PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

BRUNA MARTINI DALMORO

ASSESSING PERCEPTUAL DATA IN IMAGES

A COMPUTATIONAL AESTHETICS APPROACH

Porto Alegre
2021

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFICAL CATHOLIC UNIVERSITY OF RIO GRANDE DO SUL
SCHOOL OF TECHNOLOGY
COMPUTER SCIENCE GRADUATE PROGRAM

ASSESSING PERCEPTUAL DATA IN IMAGES: A COMPUTATIONAL AESTHETICS APPROACH

BRUNA MARTINI DALMORO

Master Thesis submitted to the Pontifical Catholic University of Rio Grande do Sul in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Prof. Dr. Soraia Raupp Musse

Porto Alegre
2023

Ficha Catalográfica

D148a Dalmoro, Bruna Martini

Assessing perceptual data in images : a computational aesthetics approach / Bruna Martini Dalmoro. – 2021.
76 p.
Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Soraia Raupp Musse.

1. Computational Aesthetics. 2. Perception. 3. Image analysis. I. Musse, Soraia Raupp. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

BRUNA MARTINI DALMORO

ASSESSING PERCEPTUAL DATA IN IMAGES: A COMPUTATIONAL AESTHETICS APPROACH

This Master Thesis has been submitted in partial fulfillment of the requirements for the degree of Master in Computer Science, of the Computer Science Graduate Program, School of Technology of the Pontifical Catholic University of Rio Grande do Sul

Sanctioned on August 13th, 2021.

COMMITTEE MEMBERS:

Prof. Dr. Marcelo Walter (Inf/UFRGS)

Prof. Dr. Isabel Harb Manssour (PPGCC/PUCRS)

Prof. Dr. Soraia Raupp Musse (PPGCC/PUCRS - Advisor)

I dedicate this work to every Thursday that will come.

“Nunca pinto sueños o pesadillas. Pinto mi propia realidad.”
(Frida Kahlo)

ACKNOWLEDGMENTS

I would especially like to thank Prof. Dr. Soraia Musse, my advisor, for all her work and dedication to her students, and for her contagious love of research and science. Special thanks to our colleagues Victor Araújo, Greice Dal Molin and Felipe Nomura, and all VHLab colleagues, for their support and for the moments where we shared so many ideas. I would also like to thank the professors and the team of PPGCC at PUCRS, which do a great job for the education of their students, regardless of adversity.

I thank Prof. Dr. Charles Monteiro, for having collaborated in the development and discussion of ideas related to art history.

I thank Prof. Dr. Isabel Manssour for having accompanied this work in the many deliveries during the master's degree (in the courses, PEP, and SA), and for having accepted to be part of the jury of the presentation. I thank Prof. Dr. Marcelo Walter for being part of the jury of this work.

For his patience, understanding, and all the support received, I thank my partner Guilherme Mansson, as well as my parents Dilvo and Lourdes Dalmoro and my sister Júlia Dalmoro. You have all my heart.

To my co-workers and friends, especially Daniela, Gabriel, and the Data Science Chapter, I thank you for welcoming me both in my moments of excitement and anguish. To Mariuá, I thank you for asking so many questions that led me to several important answers.

I thank UOL EdTech for financially supporting this research.

AVALIANDO DADOS DE PERCEPÇÃO EM IMAGENS: UMA ABORDAGEM DE ESTÉTICA COMPUTACIONAL

RESUMO

A percepção humana é o processo que captura estímulos físicos mensuráveis e os converte em compreensão do mundo. O estudo da percepção humana, no que tange estímulos visuais, é uma ampla área de pesquisa que tem sido estudada de forma multidisciplinar. Com os diversos avanços da computação e da capacidade de processamento e análise de imagens, a percepção humana passou a ser estudada computacionalmente. Uma das áreas que aborda essa discussão é a área da estética computacional, um subcampo da visão computacional, que visa pesquisar métodos computacionais que tomem decisões estéticas semelhantes às dos humanos. Uma das aplicações da estética computacional hoje é a previsão de classificações de imagens e vídeos e sua popularidade. Outra área muito explorada pela estética computacional é a área de análise de artes e pinturas. Para construir esses algoritmos, características (*features*) visuais extraídas de imagens são usadas como forma de descrever seu conteúdo. A interpretabilidade dessas *features* é de grande valor para áreas como a estética empírica e experimental, assim como para gerar insights a partir dos resultados encontrados. No presente trabalho, exploramos três problemas diferentes com a abordagem de estética computacional. No primeiro problema, desenvolvemos um modelo para prever a popularidade de vídeos postados no Facebook usando um conjunto de dados de *features* visuais. No segundo problema, usamos também *features* visuais e informações de categoria de imagem (animação ou live-action) para criar um sistema de recomendação de filmes, baseado em conteúdo. No terceiro problema, propomos uma metodologia para identificar e sugerir relações de influência entre pintores a partir de *features* visuais extraídas das faces de suas obras de arte. Nossos principais objetivos neste trabalho são: explorar diferentes problemas envolvendo diferentes tipos de imagens do ponto de vista da estética computacional; usar apenas *features* visuais para resolver problemas como forma de testar o poder e a utilidade dessas informações em diferentes aplicações; e usar apenas *features* visuais interpretáveis para gerar insights sobre a área de estética e áreas relacionadas. Os resultados encontrados neste trabalho sugerem que as *features* visuais, extraídas de imagens e vídeos, são recursos importantes para a solução dos problemas propostos. Além disso, os resultados indicam que as metodologias propostas são promissoras em tentar responder matematicamente em acordo com a percepção humana, conforme pretende a área de estética computacional, além de permitir gerar *insights* para pesquisas estéticas quando as *features* visuais são interpretáveis.

