Procedia Computer Science
Volume 88, 2016, Pages 528–533

7th Annual International Conference on Biologically Inspired
Cognitive Architectures, BICA 2016

On Virtual Characters that Can See

Eugene Borovikov[1,2] and Sergey Yershov[1]
[1]*PercepReal, USA*
[2]*National Institutes of Health, USA*
{Borovikov, Yershov}@PercepReal.com

Abstract
A virtual character (VC) acts within its virtual world boundaries, but with vision sensory capabilities, it may be expected to explore the real world and interact with the intelligent beings there. Such a VC can be equipped with algorithms to localize humans, recognize and communicate with them. These perceptual capabilities prompt a sophisticated cognitive architecture (CA), enabling our VC to learn from intelligent beings and perhaps reason like one. This CA needs to be fairly seamless, reliable and adaptive. Here we explore a vision-based human-centric approach to the VC design.

Keywords: cognitive architecture, virtual character, computer vision

1 Introduction

A typical virtual character (VC) is usually confined to its virtual world, but if we equip it with some visual/audio sensors, provide basic algorithms for object recognition and tracking, and add to it some machine learning (ML) based capabilities for it to learn and reason, our VC will be able to evolve into an intelligent virtual being. VC's perception relies on the given sensory capabilities, e.g. video cameras for its eyes or microphones for its ears. Those sensory streams should have enough of the signal resolution to distinguish among the important features of the objects and beings that our VC would need to interact with. Such a VC would need to use its evolving cognitive architecture to decide on a winning combination of the important features characterizing the real-world objects.

Communications between VC and humans are of the most interest to this study, and thus a virtual character should be able to localize and track humans (e.g. via non-rigid 2D or 3D models), recognize them (e.g. by their faces and/or voices) and communicate with them via natural interfaces (e.g. a human-like avatar). Our VC needs to work in visually unconstrained environments, perform its sub-tasks in interactive time, and constantly learn from its experiences with virtual and real worlds. Such interactions should result in a gradual development of that VC, leading in a highly realistic virtual or mixed reality experience for the humans. We propose a vision-based human-centric approach to the virtual character design by equipping the prospective VC with visual sensors and targeting arguably the most visually expressive and natural human real-world manifestations: face and body.

Selection and peer-review under responsibility of the Scientific Programme Committee of BICA 2016

doi:10.1016/j.procs.2016.07.475

2 Background

A cognitive architecture (CA) is a system that usually comprises multiple computational modules that, working as a whole, attempts to approach human-level intelligence (Goertzel, Lian, Arel, De Garis, & Chen, 2010). It is therefore not surprising that the use of cognitive architectures to generate more humanlike virtual characters has attracted considerable attention.

The biologically inspired (Hubel & Wiesel, 1968) convolutional neural networks (CNN) as introduced as *neocognitron* (Fukushima, 1980) and then improved, generalized and simplified (Simard, Steinkraus, & Platt, 2003), have seen a spectacular renaissance in the recent decade (LeCun, Bengio, & Hinton, 2015) due to the emergence of the affordable GPU computing power, which made the non-trivial image processing tractable for many visual tasks (Abdolali & Seyyedsalehi, 2012; Fan, Xu, Wu, & Gong, 2010) that may be considered quite important for visually capable virtual character adaptive cognitive architecture that needs to learn important features straight from sensors. Modern deep learning (Yue-Hei Ng, Yang, & Davis, 2015) content based image retrieval (CBIR) techniques (Wan et al., 2014) could also help with the VC's robust long-term memory sub-system development, e.g. by transferring deep networks trained on image classification to image retrieval tasks. Deep learning models are an essential part of the approach that we envision for integrating the vision modules described in this paper.

