OUTCOME Virtual: a tool for automatically creating virtual character animations' from human videos

Auriane Boudin, Ivan Derban, Alexandre D'ambra, Jean-Marie Pergandi, Philippe Blache, Magalie Ochs

▶ **To cite this version:**

Auriane Boudin, Ivan Derban, Alexandre D'ambra, Jean-Marie Pergandi, Philippe Blache, et al.. OUTCOME Virtual: a tool for automatically creating virtual character animations' from human videos. IVA '24: ACM International Conference on Intelligent Virtual Agents, Sep 2024, GLASGOW United Kingdom, United Kingdom. pp.1-3, 10.1145/3652988.3696196 . hal-04859028

HAL Id: hal-04859028
https://hal.science/hal-04859028v1

Submitted on 18 Feb 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

OUTCOME Virtual: a tOol for aUTomatically creating virtual Character animatiOns' from huMan vidEos

Auriane Boudin
auriane.boudin@univ-amu.fr
Aix Marseille Univ, CNRS, LPL, LIS

Ivan Derban
ivan.derban@etu.univ-amu.fr
Aix Marseille Univ, CNRS, LIS

Alexandre D'Ambra
alexandre.D-AMBRA@univ-amu.fr
Aix Marseille Univ, CNRS, CRVM

Jean-Marie Pergandi
jean-marie.pergandi@univ-amu.fr
Aix Marseille Univ, CNRS, CRVM

Philippe Blache
philippe.blache@univ-amu.fr
Aix Marseille Univ, CNRS, LPL

Magalie Ochs
magalie.ochs@lis-lab.fr
Aix Marseille Univ, CNRS, LIS

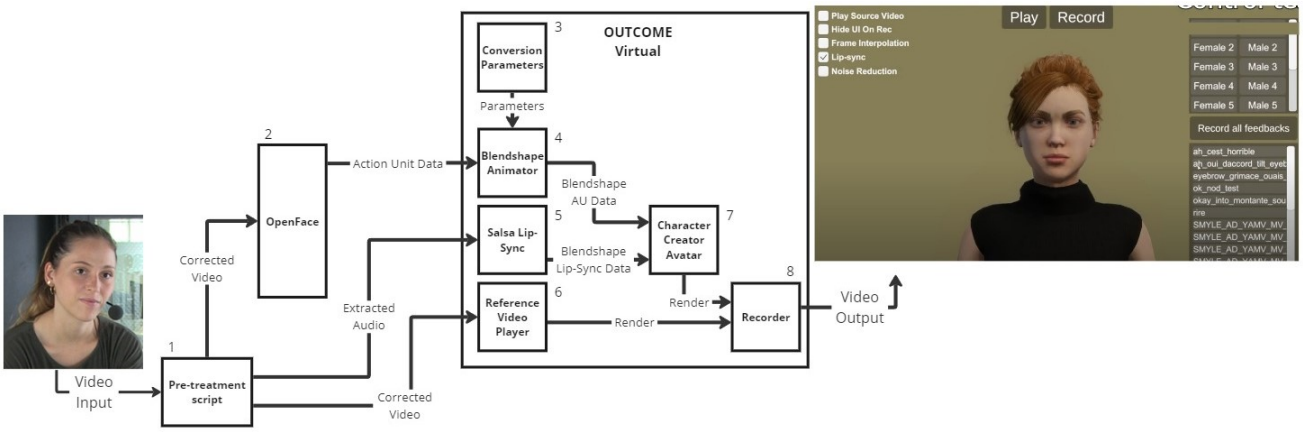

Figure 1: OUTCOME Virtual system architecture.

Abstract

In this article, we present a new open-source tool, called *OUTCOME Virtual*, to automatically replicate humans' behaviors on virtual characters. The tool takes as input a human video and extracts action units, head movements, and gaze direction to animate a virtual character's face and head with the synchronized speech in Unity. The tool is designed to be easily configured, in particular concerning the association between actions units and blend shapes, and with a user-friendly interface to select the appearance of virtual characters, the lip synchronization and the noise-correction option. The tool is illustrated in this paper with the generation of listener feedback.

