GAB - Gestures for Artworks Browsing

VALENTINE BERNASCONI, Digital Visual Studies, University of Zurich, Switzerland

Hands are an important tool for our daily communication with our peers and the world. They allow us to convey information through particular gestures that are either the product of social conventions or personal expressions. Thanks to the sophistication of sensing and computer vision technologies over the past decade, automated hand recognition can now be more easily used and integrated in simple web applications. In a context of digital artworks collections, it means that gestures can now be envisioned as a new browsing tool that goes beyond simple movements to navigate through a 3D digital space. The paper presents Gestures for Artwork Browsing (GAB), a web application which proposes to use hand motions as a way to directly query pictorial hand gestures from the past. Based on materials from a digitized collection of Renaissance paintings, GAB enables users to record a sequence with the hand movement of their choice and outputs an animation reproducing that same sequence with painted hands. Fostering new research possibilities, the project is a novelty in terms of art database browsing and human-computer interaction, as it does not require traditional search tools such as text-based inputs based on metadata, and allows a direct communication with the content of the artworks.

CCS Concepts: • **Human-centered computing** → *Interactive systems and tools*; • **Information systems** → **Search interfaces**.

Additional Key Words and Phrases: user interface, cultural heritage, browsing tool, hand gesture

ACM Reference Format:
Valentine Bernasconi. 2022. GAB - Gestures for Artworks Browsing. In *27th International Conference on Intelligent User Interfaces (IUI '22 Companion), March 22–25, 2022, Helsinki, Finland.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3490100.3516470

1 INTRODUCTION

Thanks to numerous digitization campaigns taking place in art institutions such as libraries and museums, an important number of artworks collections are made available online. Simultaneously, the rise of new technologies, sophisticated machine learning models and computer vision techniques allows us to envision this data under new perspectives. There is the possibility to extract similar patterns [10][11] or body poses [5] from a large corpus, and to propose innovative experiences to the user as well as new research tools for art historians. However, most browsing tools, especially in the case of scientific research[1], do not exploit the potential of these technologies and often use classic user interfaces, with text-based input fields and research performed on the metadata of the images [9]. Such interfaces require some domain-specific knowledge in order to query the collection efficiently, such as the name of an artwork, an artist or the year of creation. The interaction with data remains on the level of the metadata, while the content of the images is ignored. There is little room for new discoveries among the data and new research perspectives. On the other hand, content-based browsing tools for large collections of images usually propose a navigation through a grid or cluster display based on similarity content [6] [1], which allows a better overview of a collection. However, their query system either relies on keywords or image input [14], and the cost for the creation of sketch-based image retrieval remains for the time being an important barrier for real applications [4]. Regarding systems proposing gesture based interactions with image collections, they raise the problem of a definition of a language to be learned to interact with the system

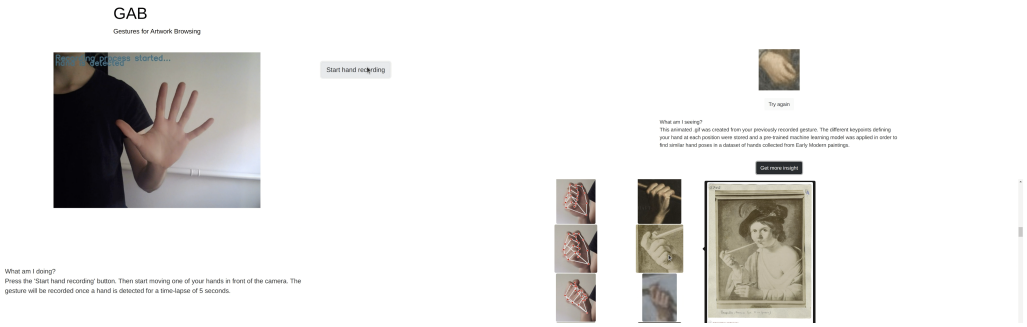
[1]For examples see https://www.biblhertz.it/it/photographic-collection and https://www.europeana.eu/fr/collections

Manuscript submitted to ACM

or to navigate through the data [7] [3] [12]. Such installations also require specific sensing technologies [8] which can be complex for other institutions to access and put in place. Hence, there is a need to propose new browsing mechanisms based on image content and more accessible interfaces in order to benefit from new computational methods and hardware, and foster new research perspectives in the domain of art history. Based on these observations and in a context of a research project on computational analysis of hand gestures in early modern art, the need to rethink the access to images in a large collection emerged. The GAB project, *Gestures for Artwork Browsing*, is an application based on a collection of hands automatically extracted from paintings that can be queried through the recording of the hand gesture of the user. The project not only exploits the possibility to access data based on innovative input, but also the possibility to easily create and deploy such interfaces. The present paper introduces the GAB application, its overall functioning and the different implementation details. It finally discusses the possible impacts and further improvements of such an application.

