2015 15th International Conference on Control, Automation and Systems (ICCAS 2015)
Oct. 13-16, 2015 in BEXCO, Busan, Korea

A Psychological Framework to Objectively Evaluate Socially Capable Robots for Interactive Tutoring Systems

Barnabas Takacs[1,*],Klara Csizinszky[1],Daniele Mazzei[2],Lajos Simon[1]

[1] Department of Psychiatry, Semmelweis Medical University,
Budapest, 1083, Hungary (btakacs@digitalCustom.com) * Corresponding author
[2] Research Center "E. Piaggio" , University of Pisa,
Largo Lucio Lazzarino 1, 56122 Pisa, Italy (daniele.mazzei@centropiaggio.unipi.it)

Abstract: We introduce a novel evaluation methodology to establish psychometrically validated measures to objectively evaluate socially capable robots. Our methodology involves first creating a digital computer generated face model designed to replicate the facial expression space of the robot with the highest accuracy, and second, using this model to render test sequences, which are in turn analyzed with independent facial metrics software. We compared three different facial modelings techniques to approximate our robot's face and achieved 98.86% accuracy in replicating the facial appearance as measured by facial metric software. This digital face model can now be used to create arbitrary expressions for interaction purposes and for a more detailed analysis of low-amplitude micro-expressions, which are critical for proper social communication with children in a virtual tutoring context. We argue that our methodology is a first step towards objectively assessing the believability of commercially available robots.

Keywords: Humanoid-robots, Psychological assessment, Virtual Tutors, FACE robot, Facial animation

1. INTRODUCTION

Humanoid robots have been gaining increased attention as a tool to interact in a personalized manner with children in an educational context [1,2]. With more-and-more realistic facial movements and capabilities for emotional display they represent new challenges and possibilities. Emotions, interaction and particularly attention, indeed play a vital role in the process of turning information into knowledge. Therefore, the goal of a social robot-based tutoring system is not only to present students with the material they need to learn, but to constantly and interactively track and enhance their performance and engagement in response to the multi-modal stimuli they are presented with.

In this paper we argue that the same paradigm holds true for the robot as well and that arbitrary virtual facial expressions may mislead the learner and reach unwanted effects while undermining the very outcome of the tutoring process. Specifically, we must carefully evaluate the psychological meaning and validity of robotic facial displays before deploying them in real world context. To address these issues, we reach for methodologies in a parallel field of research, namely virtual humans [3,4] and Animated Conversational Agents [5] to compare physical robots with their digital representations in a classical psychological framework, where videos (recorded or rendered) are assessed for their emotional validity seamlessly bridge the gap between virtual and real.

The advantage of using a digital representation to evaluate performance lies in the added benefit of dynamically changing display and viewing parameters as well as the range and amplitude of expressions to pinpoint fine transitions, timing in low-intensity expressions. Furthermore, studies can be extended to large populations in laboratory conditions without actual access to the physical robot and therefore prior to in-vivo and in-situ deployment.

We present early results from a comparison study that evaluates the performance of facial robots by comparing them with their digital counterparts. The remaining of this paper is organized as follows. In Section 2 we briefly introduce the Affective States of Learning as a prime mechanism behind learning dynamics. In Section 3 we describe the key factors influencing believability of a virtual agent (physical robot or digital model). Section 4 introduces our methodology for a comparative psychological study, and in Section 5 results are presented followed by our Conclusion in Section 6.

2. AFFECTIVE STATES OF LEARNING

Affective states in learning (like interest, boredom, excitement, confusion, fatigue, etc.) are accompanied by different patterns of facial expression, eye-gaze, head nod, hand movement, gestures and body posture. Interpreting these non-verbal signals in the context of face-to-face communication with a robotic agents is to spot students' performance difficulties and detect possible motivational problems in order to adapt their behavior to overcome frustration and the fear of making mistakes. The classical learning paradigm is based on what is often called *"learning by failure"*. Students' emotional states interweave with the *cognitive dynamics of the learning process* as shown in Figure 1. The emotional state is represented on the horizontal axis, while the vertical axis depicts the students' relationship to learning in a given moment. The students' state-of-mind moves continuously in this space in a spiraling motion progressing in time. Frequently, students begin in Quadrant I being curious and fascinated regarding a