Palavras-Chave: computação estética, percepção, features visuais.

ASSESSING PERCEPTUAL DATA IN IMAGES: A COMPUTATIONAL AESTHETICS APPROACH

ABSTRACT

Human perception is the process that captures measurable physical stimuli and converts them into understanding information about the world. The study of human perception, regarding visual stimuli, is a wide area of research that has been studied in a multidisciplinary way. With the various advances in computing and the capacity for processing and analyzing images, human perception began to be studied computationally. One of the areas that addresses this discussion is the area of computational aesthetics, a subfield of computational vision, which aims to research computational methods that can provide aesthetic decisions similar to those of humans. One of the applications of computational aesthetics today is the prediction of image and video ratings and their popularity. Another area much explored by computational aesthetics is the area of art and painting analysis. To build these algorithms, visual features drawn from images are used as a way to describe their content. The interpretability of these features is of great value for areas such as empirical and experimental aesthetics, as well as for generating insights from the found results. In the present work, we explore three different problems contextualized in the computational aesthetic areas. In the first problem, we developed a model to predict the popularity of videos posted on Facebook using a dataset of visual features. In the second problem, we also use visual features and image category information (animation or live-action) to create a content-based movie recommendation system. In the third problem, we propose a methodology to identify and suggest influencing relationships between painters based on visual features extracted from the faces of their artworks. Our main objectives in this work are: to explore different problems involving different types of images from the point of view of computational aesthetics; to use only visual features to solve problems as a way to test the power and usefulness of this information in different applications; and to use only interpretable visual features to generate insights into the area of aesthetics and related areas. The results found in this work suggest that visual features, extracted from images and videos, are important resources for solving the proposed problems. In addition, the results indicate that the proposed methodologies are promising in trying to answer mathematically, in accordance with human perception, as intended by the area of computational aesthetics, questions about perception analysis. In addition, our methods allow to generate insights for aesthetic research when visual features are interpretable.

Keywords: computational aesthetics, perception, visual features.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1. INTRODUCTION

Perception is the process that captures external physical stimuli and converts them into awareness of the world around us, connecting our mind with reality. All stimuli received can be measured in some way, but what is actually perceived by humans is purely psychological [79]. Thus, perception is more than just measures of external physical stimuli, but the mental construction of understanding about the world in a complex way. The study of human perception of different visual stimuli is a broad research area that has been studied multidisciplinary, as understanding what the perceptual, cognitive, aesthetic, and emotional response to an image or object brings benefits to the areas of design, photography, advertising, artists and painters, among others [47] [12].

With several advances in computing for extracting information from images and videos at different levels, in addition to the development of algorithms and increased capacity to process and analyze images in general, human perception began to be studied computationally. More specifically, one of the areas that have been developing recently is the area of computational aesthetics, a subfield of computer vision, which aims to research computational methods that make aesthetic decisions similar to humans, for different applications [43].

The main challenge in the field of computational aesthetics is to identify the human aesthetic perception of images through visual information, contained in the images themselves. In general, this task is translated into the generation of a rating or classification for the images based on their aesthetics, which can become more complex by exploring other types of applications, such as prediction of the popularity of a video or image on social networks [49] [78], style classification from a artwork [65], identification of authorship of paintings [80], and even generative art [33] [10].

One way to build these systems is using information extracted directly from images, called visual features. Visual features can describe information from the low-level, exploring information from each pixel, to the semantic image information [83]. Another way to obtain meaningful information from the image is by extracting visual features with the aid of deep learning. These last approach generally generate good results in aesthetic evaluation, but they lose interpretability, which is of great value to areas such as empirical and experimental aesthetics, to generate insights based on the results found [12].

Much of the work in computational aesthetics, currently, is focused on predicting image and video ratings. With more and more social networks including images and videos that have user feedback, companies have been interested in generating algorithms that identify more attractive posts [47]. As aesthetic perception is so complex to assess, it is necessary to find ways to quantify these responses. That's why these feedbacks have been increasingly important. According to Joshi et al. [47], the main advantage of using community feedback is that the subjective patterns must be captured as a whole, i.e., they are using the “wisdom of the crowd”.

Another area that has been much explored is the analysis of paintings. Identification of authorship of paintings, style prediction, and influence analysis have been recurrent themes in work in the field of computer science and obtained interesting results [12] [73]. However, the lack of cooperation between computer science and art history has generated quite pertinent questions regarding the applicability of these work.

