mar. 19, 2021. [Online]. Available: <https://www.pixar.com/feature-films/monsters-inc>
- [17] E. Heller, "Psicología del color, Cómo actúan los colores sobre los sentimientos y la razón", Gustavo Gili, Barcelona, 2004.
- [18] J. Guzmán Ramírez, "Una metodología para la creación de personajes desde el diseño de concepto", *Iconofacto*, vol. 12, no. 18, pp. 96-117, 2016, doi: 10.18566/v12n18.a06.
- [19] I. Revina and W. Emmanuel, "A Survey on Human Face Expression Recognition Techniques", *Journal of King Saud University - Computer and Information Sciences*, 2018, doi: 10.1016/j.jksuci.2018.09.002.
- [20] M. E. Holzschlag, "Color para sitios web", McGraw Hill, México, 2002.
- [21] A. Frutiger, "Signos, símbolos, marcas, señales", Gustavo Gili, México, 2007.
- [22] C. Harmon-Jones, B. Bastian and E. Harmon-Jones, "The Discrete Emotions Questionnaire: A New Tool for Measuring State Self-Reported Emotions", *PLOS ONE*, vol. 11, no. 8, p. e0159915, 2016, doi: 10.1371/journal.pone.0159915.
- [23] I. Bhattacharya, M. Foley, N. Zhang, T. Zhang, C. Ku, C. Mine, and R. Radke, (2018). "A multimodal-sensor-enabled room for unobtrusive group meeting analysis, in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 347-355.
- [24] G. Trotta et al., "A neural network-based software to recognise blepharospasm symptoms and to measure eye closure time", *Computers in Biology and Medicine*, vol. 112, p. 103376, 2019, doi: 10.1016/j.compbimed.2019.103376.
- [25] Y. Wang, R. Huang and L. Guo, "Eye gaze pattern analysis for fatigue detection based on GP-BCNN with ESM", *Pattern Recognition Letters*, vol. 123, pp. 61-74, 2019, doi: 10.1016/j.patrec.2019.03.013.
- [26] Y. Kuo, J. Lee and M. Hsieh, "Video-Based Eye Tracking to Detect the Attention Shift", *International Journal of Distance Education Technologies*, vol. 12, no. 4, pp. 66-81, 2014, doi: 10.4018/ijdet.2014100105.
- [27] M. Shakeel and K. Lam, "Deep-feature encoding-based discriminative model for age-invariant face recognition", *Pattern Recognition*, vol. 93, pp. 442-457, 2019, doi: 10.1016/j.patcog.2019.04.028.
- [28] J. Abeßer, "A Review of Deep Learning Based Methods for Acoustic Scene Classification", *Applied Sciences*, vol. 10, no. 6, p. 2020, 2020, doi: 10.3390/app10062020.
- [29] Y. Li, Q. He, S. Kwong, T. Li and J. Yang, "Characteristics-based effective applause detection for meeting speech", *Signal Processing*, vol. 89, no. 8, pp. 1625-1633, 2009, doi: 10.1016/j.sigpro.2009.03.001.
- [30] Y. Belinkov and J. Glass, "Analysis Methods in Neural Language Processing: A Survey", *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 49-72, 2019, doi: 10.1162/tacl_a_00254.
- [31] N. Saleem, M.I. Khattak, "Deep Neural Networks for Speech Enhancement in Complex-Noisy Environments", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 84-90, 2020, doi: 10.9781/ijimai.2019.06.001.
- [32] N. Saleem, M.I. Khattak, E. Verdú, "On Improvement of Speech Intelligibility and Quality: A Survey of Unsupervised Single Channel Speech Enhancement Algorithms", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 2, pp. 78-89, 2020, doi: 10.9781/ijimai.2019.12.001.
- [33] A. Torfi, R. Shirvani, Y. Keneshloo, N. Tavvaf, and E. Fox, (2020). "Natural Language Processing Advancements By Deep Learning: A Survey". *ArXiv*, Vol. abs/2003.01200, n. pag., 2020, available at: <https://www.arxiv-vanity.com/papers/2003.01200/>
- [34] A. Guzman, "Voices in and of the machine: Source orientation toward mobile virtual assistants", *Computers in Human Behavior*, vol. 90, pp. 343-350, 2019, doi: 10.1016/j.chb.2018.08.009.
- [35] M. Gurban, "Multimodal feature extraction and fusion for audio-visual speech recognition". Lausanne, EPFL, 2009, doi: 10.5075/epfl-thesis-4292.
- [36] Q. McNamara, A. De La Vega and T. Yarkoni, "Developing a comprehensive framework for multimodal feature extraction", in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, 2017, pp. 1567-1574.
- [37] X. H. Wang, D. Q. Zhang, T. Gu, and H. K. Pung, (2004, March). "Ontology

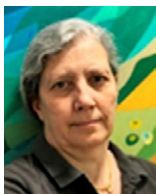
based context modeling and reasoning using OWL”, in *IEEE annual conference on pervasive computing and communications workshops*, Orlando, FL, USA, 2004, pp. 18-22.