2 THE GAB APPLICATION

2.1 Overview of the application



(a) Front page (b) Results page

Fig. 1. GAB web application

GAB is a web application that proposes a new approach to query painted hand gestures from the past. The current prototype is performing with a collection of painted hands extracted from paintings from the 15th to the 18th Century. The design for the web application was kept simple and is composed of two pages, namely a main page and a result page. The main page has a central rectangle, displaying the live stream from the webcam of the computer. Next to the frame, a record button titled *Start hand recording* and a short explanatory text below invite the user to record a short hand gesture (see fig. 1a). The button is used to mark the beginning of the interaction. By pressing it, the user launches a system which starts recording a sequence of five seconds once a hand is detected in front of the camera. The latter is indicated by a short text on the video stream (see fig. 1a). Once five seconds of hand gesture are captured, the live streaming disappears to give way to a spinning wheel, indicating to the user that computation is in process. In the background, a .gif animation is produced, mimicking the recorded hand gestures with painted hands. The animation is displayed on top of the result page, which displays the .gif file, a *try again button*, which allows the user to go back to the main page, and a *Get more insight* button. When clicked, a section at the bottom of the page collapses and shows some details of the process used to create the animation. For each frame, a cropped image of the detected hand is

displayed. The hand is overlapped with the different keypoints detected by the machine and their edges, shaping a skeleton (see fig. 1b). On the right hand side, the corresponding painted hands retrieved from the collection is displayed. Finally, by hovering over the image of the painted hand with the cursor, the user can reveal the full painting it was retrieved from and its corresponding metadata. The metadata includes the title of the painting, its author and year of creation. When data is unavailable, *unknown* is displayed instead.

The GAB application removes obstacles that traditional interfaces put between the user and visual data such as textual input and thus proposes a more direct access to specific information. The user is invited to envision knowledge through their body and to develop, with some practice, a language tailored to the data, without the need for a primary training session to communicate with the machine. The goals for the user are to learn from the data and understand over time of querying, what gestures provide more results than others. Furthermore, the application works with low material requirements, no specific configurations from the user and is non intrusive.

2.2 Implementation process

The project is based on the Fototeca collection of the Bibliotheca Hertziana – Max Planck Institute for Art History[2], and uses a subset of the collection, namely paintings from the Early Modern time, which are more figurative representations. Based on this dataset of paintings, the OpenPose model[3] [2] was used in order to produce a collection of hands. OpenPose consists of a real-time multi-person system to jointly detect human body, hand, facial, and foot keypoints on single images given as inputs. The pre-trained model was used on each painting and based on the resulting keypoints, bounding boxes were then automatically computed around detected hands in order to crop them and create a collection of hands (see fig. 2). However, as OpenPose was trained on real images, the results underwent a low accuracy and had to be manually cleaned. In the end, from the 18'641 hands detected out of the 5'234 images, a collection of about 5'993 accurate hands was gathered, with an estimated detection accuracy of 32 %. The images are all stored under the name of an identification number, which allows to easily link them to their metadata available in a database. A simple unsupervised k-nearest neighbor model (k-NN)[4] was then fitted with features extracted from the 21 hand keypoints detected by the OpenPose model. The features correspond to the joint angles and unit vectors from the vertices and are used to describe the direction and poses of the different fingers.

The further implementation of the web application was scripted in Python, using a Flask framework[5] in order to process the client's input and render the results. These results consist of a path to the .gif animation, two lists holding the paths to each image used and their corresponding painting, and a list holding the complementary metadata (title, author and year of creation). The acquisition of this material is performed through the following main steps:

(1) First, in order to detect hands from the webcam images, MediaPipe[6] [13] is used. The framework offers a Python library with tools for live hand detection. It detects hands on each recorded frame and provides their corresponding keypoints. These keypoints, which correspond to the standard of 21 keypoints used in OpenPose, are then stored and processed to retrieve similar hand poses from the hands collection.
(2) Once the recording process ends, the retrieval of similar hand poses is performed with the pre-trained k-NN model. It outputs a list of the five painted hands closest to the hand recorded. To avoid redundancy, the two previous images used in the sequence are looped through in order to check if a hand was already used in two

[2]https://www.biblhertz.it/it/photographic-collection
[3]https://github.com/CMU-Perceptual-Computing-Lab/openpose
[4]The sickit-learn implementation was used https://scikit-learn.org/stable/modules/neighbors.html
[5]https://flask.palletsprojects.com/en/2.0.x/
[6]https://google.github.io/mediapipe/

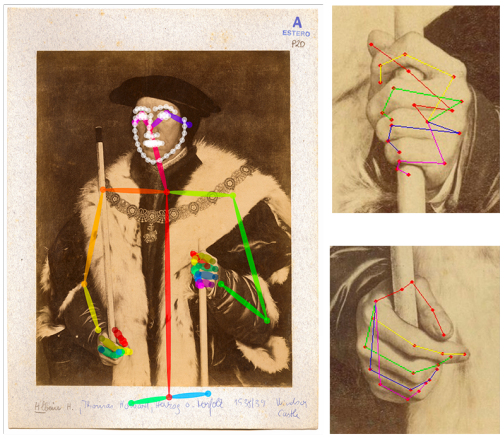


Fig. 2. Example of OpenPose results and extracted hands from a painting, ©Bibliotheca Hertziana, Max-Planck-Institut für Kunstgeschichte, Rom

previous frames. In case of repetition, the k-NN model outputs the list of painted hands for the next recorded hand in the sequence. The selection of a new painted hand is then performed from the intersection of this last list and the one of the present frame. Choosing an image close to the current and next frame aims to allow a smoother transition between images in the animation.

