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

Towards a Unified Control Framework for Humanoid Robots and their Virtual Avatars in Physical and Virtual Reality-based Interactions

Barnabas Takacs[1*],Gergely Richter[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 unified modular control architecture and virtual-reality evaluation framework to extend the use of socially capable humanoid robots. Our methodology is based on a unique gesture and facial expression representation module, called cascaded Temporal Disc Controllers (TDCs), that represents all emotional expressions, gestures and time-variant actions in a normalized mathematical space designed to accept high level control commands, while being independent of actual physical robot implementations. At the same time it also provides an underlying mechanism for avoiding repetitive behaviors and increases the "humanness" by minor random perturbations during the interaction process. Our implementation combines two available physical robots (Zeno, FACE) and their virtual representations with active perception in order to drive low- and high level reactive behaviors in support of interactive educational and pedagogical goals. Our photo-realistic representations were used in comparative evaluation studies and a low-cost Augmented-Reality interface was also developed to support seamless interaction in virtual space.

Keywords: Socially capable humanoid-robots, Virtual Reality, Augmented Reality, FACE robot, ZENO robot.

1. INTRODUCTION

Virtual reality is making a comeback [1]. With the advent of low-cost and portable VR experiences running on mobile-phones and tablet devices as primary interfaces, immersion and human-centric interaction within these environments is gaining ever increasing attention [2]. While for decades, *virtual humans and avatars* have played a crucial role in creating believable experiences in the context of education, training and entertainment [3,4,5], more recently *humanoid social robots* have become widely available as a tool to interact with children in an educational context [6,7]. Therefore we argue, that a unified framework to seamlessly combine these two worlds that employ expressive agents (virtual or physical) for symbiotic education and learning is a crucial element for future applications to come [8].

In this paper we describe a unified modular control architecture that combines two types of physical robots (FACE [10,11] and Zeno [6]), and their virtual representations with active perception in order to drive low- and high level reactive behaviors for supporting educational and pedagogical goals. The model involves a "black box" architecture, where independently operating elements are glued together via a YARP [9] communication interface and provide input to a central module, called *Temporal Disc Controllers* (TDC), for mixing high level robot behaviors communicated towards the physical and virtual manifestations simultaneously. This integration strategy allows for the easy extension of the proposed core control architecture to various future robotics platforms without modifications of the core infrastructure.

The remaining of this paper is organized as follows. In Section 2 we briefly introduce our Methodology followed by a brief section on Implementation in Section 3 with a special emphasis on Virtual- and Augmented Reality in Section 4 followed by our Conclusion in Section 5.

2. METHODOLOGY

The main interface between the heart of our robotic tutoring system the End-User are the selected socially capable robots (FACE [10,11] and Zeno [6]) and their virtual models. Our efforts therefore focused on the development and integration of these physical as well as synthetic agents, their behaviors and presentation strategies. Specifically, these strategies are driven by a closed iterative perception and reaction loop designed to project the sense of a humanness associating emotionally engaging behavior, non-verbal cues and subliminal facial feedback. This in turn creates a positive emotional signal that helps turning the information being presented into deep knowledge. This unified framework includes a real-time reactive animation system that builds on the power of animation to form a hypothesis of the learner's behavior and test that hypothesis by actively acting to verify it, e.g. by measuring the attention level of a student and changing the content being presented adaptively [4,8].

The overall system architecture is shown in Figure 1. "Blackbox" modules operation on their respective provide input/output streams to provide services, such as scene analysis, behavior planning, physical robot control or virtual renderings, by subscribing to a central YARP server (shown in blue with arrows indicating data bottles in the figure). More specifically, End-User interactions are perceived via visual and auditory channels in the *Scene Analyzer* module, which employs a Kinect2 device to extract up to 6 subjects and their body motion plus gestures from the scene and trigger facial feature analysis to derive features such as gender, age, emotional state, speaking probability, sound angles, etc.

This information is relayed to the *Behavior Control* module, used to generate multi-modal behavior for continuous interactions between a human and a virtual or physical interlocutor [8]. In other words, this module is aware of the teaching context, interaction history and pedagogical goals and issues high level control commands, such as "wave", "gesture yes" or "greet" to the central controller unit (TDCs), which is the link between the abstract representation and physical realizations in the form of robots and their virtual models.