The cognitive architecture ACT-R/E (Trafton et al., 2013) was used to improve human-robot interactions (HRI) where humans and robots share the same virtual or physical environment and where actions of each participants affect those of others. Such HRI often consider the human participant to be more of a perfect machine than a human, i.e. making no mistakes, fully predictable, and is never affected by fatigue or negative emotions. ACT-R/E focused on addressing this gap by developing a CA that is capable of deeper modeling of human cognition to enable the robot participant to recognize when its human counterpart does something wrong. While the focus of our paper is narrower, we believe that the assumption that a physical character from which a virtual one attempts to learn is not perfect is critical to development of robust VC cognition.

For a humanoid VC to be successful in interactions, it is important to understand the human behavioral traits, capturing different personalities for modeling realistic behavior of a VC as suggested by (Saberi, Bernardet, & DiPaola, 2014), where the authors present a hybrid cognitive architecture that combines the control of discrete behavior of the VC moving through states of the interaction with continuous updates of the emotional state of the VC depending on the feedback from the environment. While testing their approach using turn-taking interaction between a human and a 3D humanoid VC, the authors noticed more individualized and believably humanoid artifacts in their VC's behavior.

3 Vision based development of a Virtual Character

Provided with a vision-based cognitive architecture (CA), a VC should be able to utilize its knowledge of the virtual world (e.g. 3D object models and hierarchical infrastructure) for its real-world perception tasks via the given visual sensors. Once aware of the real-world objects and possibly of their inter-relationships and inter-actions, our VC should be able to bring the learned concepts as models to the virtual world, reason about them, emulate some of them, and possibly share them with other intelligent agents, virtual or human.

We argue that the vision-based human-centric approach (Buxton & Fitzmaurice, 1998), involving face and body modeling, can provide the necessary foundation for introducing a sound cognitive architecture to the VC design, especially if it encompasses a human-like avatar, because

- fusing several modalities (texture, color, depth) should result in finer virtual models,
- VC can learn to distinguish between general and person-specific models, and
- real-time interactions promise a natural and incremental VC development.

One could measure the accuracy of the VC design by comparing the target appearance or requested behavior to the captured data, but the subjective judgment of how naturally the evolved VC looks or behaves is likely to be made by a human. The envisioned VC's development steps include:

1. distinguish important (statistically significant) visual objects from their surroundings,
2. track the objects to learn the basic physical concepts (e.g. continuity, gravity, elasticity),
3. apply object and motion models in the virtual world to emulate their physical counterparts,
4. distinguish and track the human users, recognize the most important ones for interaction,
5. learn detailed traits of natural interaction and attempt to emulate them the best way possible.

Given a rich enough representation and computing resources, the cognitive architecture of such a virtual character should allow it to imitate virtual or human personalities it interacts with, adapt the observed traits to is virtual environment, and then eventually develop a non-trivial virtual personality of its own via exploration and experimentation with different behavior patterns and analysis of the implicit or explicit feedback from the humans.

The search for a general vision solution in the context of a virtual character design is beyond the scope of this paper, but since a vision-capable VC is likely to see and interact with humans, we would like to provide our view of a human-centric approach to the VC design, exploring specifics of visual perception regarding

- adaptive color and texture image segmentation, e.g. foreground-background separation,
- real-world objects detection and classification,
- most expressive body parts (face+landmarks, hand+fingers) detection and tracking,
- gesture recognition and interpretation,

and discuss their impact on the intelligent virtual character development that needs to smoothly interact with both natural and virtual intelligent agents. Such a cognitive architecture calls for VC to mentally separate the real and virtual worlds, treating the window to the real world (given by its visual sensors) as an exploratory tool for creating abstract object concepts (e.g. rigid vs. articulate), thus employing high-level reasoning about objects and their relationships, which could be applied in both sides of the virtual boundary.

For example, a virtual character can learn the specifics of a real human face and body motion and attempt to mimic them in virtual character representation by slightly altering its underlying model, hence enhancing the virtual reality user experience. At the same time, in some of the augmented reality (AR) scenarios, virtual objects and characters can be inserted into the real-world video streams and allowed to interact with some real objects in real time, building-up on the VC physical world experience, and ultimately enhancing the AR user experience.