new topic of interest (Quadrant I) or they might be puzzled and motivated to reduce confusion (Quadrant II). In both cases the top half of the space focuses on constructing or testing knowledge. As learning proceeds movements in this space occur. As an example, a student may get an idea how to implement a solution or build something but when the result fails the need to revise the original idea and analyze mistakes commonly moves them to Quadrant III where emotions may be negative and the initial focus on learning changes to eliminating some misconceptions. As time goes by the student consolidates his/her knowledge and gradually moves to Quadrant IV from where a fresh idea propels them back to the upper Quadrant I again. In the context of robotic tutors this closed loop dialogue model of student-robot interaction, combines the power of high fidelity facial animation and body language delivered by the robot with advanced perceptual techniques to understand student's behaviour and reactions to create a truly bi-directional interface, where the robot is perceived at a partner level. However, for this to happen in real life, a number of factors influencing believability must be considered first. These are discussed in the following section.

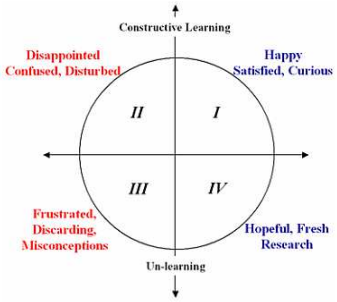


Fig. 1 Relationship of emotions to the learning process [6].

3. FACTORS OF BELIEVABILITY

To create a believable social robotic agent a number of psychological phenomena must be first addressed. A well known example is the 'uncanny valley' effect whereby a near human-looking artifact can trigger negative feelings. While this has been addressed in a mathematical framework [7] there is emerging evidence that robots that "try too hard" may achieve a contrary effect [8]. This phenomena has its roots in perceptual cues, their coherency and visual quality as well as the capability for consequent categorization. As such, it has direct implication how these robots should behave and express their feelings via body language, non-verbal cues and facial displays of affect. The following chain of logic, i.e. how *interaction*, *behavior*, *perception*, *closed-loop dialogue*, *body-language*, *believability* and *visual fidelity* are building upon each other, are explained in more detail next. It is based on our past research and development and production experiences in the field of virtual humans.

Interaction: Input from the student prompts the virtual humanoid robot to generate actions. These actions may be partly implemented as scripted reactions, but for full credibility and believability they need to be generated on-the-fly requiring a form of autonomous *behavior* ⇒.

Behavior: Artificial intelligence and artificial emotion techniques are necessary to create the reactions of a virtual humanoid to its changing dynamic environment. These include time-varying input and reactions from the user, pedagogical context and goals. To simulate believable reactiveness the humanoid robot is governed by the "stimuli" it receives from its environment in the form of digital *perception* ⇒.

Perception: Much like natural humans, humanoid robots and their virtual counterparts possess multiple sensory channels to perceive their environment. These channels primarily include vision (web camera), hearing (microphone) and tactile input (collisions, external forces, touch screen). Based on these modalities the virtual humanoid is capable of recognizing emotions and gauging the attentive state or mood of the student and generate its own responses accordingly. Eye-contact and proper facial expressions that are consistent with the presented content are of critical importance in the emergent role-taking. Thus, interaction leads to what is often called a *closed-loop-dialogue* ⇒.

Closed-Loop-Dialogue: Having a dialogue with a robotic humanoid assumes it has a personality designed specifically to meet the needs of the application in hand. Personality is also reflected in the general speed of reaction and the precision of associated motions − as well as moods and emotions − It is primarily communicated towards the user via non-verbal signs, also called the robot's *body language* ⇒.

Body-Language: Body language speaks volumes regarding the internal state, personality and intent of the socially capable robot. It comprises of subtle events requiring high fidelity and frequently subliminal motion data that is difficult to animate or even capture. In addition, culture specific aspects including conventions for hand signals and socially governed behaviors also need to be layered seamlessly on top of functional motion such as the actions of the body and/or speech taking place in parallel and therefore dividing students' attention. The timing and quality of these layered and integrated motions − most frequently implemented via Behavior Markup Language (BML) [9] − ultimately affect the virtual humanoid's *believability* ⇒.

Believability: Life-like virtual humanoids hold the promise of blurring the boundary between real and virtual by being able to create interactive experiences similar to those between two people. For this the notion of presence and believability are of critical importance. The believability of a robotic humanoid (that would pass as another person, similar to a "visual Turing Test") largely depends on the variability and consistency of its responses, the speed of feedback and reactions it provides. In short, despite the uncanny valley phenomena, it ultimately depends on its high quality and fine details collectively termed as *visual fidelity* ⇒.

