Studies in Communication Sciences (2023), pp. 1–18

Medals and likes: A methodology for big data image dataset analysis of Olympic athletic beauty on Instagram

Carlos Roberto Gaspar Teixeira*, Pontifical Catholic University of Rio Grande do Sul, PUCRS, School of Communication, Arts and Design, Graduate program on Social Communication, Brazil
Roberto Tietzmann, Pontifical Catholic University of Rio Grande do Sul, PUCRS, School of Communication, Arts and Design, Graduate program on Social Communication, Brazil
*Corresponding author: carlos@ctexdesign.com

Abstract
This article seeks to understand how the Cultural Analytics' methodological approach and computational tools help interpret large image datasets. A set of 87 730 images of 389 Olympic athletes was collected from Instagram and analyzed, featuring a timespan from September 2011 to November 2020. The image set was structured and organized using computer vision processing combined with interactive visualization tools (Google Vision, PixPlot, Image Network Plotter). The analysis, mixing quantitative and qualitative methods, identified patterns represented as image clusters. Regular personal computers served as the hardware platform. Approximately 60 % of the athletes' posts were related to non-sports topics, highlighting common characteristics of the visual culture disseminated on Instagram, such as selfies, lifestyle, leisure, travel, and food. Images of sports content, considered a central aspect of the research, had a lower frequency of publications featuring topics such as competitions, training, exercises, and sports practices in general. Beyond this result, the study offers a possible technical framework for similar researchers using large image datasets.

Keywords
Olympics, Instagram, Cultural Analytics, computer vision, data visualization

1 Introduction: Olympic beauty, Digital Humanities and Cultural Analytics

The Olympic Games are one of the most significant global media events. Since their revival in 1896, they have been a hallmark of tests and the introduction of innovative communication technologies for larger audiences ever since. More recently, with the growth of the mobile broadband Internet and social networks, visual content production has risen enormously, with viewers, press, athletes, and delegations managing their channels and publishing images and videos. The abundance of posts presents a problem for media communication analysts and scholars: How can they identify patterns and trends among thousands of images?

These image collections under discussion come from the Olympic cycle: the preparation, training, team selection, and competitions that move such sports every four years. The iconography of athletes in competition, overcoming difficulties and limitations and eventually reaching victory operates as a connective thread between past and present times and features some constant aesthetic values. The concept of Olympic athletic beauty articulates two main sources: Eco's (2010) reasoning that beauty is relative to time and culture, achieving what Gumbrecht (2006) identifies as athletic beauty, arising from the unique and incomparable ability that sports have to captivate spectators through the abundant production of images propagated by the media.

Digital Humanities (DH) is a broad field of knowledge involving several areas, blending computational tools to aid understanding problems derived from the Humanities (Evans & Rees, 2012; Liu, 2012; Svensson, 2016). The field "(…) marks a move beyond a privileging of the textual, emphasizing graphical methods of knowledge production and organization, design as an integral component

https://doi.org/10.24434/j.scoms.2024.01.3886

of research, transmedia crisscrossings, and an expanded concept of the sensorium of humanistic knowledge" (Burdick, Drucker, Lunenfeld, Presner, & Schnapp, 2012, p. 122). According to Berry (2012), DH approaches does not only propose quantifying traditionally qualitative research. Instead, they prioritize ideas in which computational tools can stir curiosity, reveal evidence, suggest patterns and structures, or reveal trends, allowing them to understand and identify cultural, social and political processes on a large scale, thus making visual data analysis readily available. For Burdick et al. (2012), the screen culture of the 21st century allows the visual to become increasingly fundamental to DH.

From this context, Lev Manovich developed the concept of Cultural Analytics (Hochman, Manovich, & Chow, 2013; Manovich, 2009, 2012, 2013, 2017a, 2017b, 2018), expanded in his eponymous book (Manovich, 2020). CA (Cultural Analytics) emerged from DH, based on the need to analyze patterns and trends of the contemporary digital culture, covering different types of media (in addition to textual media) using computational tools.

It is possible to summarize the concept from Manovich's (2020) perspective as the author advocates the utilization of various computational and design methodologies. These methods include data visualization, media, interaction design, statistics, and machine learning. These tools are used to probe and analyze contemporary culture on an extensive scale. The objective of this approach is to glean insights into the creative output, imaginative processes, and values of hundreds of millions of individuals worldwide. These insights are extracted from the content these individuals share online, in both professional and personal contexts.

Visualization methods related to CA enable the exploration of extensive collections of visual cultural data, without necessarily using metric or statistical measurement. Through computational procedures, it becomes possible to observe trends that cannot be identified by reading, viewing, or to interact individually with each of the artifacts or even small sets. This technique can be applied to vast media universes, operating at much faster playback speeds than customarily intended. It is necessary "to compress massive media universes into smaller observable media landscapes compatible with human information processing rates" (Manovich, 2020, p. 255) while maintaining sufficient detail to identify subtle patterns. The author emphasizes the idea of agility and scalability in data interpretation. As an example, Manovich (2020) postulates that one should be able to discern patterns in a million images as fast as it would take to do so in a single image, enabling for the insightful understanding of large datasets.

Image sets, such as the ones analyzed here, can be arranged in a variety of settings according to their metadata (such as authors, dates, likes, hashtags), content properties (the presence of faces, logos, objects, for example) or visual properties (such as dominant colors, amount of texture, number of shapes, among others). The key to understanding this methodological strategy is that it is based on a qualitative rather than a quantitative approach, allowing the researcher to work with large datasets. Previous methodologies for analyzing image groups usually involved manual counting of the observed features (e. g., the average shot length in a movie, how many shots are close-ups, how many feature Coca-Cola bottles), demanding the researcher to consider each item subjectively, albeit guided by the assigned methodology and translate the potential richness of the image datasets usually into a spreadsheet for further calculation and summarization.

Manovich (2020) acknowledges the potential of media visualizations, but also recognizes their limitations in disclosing all possible patterns within a collection of images. These visualizations are especially beneficial when the amassed images share common characteristics, which underscores the significance of sample selection. Despite its limitations, Manovich (2020) defends the methodological validity of this approach. He justifies its labeling as visualization by focusing on one of its key operations: the arrangement of elements in a manner that facilitates pattern recognition for the user, particularly for ones that might otherwise be challenging to detect. This approach does not delegate the complete interpretation of the dataset to machine learning engines, opting to empow-

er the human observer / analyst instead. The concept of media visualization features three operations:

1. Zoom to see the entire collection (image montage)
2. Temporal and spatial sampling
3. Remapping (rearranging the media samples into new configurations)

In this sense, when thinking about image analysis capabilities and the construction of data visualization models, Manovich (2020) emphasizes the importance of computational tools in this development. He outlines three pivotal decisions that drive the creation of data representations, which in turn enable these representations to be computed, managed, understood, and disseminated through data science techniques. Manovich (2020) underscores the critical role of computational tools in the development of image analysis capabilities and the construction of data visualization models. These key decisions involve the selection of the objects, the resources used, and the coding involved, emphasizing that the objects chosen, the resources selected, and the coding involved are the "(…) three decisions responsible for creating data representations and, consequently, making them computable, manageable, knowable, and shareable through data science techniques" (p. 131).