Keywords

Virtual Characters, Facial Expressions, Head Movements, behaviors animations

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
IVA 24, September 16–19, 2024, Glasgow, UK
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Auriane Boudin, Ivan Derban, Alexandre D'Ambra, Jean-Marie Pergandi, Philippe Blache, and Magalie Ochs. 2018. OUTCOME Virtual: a tOol for aUTomatically creating virtual Character animatiOns' from huMan vidEos. In *Proceedings of Intelligent Virtual Agent (IVA 24).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The variability and credibility of agent behaviors are crucial for user engagement in human-machine interactions and virtual characters. Various methods exist for creating libraries of animations for virtual characters, including rule-based procedural approaches [9], machine learning methods [4, 10, 16], and motion capture techniques [1, 11], for a review see [12]. In the context of motion capture, tools exist to automatically extract Action Units (AUs) and head movements from human video. However, a significant challenge is that animation software like Unity does not handle AUs. Some tools[1] propose to replicate the outputs of OpenFace on avatars. However, the functionalities remain limited concerning the possibility, for instance, to vary the coefficients between AUs and blend shapes, to integrate lip synchronisation with speech or to simulate the behavior on a wide range of virtual characters. To address these issues, we developed a new tool that bridges this gap.

Our tool processes face videos of humans to generate virtual characters that replicate the captured behaviors. In this article, we

[1] as for instance https://github.com/alexismorin/OpenFace-FACS-Unity-Facial-Animator

illustrate the functionalities of the tool using feedback sequences [3, 7, 13] from listeners in real and spontaneous face-to-face interactions from the SMYLE corpus [6, 8]. The tool facilitates the creation of a library of various feedback behaviors for virtual characters by automatically extracting AUs, gaze direction, and head position using OpenFace [2] and converts them into blend shapes in Unity. This tool is composed of several configurable modules, as illustrated in Figure 1.

2 The OUTCOME Virtual Tool

The *OUTCOME virtual* open-source tool has been developed in Unity using the High Definition Render Pipeline (HDRP).

Video pre-processing. A first step consists in pre-processing the input video. The module 1 (Figure 1) takes the interlocutor's video as input and performs initial processing using the FFmpeg package [14][2]. This step extracts both visual and audio data. Subsequently, the module 2 (Figure 1) extracts OpenFace [2] data (AUs, head positions, and gaze directions). To reduce noise, we propose to apply a Python median filter smoothing function from the SciPy library [15]. Module 1 outputs two CSV files, providing users the choice between raw OpenFace data (without noise reduction) or noise-reduced data for the generation of virtual characters' animations.

Facial behavior generation. In Unity, the virtual characters' facial expressions are controlled by blend shapes. Table 1 presents the correspondence between OpenFace AUs and blend shapes used in Unity.To replicate the extracted AUs, a conversion to blend shape is required. This conversion involves applying coefficient to scale AUs from 0-5 to blend shapes' 0-100 range. The coefficients have been defined empirically. However, these coefficients are stored in a user-configurable file, allowing for easy adjustments by users (Module 3, Figure 1). Before generating animations, users can choose from two additional option (Module, options depicted in the top left corner of Figure 1): *Noise reduction*, which replaces outlier AU values with averages from neighboring frames, and *Frame interpolation*, which improves animation smoothness by generating intermediate blend shape values between actual video frames. For the animation of the virtual character' head and gaze, since OpenFace records the head rotation relative to the camera and the gaze direction relative to the head, these values are adjusted by subtracting initial frame values to align with the character's default pose. Finally, a reverse smooth transition technique ensures animation ends smoothly.

Lip synchronisation. Lip synchronization for virtual characters is challenging with OpenFace providing only mouth openness data. To address this issue, the tool integrate the module Salsa liSync[3] The module Salsa (Module 5, Figure 1) processes audio to generate the blend shapes corresponding to the visemes. Salsa also enhances animation realism with random blinking and gaze shifts. Integrating Salsa posed challenges as it overrides the output of OpenFace concerning the gaze and blinking blend shapes. To manage this issue, we automatically disable Salsa's eye and lid processing when

ID	OpenFace Action Units	Arkit Labels	Character Creator Blend shapes
AU01	Inner Brow Raiser	Brow inner Up	Brow raise inner
AU02	Outer Brow Raiser	Brow outer Up	Brow raise outer
AU04	Brow lowerer	Brow down	Brow drop
AU05	Upper lid raiser	Eye wide	Eye wide
AU07	Lid tightener	Eye squint	Eye quint
AU09	Nose wrinkler	Nose sneer	Nose sneer
AU10	Upper lip raiser	Mouth upper	Mouth up upper
AU12	Lip corner puller	Mouth smile	Mouth smile
AU14	dimpler	Mouth dimple	Mouth dimple
AU15	Lip corner depressor	Mouth frown	Mouth frown
AU17	Chin raiser	Mouth shrug	Mouth chin
AU20	Lip stretcher	Mouth stretch	Mouth stretch
AU23	Lip tightener	-	Mouth tighten
AU25	Lips part	-	V lip open
AU26	Jaw drop	Jaw open	-
AU28	Lip suck	Mouth roll upper	mouth roll in upper
AU28	Lip suck	Mouth roll lower	Mouth roll in lower
AU45	Blink	Eye blink	Eye blink

Table 1: Correspondance of OpenFace Action Unit, Arkit labels and blend shapes.