Visual Fidelity: The new generation of life-like humanoid robots are built with photo-realistic skin, fur, hair. Secondary surface deformations such as micro-expressions, involuntary muscle twitches and low-level reflexes can be readily programmed. In the near future, these can be combined with tertiary texture effects for blushing, goose bumps or glinting eyes showing interest or boredom. Still, even at this level of fidelity maybe perfect in a static or stationary state, but quickly falls apart when motion is present. Therefore facial expressions and their meaning must be researched and evaluated as a dynamic system using *psychological principles⇒*.

Psychological Evaluation: The standard method of understanding and analyzing facial expressions is to characterize emotive facial expressions using the Facial Action Coding System (FACS) [10,11], where all facial muscles are grouped into 58 so called Action Units (AU) and subsequently used to encode all possible movements. Emotions represent a sub-space of valid facial expressions and can be compared via statistical analysis. If a given facial display is recognized as one expression with high statistical power, it is accepted as a believable and credible visual representation of that emotion. Based on the above chain of logic, we set up an evaluation framework, where a robot's emotional facial expression repertoire is studied as if it was a real human and subjected to standard metrics as discussed in the following section.

4. METHODOLOGY

To objectively evaluate a humanoid-robot's facial expression set and its emotional expressiveness from a psychological perspective, we compared renderings of 3D virtual model representations of the robot head with video footage obtained directly from the robot itself with the help of advanced *facial analytics tools* [11] dedicated for the analysis of facial emotions. Figure 2 demonstrates the concept of how digitally animated and rendered virtual humans (above) can be used as a tool to compare their performance with humanoid facial robots. Our basic experimental set up consisted of the following

- STEP#1: Create high fidelity digital face models of a humanoid face robot using multiple modeling and rendering techniques.
- STEP#2: Develop a unified control architecture to match the facial motion of the physical and virtual representations and to create animations
- STEP#3: Compare the quality and output of the digitally created models and rank them according to their visual performance using automated facial analytics tools.
- STEP#4: Select the best digital representation model and compare its performance with video footage from the physical robot.

Fig. 2 The technology of high fidelity animated digital faces (Above) is used to evaluate an emotionally expressive physical robot (Below).

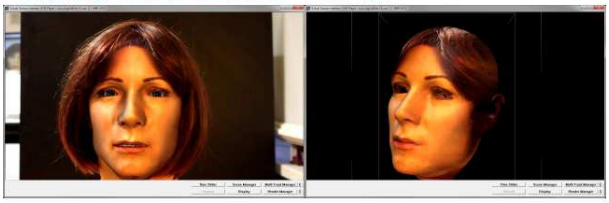
STEP1: In our experiments we used a robot called FACE [12] and created 3D representations of its head shape from a series of photographs selected from recorded reference videos as shown in Figure 3. Next, we developed three alternative digital representations aiming to match the quality of the original robot. These are demonstrated in Figure 4. The first one (*Model1* - above) was hand modeled by an artist and created face shapes as morph targets for a real-time rendering in Unity3D engine. *Model 2* (center) was created via parametric head shape estimation in the Virtual Human Interface (VHI) system [13], which compares the input photograph to a data set of digitized head shapes to estimate the visual appearance of a person. Finally, *Model 3* (bottom) is an Image-based rendering technique, also in the VHI [13], that creates a circular 2D morph space to animate continuously changing facial expressions derived from a set of photographs. This space is used to generate dense samples that maybe rendered with low computational overhead using Flash or Java for mobile platform (IOS and/or Android).

STEP2: The control architecture to animate both the robot and the digital virtual models is provided through a psychologically validated unified common control framework, called Temporal Disc Controllers (TDC). The TDC is based on the well known circumplex model of affect [14] essentially defining a unit circle over a Euclidean Cartesian Space (ECS) that represents a two-dimensional model of the affective experiences. As a result, each emotive expression is a linear combination of two independent variables (valence, arousal) and corresponds to a point in the valence-arousal plane. This method has been chosen since people tend to recognize emotions as ambiguous and overlapping experiences rather than discrete entities [15]. Indeed the *ECS* provides a consistent way to synthesize novel facial expressions that are intermediate/interposed/mixed among the universal ones [12] .

Fig. 3 Reference photos used to create 3D models of the FACE robot [12]. The two versions had different hair style and make up to simulate variations in appearance.

MODEL1 - Hand modeled by an artist and rendered in Unity3D

MODEL2 - 3D face shape generated and rendered in the VHI



MODEL3 - Image-based morph model (R50) created in the VHI and rendered in Flash

Fig. 4 Three digital representations of the FACE robot [12] used in our experiments (see text).