In this work we explore three different problems from the perspective of computational aesthetics. In the first problem, using a dataset of visual features extracted from videos posted on Facebook, we developed a model to predict the popularity of videos. In the second problem, we use visual features and video image category information (animation or live-action) to create a content-based movie recommendation system. Finally, in the third problem, we explore the field of arts, creating a methodology to identify and suggest influencing relationships between painters based on visual features extracted from the faces of their artworks. Thus, the main objectives of this work are:

- **To explore different problems involving different types of images from the perspective of computational aesthetics**. Throughout this work we explore problems related to videos with real images, live-action and animation/computer graphics movies, as well as painting images, with the objective of solving issues related to aesthetic perception, such as popularity prediction, movie recommendation and identification of influences;

- **To use only visual features to solve problems as a way to test the power and usefulness of this information in different applications**. The problems we address in this work, in general, can be dealt with in different ways, for example, using temporal features collected after the video is published to predict the popularity or using the movie preferences of similar users to make recommendations. However, to assess the power and usefulness of the information extracted from the images, we only use visual features in the development of our work;

- **To use only interpretable visual features to generate insights for the area of aesthetics and related areas**. All information utilized in our work is interpretable. The interpretability of visual features is associated with the capacity to comprehend the representation and correlation of these features with perceptual characteristics found in images, such as color information, the presence of text and faces, scene transitions, details about facial proportion and symmetry, position, gaze, among others. By employing interpretable visual features, one can enhance their understanding of the obtained results and derive meaningful insights from them.

The results found in this work suggest that visual features extracted from images and videos are an important input for solving the proposed problems, indicating that the computational aesthetic approach helps to achieve good results, and can also generate insights for aesthetic research when visual features are interpretable.

This work is organized as follows. In Chapter 2 we will explore the development of computational aesthetics, and how visual features are important in this area of research. In Chapter 3 we explore the first problem, the prediction of popularity of videos posted on social networks, expanding the concept of aesthetic exploration from a singular image to a sequence of images related by time. In Chapter 4 we continue to explore the scope of videos, but now considering different types of images, to make movie recommendations based on users' aesthetic tastes. In Chapter 5, we enter the field of arts and explore the complex task of identifying influence between artists, based on the way they represent the faces present in their works-of-art. Finally, in Chapter 6 we discuss the results and our findings.

2. COMPUTATIONAL AESTHETICS AND VISUAL FEATURES

In this chapter we will explore the evolution of aesthetic computing, its concepts and the importance of visual features for this area.

2.1 Computational Aesthetics

The word “aesthetics" comes from the Greek, derived from the concept of “to perceive, to feel" [43]. As a branch of philosophy, aesthetics is defined as the study of what is beautiful, often related to the arts, and has attracted philosophers and researchers from other fields for many centuries [12]. In psychology, it is accepted that aesthetic experience is a result of the interaction between perception, cognition, and emotion [47].

Experimental studies in aesthetics were firstly developed by Fechner in his publications of 1871 [28] and 1876 [29], where the author argued that objects have physical properties linked to the concept of beauty that could be measured objectively, and that through these properties it was possible to study the observer's emotional response. Fechner is also credited with designing the area of psychophysics [27], which directly relates physical stimuli to human perceptions.

The study of aesthetics from a computational angle began with Birkhoff, in work published in 1929 [8]and formalized in his book in 1933 [9]. His work aimed to propose a way to measure the aesthetic value of certain shapes and objects mathematically. According to Birkhoff, aesthetic quality is related to the amount of attention required when observing an object, called complexity and denoted by C, weighted by the notion of order of this same object, denoted by O, composed of elements mainly related to symmetry and repetition.

So, to have a good aesthetic quality, as the complexity of an object grows, the need for order also grows, therefore, the aesthetic value, denoted by M, would be a function of the ratio between order and complexity, according to Equation 2.1. This measure was applied to several objects in Birkhoff's work, from polygons and vases to music and poetry.

$$M = f(O/C). \tag{2.1}$$

The notion of order of an object, according to Birkhoff, has two types of association: formal and connotative. The formal association is related to object properties, such as symmetry, repetition, contrast, etc. Connotative associations, on the other hand, are associations that are not related to the form itself, being more subjective, such as the utility of the object, the relationship of the observer with the object, cultural relations, etc. For the calculation of the aesthetic measure M, only formal associations are taken into account, disregarding all types of connotative associations that may exist.

Early on, his work attracted the attention of researchers in the field of psychology, who start a series of experiments testing Birkhoff's theory and trying to prove or refute. Therefore, the results that were being obtained could not prove Birkhoff's theory. According to McWhinnie [58], in a review of the psychology literature on Birkhoff's theory, the experiments carried out by the field of psychology are not interested in the aesthetic judgment of the object itself, as Birkhoff's aesthetic measure is, but in the aesthetic preferences of the observer. Precisely for discarding connotative associations, focusing only on the object's formal characteristics, and for not taking into account the observer's context, the psychology experiments based on Birkhoff's theory did not obtain good results [58]. Despite this, to this day there are work in the field of psychology that make use of this theory to create a way to quantify beauty or aesthetics, as will be discussed next.