In this text, we question how the methodological approach of Cultural Analytics in the field of Digital Humanities helps interpret large image datasets with computational tools, identifying which patterns are present in the representation of Olympic beauty as collected from contemporary athletes' Instagram posts and indicating a potential framework for similar research endeavors. The text is organized as follows: After this introduction, we present an overview of the methodology, explain each of the main tools used on its features and results, and discuss the research outcomes in the final considerations.

2 Methodological selection

This article revisits the results from a doctoral thesis questioning how the legacy of Olympic athletic beauty represented a visual dialogue between the records of antiquity and the contemporary on Instagram (Teixeira, 2022). It is, hence, thematically aligned with the thesis and reflects the path developed there, but it presents a different experience in the points that we highlight below:

1. The analyses come from the database collected on Instagram, not considering antique civilization sources as did the thesis.
2. The emphasis is on the stages of the analysis process rather than discussing the cultural implications of the results.
3. Technical-methodological aspects are highlighted.
4. The revision of graph and figures.
5. The emphasis on the clusterization of images to identify groups and patterns.

The choice of DH and CA techniques emanated from an ambition to identify patterns more comprehensively, and in this section of the text we present the steps taken in detail. The process is summarized in four stages: the definition of the research corpus, the collection of records, the application of computer vision techniques, and the creation of visualizations.

Instagram was chosen for the analysis because it is the leading online source on how the visual representations of this contemporary Olympic Athletic Beauty are defined and updated continuously. The timespan analyzed in the investigations considers the emergence of this platform in 2010 and its development until mid-2020. Photography was still the focus of social media throughout this timeframe, while video formats were advancing among audiences.

Several criteria were adopted to define the sample used in the study. When considering the quotas for individual competitors for the Tokyo 2020 Games, two main sports fields were selected: track and field athletics and wrestling, adding up to a share of 19.7 % of the competitors. A historical criterion was considered, as both are some of the oldest competition modalities, dating back to Ancient Greece. Also, individual sports that held the most extensive number of qualifying events until the time of collection were prioritized, considering that the postponement of the

Games also delayed the selection of the competitors.

Consequently, the selected sports of the Summer Olympic Games were:

- Athletics – Running: 100 m, 400 m, and marathon events
- Athletics – Jumps: high jump, pole vault, long jump, and triple jump
- Athletics – Throwing: discus and javelin
- Wrestling

Considering the qualifying slots available for the Tokyo 2020 Games, the total number of competitors in the Olympics was 11 315 athletes,[1] between individual and collective modalities, from 206 countries, distributed in 339 events (International Olympic Committee, 2020). The next step to define the corpus of athletes was a criterion based on the statistical relevance of the data, applied to help reduce the sample size. Based on a sample calculation (Gil, 2008), considering a confidence level of 95 % and a margin of error of 5 %, a minimum sample of 372 individuals was indicated. The best performance indices obtained for track and field athletics between May 1, 2019 and June 29, 2020 were then collected (World Athletics, 2021); and for wrestling, between September 2019 and May 2020 (United World Wrestling, 2021).

Once the sports modalities were defined, the Instagram profile selection began. This step was performed manually from the preliminary qualifying listing for Tokyo 2020 (United World Wrestling, 2021; World Athletics, 2021). The names of the athletes who would have their Instagram profiles previously analyzed were defined to validate the final set, considering those with the best results at that moment, thus representing the potential favorites for a spot in the Olympic Games. The pre-analysis and selection of athletes exclusively considered public / open profiles, in addition to prioritizing those verified by the platform.

The final sample listing contained 389 profiles of athletes, both men and women, from 86 different countries. It was a deliberately diverse dataset, based on the supposition that such diversity of contexts and backgrounds would reveal transnational visual patterns stemming from the Olympic imagery. Basic information was gathered with the Phantom Buster tool, recording general profile data, such as profile name, biography description, total of followers, total of posts, and whether it is a verified account or a business account.

The first collection of the data contained in each post came later, using a data scraping script written in Python by user ARC298, currently unavailable, between December 4 and 7, 2020. This script ran the entire timeline of a listing of profiles, storing information from publications such as date, username, caption text, the total number of likes, and the total number of comments. In addition, the tool also saved the first post, either image or video, disregarding sequences, or carousel posts. The complete collection featured 115 204 publications from the 389 selected profiles, with 86.1 % of images and 13.9 % of videos. The significantly greater volume of static images, which exceeds the number of videos by over six times, justifies the exclusion of posts containing moving content. Only the entries with photographs (static images, visual combinations of photography, and design and graphic cards) were considered in the analysis. Out of these criteria, a dataset of 87 730 images was used for processing, visualization and analysis. This image bank included posts from September 19, 2011, the date of the first publication collected, until November 30, 2020.

3 Computer vision tools dedicated to the development of an image analysis method

Based on the research steps presented, a series of computer vision tools were used to construct and validate a valid methodological path that could indicate imagery analysis and interpretation processes based on a prior ordering of an extensive database using computational tools. In this sense, the proposed path is described below, followed by the respective tools.

1 Number referring to positions open in 2020. With the postponement of the Olympic Games due to COVID-19, the number was subsequently changed.

1. Computer vision processing: finding labels with Google Vision
2. Interpreting and understanding the identified labels: GENUS
3. Building views to find and analyze clusters:
   a) Image cluster viewing with ImageNet: PixPlot
   b) Image cluster viewing with Google Vision: Image Network Plotter

The above order mirrors the one developed in the thesis, which was this text's starting point. Other work pipelines are possible, either shorter or longer, incorporating updated tools or merging them. The following sections detail the tools with the respective analyses made possible by each based on the theme and research previously discussed.