OpenFace detects eyelid-related Action Units. Concerning the possible conflict between the Salsa lip animation and the output of OpenFace for the lip, by default, we use the Salsa lip synchronisation. However, note that the users have the option to disable Salsa through the interface (Figure 1, option 'Lip-sync' on the video output) to give the priority to the extracted AUs of OpenFace. In that case, the lip synchronization is then managed using only the data from OpenFace.

Visualization and Recording. For easy comparison, the *Play Source Video* option (Figure 1, option 'Play Source Video' on the video output) in Module 6 allows to display the source video in the background. Module 7 uses Character Creator 4 to integrate a selection of 10 male and 10 female virtual characters, offering users a wide variety of virtual character's appearance. Module 8 handles video recording, allowing the users to export selected or all videos contained in a folder (list of the names of the videos displayed on the bottom left of the interface, Figure 1).

3 Conclusion

By using multimodal humans video data from real interactions and converting it into realistic virtual character' animations, OUTCOME allows to generate complex and realistic virtual characters' behaviors. Our system's ability to generate a library of behavior is particularly useful for example to construct stimuli for perceptual experiments [5]. For example, researchers can manipulate the form and placement of feedback to study how these changes are perceived, or to assess listener engagement based on feedback complexity. Future work will integrate body postures and movements to further enhance the believability of the agents.

Acknowledgments

This work, carried out within the Institute of Convergence ILCB, was supported by grants from France 2030 (ANR-16-CONV-0002) and the Excellence Initiative of Aix-Marseille University (A*MIDEX) and by the Laboratoire Parole et Langage.

[2]MP4 videos are processed with parameters: vcodec='libx264', profile='high', preset='medium', r=25, g=25, video$_b itrate =' 20M', acodec =' aac', audio_b itrate =' 192k'$

[3]https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442

References

[1] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 487–496.

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 59–66. https://doi.org/10.1109/FG.2018.00019

[3] Janet B Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of personality and social psychology* 79, 6 (2000), 941.

[4] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*. IEEE, 1–10.

[5] Dario Bombari, Marianne Schmid Mast, Elena Canadas, and Manuel Bachmann. 2015. Studying social interactions through immersive virtual environment technology: virtues, pitfalls, and future challenges. *Frontiers in Psychology* 6 (2015). https://doi.org/10.3389/fpsyg.2015.00869

[6] Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Matthis Houlès, Thierry Legou, Magalie Ochs, and Philippe Blache. 2023. SMYLE: A new multimodal resource of talk-in-interaction including neuro-physiological signal. In *Companion Publication of the 25th International Conference on Multimodal Interaction*. 344–352.

[7] Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Magalie Ochs, and Philippe Blache. 2024. A Multimodal Model for Predicting Feedback Position and Type During Conversation. *Speech Communication* (2024), 103066. https://doi.org/10.1016/j.specom.2024.103066

[8] Auriane Boudin, Stéphane Rauzy, Roxane Bertrand, Magalie Ochs, and Philippe Blache. 2024. The Distracted Ear: How Listeners Shape Conversational Dynamics. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 15872–15887. https://aclanthology.org/2024.lrec-main.1379

[9] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '94)*. Association for Computing Machinery, New York, NY, USA, 413–420. https://doi.org/10.1145/192161.192272

[10] Mireille Fares, Catherine Pelachaud, and Nicolas Obin. 2023. Zero-shot style transfer for gesture animation driven by text and speech using adversarial disentanglement of multimodal style encoding. *Frontiers in Artificial Intelligence* 6 (2023). https://doi.org/10.3389/frai.2023.1142997

[11] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 79–86.

[12] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 569–596.

[13] Emanuel A Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk* 71 (1982), 71–93.

[14] Suramya Tomar. 2006. Converting video formats with FFmpeg. *Linux Journal* 2006, 146 (2006), 10.

[15] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[16] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 736–747. https://doi.org/10.1145/3536221.3558058

A Online Resources

The open-source tool : OUTCOME Virtual
`https://github.com/MagalieOchsLIS/OUTCOME_Virtual`

A video presentation of the OUTCOME virtual tool:
`https://www.youtube.com/watch?v=Rngc`$_g YUaUA$