(3) All selected images from the hand collections are then used to create the short .gif animation that is temporarily stored on the machine running the script [7].

(4) The identification number of each painted hand from the sequence is then used to retrieve the path to the full image it belongs to and to its metadata stored in the database. This information is then passed on to the template rendering the results.

3 CONCLUSION

The creation of the GAB application shows the strong potential of recent technological breakthroughs to easily build systems that allow us to rethink traditional user interfaces for browsing tools. The project avoids the usual need to verbalize a specific research and brings the user back to a more fundamental and instinctive way of communicating with its environment through his own gestures. Furthermore, the low requirements in terms of technological devices, such as the use of the webcam incorporated in almost every personal computer, and the use of open-source models for body pose estimations, allows a quick installation of such systems for any digital collection as well as an easy integration for a large audience. However, future works will consist of re-training the OpenPose model in order to get better performances on paintings, as the cleaning process of the results is a time consuming task. Another improvement will be an expansion to other body parts, such as facial expressions, leg poses or the full body. Future implementations of these extensions can take the shape of a modular application, where the user is invited to select the type of research he desires to perform. Finally, the project also fosters further conversations on the possibility to access and gain knowledge through the body, starting from a personal gestural language rather than a language established specifically for the computational system.

[7] The Python library imageio, https://imageio.readthedocs.io/, was used

ACKNOWLEDGMENTS

I would like to thank Dr. Darío Negueruela del Castillo for the many discussions that lead to this project.

REFERENCES

[1] Kai Uwe Barthel, Nico Hezel, and Klaus Jung. 2017. Visually Browsing Millions of Images Using Image Graphs. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval (ICMR '17)*. Association for Computing Machinery, New York, NY, USA, 475–479. https://doi.org/10.1145/3078971.3079016

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1 (Jan. 2021), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257 Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[3] Lins Derry, Douglas Duhaime, Jordan Kruguer, Dario Rodighiero, Jeffrey Schnapp, and Christopher Pietsch. 2021. Surprise Machines. https://vimeo.com/595473865

[4] Conghui Hu, Yongxin Yang, Yunpeng Li, Timothy M. Hospedales, and Yi-Zhe Song. 2021. Towards Unsupervised Sketch-based Image Retrieval. arXiv:2105.08237 [cs.CV]

[5] Leonardo Impett. 2020. Analyzing Gesture in Digital Art History. In *The Routledge Companion to Digital Humanities and Art History*. Routledge. Num Pages: 22.

[6] Yanir Kleiman, Joel Lanir, Dov Danon, Yasmin Felberbaum, and Daniel Cohen-Or. 2015. DynamicMaps: Similarity-based Browsing through a Massive Set of Images. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 995–1004. https://doi.org/10.1145/2702123.2702224

[7] Panayiotis Koutsabasis and Chris K. Domouzis. 2016. Mid-Air Browsing and Selection in Image Collections. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '16)*. Association for Computing Machinery, New York, NY, USA, 21–27. https://doi.org/10.1145/2909132.2909248

[8] Sreejith M, Siddharth Rakesh, Samik Gupta, Samprit Biswas, and Partha Pratim Das. 2015. Real-time hands-free immersive image navigation system using Microsoft Kinect 2.0 and Leap Motion Controller. In *2015 Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*. 1–4. https://doi.org/10.1109/NCVPRIPG.2015.7489999

[9] Vivien Petras, Timothy Hill, Juliane Stiller, and Maria Gäde. 2017. Europeana – a Search Engine for Digitised Cultural Heritage Material. *Datenbank-Spektrum* 17, 1 (2017), 41–46. https://doi.org/10.1007/s13222-016-0238-1

[10] Benoit Seguin. 2018. The Replica Project: Building a visual search engine for art historians. *XRDS: Crossroads, The ACM Magazine for Students* 24, 3 (April 2018), 24–29. https://doi.org/10.1145/3186653

[11] Xi Shen, Alexei A. Efros, and Mathieu Aubry. 2019. Discovering Visual Patterns in Art Collections With Spatially-Consistent Feature Learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 9270–9279. https://doi.org/10.1109/CVPR.2019.00950

[12] Tijana Vuletic, Alex Duffy, Laura Hay, Chris McTeague, Gerard Campbell, and Madeleine Grealy. 2019. Systematic literature review of hand gestures used in human computer interaction interfaces. *International Journal of Human-Computer Studies* 129 (2019), 74–94. https://doi.org/10.1016/j.ijhcs.2019.03.011

[13] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. MediaPipe Hands: On-device Real-time Hand Tracking. *arXiv:2006.10214 [cs]* (June 2020). http://arxiv.org/abs/2006.10214 arXiv: 2006.10214.

[14] Wengang Zhou, Houqiang Li, and Qi Tian. 2017. Recent Advance in Content-based Image Retrieval: A Literature Survey. arXiv:1706.06064 [cs.MM]