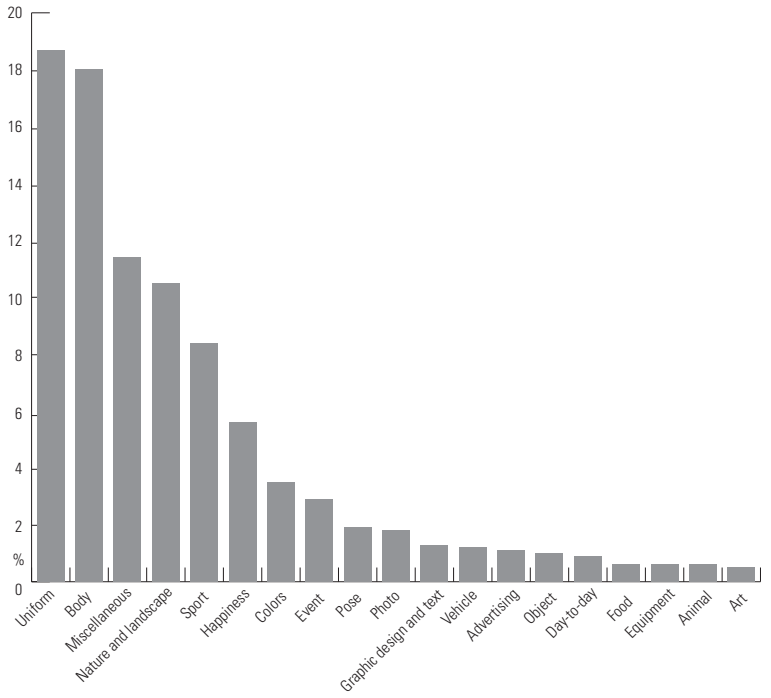
3.1 Google Vision

Google Vision was the first computer vision technology used in this research. This platform contains tools that detect the presence of objects, logos, types of visual composition, photographic techniques, or abstract concepts. Vision AI's artificial intelligence allows automated image analysis, using previously trained models to detect texts, objects, brands and emotions, among other possibilities (Google, 2021), while involving less technical complexity than other competitors and open-source projects.

Three different computational tools were applied to analyze big data image banks. For processing the Google Vision API, a Python script provided by Mintz (2019b) enabled the labeling of images in bulk form, containing the data extracted from the application relationally, allowing the definition of a minimum confidence index of 50 % in the results. With this initial processing, it was possible to generate different visualization types of images to identify possible groupings, similarities and differences between them.

The Google Vision API processing identified approximately 875 000 tags on the image dataset, having determined 3101 unique labels, about ten per image. From the labels'

Figure 1: Sets of labels by frequency



Source: Teixeira (2022).

analysis, it was possible to establish some groups based on those containing a definition considered easy to interpret based on the analysis of the images. More than 70 labels were categorized as "miscellaneous," as these contained images with numerous – often conflicting – contexts and interpretations. For example, the API tagged the label "asphalt" in a pattern emphasizing the differences in perception between humans and machines. Images of roads, athletic racetracks, and concrete floors received the label, ignoring a contextual clue, seeing that track sports events do not use such material, even if they resemble a minor road. Figure 1 presents the 20 identified sets, ordered by the percentage of labels in the image dataset.

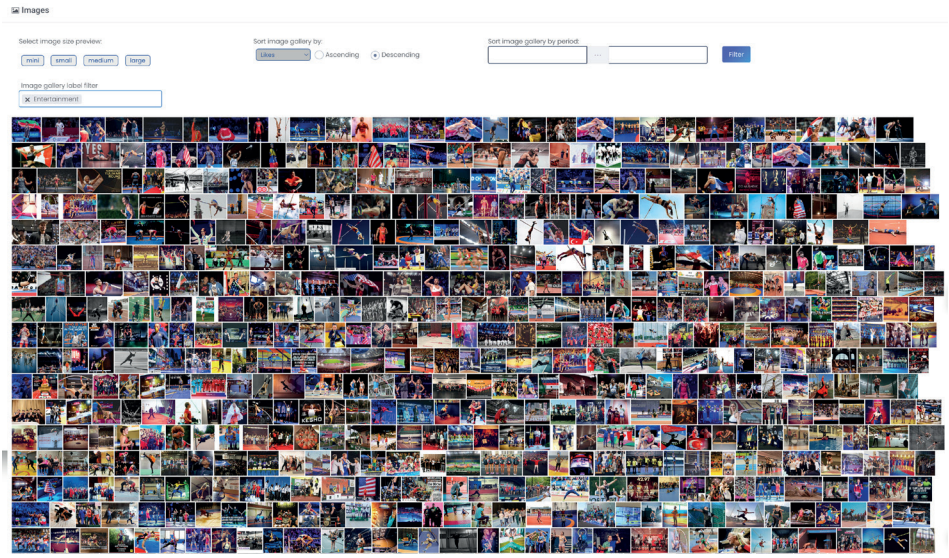
The definition of these groupings conveyed a first understanding of how the computer vision process works. At first, the machine vision's denotative, descriptive logic caught our attention. As Manovich (2020) asserted, computational tools are highly capable of detecting objects in images, even if they cannot grasp context and coherence. In this sense, the detection of clothing items (clustered in the "uniform" category) showed a high frequency of markings, with images containing some sportswear such as shorts, wrestling singlets or bathing suits. Similarly, the identification of parts of the human body was also recurrent.

Sets of categories considered subjective were also verified. In the "happiness" category, subjectivity appeared more superficially because, despite featuring a qualitative character, labels such as "smile" and "happy" were grouped, which tend to be easily identified by the machine. The same happened with the "event" category, where it was possible to verify competitions, sporting events, or training in environments characteristic of these practices, covering labels such as "championship," "competition event," and "entertainment." Despite the greater subjectivity, the "day-to-day" category was valid when analyzing images with labels such as "daytime," "kitchen," "beach," and "field house," among others.

3.2 GENUS

The author developed the second tool, GENUS (Teixeira, 2021). This cloud-based tool employs HTML, PHP, and MySQL database structuring languages. Its main objective is to help understand and interpret found labels and facilitate categorization methods, dynamic filtering, and image visualization.

Figure 2: GENUS interface



Source: Teixeira (2022).

The tool also helps the researcher to organize an assortment of details about the database that can be adapted and edited freely. In this sense, the platform seeks to solve one of the main limitations of public institutional image collections and clusters available online: the limited editing possibilities of categories, metadata, and tags initially inserted by their administrators or database creators, where adaptations and the creation of "new forms of organization" are not allowed (Manovich, 2020, p. 254). Such limitation is problematic because it restricts the construction of new interpretations that stem from the contact of established datasets with innovative techniques and research questions. The GENUS platform made possible the filtering and interactive visualization of the images from the labels in an orderly and distributed way by date, by likes or by users, presenting the mosaic format shown in Figure 2.

This interface automated the data organization, helping the qualitative analysis and the interpretation and understanding of how the automated labeling on the images occurred. Many labels could only be adequately understood when compared to the mosaic images. Images that depended on a specific context or connotation for interpretation, such as subjective meanings, visual gags, language puns, or in-jokes, were not fully labeled. Figure 2 represents a mosaic of 1 328 images selected from the "entertainment" label and arranged in descending order by number of likes. In this example, it is possible to understand better what type of image the computer vision defined as entertainment within the sports-themed image set. The mosaic view allowed inferring that, as verified, Google Vision labels images of events and athletic sports practices as entertainment. Furthermore, different from all other tools, GENUS considers image engagement in the form of likes and comments, thus allowing a better interpretation of the images that present labels with potentially greater engagement capacity.