According to Douchová [22], unlike the first studies carried out in the field of psychology, Birkhoff's work is currently used with more flexibility, inspiring new researchers based on the central idea that aesthetics is related to order and complexity and that it can be measured by someway. Douchová [22] points out that some work see the formula created by Birkhoff as a measure of aesthetic efficiency, and suggests that its result would be just an input of a function g that generates the measure of aesthetic judgment, denoted by A, which can be defined as Equation 2.2.

$$A = g(M, ...) \tag{2.2}$$

Greenfield [39], in 2005, for the First Eurographics Workshop on Computational Aesthetics in Graphics, Visualization and Imaging [59], presents a 75-year timeline of the evolution of the term “computational aesthetics" since Birkhoff's book, considered the birth of the field. The first formal appearance of the term “computational aesthetics" was in 1993 in the work of Scha and Bod [66]. However, from Birkhoff to 2002 several work related to aesthetics were published, calling the area by different names – informational aesthetics, algorithmic aesthetics, exact aesthetics... –, but all with the aim of finding an aesthetic measure for the object of study.

Finally, in 2005, also motivated by the Eurographics Workshop [59], Hoenig [43] created the definition of the discipline of computational aesthetics intending to motivate the continued development of the area. According to Hoenig (2005, p. 16): "Computational Aesthetics is the research of computational methods that can make applicable aesthetic decisions in a similar fashion as humans can". This definition is intended to emphasize two aspects: one is the use of computational methods and the other is the enhancement of applicability. Hoenig also draws attention to how the field of computational aesthetics has turned almost entirely to visual aesthetic evaluations and is even considered today as a subfield of computational vision [12].

Over the years, the techniques used to explore aesthetic issues have been transformed, as well as the objects studied. According to Greenfield [39], the first work that explores computational aesthetics with the use of algorithms is the work of Baluja et al. [6], in 1994, where the main objective of the study was to learn user preferences via artificial neural networks and apply this knowledge to evolve aesthetically pleasing computer-generated images.

Since then, the exploration of aesthetic problems has developed more and more, dominating different areas, such as photography [49] [51] [26], computer graphics and rendering [57] [4] [18], architecture [46] [1], human-computer interaction [56], and arts [69] [65] [40] [31]. In addition, its applications also diversified, starting not only to predict the human aesthetic evaluation of a given object but to make different and increasingly complex decisions based on this evaluation. Some of the popular applications are related to measuring the aesthetic value of photos, proposing automatic corrections and cropping in photos, or selecting the best photos among many [73][12], and evaluating the quality of videos and images, identifying distortions and assigning quality scores to images [48] [36] [16].

Brachmann and Redies [12] highlights rating prediction as one of the applications of computational aesthetics generally used to compare professional and non-professional photographs, or to predict user ratings of photographs posted on social networks. In the next two chapters, we will address issues that can be seen as an extension of predicting photo ratings: predicting the popularity of videos (Chapter 3) and creating a movie recommendation system (Chapter 4), both of which we treat as an aesthetic problem, seeking its resolution through information present in the image sequences.

Another area much explored by computational aesthetics is the area of the arts. The analysis of the aesthetics of paintings has been carried out in different ways, two of the most popular applications being style prediction and identification of authorship of works of art [12]. Applications within the field of arts diversify much further, addressing problems that art history has been exploring empirically for many years. In Chapter 5 we will present an analysis of the influence of artworks from an aesthetic perspective of how faces are portrayed by artists.

2.2 Visual Features

At the beginning of computational aesthetics, Birkhoff [9] used information on vertical symmetry, balance, rotational symmetry, horizontal-vertical network, and non-pleasing form as input in the calculation of the aesthetic measure. This information was mainly related to movements in the plane and was not necessarily information that we can call numerical, being used as in a point system where the more positive the value, the better. More details on the calculation can be found at in [22] and Birkhoff's work [9].

Over time, the information used to create the aesthetic measure was improved, contemplating more aspects of the objects and naturally becoming numerical measures. Today, as aesthetic computing is a subfield of computer vision, the main way to obtain these quantitative data from images is through visual features [12] [73].

Visual features are based on features that can be observed in images or image sequences (videos). The first studies of human visual perception date back to ancient Greek theories about how vision is realized. Since then, many investigations have produced improved insights into human vision [83].

The need to adequately describe visual characteristics arose in the 1920s in the field of visual perception. Thereafter, in the field of computer science, different methods to extract these characteristics were proposed. One of the first researches on the extraction of visual features was developed in 1969 [55]. Since then, many approaches to detect different types of features have been presented [83].

In general, visual features describe a property by which real or abstract elements or objects can be distinguished, providing a compact representation of the image's content. Visual features can describe a property of an image as a whole or an object within the image, i.e., it can be a local property or a global characteristic of the image [83]. Another difference can be made regarding the level of abstraction: low-level features describe basic features such as colors and borders, while high-level features can describe more abstract image content, as faces or object detection [12].

According to Brachmann and Redies [12], in recent years, computational aesthetics has moved from the conception of hand-crafted features, developed especially to describe visual and perceptual information from the image of an aesthetic vision, to the use of generic features that were developed for other purposes in computer vision. These generic features were generally developed for object detection and classification, scene comprehension, or image retrieval, and mostly use deep neural networks. However, interpretability is lost with generic features while using hand-crafted features it is possible to reach a conclusion about which features contribute to the aesthetic value of an image.

In cases where the purpose of visual features is simply to classify the images aesthetically for making some decision as the ultimate goal, the lack of interpretability of generic features is not necessarily a problem. However, for aesthetic researchers, applications are not the focus of their research. Rather, the goal is to discover what visual information influences human aesthetic judgment and thus gain a better understanding of the aesthetic experience [12].