- [38] R. Speer, J. Chin, and C. Havasi, “Conceptnet 5.5: An open multilingual graph of general knowledge”, in *Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, 2017, pp. 4444-4451.
- [39] E. Cambria, “Affective Computing and Sentiment Analysis”, *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102-107, 2016, doi: 10.1109/mis.2016.31.
- [40] R. Picard, “Affective computing: challenges”, *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 55-64, 2003. doi: 10.1016/s1071-5819(03)00052-1.
- [41] M. Dragoni, S. Poria and E. Cambria, “OntoSentNet: A Commonsense Ontology for Sentiment Analysis”, *IEEE Intelligent Systems*, vol. 33, no. 3, pp. 77-85, 2018. doi: 10.1109/mis.2018.033001419.
- [42] R. Sutton, F. Bach and A. Barto, “Reinforcement Learning”, Massachusetts: MIT Press Ltd, 2018.
- [43] P. Hohenecker and T. Lukasiewicz, “Ontology Reasoning with Deep Neural Networks”, *Journal of Artificial Intelligence Research*, vol. 68, 2020, doi: 10.1613/jair.1.11661.
- [44] K. Arulkumaran, M. Deisenroth, M. Brundage and A. Bharath, “Deep Reinforcement Learning: A Brief Survey”, *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26-38, 2017. doi: 10.1109/msp.2017.2743240.
- [45] J. Luketina, N. Nardelli, G. Farquhar, J. Foerster, J. Andreas, E. Grefenstette, S. Whiteson, and T. Rocktäschel, “A Survey of Reinforcement Learning Informed by Natural Language”, *ArXiv*, Vol. abs/1906.03926, 2019, available at: <http://arxiv.org/abs/1906.03926>



Salvador Cobos Guzman

He is professor of Robotics, Cyberphysical Systems, and Artificial Intelligence in the Faculty of Engineering at the Universidad Internacional de La Rioja, Spain. He teaches in the master’s degree of Industry 4.0. He received his first B.Sc. in Industrial Robotics from National Polytechnic Institute (I.P.N.), Mexico, in July 2003. The second Engineering degree received was in Automation and Industrial Electronics (A.I.E.) from Polytechnic University of Madrid (U.P.M), Spain, in 2009. Also, he received an Advanced Studies Diploma (A.S.D.) in Robotics and Automation from Polytechnic University of Madrid (U.P.M), Spain, in 2007, and his Ph.D. with honors (“Sobresaliente Cum Laude”) in Robotics and Automation from Polytechnic University of Madrid (U.P.M), Spain, in 2010. The Ph.D. thesis was focused on obtaining virtual human hand models for virtual manipulation tasks. Dr. Cobos worked as a postdoc in the UMI 375-LAFMIA laboratory for one year from 2011 to 2012. During this period, he was working in the design of underwater robots. Also, he was a Senior Research Fellow at The University of Nottingham working in the area of hyper-redundant robots for in-situ inspection from 2012 to 2016.



Silvia Nuere

She is a professor in the Department of Mechanical, Chemical and Industrial Design in the Technical School of Engineering in Industrial Design at the Universidad Politécnica de Madrid, Spain. She teaches Artistic Drawing, Basic Design, Graphic Design and Visual Communication. She received her bachelor’s in fine arts in 1989 and her PhD in 2002 from the Universidad Complutense de Madrid, Spain. Her research interests include, teaching and learning methods based on Project Oriented Learning and in the need of mixing different fields of knowledge as art, design and engineering. She has promoted this approach to education through more than 30 Innovation Education Projects and from 2011 she is the Creator and Director of the scientific journal *ArDIn* (Art, Design and Engineering). She is author and co-author of more than 50 publications about artistic learning methods and humanistic approach to education. She has also, as an artist, took part in more than 20 collective exhibitions and made several illustrations for the Scientific Magazine “Investigación y Ciencia”, the Spanish edition of the Scientific American Magazine.



Laura de Miguel

She has a Bachelor of Fine Arts from UCM in Image Arts and PhD in Fine Arts from the Universidad Complutense de Madrid, Spain. She is a professor in the Higher School of Engineering and Technology, at the Universidad Internacional de la Rioja (UNIR), Spain. With more than 15 years in the university field, she has specialized in being a teacher and author of content in areas of Graphic-artistic

Expression and Design (graphic, industrial, fashion). She teaches subjects such as: creativity, drawing, analysis of the form or projects. She has numerous publications and has participated in artistic and academic outreach activities. Furthermore, in parallel to these activities, she has always kept her facet as a creator through the generation of multidisciplinary workshops (painting, drawing, engraving or short film.) shown in individual and group exhibitions. She has also been curator of exhibition projects, designed and directed art workshops in spaces for cultural dissemination, the web or congresses. Her lines of research focus on creative processes in Art-Design, relationship between Art-Design-Society, educational innovation in creative training, methodological exploration for the teaching of graphic representation, evaluation systems in areas of graphic expression, direction of audiovisual productions as an awareness tool.



Caroline König

She graduated in 2007 in Informatic Engineering at the Faculty of Informatics of the Polytechnic University of Catalonia and worked as software engineer during several years. In 2012 she received a master’s degree in advanced Methods in Artificial Intelligence from the Universidad Nacional a Distancia (UNED). From 2013 - 2018 she was a predoctoral researcher of the PhD program of Artificial Intelligence of the UPC of the Soft Computing (SOCO) research group. In 2018 she received her PhD in Artificial Intelligence from the UPC. Since 2019 she is teacher of the Computer Science Department at Universitat Politècnica de Catalunya, UPC and postdoctoral researcher of the ‘ML-PROMOLDYN’ project (2020). She is involved in the development of artificial intelligence applications in the field of industry, bioinformatic and robotics. Her research interests are on deep learning approaches for automatic feature extraction from multimodal sequential data, anomaly detection and explainability of machine learning models.