This example illustrates one of the features of GENUS. The image set was created from the collected data and labeled by image processing with Google Vision, which helped understand image clusters without requiring a manual categorization. The grouping / consolidating of categories can be automated using the labels previously identified by computer vision APIs such as Google Vision. In this sense, it was possible to test interactive, organized visualizations that allowed more possibilities of interpretation of the dataset. This tool seeks to compose the methodological process, offering flexibility in manipulating the views, having an important role in the process of qualitative analysis and interpretation of images, which is often lost with other computational applications.

3.3 PixPlot

Developed by the Yale Digital Humanities Laboratory (DHLab, 2021), PixPlot is an application for dynamically exploring thousands of images using a pre-trained neural network (Szegedy et al., 2015) for image tagging and visualizations generation, made available by an open-source Python code. According to Manovich (2020), this tool helps to recognize emerging subjects in photographs, organizing them by similarity.

PixPlot uses a neural network trained on ImageNet. ImageNet is an ongoing research project, founded in 2009, that provides researchers worldwide with image data for training large-scale object recognition models (ImageNet, 2021). According to its official website (DHLab, 2021), PixPlot processes one of the layers of this pre-trained network to derive descriptions of images in a multidimensional space, transforming them into a map distributed on the computer screen, preserving the clusters, their locations, and an interpretable global form. Therefore, the resulting visualizations project the images grouped by similarity in a two-dimensional way, allowing the user to navigate through the spaces, enlarging, reducing, or selecting image groupings identified by the tool, as well as creating their groupings or selecting and viewing other specific observed sets.

This previous analysis showed evidence of specific patterns used in these visual collections. By closely observing the groupings, it was possible to establish an initial hypothesis, indicating that there seems to be a polarization between the image types in two aspects: athletic sports and "instagrammable," a current term in everyday life, adapted from the concept created by Manovich (2016, p.

73), "Instagramism", with a similar meaning of images with the potential for higher online circulation. Therefore, it is possible to argue that this social media behavior presents points of similarity, with common aspects repeatedly explored by its users. Likewise, the division between sports / competition and non-sports images was evidenced, highlighted in the discussions on athlete branding (Arai, Ko, & Kaplanidou, 2013; Doyle, Su, & Kunkel, 2022; Geurin-Eagleman & Burch, 2016; Smith & Sanderson, 2015).

When segmenting the analyses from the profiles of Olympic athletes, the initial expectation suggested that sports could be the primary theme based on the labels' clustering. However, when separating the categorical groupings, about 40 % of the images contained other labels with sports aspects (those with essentially sports categories: "uniform," "sport," "event," or "equipment"). On the other hand, groupings with similar characteristics to what is constantly shared on Instagram (Arai et al., 2013; Geurin-Eagleman & Burch, 2016; Hu, Manikonda, & Kambhampati, 2014; Manovich, 2017b), such as "body," "photo," "vehicles," "pose," "day-to-day," "food," "animals," "art," "nature," and "landscape" represented almost 60 % of the labels identified in the dataset.

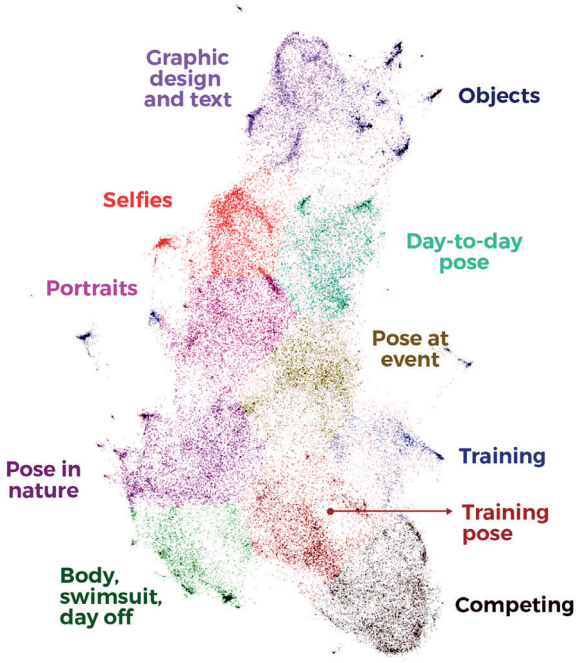
This first finding indicated that the human analytical presence is fundamental for a contextual understanding of what is represented in the images. The use of machines could benefit from going beyond denotative recognition toward a deeper contextual, connotational understanding of the images, especially regarding athletic representations on digital social networks. However, as argued by Manovich (2020), computational processing helps organize data while emerging possible aspects and patterns need to be further explored by human analysts. Therefore, in possession of these first data, diagnoses and hypotheses, intelligent applications were tested, generating mass visualizations of the images in a grouped way to corroborate, complement or contradict the previous textual analysis, adding visual and interpretive aspects intrinsic to the theme.

Before using the labels found in Google Vision image processing to generate visualizations, an additional analysis was proposed using another computer vision application: PixPlot. This approach sought to complement the previous analyses, based on another type of processing (other than the use of Google Vision labeling), allowing comparison and validation of the results obtained, generating an image map of the same image dataset. In addition to creating a map distributing the images geographically by similarity, it automatically identified ten prominent clusters (hotspots) from the processing. By qualitatively observing the images of these sets, it was possible to name them according to the main characteristics identified. For a better static visualization (non-interactive), Figure 3 was created by manipulating the initial output in Photoshop and coloring the clusters.

The position of the groups on the map (Figure 3) displayed a coherent logic. At the top, the images that involved aesthetic elements clustered where the presence of the athletes was not the central visual aspect identified. Graphic design and text featured posts with graphic manipulation, whether inserting texts, modifying photographs or creating some graphic layout. The "objects" set contained images with non-human elements (Latour, 1993), which may be related, in some way, to the sport (such as medals, trophies, equipment, or uniforms) or not (animals, drinks, food, among others). These two sets were easily identified because they deal with objective topics, with visual elements that machines can detect with a low margin of error.

The other groups presented a more significant definition of complexity because they dealt with images with subjective, interpretative themes and contents that are difficult to detect. Despite the difficulty, it was possible to identify general patterns in each set. Following the map (Figure 3), the "selfies" and "portraits" sets contained images focused on people, usually with a close-up, emphasizing the athlete, with the framing being its central differential aspect. Selfies were not necessarily restricted as those photographs taken by the photographers themselves, but by those that presented an angle and framing that visually referred to this aesthetic. This logic is similar to the one used in the Selfiecity project (Manovich & Tifentale, 2015), in which selfies were identified as images with frames close to the close-up as used in film productions (Gerbase, 2012). Portraits, on the other

Figure 3: PixPlot: general analysis (sets)



Source: Teixeira (2022).

hand, showed images with more open framing, reminiscent of cinematographic medium shots (Gerbase, 2012), emphasizing the face to the detriment of the body. Groups of images defined as "poses" were found, where the open shots, in addition to showing some behavior or body expression, permitted a more subjective and contextual analysis of what was being represented, advancing beyond the athlete's image exclusively.