Still in [12], Brachmann and Redies note that with the introduction of deep learning in the area of computational aesthetics, it becomes more difficult to share knowledge between computational aesthetics and experimental aesthetics. At first, insights from the field of experimental aesthetics were input to computational aesthetics, and empirical aesthetics also profited from computational methods that have the power to evaluate large data sets, rather than the small number of images tested in psychological experiments with human observers. However, with deep learning, it has become more difficult for empirical aesthetics to keep up with computational approaches.

With this in mind, the experiments developed in this work mainly use hand-crafted features, preserving the interpretability of the results. So, in the context of computational aesthetics, the aim of our studies is to investigate the problem of predicting popularity of videos (Chapter 3), the movies recommendation (Chapter 4) and the identification of influence between artists (Chapter 5). We propose to use visual features in the three domains and compare our methods with competitive techniques in order to discuss the challenges and possibilities.

3. PREDICTION POPULARITY OF ONLINE VIDEOS

The sharing of content on social networks has been the reality of a large part of the population. Nowadays, there is a great facility to publish videos on the web, mainly through social networks, such as YouTube and Facebook, for example. On YouTube, about 400 hours of videos per minute are published, accessed by 2 billion monthly users who generate billions of views daily [86]. Regarding content creators, the number of channels with more than one million subscribers grew by more than 65% per year. When it comes to revenues, the number of channels that had six digits annual revenue on YouTube, grew by over 40% per year [86]. Part of these revenues can come from ads [87], as it also happens on Facebook [25]. It is advantageous to advertise on social networks: on mobile devices alone, YouTube reaches more people between 18 and 34 years old in the USA than any other TV channel [86]. Yet, the opportunity to use content sharing platforms as digital advertising channels was identified. So, understanding what makes a video popular and being able to predict its popularity is a problem that companies like Facebook and Netflix have invested in solving. This predictive power is useful both for advertisements, since they can be directed to videos of greater reach, and for content creators, with regard to the management and production of content based on characteristics that generate more views. It has already been explored in literature [50], [78], [49].

However, the best scenario is to be able to predict the popularity of content before it is published.In the context of images, Khosla et al. [49] used visual features to make popularity prediction. Trzciński and Rokita [78] produced a work aiming the same goal, but with video content. In both work, Support Vector Regression with Radial Basis Function using Gaussian kernel were used. Trzciński and Rokita using data collected from Facebook pages, propose a method called Popularity-SVR, that predicts popularity of an online video using Support Vector Regression (SVR) [78]. The Facebook video data included visual features and temporal features, that is, features captured soon after the content was published, such as number of views over time. To assess the performance of the proposed predictive model, they used Spearman's correlation [71], a non-parametric measure of statistical dependence between two variables. This measure ranges from -1 to 1, where -1 indicates a perfect inverse relationship between these variables, 1 indicates a perfect positive relationship and when the relationship between them is closer to 0, the relationship between them is smaller. When it comes to visual features, the Popularity-SVR shows that, individually, deep features provide the highest Spearman correlation value with video popularity (0.13), followed by the feature groups Clutter (0.12) and Scene Dynamics (0.08). Overall correlation value using all visual features reached over 0.23. However, the best results were obtained when visual features were combined with temporal features, where the Spearman correlation reached over 0.94.

From a computational aesthetics point of view, the advantage of using data extracted from the internet is that in this way it is possible to take numerical measures of human aesthetic judgments, using "crowd knowledge", preventing very discrepant assessments from influencing the results [47]. A view does not necessarily happen due to the aesthetic appeal of the video, the specific and social context, the subject, among other subjective factors are important. However, when using a large data set of social media user behaviors, the more general aesthetic preference pattern must be observable.

As discussed in the previous chapters, in the present work we investigate the use of visual features extracted from videos posted on Facebook as an input for predicting the popularity of online videos, assuming that the number of views can be a response of human aesthetic perception about videos and that visual features can bring a good idea about it. We use Support Vector Machine with Gaussian Radial Basis Function to classify these videos into two groups: the most and the least popular, according to the number of views. In the next sections we will present the dataset and the visual features used (Section 3.1), the methodology applied (Section 3.2) and the results obtained (Section 3.3).

3.1 Dataset and Features

The dataset[1] used in this work was available by the authors Trzciński and Rokita [78]. The available file contains features extracted from 1,820 videos published on Facebook between August 1st and October 15th 2015 from pages such as AJ+[2] and BuzzFeedVideo[3]. Two types of features were considered in the data extraction: temporal and visual features. The next sections describe some details about both features data.

3.1.1 Temporal Features

After the video is published, the temporal features show the number of views, likes, comments, and shares every hour, for seven days after posting, collected by the URL scraper on the posting page [78]. In the present work, we used only visual resources as predictor variables, while temporal resources, such as number of likes, comments, and views were excluded from the analysis. However, we use the number of views at the end of the seven-day period after publication as a response variable. Therefore, the main reason for discarding temporal data in this work is that we want to analyze and investigate only visual data, so that this analysis can be produced before publication.