Temporal Disc Controllers are responsible for mixing the behavior commands into low-level motor controls to implement time-dependent gestures and actions-reactions as required by the interaction process. Their primary contribution is to represent these complex behaviors over a limited set of normalized unit circles, which allow for continuous animation and avoiding repetitious behaviors. As such, they replace the previously used techniques of triggering 'canned' animations the number of which can grow exponentially if a broad repertoire of reactions is required. TDCs also have a close tie to psychologically validated emotional representations and as such they are ideal for socially capable robots. A single TDC is a circular normalized interface, where robot control parameters are grouped and mapped onto the circumference of a unit circle as fixed points (called expressions and poses), with the center of the circle representing a neutral position or the output of another TDC. A coordinate sequence in this 2D normalized space [0,1] encodes a transition determined by the center and two closest fixed points forming a sector. Behaviors are triggered as transitions of these coordinates from the current state of the system to a target state over a number of steps (n) in a given time (t1) and subsequently returning during in the span of another time frame (t2). Within a sector this transition path is generated randomly resulting in a slightly different gesture sequence for the same high level command, thereby making reactions more "human-like". Multiple TDCs are cascaded to create a complex behavior animation model involving emotional reactions, body poses and gestures. As a result, this normalized animation space represents robot-actions independent of their physical manifestation and thus separates the abstract action representations from the physical capabilities of the robots.

The final elements of our architecture involve the actual physical robots and their virtual counterparts. These modules receive the same YARP data streams and execute the given instructions taking into account their specific capabilities. This integration strategy forms the foundation of easily extending of the proposed core control architecture to various robotics platforms with each future robot models to develop their own behavior prototype without modifications to the central architecture or loosing its general nature.

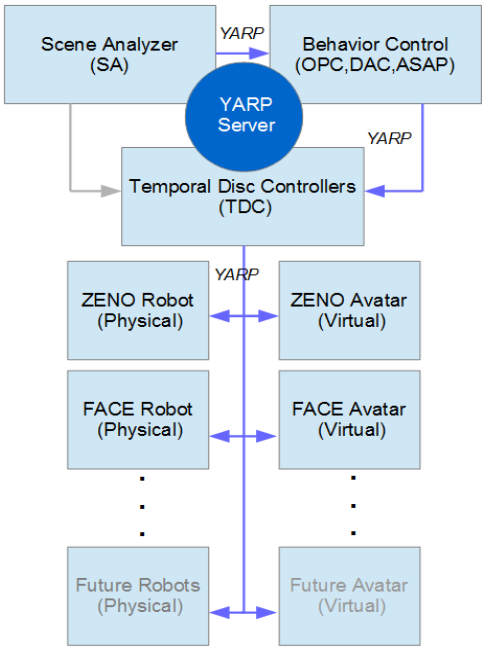

Fig. 1 Overall system architecture (see text).

3. IMPLEMENTATION

The architecture described in the preceding section was implemented to demonstrate the easily extendable nature and powerful operation of the TDC control methodology. For the Scene Analyzer we used Visual Studio 2012 and the .NET 4.5 framework in combination with the Kinect's own SDK, The central TDC and the virtual avatar modules were developed in the Unity3D engine in combination with HTML5 and various high level languages, such as Java and C# were used to control Zeno and FACE physical robots, respectively.

Our two prototype implementations involved a highly detailed facial robot head and upper body capable of expressing fine shades of humanoid emotions (FACE) and a lower fidelity, but fully embodied ideally suited and frequently used for pedagogical tasks (ZENO). Using YARP as a "glue" to tie architectural elements together also allowed for the dislocation and distribution of system components over different locations. As an example, the physical robots and the Scene Analyzer operating in close proximity of the students such as a museum, school setup or university lab could be connected to the central YARP server hosted at another location with additional third parties participating via the virtual robot models from yet another set of locations. To create the virtual robot models we have used two methodologies, the first one being manual 3D modeling from reference photographs, which resulted in two fully animatable 3D models and second, image-based approach which resulted in virtually indistinguishable photo-realistic output used for frontal interaction.