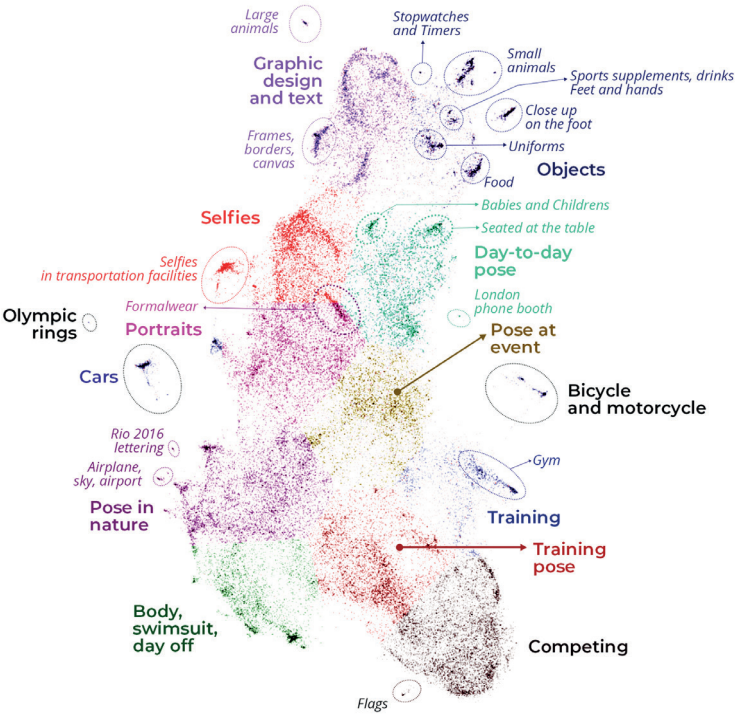
"Day-to-day poses" presented images of athletes performing activities, for the most part, outside of sports practice, showing scenes of their daily routine, such as sitting on a couch or in other environments, sometimes accompanied by other people, such as family and children. These images, in general, represented the life of athletes external to the practice of sports and their profession. This set is located geographically close to the "objects" set as we found, many times, some element of this group was identified in the photos. "Poses in nature," on the other hand, appeared next to the "portraits" on the opposite side of the map, considering that most of these images were representations of athletes in some place, usually outdoors, with lots of green or water, standing in front of the camera, with a tangential relationship with sports practice. In the same way, the images of the "body bathing suit" were left off. This set, a little wider, was more subjective. However, most of the images invariably showed a combination of the athletes' bodies, with few clothes (sometimes bathing suits), posing in environments with water, for example, a beach – justifying the proximity to the group of poses in nature – or a pool, emphasizing moments of relaxation (equally unrelated to sports practice), thus representing, in the same way, an idea of a vacation or a break. These aspects approach the dimension "offstage," proposed by Doyle et al. (2022).

Finally, the last four thematic sets represented the athletes' sports practice and activity. This category meant poses at events, that is, images of athletes, usually accompanied by coaches or colleagues at sporting events in general, ranging from competitions or awards to training, or unofficial events. They were, for the most part, representations of athletes related, even if superficially, to some

Figure 4: PixPlot: general analysis (examples of grouped sets)

Source: Teixeira (2022).

Figure 5: PixPlot: general analysis (subsets)



Source: Teixeira (2022).

event associated with their sports activity. "Training poses" contained photographs of athletes in gyms or athletics tracks, usually alone, not wearing "formal" uniforms (with a competition aspect), without many people around or without some element that would label the activity as official / institutional. This thematic group represented specific types of training, that is, those directly linked to the modality practiced by the athlete.

On the other hand, "training" featured more generic physical activities, primarily in gyms, where the athlete performed a preparation simulating their sport or simply a physical exercise, such as bodybuilding or weightlifting. In the "competition" set, images showed athletes in uniform, practicing physical activities in specific environments (gyms or stadiums), usually with an audience or some infrastructure that indicated that it was a possible sporting event or organized competition.

Figure 4 shows examples of posts that illustrate the types of images in each set. It is essential to highlight that these groupings indicate specific interpretive patterns, in which, invariably, there are divergent images that, from the qualitative visual analysis, may only fit into a variety of sets. However, it was possible to identify patterns and common elements in the groups that indicate trends, allowing the naming of their themes.

In addition to the groupings automatically identified by the tool, the process of interactive visual analysis on the platform allowed for finding new thematic sets and their development into subsets. By manipulating the image map in Photoshop and coloring the clusters to differentiate them, it was possible to verify some dense point groupings. This was a clue indicating that the more similar the images, the greater their concentration, as shown in Figure 5, which allowed the discovery of unexpected interpretations. PixPlot provides an image selection and visualization tool in a specific area, thus allowing qualitative visual interpretation and the identification of new themes in the groups.

Among the non-sports explored aspects, elements such as animals, food, means of transport (cars, bicycles, motorcycles, planes, maritime transport) and tourist monuments, among others, are worth mentioning. Along with the sporting elements, timers, uniforms, the Olympic rings, the Rio 2016 sign and the gym, among others, were found.

3.4 Image Network Plotter

The last visualization method chosen was the image network plotter algorithm from the analysis of optical networks, previously explored by one of the authors (Teixeira & Silva, 2020). In this computational methodological approach, the labels obtained by Google Vision are considered vertices of a bi-modal network through tools and metrics of network analysis and graph theory embedded in the Gephi software (Bastian, Heymann, & Jacomy, 2009). Their approximations or distances can associate each image through the co-occurrence (or absence of co-occurrence) between the labels, where two images with the same label, for example, are connected. By using network layout algorithms such as ForceAtlas2, it is possible to distribute the network nodes – images and labels. These algorithms are responsible for spatially organizing the network data (Jacomy, Venturini, Heymann, & Bastian, 2014), generating a distribution of images and labels according to the adequate connections between them. Thus, it was possible to plot these referenced images in an image graph, in a two-dimensional space, generated from the Image-network plotter script (Mintz, 2019a), working with the transposition of textual information to a visual network that amplifies the analyses.

From the interpretation of Google Vision labels and the PixPlot image map, the last visualization was performed using image graphs from Google Vision results, aiming to validate and complement the previous findings. From the initial graph derived from Google Vision processing, presented in the methodological proposal by Mintz (2019b), it was possible to infer the distribution percentage of each group detected, which allowed us to find eight clusters in the network. The groupings were automatically identified, having their visual validation for identifying and defining the thematic groups. When considering "sports" and "gym" as classes with a direct sports relationship, less than 30 % of the athletes' images referred more objectively to sports. Despite a different format from PixPlot, the sets found in the graph presented similar characteristics to those previously investigated. Figure 6 shows the developments identified from plotting the images within the graph.

