Collection to Creation: Playfully Interpreting the Classics with Contemporary Tools

Ana Herruzo^{1,2}(✉) and Nikita Pashenkov¹(✉)

¹ Woodbury University, 7500 N Glenoaks Blvd, Burbank, CA 9150, USA
{Ana.herruzo,Nikita.Pashenkov}@woodbury.edu,
ae.herruzo@alumnos.upm.es

² Universidad Politecnica de Madrid, Pº Juan XXIII, 11, 28040 Madrid, Spain

Abstract. This paper details an experimental project developed in an academic and pedagogical environment, aiming to bring together visual arts and computer science coursework in the creation of an interactive installation for a live event at The J. Paul Getty Museum. The result incorporates interactive visuals based on the user's movements and facial expressions, accompanied by synthetic texts generated using machine learning algorithms trained on the museum's art collection. Special focus is paid to how advances in computing such as Deep Learning and Natural Language Processing can contribute to deeper engagement with users and add new layers of interactivity.

Keywords: Machine learning information visualization · Real-time generative graphics · Creative technology

1 Introduction: Generations to Generative

Last academic year, Applied Computer Science program at Woodbury University was invited by The J. Paul Getty Museum, a premier art institution in Los Angeles, to design an installation for the College Night event and exhibition scheduled to take place in April 2019. In conversations with educational specialists at the museum, interest was shown in exploring human emotions as a thematic element in a project that would merge art and technology, which successfully aligned with the undergraduate program's mission.

To approach the design and development of this project we created a collaboration between two sophomore classes and designed a new syllabus for each course to include dialogue with the museum. The syllabus also included guided visits to the museum to study some of the art works in its collection, learning about the artists' intentions, and what the works are communicating in their portrayal of emotions and facial expressions.

While studying the museum's art collection [5] that comprises Greek, Roman, and Etruscan art from the Neolithic Age to Late Antiquity; and European art from the Middle Ages to the early 20th Century; with a contemporary lens, questions arose regarding static and finished pieces of art, in contrast to interactive and responsive artworks [6]. In our attempt to find ways to connect the student project to the museum and its art collection,

we decided to explore solutions that would allow for creation of dynamic generative visuals in combination with content somehow derived from the existing collection.

A key observation guiding our concept was that artworks in exhibitions and museums are typically accompanied by a title and a brief description. In our case, unique visuals would be generated based on user interaction, so we proposed the creation of new synthetic titles and descriptions to accompany each user engagement. It seemed fitting that the textual output would be generated by machine learning algorithms trained on existing texts from the museum's art collection, to create a unique connection bridging carefully curated static content with dynamic generated visuals.

The resulting installation utilizes a Kinect sensor to analyze the users' movements and a separate camera to read facial expressions [3] via computer vision and Deep Learning algorithms, using their outputs in the next stage as a basis for text generation based on natural language models trained on the text descriptions of the Getty Museum's art collection.

2 Process: Beyond Codified Interaction

Our goal was to create a new media art [4] piece that has an intimate connection with the users while simultaneously generating new periodic content that is always evolving, changing; never the same.

The installation consists of a vertical video wall composed of three landscape-oriented screens, with two sensors embedded on top: Microsoft Kinect ONE and a USB web camera. We used two software platforms, PyCharm as an integrated development environment for the Python programming language; and Derivative TouchDesigner as a real-time rendering, visual programming platform. The two platforms communicated with each other via TCP/IP sockets over the network.

The title of the project, "WISIWYG", is a play on the popular acronym "What You See Is What You Get" (WYSIWYG), based on the idea that the installation incorporates computer vision processing and machine learning algorithms, in a sense allowing it to generate outputs according to what it sees from its own perspective.

The user experience was a central component of the project and was carefully crafted. After several sessions of user testing to determine the ideal flow for the installation, we designed a sequence of animated events, to successfully guide the user through the experience lasting a total of about two minutes (Fig. 1).