[1] http://ii.pw.edu.pl/˜ttrzcins/facebook_dataset_2015.csv
[2] https://www.facebook.com/ajplusenglish
[3] https://www.facebook.com/BuzzFeedVideo

3.1.2 Visual Features

In this work, we hypothesize that only visual features can be used to predict popularity of video content with a certain accuracy. The visual features were collected directly from the video, using various computer vision algorithms, as described in [78]. The list of available visual data that was used in this work is:

1. **Video characteristics:** This class states for general video information, such as duration, frames per second, number of frames, and frame dimensions of the analyzed video.

2. **Dominant color:** The color space of the video was divided into 10 classes (black, white, blue, cyan, green, yellow, orange, red, magenta, and other) and each frame of each video was assigned to one of these classes. In addition, the data set contains information about which class of colors is dominant and what proportion of each color is present for each video.

3. **Face detection:** Presents information about the presence of faces in the video, such as the average number of faces per frame, the proportion of frames with faces, and the average proportion of the face size in relation to the size of the frame.

4. **Text detection:** Similar to face detection, it concerns information about the presence of text in the video, such as the proportion of frames with text and the average proportion of the text size in relation to the size of the frame.

5. **Scene dynamics:** It regards information about the number of shots in the video and classification of the shots as hard and soft cuts.

6. **Speed:** They provide information about the average video speed, a clutter metric, and a metric that specifies the video rigidity (average number of frames where are homography between current and previous frames).

While Trzciński and Rokita propose the Popularity-SVR using Support Vector Regression [78], in our work we use Support Vector Classifier, as described in the next section. Indeed, when we consider this question as a classification problem, we understand that some milestones in terms of video visualizations are more interesting, and probably relevant. For example, 100,000 visualizations or 1 million are maybe good milestones.

3.2 Methodology

In this experiment, our goal is to predict whether a video will be popular or not based on number of views, given its visual features computed using computer vision algorithms. This section presents the pre-processing phase executed on available data [78], the model tuning to configure SVM hyperparameters, and details about the used SVM classifier model. We performed all analysis and modeling using R software version 3.6.3 [63], through caret package version 6.0-86 [52].Different from the work proposed by Trzciński and Rokita [78], in our method, the popularity prediction is treated as a classification problem. In this case, we do not want to provide the exact number of views for a given video, but to identify whether the video will have more views than a certain pre-established milestone, 7 days after its publication. In this work, we tested 5 different milestones according to the number of views: **10,000**, **100,000**, **500,000**, **750,000** and **1 million** views. We named videos that have reached the milestone as *successful-videos*. Further details are presented in Section 3.3.

3.2.1 Pre-processing

Pre-processing of data is essential in building a statistical model. In this phase, it is possible to identify missing values or the relationship between the variables that can harm the modeling process. It is during pre-processing that the addition, deletion, or transformation of the dataset is done. According to Kuhn and Johnson [53], data preparation can create or break a model's predictive ability. In the pre-processing of our method, the missing values, zero- and near zero-variance feature predictors were analyzed, identifying correlated feature predictors and linear dependencies.

We identified four correlated feature predictors: number of frames is highly correlated with video duration, frame width is highly correlated with frame height, average proportion of frames with faces is highly correlated with average number of faces per frame, and two features about soft and hard cuts are complementary, so they have a -1 correlation. Consequently, the features about the number of frames, frame width, average proportion of frames with faces, and one of two features about shot cuts have been removed from the features list. Regarding the linear dependencies, QR decomposition [37] is used to determine whether features are linearly independent and then identify the sets of features involved in the dependencies if any. There was no need to treat missing values since the dataset has complete information for all videos. It was tested for feature predictors with zero- and near zero-variance, but none were identified, so there was no need to remove resources in this case. The train and test sets are splitted in proportion of 70% and 30%, respectively, preserving the overall class distribution of the response variable. Finally, we centered and scaled the data, to improve the numerical stability of some calculations [53].

3.2.2 Support Vector Machine Classifier

The technique used for predictive modeling of popularity of videos is a Support Vector Machine with Gaussian Radial Basis Function classification model. The Support Vector Machines classifier is a binary classifier algorithm that looks for an optimal hyperplane as a decision function in a high-dimensional space [11]. Given a set of labeled training patterns $(\mathbf{x}_i, y_i)$, $i = 1, ..., l$ where $\mathbf{x}_i \in \mathbb{R}^n$ e $y \in \{1, -1\}^l$, the algorithm finds the parameters of the decision functions $D(\mathbf{x})$ during a learning phase. The decision function has the following form: $D(\mathbf{x_i}) = \sum_{k=1}^{p} \alpha_k K(\mathbf{x}_k, \mathbf{x_i}) + b$, where $\mathbf{x}_k$ are the support vectors returned by algorithm, α_k are the coefficients, x is a feature vector for a video, the function K is a predefined kernel and b is the intercept. For non-linearly separable problems, Support Vector Machines can not find a separation hyperplane that provides a good generalization. For that, a kernel can be used to transform the data to a higher-dimensional space and thus a linear hyperplane can be obtained to proper separate the different classes. We used the Gaussian radial basis function kernel as follows:

$$K(\mathbf{x}_i, y_i) = exp(-\frac{\|\mathbf{x}_i - y_i\|^2}{\sigma^2}), \tag{3.1}$$

where $\sigma > 0$ is a parameter from Gaussian kernel. The model hyperparameters are the cost C, from Support Vector Machine, and σ from kernel.