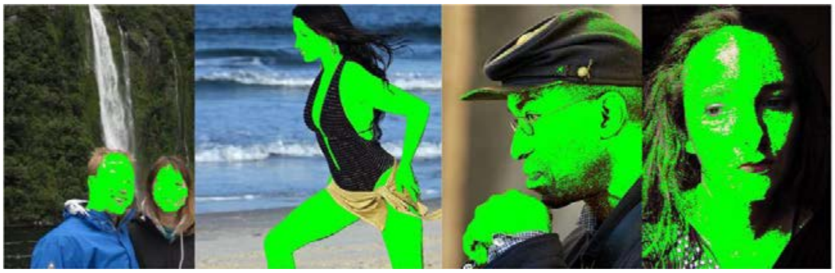

Figure 1: skin mapping using ANN - green pixels denote skin above the threshold

3.1 Adaptive Image Segmentation

One of the important vision tasks for any perceptual system is input image segmentation, that is breaking it into distinct blobs (objects vs. background) using visual cues, e.g. color, texture and motion (Bouwmans, 2014). An important application is color-based human skin mapping, determining a skin likelihood map over any image with an appropriate threshold for skin blob detection that can be determined either theoretically or empirically from experiments.

Among the biologically inspired methods, we suggest the artificial neural network (ANN) that can learn (Kim & Adali, 2002) a skin likelihood maps using a labeled dataset (Jones & Rehg, 2002) with skin and non-skin pixels. A fully connected multi-layer perceptron with a single hidden layer can model the binary decision surface in Extended Color Space (ECS), which is a composite multi-dimensional color space (Malacara, 2002), e.g. [RGB, HSV, Lab] as in our case.

Figure 1 illustrates the ANN-based skin mapping labeling skin pixels as green masks showing the classifier robustness to various image conditions, e.g. outdoors, skin-like backgrounds, and uneven illumination. The resulting GPU-aware ANN-based classifier can be trained and used as an efficient general-purpose image filter, responding to pixel-wise features. The resulting computational speed-ups allow training custom classifiers in several hours, rather than in several days, and the GPU-based deployment allows for fast (near real time) image filtering. For instance, a GPU-based skin mapper processes one image in about 0.1 seconds, while its CPU implementation takes about one second.

This adaptive foreground/background mapping using color can be utilized for more advanced vision stages, e.g. face and hand detection. When coupled with the texture and other image features, it could be used for more general blob/object segmentation in the *early* vision modules of the proposed human-centric cognitive architecture, enabling our intelligent VC to pay attention to the important content in the live video feed, and build the set of robust probabilistic color/texture classifiers based on the important features, forgetting about the non-essential ones.

3.2 Human Face Recognition

In a human-centric cognitive architecture, it is important to have a reliable face recognition (FR) module to distinguish among the humans, and identify some important people a VC interacts with. A robust FR module needs to be fast and robust to adapt to the video signal noise, variations in the visual environment and human appearances due to occlusions, facial hair, jewelry, make-up, etc.

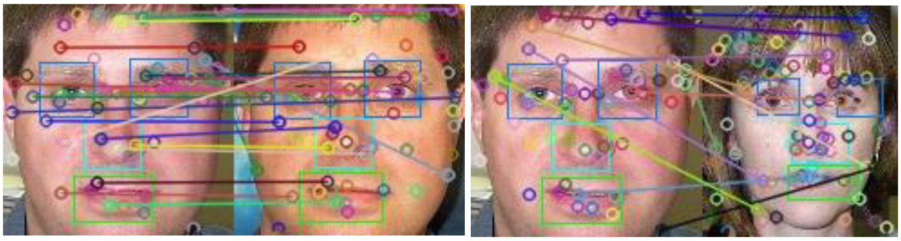

Figure 2: key-spot face image matching - same person vs. different using CalTech faces set

A vision-capable VC needs to recognize faces, matching each against open set collections with very few shots per person, which implies no person-specific model training, but rather relying on a robust ensemble of weighted image descriptors, as in Figure 2: same person gets fewer key-spot mismatches, for content based face image retrieval (Jain, Klare, & Park, 2012; Borovikov, Vajda, Lingappa, Antani, & Thoma, 2013).