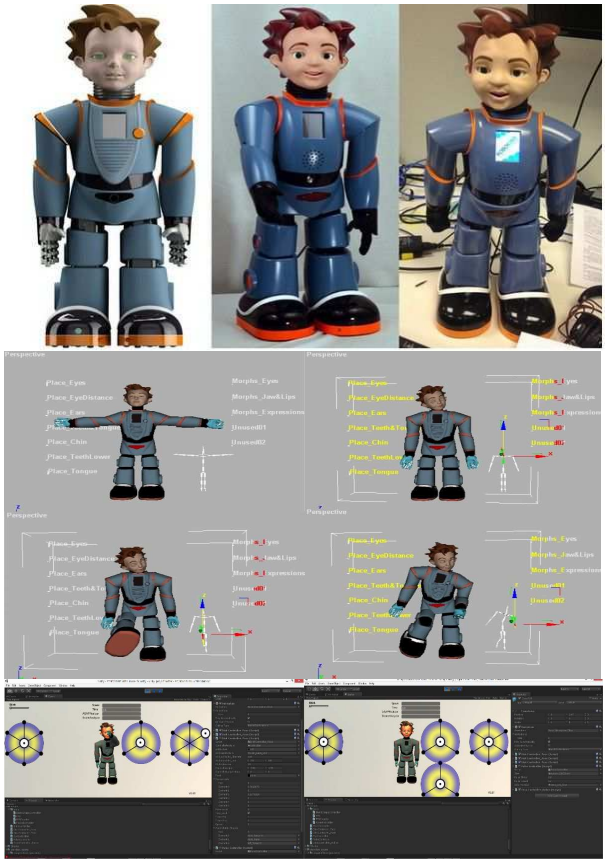

Fig. 2 Virtual Zeno robot with multiple disc controllers attached (see text).

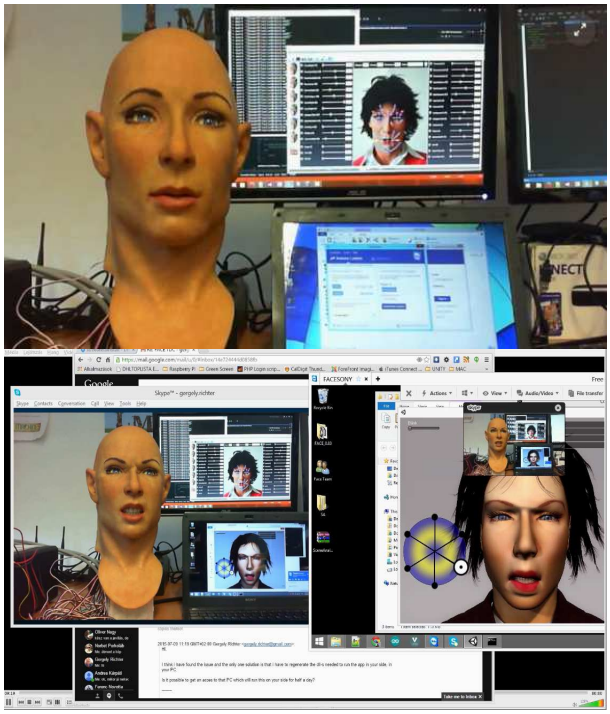

Fig. 3 Remotely controlled high fidelity facial robot (FACE) and its avatar showcasing the full distributed operation of the architecture.

Figure 2 demonstrates our virtual ZENO robot, which is one of the primary platforms we use to interact with children (above: the original photos, center: virtual interactive model, below: Unity3D interface with multiple TDCs attached for emotional facial expression and body pose control). Our second robot (FACE), is shown in Figure 3 while being controlled by the central TDC module and its virtual avatar reacting in the same manner, as described above. Finally, Figure 4 shows an example of our photo-realistic facial avatar representation as an on-line demo used in our first evaluation experiments to compare the efficiency of our virtual models with the real robots using advanced facial analysis software [12,16] as well as in a large public-event study to see whether people using this virtual face to express their mood and feelings could detect the fact that this is not a real human [14].

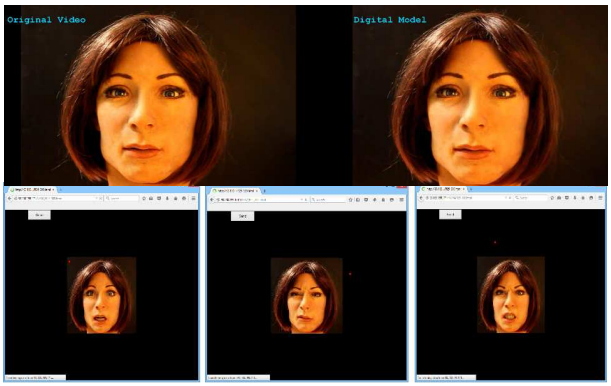

Fig. 4 On-line demonstration of a photo-realistic virtual avatar of the FACE robot for emotional interaction purposes (See **on-line demo** [14] and **comparison Video**: [17] *Left* original video, *Right* digital virtual face model).