3.2.3 Model Tuning

For each visualization milestone, we tested 21 different combinations of visual features. To search for the best hyperparameters for the model, we create a grid of values for the hyperparameter of the model. For σ hyperparameter, we define 12 possible values between 0 and 0.5, and for C hyperparameter, 7 values between 0.25 and 8. This setup results in 8,820 different models, one for each combination of visual features, according to Table 3.1, milestones and pair of hyperparameters.

Then, in the model tuning process, we use repeated 10-fold cross-validation, where three separate 10-fold cross-validations were used as the resampling scheme. For each combination of visual features and milestones, the pair of hyperparameters that generated the model with the largest Kappa [54] was selected, thus leaving 105 models.

Table 3.1: Combinations of group of features, as defined in Section 3.1.2, tested in the models and abbreviations that we use to refer to the feature setup.

Abbreviation	Combination of Visual Features
V	Video Characteristics
C	Dominant Color
F	Face Detection
T	Text Detection
D	Scene Dynamics
S	Speed
VC	V. Char. + Color
VF	V. Char. + Faces
VT	V. Char. + Text
VD	V. Char. + S. Dyn.
VR	V. Char. + Speed
VDC	V. Char. + S. Dyn. + Color
VDF	V. Char. + S. Dyn. + Faces
VDT	V. Char. + S. Dyn. + Text
VDS	V. Char. + S. Dyn. + Speed
VDSC	V. Char. + S. Dyn. + Speed + Color
VDSF	V. Char. + S. Dyn. + Speed + Faces
VDST	V. Char. + S. Dyn. + Speed + Text
VDSTC	V. Char. + S. Dyn. + Speed + Text + Color
VDSTF	V. Char. + S. Dyn. + Speed + Text + Faces
Complete Model	V. Char. + S. Dyn. + Speed + Text + Color + Faces

3.3 Results

We trained 8,820 different models with different configurations to predict whether a video will be a *successful-video* and which milestone it has achieved. After selecting hyperparameters, 105 models remained, combining different features to predict the number of views according to 5 different milestones: 10,000, 100,000, 500,000, 750,000, and 1 million views. To select the best models among the 105, three different metrics were used: Kappa, Sensitivity, and Positive Predictive Value.

Cohen's Kappa Coefficient is a statistical measure of agreement between classifications, which compares the model's classification with the response variable more robustly than accuracy since it takes into account the chance of the result being the result of chance [82]. Sensitivity is the ability of a model to identify positive cases, that is, the percentage of successful-videos correctly classified in the model among all successful-videos in the dataset. While the Positive Predictive Value measures how many true positives are actually positive, that is, how many of the videos are classified as successful-videos. Landis and Koch proposed a classification of strength of agreement for certain metric value ranges [54], as shown in Table 3.2. We consider models that have resulted in Kappa with moderate strength as agreement or more, that is *Kappa* $>=$ 0.41, and sensitivity and positive predictive value of at least 0.5. Based on these metrics, 21 models were selected, shown in Table 3.3.

Table 3.2: Agreement strength classification for Cohen's Kappa Coefficient proposed by Landis and Koch in [54].

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost perfect

Table 3.3: Results and configurations of the 21 best models among the 105 models described in Section 3.2.3, selected based on the Kappa, Sensitivity, and Positive Predictive metrics. The model with the best overall rating is highlighted in bold.

Features	Views	Kappa	Sensitivity	Pos Pred Value	Sigma	C
V	*100k*	0.7288	0.8899	0.8926	0.08	0.5
D	*100k*	0.4317	0.9174	0.7282	0.5	8
S	*100k*	0.4266	0.8746	0.7371	0.5	2
VC	*100k*	0.6892	0.8838	0.8705	0.03	8
VF	*750k*	0.5345	0.5000	0.7576	0.25	5
	100k	0.7144	0.8777	0.8913	0.25	5
VT	*1m*	0.5165	0.5422	0.6338	0.5	8
	750k	0.5386	0.5100	0.7500	0.5	5
	100k	0.7292	0.8869	0.8951	0.5	5
VD	*100k*	0.6856	0.8807	0.8701	0.5	8
VR	***100k***	**0.7324**	**0.8930**	**0.8930**	**0.04**	**2**
VDC	*100k*	0.6503	0.8716	0.8533	0.03	8
VDF	*100k*	0.7126	0.8899	0.8818	0.25	5
VDT	*100k*	0.6978	0.8807	0.8780	0.25	8
VDS	*100k*	0.7009	0.8869	0.8761	0.08	8
VDSC	*100k*	0.6508	0.8685	0.8554	0.02	8
VDSF	*100k*	0.7140	0.8807	0.8889	0.06	8
VDST	*100k*	0.6987	0.8746	0.8827	0.25	5
VDSTC	*100k*	0.6877	0.8930	0.8639	0.01	8
VDSTF	*100k*	0.7131	0.8869	0.8841	0.04	8
Complete model	*100k*	0.7009	0.8869	0.8761	0.02	5