STEPS 3&4: To establish credible psychometrically validated measures and compare our different display models of facial affect we created a set of videos showing timed emotions and their transitions and used well established *facial analytics tools* [11] to provide comparative metrics. The facial analytics system analyses, frame by frame, the input video, where it first detects faces, next builds active appearance models (AAM) to synthesise an artificial face model, which describes the location of hundreds of key points in the face and the facial texture of the area entangled by these points. Next, for expression classification, this model is compared to a large database of manually annotated images using statistical and machine learning methods in order to calculate the main sources of variation found in the images to classify the neutral plus six basic or universal emotions [10] as well as a number of FACS codes. The system is capable of detecting subtle changes in the emotional expressions (micro-expressions) of the face with high fidelity across various populations and age ranges. Beyond this psychometrically validated approach it can can further be used to detect facial states the test participant's global gaze direction to measure attention, track the head orientation or indicate the person's gender, age, ethnicity.

It is important to note that no facial analysis software provides 100% correct output. The selected platform guarantees 95% accuracy, which is currently the best in the World. Yet, the methodology we introduced here offers an objective and repeatable way to measure facial affect in social robot and establish base-line performances through experimental results. This is the subject of the following section.

5. EXPERIMENTAL RESULTS

Using the evaluation framework described in the preceding section we first evaluated the performance of our three digital face models to select the best performing one (*Experiment#1*), and next we compared its performance with original video footage from the FACE robot (*Experiment#2*).

Experiment#1: Figure 5 shows the outcome for Model1. The continuous lines indicate the recognized expressions (above) as output. The marks indicated the accuracy of the recognition **+** for correct result, ***** for misses and **o** indicating mismatches. The semitransparent bars indicate the relative probabilities allowing us to pinpoint when expressions were recognized, but fell below detection threshold. The Ground Truth data (below) with the correct expressions indicated with letter codes for better visibility. As demonstrated in Figure 5, Model1 does rather poorly. This is due to the fact, that hand modeled facial expression morph targets represent a linear expression space and do not always match psychologically valid facial expressions, as well as the fact that the real-time render quality and the architecture of Unity3D is not suited for creating interactive photo-realistic humans.

Figure 6 shows the results using the Parametric 3D model of the VHI. As we can see that performance is slightly improved with 4 emotions detected correctly but only 2 recognized with high confidence.

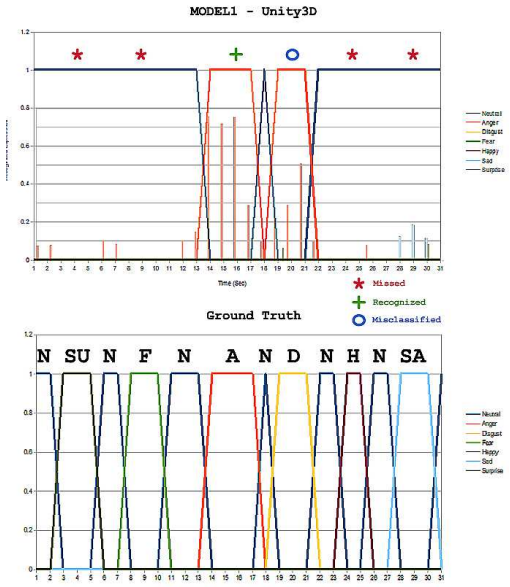


Fig. 5 Recognition results for MODEL1 (see text).

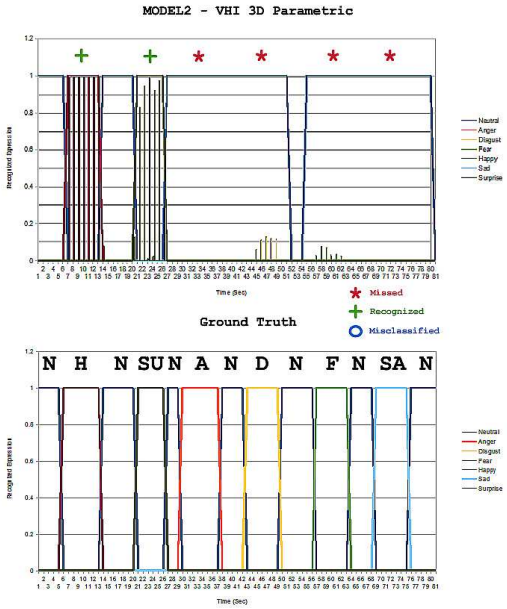


Fig. 6 Recognition results for MODEL2 (see text).