When the number of the most frequent users interacting with the virtual character is limited, then one could envision training a classifier and perform 1-of-N face recognition with an occasional enrollment of either a new user or a new appearance of the existing one, followed by a re-training session optimizing the classification boundaries given the new data, hence facilitating the short-to-long term memory information assimilation. Such a limited circle of acquaintances may protect the developing VC from being overwhelmed, but this kind of architecture (if not allowed to change dynamically) may considerably limit that VC's cognitive development with respect to recognizing some new faces, when their set expands beyond just a few in a fairly short amount of time.

3.3 Articulate Shapes Reconstruction and Tracking

A human-centric cognitive architecture for a vision-capable virtual character may need to deal with various manifestations of the human appearance accessible non-intrusively by the vision sensors. The most expressive ones (aside from faces) could be human hands and the whole human body. All three categories call for efficient means of representing and handling non-rigid 3D shapes.

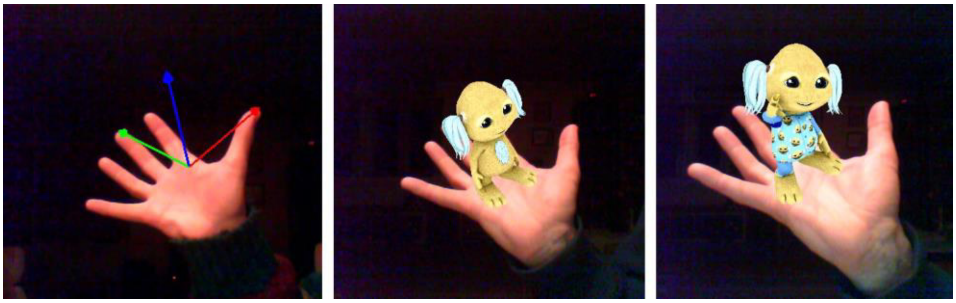

Figure 3: hand tracking and interaction with virtual characters in augmented reality settings

Dealing with non-rigid shapes from a single camera with no markers is ill-posed in general, but in some particular cases may be solvable, when such shapes are known, and their motions are fairly constrained, as it may be the case with the marker-less human hand palm detection and tracking (Lee & Hollerer, 2009). This particular system detects hands candidates by the skin color (e.g. as described in the adaptive image segmentation section), computes the hand outline and all its fingers, and fits a five-finger hand model using the constraints on the natural hand to estimate its pose with respect to the camera; it then can place a virtual object/character on the hand palm, and track it with the Kalman filter algorithm (Chan, Hu, & Plant, 1979), as shown in Figure 3 (from left to right): virtual 3D axes, virtual pet on a hand, and virtual pet in pajamas pointing towards the user's face. One practical application of hand tracking for a humanoid VC could be mastering a sign language, especially in the conjunction with the face tracking and lip motion interpretation.

4 Conclusion

We have proposed implementation of several essential components of a flexible, real-time and continuously learning human-centric VC development system. These include a Bayesian framework for adaptive foreground/background separation, an ensemble method combining several heterogeneous algorithms for face detection and matching, and 3D model based non-rigid body parts detection and tracking. The discussed vision components are expected to compose the core of the human-centric VC cognitive architecture that when combined with the modern machine learning frameworks (e.g. deep neural nets, or DNN) should allow a virtual character to evolve into an intelligent virtual being.