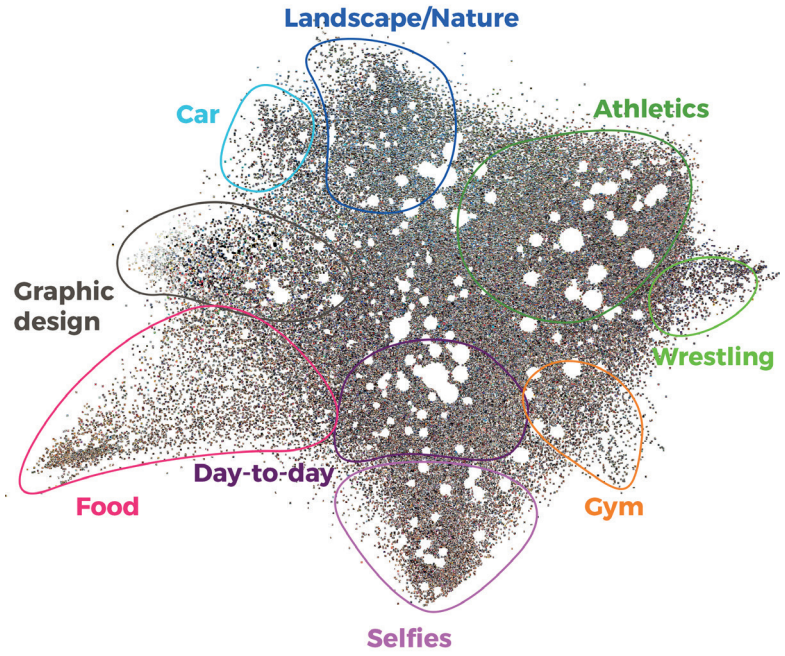
Unlike the map generated by PixPlot (Figure 3), the framing of the photos did not appear as a relevant criterion in the organization and distribution of the graph. When considering the Google Vision labels, the theme established from the types of content in the images determined the groupings. However, the overall result showed similar patterns to the previous analysis. The sets considered sports ("athletics," "wrestling," and "gym") were minor compared to the "Instagrammable" and non-sporting aspects, such as "selfies," "daily life," "food," "graphics," "cars," "landscapes," and "nature."

Therefore, the general analysis indicated the existence of two distinct visual lines: a) "athletic," focusing on a predominantly sports aesthetic, with images of competitions, training, exercises, or sports practices; b) "Instagrammable / non-sports," in which Instagram reinforced its own culture, which seemed to be replicated in the behavior of athletes' posts on the platform, with images of selfies, food, lifestyle, and other aspects about the Instagram culture. In the end, it was possible to infer an overlapping trend of "Instagrammable aspects" about athletic aspects, reinforcing similar results found in other studies that relate sports, athletes, and Instagram (Arai et al., 2013; Doyle et al., 2022; Geurin-Eagleman & Burch, 2016; Smith & Sanderson, 2015).

4 Final considerations

From the methodological point of view, the presented approach proved adequate, elucidating a series of developments and reflections about a large volume of images. Cultural Analytics, together with DH, combined with methods of analysis and computer vision, was considered efficient for the orga-

Figure 6 : Image graph: general analysis (sets)



Source: Teixeira (2022).

nization and structuring of the visual corpus of the research. As expected, the complexity and subjectivity in the published images demand a qualitative methodological strategy. Based on the structuring of the clusters, a series of subsequent interpretations were made possible.

Social media occupied the role of digital media, causing the productive spectrum and its emitters to be amplified again. Every individual can now produce and disseminate content, achieving the same scalability previously only possible by the traditional mass media. Digital platforms have created an environment that allows for different possibilities, such as massive reach for all users, expressive financial gains and integration between bodies, institutions, committees, companies, people, and athletes, among many other parts.

Athletes were inserted in this context, where they also started to produce, share and commercialize their images for a specific niche and in a generalized and globalized way. The current Olympic situation thus constitutes a feedback system. The media coverage of the Olympic Games carried out by traditional media has amplified the reach of the athletes' image, boosting the growth and relevance of their official profiles on social media. The extent of this repercussion depends on the competitive results obtained during competitions, broadcast on a global scale, in which winning athletes have potentially more opportunities for media exposure than non-winners. The growth of digital profiles, aided by this media exposure and achievements, can be pointed out as one of the primary sources of monetary resources for modern Olympic athletes (Arai et al., 2013; Geurin, 2017; Geurin & McNary, 2021; Hayes, Filo, Geurin, & Riot, 2020).

In this context, Instagram is one of the leading social media platforms used by athletes to reflect their athletic and professional image. This platform combines a series of aspects and functionalities that make it a catalyst for athletic beauty (Gumbrecht, 2006), mainly through the exploration of the aesthetic visuality of the content, and, in the

case of the present analysis, of sport and the Olympic Games. In this way, analyzing the posted images and their patterns allows a deeper understanding of how Olympic athletes use this digital ecosystem in the media.

In the visual analysis of the collected images, a primary division was observed between the types of images published by the athletes, reinforcing relevant aspects pointed out in the previous discussion. Most of the photographs posted (60 %) referred to non-sporting, informal, offstage, or "Instagrammable" aspects, that is, those that highlighted common characteristics of the visual culture disseminated on Instagram, such as selfies, lifestyle, leisure, travel, and food, among others. Images of a sports nature, considered a central aspect of the research, had a lower frequency of publications, using topics such as competitions, training, exercises, and sports practices in general. However, by emphasizing a predominantly athletic visual representative aesthetic, these specific contents allowed a series of other considerations.

The sports visual representations could be divided into two main aspects: training and competition. In the representations that emphasized the training, the images showed a bias that highlighted the exposure of the athletes' muscular bodies, whether in gyms or other informal training environments. These images also tended to accentuate an idea related to the effort and dedication of athletes, expressing a concept close to Hellenic sports and agonistic idealism. On the other hand, the images that represented competitions, with official and institutional bias characteristics, often presumed a relationship between happiness, conquest, victory and commercial aspects. The visual sporting representations denote an economic and commercial character along with the victory. Images on Instagram are the means that justify an end, their main purpose being, in most cases, commercial results.