This is due to the fact, that the 3D model was estimated from a single photograph, which is the first step of the 3D modeling process. In fact, in the VHI capture protocol we normally record up to 60 different scanned facial expressions, which allows the system create a complete representation of the face space. Our future work includes plans to scan the robot head and deliver a more detailed 3D model for evaluation. For now, we turn our attention to our Image-based 2D Morph model , also created in the VHI. Figure 7 summarizes these results for Model3. 4 emotions are successfully detected and 1 is indicated but with low confidence, while only one expression is missed. In fact, as we will see in experiment #2, this misses is due to the fact that the robot's facial expression (Sad) itself is incorrect.

Experiment#2 In the following step we analyzed the source video (H2015) showing the FACE robot in a frontal position exhibiting the same series and timing of emotive expressions as we have already demonstrated in Fig 7 for Model3. As it can be seen in Figure 8, the outcomes are qualitatively identical, with only minor differences in probabilities, thus indicating that the facial expressions of the robot are not perfect. Finally, to qualitatively assess the recognition accuracy achieved with Model3, we plotted the difference of recognition results from the H2015 video and those with Model3.

The 7 expressions over 50 seconds provided 350 samples to compare, of which in only 4 cases the recognition results were incorrect, thus achieving 98.86% accuracy.

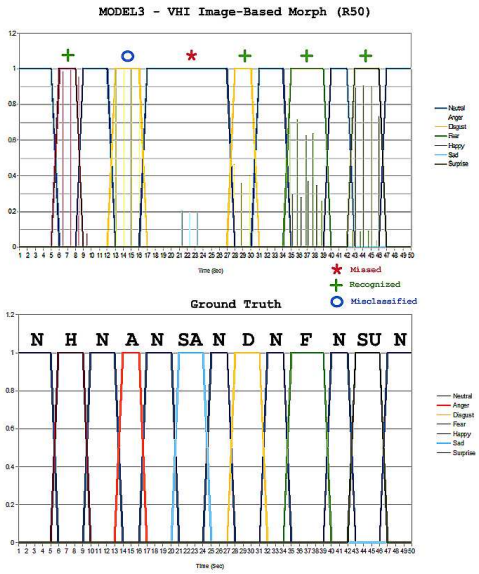


Fig. 7 Recognition results for MODEL3 (see text).

This underlines our key point, that digital virtual faces in combination with automated psychological facial affect analysis can be effectively used to evaluate the emotive performance and believability of future generation robots that interact with people and children. As a final step, since the robot's "SAD" expression was misclassified in each of our experiments, we decided to have a closer look at the phenomena. Figure 10 shows the original frame from the H2015 video (left), and the output of the digital Model3 (right). The two expressions appear identical to human observers, but from a static image it is ambivalent to assess the true emotional value. The entire comparison video is available from here [16].

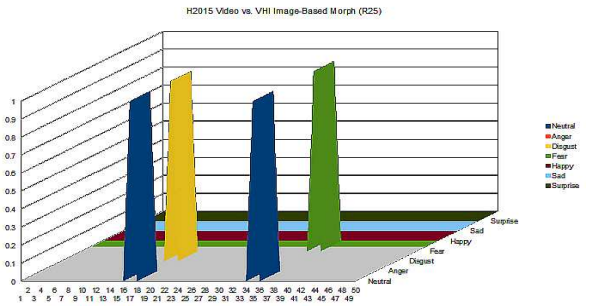


Fig. 8 Recognition results on a video showing the FACE robot are identical to the results achieved with the Digital Model 3.

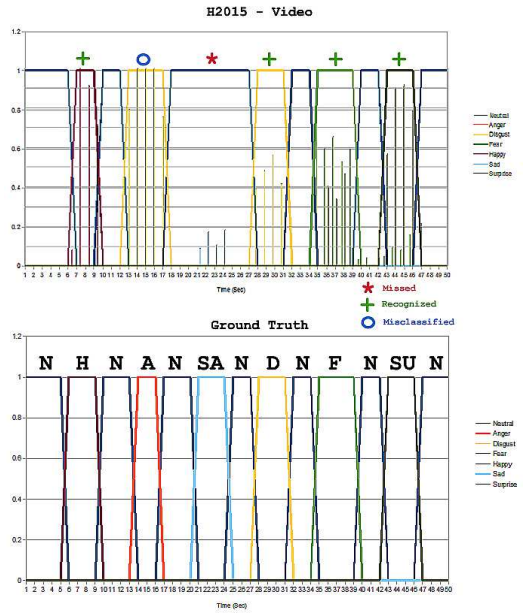


Fig. 9 Comparing recognition results on H2015 FACE robot video with those obtained on digital Model3 missing only 4 samples from 350.