Having created high fidelity virtual models of our socially capable robots, that move and act as their physical counterparts, the final step in our unified control architecture is to create a user interface that extends the natural world by augmenting the users' surroundings and therefore allows them to freely interact with these robots just as if they were present in the room. This is discussed in the following section.

4. VIRTUAL- AND AUGMENTED REALITY INTERFACES FOR ROBOT INTERACTION

The primary purpose of our work is to create a unified control architecture that can uniformly control physical and virtual robots in a social context whereas users can seamlessly interact with the virtual robots as if they were physically existing. The tools to achieve these goals involve virtual reality and augmented reality. Augmented and Mixed Reality (AR/MR) is a technique that seamlessly overlays computer generated digital imagery (such as 3D objects, animated models, information tags, video and new-media) on live-video

captured by a camera. While the user moves through the environment these visual elements are continually updated to appear as if they were part of the same scene. This is achieved by *tracking* the camera's motion in six degrees-of-freedom (6DOF = 3 translational axis and 3 rotational yaw,pitch,roll parameters) to compute the intrinsic parameters needed to calculate how each point of a virtual object is mapped, and overlaid onto the screen. AR had been in the research arena for several decades focusing on experiences via *see-through AR*, where the camera is attached to the user's head and a tiny display in front of their eyes displays the combined images. The accuracy of position and orientation estimation remained critical and, to a large degree, determined what type of practical applications the technology could solve. To ease the computational problem of camera position and orientation estimation, *printed markers* of known size and patterns were introduced, which in recent years were extended to allow the use of *natural markers,* which are essentially photographs with high density good features in them. This is a key enabler for our AR/VR solution.

Our prototype implementation combines an Android-based mobile phone (Samsung Galaxy S6), running the Zeno virtual avatar implemented in Unity3D and integrated with *Vuforia* [15] to detect landmark points and natural markers in the scene. This detection allows the application to seamlessly place virtual Zeno in the scene as if it was there. Intricate patterns, with lots of texture details, such as a rug or painting, may be used as good references in this context. Because Yarp is not compatible with Android an additional data path was implemented through witch commands are relayed to the mobile device. An additional wireless controller to extend the 3D navigation and interaction capabilities was also added. Figure 5 shows the TDC module controlling the Zeno robot in AR mode. Figure 6 shows a user interacting with the model wearing a low-cost, branded paper-based Virtual Reality headset that holds the mobile phone in place.

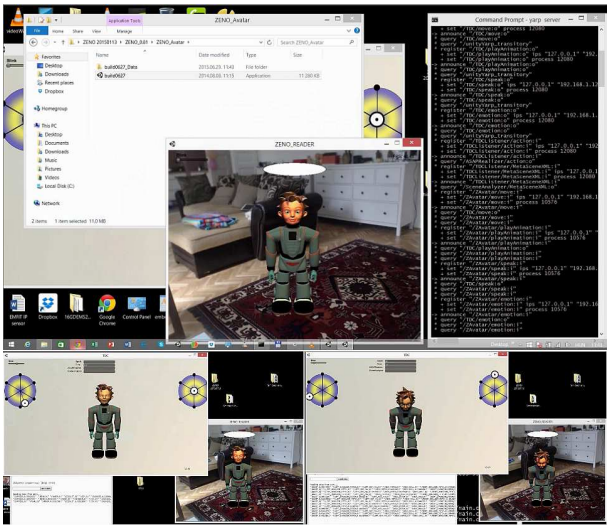

Fig. 5 Augmented Reality interface for the Zeno robot.

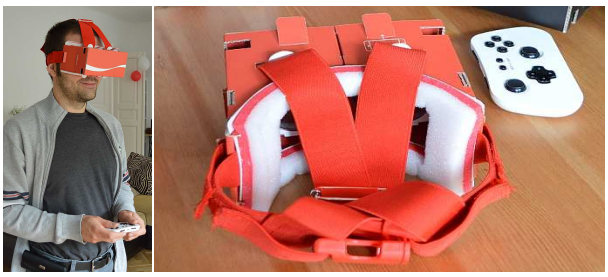

Fig. 6 User wearing a high quality, low-cost Augmented Reality paper viewer (from [13]) for interacting with the virtual Zeno robot in digital space.