The tools presented contribute in different ways to this process. Google Vision offered labeling based on a model that prioritizes image annotation, being fundamental in the detection and description of objects and elements present in the images. This characteristic helps the researcher to make inferences according to certain types of recurring objects and other visual elements. By associating the labeling of diverse objects in a photo with the analytical insight of the researcher, a range of possible inferences was found. GENUS, in this sense, helps in interpreting the labels and filtering image types associated with their detection. Thus, it is up to the researcher to qualitatively analyze the results derived from such filtering. Finally, the computer visualization tools tested in this research (PixPlot and Image Network Plotter) order and structure this tangle of data-related images, visually organizing the image elements and allowing an analysis of the whole from the parts, making it possible to determine potential groupings. Such clusters need dedicated interpretation and individual analysis by the researcher, as well as their internal relationship and association with the other clusters. Again, the qualitative bias and the possibilities of filtering presented by GENUS showed to be critical.

From the point of view of usefulness, both visualization tools proved to be valid. Two different processes were used to create the graphs to compare and validate the methods. Despite being different, they presented similar results, which was considered a positive point, especially considering that the processed image dataset was the same. However, despite the similarities, the models had some matching points, indicating the use of both. PixPlot stood out for being a faster and easier-to-navigate application, with the great advantage of navigation tools within the graph, such as zoom, selecting images from a pre-selected area, and other visualization modes. In addition, computer vision processing considered aspects of the image, such as framing and angle of the photos, something that did not happen in the Image Network Plotter.

On the other hand, the second tool, using information from Google Vision, allows the export of the list of labels used, expanding its use on other platforms, as was the case with GENUS, for example. From these data, together with the graph processing, it was possible to have percentage quantitative information on the size of each cluster or label within the database. In addition, identifying objects and context within the images proved

to be more satisfactory than the first tool, with more depth, variety, and assertiveness in the labels.

Among other learning experiences, the computational capacity for data processing should be considered part of the research planning and design. When working with images, the processing requirement is considerable. When viewing works such as the Selfiecity project (Manovich & Tifentale, 2015), where millions of images were analyzed, a bank with less than 100 000 photographs does not depend on a very arduous computational task. However, this hypothesis needed to be revised when analyzing the present research. At various times throughout the process, it was necessary to establish new cuts in the dataset size to reduce the computational demand. It is worth emphasizing that the results presented ran on "conventional" computers, widely available to consumers and researchers alike.

The main contribution of this article is to offer a methodological path where it is possible to explore and relate some of the numerous computational analysis tools currently available. By showing, comparing, and discussing some of these methods, it was possible to determine the importance of computational tools in ordering and structuring large volumes of images.

Manovich (2020) emphasizes the limitations of relying solely on the quantitative insights provided by computer vision for image analysis. While purely quantitative methods can identify patterns derived from simple image attributes (such as color palettes, size, and metadata) or more intricate aspects through computer vision techniques, a comprehensive understanding of the subject under scrutiny necessitates a qualitative approach. This approach often requires the involvement of one or more human researchers, despite recent advancements in artificial intelligence. To illustrate, discerning which sports brands were most frequently tagged in each image dataset can contribute to a wider discussion on national team funding and the economics of sport – a level of analysis that, for the time being, no algorithm can accomplish single-handedly.

While acknowledging the necessity of a complementary approach, researchers in this field must also recognize that machine learning tools often lack transparency regarding their use of training datasets. Denton, Hanna, Amironesei, Smart, and Nicole (2021) argue that machine learning systems often miss marginalized communities due to the underrepresentation or misrepresentation of these groups in the data upon which these systems are built, and that a straightforward technological problem-solving mindset is unable to address these issues. They advocate the need to critically examine the creation and use of machine learning datasets, considering both their infrastructural, workforce and symbolical – genealogical aspects. Crawford and Paglen (2021) echo this discussion, emphasizing the need to scrutinize the training sets used in artificial intelligence. They particularly stress the importance of examining datasets containing images of people, as these datasets, in turn, shape our understanding and perceptions. In the research on Olympic images that formed the basis of this text such complex issues were not considered on its forefront but fomented the use of more than one tool to organize the visualizations. The reasoning was not to rely naively and automatically on the results brought by only one tool, but cross-check it with competing ones. Besides that, the observation of Olympic sports contents as seen by athletes' Instagram feeds featured in itself a more egalitarian approach that sits on the cultural mainstream.

The transparency of training datasets and its implications has become a point of discussion within creative communities following the rise of generative image AI platforms like OpenAI's Dall-E and Stability AI's Stable Diffusion, which became popular in late 2022. These platforms have been criticized for allegedly using training data sourced from unauthorized materials created by artists and owned by stock image dealers. While this topic may seem tangential to the focus of this text, it underscores the point that no dataset or training model achieves the utopian vision depicted by Argentine writer Jorge Luis Borges in his short story, "The Aleph", first published as a short story in 1945 and later compiled in an eponymous book. In Borges's story (Borges, 2000), the Aleph is a point in space that contains all

other points, allowing anyone who investigates it to see everything in the universe from every angle simultaneously, without distortion, overlap, or confusion. This concept can be seen as an idealized, unbiased model for analysis – a utopia that, regrettably, remains within the realm of fiction.

After all, Digital Humanities (DH) methodologies strive to progressively enhance automated categorization and analysis with manual, qualitative, and interpretive perspectives. This synergistic strategy leverages both the expansive capabilities of digital tools and the unique potential for insightful discoveries inherent to human researchers.

Acknowledgements

The authors would like to acknowledge that the research and content developed for the doctoral thesis of Dr. Teixeira, under the supervision of Dr. Tietzmann, formed the basis upon which this article was written. Mr. Teixeira developed his doctoral research with funding by the Brazilian agency CAPES – Coordination of Superior Level Staff Improvement. Also, we express our sincere gratitude for the invaluable insights and observations made by the editorial team during the publishing process.

Conflict of interest

The authors declare no conflict of interest.

References

Arai, A., Ko, Y. J., & Kaplanidou, K. (2013). Athlete brand image: Scale development and model test. *European Sport Management Quarterly, 13*(4), 383–403. https://doi.org/10.1080/16184742.2013.811609

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media, 3*(1), 361–362. https://doi.org/10.1609/icwsm.v3i1.13937

Berry, D. M. (2012). *Understanding Digital Humanities*. Basingstoke, UK: Palgrave Macmillan.

Borges, J. L. (2000). *The Aleph.* London, UK: Penguin.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012). *Digital_humanities*. Cambridge, MA: MIT Press.

Crawford, K., & Paglen, T. (2021). Excavating AI: The politics of images in machine learning training sets. *AI & SOCIETY, 36*, 1105–1116. https://doi.org/10.1007/s00146-021-01162-8

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society, 8*(2), 1–14. https://doi.org/10.1177/20539517211035955

DHLab. (2021). PixPlot. *Website*. Retrieved from https://dhlab.yale.edu/projects/pixplot/

Doyle, J. P., Su, Y., & Kunkel, T. (2022). Athlete branding via social media: Examining the factors influencing consumer engagement on Instagram. *European Sport Management Quarterly, 22*(4), 506–526. https://doi.org/10.1080/16184742.2020.1806897

Eco, U. (2010). *On beauty: A history of a Western idea*. London, UK: MacLehose Press.