The applications of the human-centric VC design are numerous: VR Q&A kiosks, VR instructors, VR orchestra conductors, VR museum guides, virtual cinema/theatre, etc. We also envision the possibility of non-human-like virtual characters (e.g. AR cartoons), that would have to learn their AR world, interacting with their perceived world. The field of robotics may also benefit from the vision-based human-centric approach by letting robots better imitate human movements and face expressions. The proposed approach could also help the humans interacting with the human-centric virtual characters to start viewing and treating such VC more as intelligent (although artificial) beings.

Acknowledgements: we express our thanks to Ilya Zavorin, Szilárd Vajda and Michael Bonifant for ML R&D collaboration, as well as to Yohan Baillot and Shahar Figelman for AR R&D collaboration.

References

Abdolali, F., & Seyyedsalehi, S. A. (2012). Improving face recognition from a single image per person via virtual images produced by a bidirectional network. *Procedia - Social and Behavioral Sciences*, *32*, 108 – 116.

Borovikov, E., Vajda, S., Lingappa, G., Antani, S., & Thoma, G. (2013). Face Matching for Post-Disaster Family Reunification. In *IEEE International Conference on Healthcare Informatics* (pp. 131–140). http://doi.org/10.1109/ICHI.2013.23

Bouwmans, T. (2014). Traditional and Recent Approaches in Background Modeling for Foreground Detection: An Overview. *Computer Science Review*.

Buxton, B., & Fitzmaurice, G. W. (1998). HMDs, Caves &Amp; Chameleon: A Human-centric Analysis of Interaction in Virtual Space. *SIGGRAPH Comput. Graph.*, *32*(4), 69–74.

Chan, Y., Hu, A. G., & Plant, J. (1979). A Kalman filter based tracking scheme with input estimation. *Aerospace and Electronic Systems, IEEE Transactions on*, (2), 237–244.

Fan, J., Xu, W., Wu, Y., & Gong, Y. (2010). Human Tracking Using Convolutional Neural Networks. *Neural Networks, IEEE Transactions on*, *21*(10), 1610–1623.

Fukushima, K. (1980). Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biological Cybernetics*, *36*, 193–202.

Goertzel, B., Lian, R., Arel, I., De Garis, H., & Chen, S. (2010). A world survey of artificial brain projects, Part II: Biologically inspired cognitive architectures. *Neurocomputing*, *74*(1), 30–49.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243.

Jain, A. K., Klare, B., & Park, U. (2012). Face Matching and Retrieval in Forensics Applications. *MultiMedia, IEEE*, *19*(1), 20–20. http://doi.org/10.1109/MMUL.2012.4

Jones, M., & Rehg, J. M. (2002). Statistical Color Models with Application to Skin Detection. In *International Journal of Computer Vision* (pp. 274–280).

Kim, T., & Adali, T. (2002). Fully complex multi-layer perceptron network for nonlinear signal processing. *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, *32*(1-2), 29–43.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, T., & Hollerer, T. (2009). Multithreaded Hybrid Feature Tracking for Markerless Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics*, *15*(3), 355–368.

Malacara, D. (2002). *Color vision and colorimetry: theory and application*. SPIE Press.

Saberi, M., Bernardet, U., & DiPaola, S. (2014). An Architecture for Personality-based, Nonverbal Behavior in Affective Virtual Humanoid Character. *Procedia Computer Science*, *41*(0), 204 – 211.

Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on* (pp. 958–963).

Trafton, G., Hiatt, L., Harrison, A., Tamborello, F., Khemlani, S., & Schultz, A. (2013). ACT-R/E: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction*, *2*(1), 30–55.

Wan, J., Wang, D., Hoi, S. C. H., Wu, P., Zhu, J., Zhang, Y., & Li, J. (2014). Deep Learning for Content-Based Image Retrieval: A Comprehensive Study. In *Proceedings of the 22Nd ACM International Conference on Multimedia* (pp. 157–166). New York, NY, USA: ACM.

Yue-Hei Ng, J., Yang, F., & Davis, L. S. (2015). Exploiting Local Features From Deep Networks for Image Retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.