Fig. 10 "SAD" expression comparison. The original frame from the H2015 video is shown on the left, the digital Model3 is shown on the right (see [16]).

6. CONCLUSION

In this paper we introduced a novel evaluation methodology to establish psychometrically validated measures to objectively evaluate socially capable robots. Our methodology involves first creating a digital computer generated face model designed to replicate the facial expression space of the robot with the highest accuracy, and second, using this model to render test sequences which are in turn analyzed with independent facial metrics software. We have successfully developed three different facial modeling techniques to approximate our robot's face and achieved 98.86% accuracy in replicating the results for reference video. This digital face model of the robot can now be used to create arbitrary expression spaces for interaction purposes or for a more detailed analysis of low-amplitude expressions, which are critical in the social communication tasks the virtual tutoring robot is supposed to achieve. As these robots and their counterparts, such as the *Zeno R25* and *R50* or the *Nao* are more and more frequently used in the context of education and autism research, we argue that their facial expressions should be properly and extensively validated with tool sets available to the broad scientific community. Our current results represent a first step in this direction.

REFERENCES

[1] D. Leyzberg, S. Spaulding, and B. Scassellati. "Personalizing Robot Tutors to Individual Learning Differences," *In Proc. HRI'14*, pp 423-430, 2014

[2] Robokind Robots for Autism, accessed July 20, 2015 http://www.robokindrobots.com/ , 2015.

[3] B. Takacs, "Special Education & Rehabilitation: Teaching and Healing with Interactive Graphics", IEEE Computer Graphics and Applications,Special Issue on Computer Graphics in Education, 2005.

[4] B. Takacs, "BabyTeach: Using Ambient Facial Interfaces for Interactive Education", in ERCIM News, Technology-Enhanced Learning, 2007.

[5] J. Hyde, Jet al. "Conversing with children: cartoon and video people elicit similar conversational behaviors." *in Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2014.

[6] B. Kort, R. Reilly, "Analytical Models of Emotions, Learning and Relationships: Towards and Affect-Sensitive Cognitive Machine",in *Proc. Intelligent Tutoring Systems Conf.*, pp. 955-962, 2002.

[7] R.K. Moore, "A Bayesian explanation of the 'Uncanny Valley' Effect and Related Psychological Phenomena", *Nature Scientific Reports,* 2(864), doi:10.1038/srep00864.

[8] J. Kennedy, P. Baxter, and T. Belpaeme, "The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning", *In Proc 10th ACM/IEEE International Conference on Human-Robot Interaction, At Portland, Oregon, USA.* 2015

[10] Diprose, J. P., et al. "A human-centric API for programming socially interactive robots.", *Visual Languages and Human-Centric Computing (VL/HCC), 2014 IEEE Symposium on*. IEEE, 2014.

[11] Ekman, Paul, Wallace V. Freisen, and Sonia Ancoli. "Facial signs of emotional experience." *Journal of personality and social psychology* 39.6 (1980): 1125.

[12] EMOTIENT cloud-based facial analytics, accessed July 20, 2015 http://emotient.com/ . 2015.

[13] D. Mazzei, et al, "A Hybrid Engine for Facial Expressions Synthesis to Control Human-like Androids and Avatars", Proceedings of the IEEE RAS and EMBS International Conference on Biomedical Robotics and Biomechatronics, art. no. 6290687, pp. 195-200, 2012.

[14] B Takacs, B. Kiss, "The Virtual Human Interface: A Photo-realistic Digital Human", *IEEE Computer Graphics and Applications,Special Issue on Computer Graphics in Education,* Sept/Oct, 2003.

[15] J. A. Russell, "The Circumplex Model of Affect" *J. of Personality and Social Psychology*, 39:1161–1178, 1980.

[16] **VIDEO** , accessed July 20, 2015 http://www.digitalElite.US.com/Download/ICCA2015_BarnabasTakacs_H2015_M3R50_Comparison.zip

ACKNOWLEDGEMENT

This project has received funding from the European Union Seventh Framework Programme (FP7-ICT-2013-10) as part of EASEL under grant agreement n° 611971.