Evans, L., & Rees, S. (2012). An interpretation of Digital Humanities. In D. M. Berry (Ed.), *Understanding Digital Humanities* (pp. 21–41). London, UK: Palgrave Macmillan.

Gerbase, C. (2012). *Cinema: Primeiro Filme: Descobrindo, Fazendo, Pensando* [Cinema: First film: Discovering, making, thinking]. Porto Alegre, Brazil: Artes e Ofícios.

Geurin, A. N. (2017). Elite female athletes' perceptions of new media use relating to their careers: A qualitative analysis. *Journal of Sport Management, 31*(4), 345–359. https://doi.org/10.1123/jsm.2016-0157

Geurin, A. N., & McNary, E. L. (2021). Athletes as ambush marketers? An examination of Rule 40 and athletes' social media use during the 2016 Rio Olympic Games. *European Sport Management Quarterly, 21*(1), 116–131. https://doi.org/10.1080/16184742.2020.1725091

Geurin-Eagleman, A. N., & Burch, L. M. (2016). Communicating via photographs: A gendered analysis of Olympic athletes' visual self-presentation on Instagram. *Sport Management Review, 19*(2), 133–145. https://doi.org/10.1016/j.smr.2015.03.002

Gil, A. C. (2008). *Métodos e técnicas de pesquisa social* [Methods and techniques for social research] (6th ed.). São Paulo, Brazil: Atlas.

Google. (2021). Google Vision. *Website*. Retrieved from https://cloud.google.com/vision/docs

Gumbrecht, H. (2006). *In praise of athletic beauty*. New York, NY: Harper Perennial.

Hayes, M., Filo, K., Geurin, A., & Riot, C. (2020). An exploration of the distractions inherent to social media use among athletes. *Sport Management Review, 23*(5), 852–868. https://doi.org/10.1016/j.smr.2019.12.006

Hochman, N., Manovich, L., & Chow, J. (2013). *Phototrails*. Retrieved from http://phototrails.net/

Hu, Y., Manikonda, L., & Kambhampati, S. (2014). What we Instagram: A first analysis of Instagram photo content and user types. In Association for the Advancement of Artificial Intelligence (Ed.), *Eighth International AAAI Conference on Weblogs and Social Media* (pp. 595–598). Palo Alto, CA: The AAAI Press.

ImageNet. (2021). ImageNet. *Website*. Retrieved from https://image-net.org

International Olympic Committee. (2020). Tokyo 2020. *Website*. Retrieved from https://olympics.com/en/olympic-games/tokyo-2020

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE, 9*(6), 1–12. https://doi.org/10.1371/journal.pone.0098679

Latour, B. (1993). *We have never been modern*. Cambridge, MA: Harvard University Press.

Liu, A. (2012). The state of the digital humanities: A report and a critique. *Arts and Humanities in Higher Education, 11*(1–2), 8–41. https://doi.org/10.1177/1474022211427364

Manovich, L. (2009). *Cultural Analytics: Visualizing cultural patterns in the era of "more media"*. Los Angeles, CA. Retrieved from http://manovich.net/content/04-projects/063-cultural-analytics-visualizing-cultural-patterns/60_article_2009.pdf

Manovich, L. (2012). How to compare one million images? In D. M. Berry (Ed.), *Understanding Digital Humanities* (pp. 249–278). London, UK: Palgrave Macmillan.

Manovich, L. (2013). Museum without walls, art history without names: Visualization methods for humanities and media studies. In C. Vernallis, A. Herzog, & J. Richardson (Eds.), *The Oxford Handbook of Sound and Image in Digital Media* (pp. 251–278). Oxford, UK: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199757640.013.005

Manovich, L. (2016). *Instagram and contemporary image*. Retrieved from http://manovich.net/index.php/projects/instagram-and-contemporary-image

Manovich, L. (2017a). *Automatizando a estética: inteligência artificial e cultura das imagens* [Automating aesthetics: Artificial intelligence and image culture]. *Esferas, 11*(316), 119–126.

Manovich, L. (2017b). *Visual semiotics, media theory, and Cultural Analytics*. Retrieved from http://manovich.net/content/04-projects/103-visual-semiotics/manovich_visual_semiotics.pdf

Manovich, L. (2018). Can we think without categories? *Digital Culture & Society, 4*(1), 17–28. Retrieved from https://digicults.org/files/2019/11/dcs-2018-0103.pdf

Manovich, L. (2020). *Cultural analytics*. Cambridge, MA: MIT Press.

Manovich, L., & Tifentale, A. (2015). Selfiecity: Exploring photography and self-fashioning in social media. In D. M. Berry & M. Dieter (Eds.), *Postdigital aesthetics: Art, computation and design* (pp. 109–122). New York, NY: Palgrave Macmillan. Retrieved from http://manovich.net/index.php/projects/selfiecity-exploring

Mintz, A. (2019a). Image-network plotter (Python script). *Website*. Retrieved from https://github.com/amintz/image-network-plotter

Mintz, A. (2019b). Memespector (Python script). *Website*. Retrieved from https://github.com/amintz/memespector-python

Smith, L. R., & Sanderson, J. (2015). I'm going to Instagram it! An analysis of athlete self-presentation on Instagram. *Journal of Broadcasting and Electronic Media, 59*(2), 342–358. https://doi.org/10.1080/08838151.2015.1029125

Svensson, P. (2016). *Big digital humanities: Imagining a meeting place for the humanities and the digital*. Ann Arbor, MI: University of Michigan Press.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition*

(CVPR) (pp. 1–9). Boston, MA. https://doi.org/10.1109/CVPR.2015.7298594

Teixeira, C. R. G. (2021). GENUS. *Website*. Retrieved from https://www.ocarlosteixeira.com.br/genus

Teixeira, C. R. G. (2022). As representações visuais da Beleza Atlética Olímpica: da Antiguidade ao Instagram [The visual representations of Olympic beauty: From antiquity to Instagram] (Doctoral thesis). Pontifical Catholic University of Rio Grande do Sul, Porto Alegre.

Teixeira, C. R. G., & Silva, T. R. (2020). Comunicação, Esportes e Visão Computacional: explorando a visualidade dos Jogos Olímpicos no Instagram [Communication, sports and computer vision: Exploring the visual aesthetics of the Olympic Games on Instagram]. *Rizoma, 8*(2), 110–127. https://doi.org/10.17058/rzm.v8i2.14365

United World Wrestling. (2021). World Olympic Games qualifier entries results. *Website*. Retrieved from https://uww.org/article/world-olympic-games-qualifier-entries

World Athletics. (2021). Road to Olympic Games 2020. *Website*. Retrieved from https://www.worldathletics.org/stats-zone/road-to/7132391