5. CONCLUSION

In this paper we presented a unified control architecture and augmented/virtual-realty evaluation framework to compare socially capable robots using a “black box” approach and YARP to connect the individual components. The central contribution of our work is the development of cascaded Temporal Disc Controllers (TDCs) that hide the physical realization of different robot types and accept high level commands, while providing a mechanism for reducing repetitive behaviors at the same time. We implemented our architecture using two existing robots (Face and Zeno) used frequently in educational context. Finally, we demonstrated how photo-realistic digital robot representation open up the possibility of comparing the social and emotional performance in these robots without the need for actual physical access to them.

Because these robots are not yet affordable for most students and schools, we argue that our virtual-reality extensions and augmented-reality interfaces in combination with photo-realistic digital models represent an important step to extend the operational and application domain of these robots, while being able to carry out detailed and large-scale comparative studies on them to establish their true affective performance.

Our preliminary studies to determine the usability of our digital image-based TDC representation space in this context were carried out in a large-scale public event and demonstrated that these models are practically indistinguishable from that of the real robot videos (see links in Fig 4). Furthermore, a detailed comparative study using the digital model varieties have also been carried out [12,16]. Our future work involves the use of Augmented Reality and VR tools to further evaluate our architecture in the context of pedagogical and educational goals.

REFERENCES

[1] The Verge. "The Rise and Fall and Rise of Virtual Reality" accessed July 20, 2015 http://www.theverge.com/a/virtual-reality/oral_history, 2015.

[2] J. Cummings, Bailenson. "How immersive is enough? A meta-analysis of the effect of immersive technology on user presence". *Media Psychology*. Pp1-38, 2015.

[3] Lane, H.C., Hays, M. J., Core, M. G., & Auerbach, D. "Learning intercultural communication skills with virtual humans: Feedback and fidelity". *Journal of Educational Psychology, 105*(4), pp 1026-1035, 2013.

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

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

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

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

[8] V. Charisi, *et.al.* "Towards a Child-Robot Symbiotic Co-Development: a Theoretical Approach", *in 4th Intl. Conf. On New Frontiers in Human-Robot Interaction,* 2015.

[9] G. Metta, P. Fitzpatrick, and L. Natale. "YARP: yet another robot platform." *International Journal on Advanced Robotics Systems* 3.1, pp 43-48, 2006.

[10] A. Zaraki, D Mazzei, M Giuliani, D De Rossi. "Designing and Evaluating a Social Gaze-Control System for a Humanoid Robot. " in IEEE Transactions on Human-Machine Systems 44 (2), 157-168, 2014

[11] D. Mazzei, *et al.* "Hefes: An hybrid engine for facial expressions synthesis to control human-like androids and avatars.", *in Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS & EMBS International Conference on. IEEE*, 2012.

[12] B. Takacs, *et.al,* "A Psychological Framework to Objectively Evaluate Socially Capable Robots for Interactive Tutoring Systems", *in Proc. ICCAS, 2015.*

[13] Panocast Paper VR Viewers, accessed July 20, 2015, http://www.panocast.com

[14] FORBES Flow, accessed July 20, 2015 http://98.130.199.171/NYSM/R25_300.html http://98.130.199.171/EASEL/Forbes_EASEL_LetterofSupport.pdf

[15] Vuforia, Qualcomm, accessed July 20, 2015 https://www.qualcomm.com/products/vuforia

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

[17] **Video:** Comparison of Physical robot with Virtual image-based interactive model , accessed July 20, 2015 http://98.130.199.171/EASEL/3_EASEL_D4.2_TDC_2DImageBasedModel_Comparison.mp4

[18] **Videos:** Supporting material on YARP, TDC, Virtual Robot models, accessed July 20, 2015

http://98.130.199.171/EASEL/1_EASEL_D4.2_TDC_and_ZenoReader_YARPServer_MultipleDiscControllers_AugmentedReality_UnitySWInterface.mp4

http://98.130.199.171/EASEL/1b_EASEL_D4.2_TDC_and_ZenoReader_YARPServer_HighLevelCommands_UnitySWInterface.mp4

http://98.130.199.171/EASEL/2_EASEL_D4.2_TDC_and_FACEReader_YARPServer_MultipleDiscControllers_UnitySWInterface.mp4

http://98.130.199.171/EASEL/4b_EASEL_D4.2_TDC_RemoteSessions_FACE_PhysicalRobot_UPisa.mp4

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.