As a result, two types of media proposals were developed in parallel: media to be displayed during users' interaction, and media generated in-between interactions (Fig. 2).

Once the user experience had been designed, the following strategies were developed in order to connect and interact with the attendees.

INSTALLATION DIAGRAM

Fig. 1. Installation diagram

Fig. 2. User experience diagrams

3 User Analysis

The primary driver of the project is the camera input, processed through computer vision and machine learning algorithms. During the course of users' engagement, computer vision is first used to isolate faces and determine the number of people in camera's field of view. This step is accomplished via a traditional computer vision face tracking method using the popular Open Computer Vision (OpenCV) library. The second step uses a Deep

Learning model based on a Convolutional Neural Network (CNN) [7] programmed in Python with the help of Keras and TensorFlow frameworks.

We utilized a 5-layer CNN model made popular by the Kaggle [2] facial expression recognition challenge. The model has been trained on the Feb2013 dataset distributed with the challenge, which consisted of 28,000 training and 3,000 test images of faces stored as 48 by 48-pixel grayscale images. In order to provide the image data in the format that the CNN model expects, sub-windows of the camera feed with faces detected by the OpenCV library were scaled down to 48×48 size to be passed on in this step. The Python code to construct the facial expression detection model in Keras, as well as other Deep Learning models and associated weight parameters utilized in the project and discussed further, are available in an open GitHub code repository.

4 Media Creation

When visitors approach the installation, depending on how many people are in front of the screen; their facial expressions and estimated ages; a unique animation is generated on the video wall, comprising the following elements:

- Silhouettes: A mirror-effect reflection of users' silhouettes on the screen featuring real-time generated visuals, playing inside and outside the silhouettes.
- Scenes: Each student designed four real-time scenes with animated content, driven by hard-coded rules based on the parameters obtained through computer vision algorithms.
- Color palettes: Students generated parametrized color animations using the computer vision inputs (estimated ages, facial expressions, and number of people) (Fig. 3).

Fig. 3. Gallery showcasing several of the animations designed by the students and the different color palettes applied based on the users' facial expressions and movements.

5 Synthetic Text Descriptions

This portion of the project began by screening the Getty's art collection online, and selecting all artifacts depicting people. As the next step, a database was created by recording the artworks' titles and descriptions, as well as subjectively estimating the number and ages [10] of people featured.

In analyzing the Getty's art collection, students experimented with the Deep Learning natural language model GPT-2 [9], an acronym for Generative PreTrained Transformer, released by the non-profit foundation OpenAI in February 2019. The language model generates convincing responses to textual prompts based on set parameters such as the maximum length of response and 'temperature' of indicating the relative degree to which the output conforms to the features resembling training data.

Our project utilized the GPT-2 model with 144 million parameters, the largest made available by the OpenAI Foundation as of April 2019. The model has been "fine-tuned" by the students, a process in which an existing machine learning model was re-trained to fit new data, using descriptions of artworks on display at the Getty Center as the training dataset. To accomplish this task, we utilized the Google Colaboratory [1] notebook environment that allows Python code to be shared and executed online. A Colaboratory notebook designed for fine-tuning the GPT-2 model allowed each student to individually analyze and modify the Python code to read new text data, re-train the language model, and produce new text descriptions based on interactive text prompts.

The text prompts were pre-generated by the students based on their own analysis of artworks and consisted of short singular and plural descriptions like "sad young person" or "two happy people." The text prompts were interactively input to the GPT-2 model to generate responses that were entered into a database and associated by rows with tagged columns for age, number of people and facial expressions. The database content was then programmatically correlated with the outputs of computer vision processing and Deep Learning classification using the Pandas library in Python, by selecting a random database cell containing a response that matches detected facial expression tags. Finally, the selected response was rendered as the text description that accompanied visual output onscreen.

It was beyond the scope of the project to build students' expertise in constructing their own Deep Learning models or invest resources into training the models from scratch, especially in view of time and computing power expended in the process. Researchers at University of Massachusetts, Amherst, have shown that training the GPT-2 language model, for example, can consume anywhere between \$12,902–\$43,008 in cloud computing costs [11]. While training such computationally expensive models was beyond the scope of our project, each student had a chance to install the relevant Python code and libraries, analyze them, fine-tune and run the models on their personal computers.

In addition to facial expression detection, the project incorporated Deep Learning models for age and gender detection, though ultimately we decided not to pursue gender recognition due to non-technical concerns with bias and non-binary gender identification. The age detection was built on a successful CNN architecture based on Oxford Visual Group's (VGG) model [8] that gained recognition in the ImageNet Large Scale Visual Recognition Competition (ILSVRC) in 2014 [11]. The datasets that formed that basis of the VGG model, known as IMDBWIKI [13], consist of over 500,000 labeled faces.

In order to utilize the model effectively, we again made use of pre-trained weights found online [14], applying transfer learning techniques to work with the real-time video data in our project. As a result, each student was able to directly experiment with deploying machine learning models in Python code to process computers' built-in camera input and generate predictions for detected facial expressions and age estimates of each other, initially in the classroom environment and eventually in the public setting of the museum exhibition (Fig. 4).

Fig. 4. Example of texts generated by machine learning algorithms overlaid on the real-time generated animations.

6 Thoughts

With the ongoing evolution in computing hardware and software, machine learning is becoming more accessible and widespread. We're now able to expand the levels of interaction to more complex layers, allowing recognition of patterns, movements, facial expressions and more. Along with the rapid evolution of realtime rendering engines, programmable shaders, and new algorithms, it is now possible to effectively create real-time data-driven media at high resolutions and with great rendering quality. Advances in technology and computing exemplified by Deep Learning can contribute to deeper connections and new layers of interactivity.

Different simultaneous levels of interactivity occurred in this project, some with a direct and transparent effect, and others with more elaborate or indirect interaction. For example, we discovered that when the users recognized their own silhouettes, they would engage more actively with the installation: dancing, jumping, moving their hands or even performing backflips. The facial expression and age-based visualization parameters were

not as transparent, since it is harder for the user to acknowledge this relationship during a time-limited engagement (Fig. 5).

Fig. 5. Users experiencing the installation and playing with their silhouettes on the screen.

Even though our database was quite small, the generative algorithms provided widely variable outputs. The level of uncertainty in understanding generated textual content created a level of engagement as well, attested by visitors inquiring on many occasions why a certain title or description appeared on the screen. It is of our interest that further explorations should address a more thorough analysis of the users' feelings or personality, in search of a deeper and more profound connection between the art piece and the subject.

After creating our database, at the beginning of the user testing process, some of the students raised concerns about the use of gender in camera-based analysis and generated texts. As a result, questions came up regarding the role of gender identity in today's society and opened the door for further discussions involving human-machine interactions. We attempted to address this issue in part by avoiding the use of gender, and programmatically manipulating the generated texts, as well as manually editing our database by screening for male and female pronouns, attempting to "de-gender" it by replacing those with plural "they," or neutral "person", wherever appropriate.

The possibility of using gender detection and other forms of identification, afforded by Deep Learning software algorithms, opens the door to a host of significant questions and concerns. Shoshana Zuboff, in her book *Surveillance Capitalism* [15], raises key elements to reflect on, now that a wide range of software applications and hardware devices that we use daily, monitor, log and process a wide variety of data obtained from the user. Zuboff warns that these technologies are often designed to obtain users' data in disguise, in search of individual and society behavior modification. Users' privacy while

being monitored and issues surrounding this topic were briefly addressed in this project, but are undoubtedly significant questions to be aware of, discussed and reflected upon when designing interactive experiences.

7 Conclusion

Features that make this project unique include the combination of real-time generative graphics with exciting new machine learning models, the nature of its development within an academic environment and the opportunity for the students to exhibit at a great art institution. Trying to find ways to connect the existing art collection to our project, while addressing the student learning outcomes, resulted in a project that successfully fulfills the mission of our University program: hybrid art and technology.

By tapping into different areas of study in one experimental project, we feel that we managed to offer the students an opportunity to understand how diverse disciplines can be intertwined and relevant to each other, while the deployment of a live interactive installation at a world-renowned art institution endowed them with valuable professional production experience. The number of constraints, such as addressing multiple curriculum requirements, a collaboration between academic and art institutions, maintaining appropriate workloads and so on, turned out to strengthen and boost the level of creativity, allowing the students to become proficient in several technical skills working with advanced programming frameworks and computational models.

Perhaps our primary contribution as faculty guiding this project has been to help students synthesize and leverage the somewhat disparate technical and creative tools, frameworks and resources as part of the learning process, as well as providing a critical view towards that state of technology and computer science in today's culture. We hope this project summary provides a useful reference point for others seeking an approach in creative applications of contemporary technologies.

References

1. Google Colaboratory. Colab.research.google.com (2020). Cited 25 May 2020. <https://colab.research.google.com/notebooks/welcome.ipynb>
2. Challenges in Representation Learning: Facial Expression Recognition Challenge | Kaggle. Kaggle.com (2020). Cited 25 May 2020. <https://www.kaggle.com/c/challenges-in-representation-learning-facial-expressionrecognition-challenge/>
3. Viola, P., Jones, M.: Robust real-time face detection. *Int. J. Comput. Vis.* **57**(2), 137–154 (2004)
4. Tribe, M., Grosenick, U., Jana, R.: *New Media Art*. Taschen, Köln (2006)
5. Collection (Getty Museum). The J. Paul Getty in Los Angeles (2020). Cited 25 May 2020. <http://www.getty.edu/art/collection/>. Accessed 10 Jan 2020
6. Krueger, M.: Responsive environments. In: *AFIPS 1977 Proceedings, National Computer Conference*, pp. 423–433 (1977)
7. Lopes, A., de Aguiar, E., De Souza, A., Oliveira-Santos, T.: Facial expression recognition with Convolutional Neural Networks: coping with few data and the training sample order. *Pattern Recogn.* **61**, 610–628 (2017)

8. Parkhi, O., Vedaldi, A., Zisserman, A.: Deep Face Recognition. Visual Geometry Group - University of Oxford (2015). Cited 25 May 2020. <http://www.robots.ox.ac.uk/~vgg/publications/2015/Parkhi15/parkhi15.pdf>
9. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language Models are Unsupervised Multitask Learners. OpenAI Foundation (2019). Cited 25 May 2020. https://cdn.openai.com/better-languagemodels/language_models_are_unsupervised_multitask_learners.pdf
10. Rothe, R., Timofte, R., Van Gool, L.: Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vis.* **126**(24), 144–157 (2016)
11. Strubell, E., Ganesh, A., McCallum, A.: Energy and Policy Considerations for Deep Learning in NLP. arXiv.org. (2019). Cited 25 May 2020. <https://arxiv.org/abs/1906.02243>
12. ImageNet Large Scale Visual Recognition Competition 2014 (ILSVRC2014). Image-net.org (2020). Cited 25 May 2020. <http://www.image-net.org/challenges/LSVRC/2014/>
13. IMDB-WIKI - 500 k + face images with age and gender labels. Data.vision.ee.ethz.ch (2020). Cited 25 May 2020. <https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/>
14. Apparent Age and Gender Prediction in Keras - Sefik Ilkin Serengil. Sefik Ilkin Serengil (2020). Cited 25 May 2020. <https://sefiks.com/2019/02/13/apparentage-and-gender-prediction-in-keras/>
15. Zuboff, S.: *The Age of Surveillance Capitalism*, 1st edn. Public Affairs, New York (2019)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.