As we can see in the results presented in Table 3.3, the milestone that obtained the best results was 100,000 views. Only three of the 21 selected models generated positive results for other milestones 750,000 and 1 million. One of the possible causes of this better performance of the 100,000 view milestone is related to the balance of the sample since the other milestones are more unbalanced in the amount of successful-videos. The model performed better in the three metrics, highlighted in bold in Table 3.3, uses the 100,000 views framework and combines the visual features **Video Characteristics + Speed**. Video features include duration information, frames per second, and frame dimensions, and the Speed features provide information about the average video speed, clutter metric, and video rigidity, as defined in Section 3.1.2. The hyperparameters that generated the best model are *Sigma* = 0.04 and *C* = 2, obtaining Kappa of 0.7324, sensitivity of 0.8930, and positive predictive value of 0.8930.

The superior performance of this model in predicting the number of views may suggest that fundamental video characteristics such as duration, frames per second, and frame dimensions play a significant role. It is hypothesized that excessively long videos may fail to capture viewers' attention and consequently receive fewer views. Additionally, video quality in terms of image clarity is also deemed relevant. Furthermore, insights from the metrics of speed, rigidity, and clutter suggest that videos that are either too slow or too fast, as well as visually cluttered, may be less appealing to viewers, resulting in decreased viewership.

In [78], Trzciński and Rokita describe that when they use only visual features, they obtained better results in the complete model. The features that most contributed to the performance of their model were Deep features, Clutter and Scene Dynamics but Deep features were not present in the available dataset, used also in the present work. Therefore, this is the reason why we could not re-implement their method from scratch because we do not have all available data. Even though, our technique and their method [78] are different approaches of the same technique (regression and classification), which makes it difficult to compare the results. Anyway, we believe that we obtained better results once our complete model reached Kappa of 0.7324, sensitivity of 0.8930, and positive predictive value of 0.8930, while their complete model reached 0.23 in Spearman correlation.

Our results indicate that using visual features to predict the popularity of videos provides good results and can be used even before the videos are published. Furthermore, using interpretable visual features facilitates the understanding of the result and provides input for further studies.

4. MOVIE RECOMMENDATION

Nowadays, streaming services have grown rapidly. Netflix has approximately 195 million paid users [32] and other streaming services keep coming [81]. Amid so much content available, a challenge for users is to find movies that are relevant and can provide a good experience. To indicate relevant content to users, recommendation systems assess users' behaviors concerning items and their preferences. Recommendation systems have as a basic principle to use interaction data and feedbacks that users have given in the past for certain items to make inferences about their interests. According to Aggarwal [3], for recommendation systems, past interests and inclinations are usually good indicators of future choices. A popular type of recommendation system is collaborative-filtering, which can be item-based, about the users who rated it, or can be user-based, according to their favorite items [76]. This type of recommendation system depends on a large number of interactions of users for each item, containing rates. Items with few interactions do not bring enough information to the recommendation system and tend to be disregarded [76]. The content-based recommendation system takes advantage when it comes to new items or with sparse datasets, where the items have few evaluations because it uses information intrinsic to the item [2]. In movie recommendation systems, commonly used information is cast, genre, and director, and in some work, they already explore visual features.

Thus, several recommendation techniques have been developed over time. According to Thorat et al. [76], the techniques can be divided into collaborative-filtering, content-based, and hybrid. Collaborative-filtering is the most extensively used approach to design recommender systems. However, the limitation of collaborative-filtering indicated by Thorat et al. [76] refers to the need for a large amount of feedback data. Thus, new or unpopular items tend to perform poorly on recommendations. This limitation is not observed in the content-based recommendation system, which can provide good recommendations even when it comes to new items.

According to Deng et al. [21], most content-based recommendation systems explore textual data to make recommendations. In the context of movie recommendation, the information commonly used are cast, genre, and director, ignoring the positive effects of using visual information from the image. In this context, Elahi et al. [24] investigate the semantic gap generated by using only high-level information and also propose the use of visual features, such as brightness, contrast, and movement.

Several work already demonstrate the benefits of using visual features. Deldjoo et al. had better results in [19] when movies were recommended based on visual features such as scene duration, light, and movement than when recommended by the genre, and in [20] the best results were obtained when using visual features based on MPEG-7 and deep learning in comparison to information like genre and tag. Rassweiler et al. [30] also obtained good results using deep visual features. Qu et al. [62] explored the movie recommendation for casual users based on a combination of audio, text, and visual features.

However, none of these work attempted to provide information about the type of image that makes up the movie: real images or animations. Even though “animation" is generally considered a movie genre, it is usually combined with other genres that can confuse what animation is and what is live-action. Our hypothesis is that the movie category (animation or live-action) can help to better recommend the movies.

The objective of this part of our work is to propose visual features that can be used in the development of recommendation systems along with the movie category information (animation or live-action) as a way to improve the results of the recommendations. Our results show that information from the movie category can significantly contribute to better recommendations and the result was similar to a work in the area that uses visual features and deep learning. The main contribution of our work is the proposition of visual features and the use of the movie category explicitly to improve the results of recommendations, in addition to comparing several calculation methods for validation of recommendations and exploring their results with different combinations of features visuals as a way to optimize the recommendations.

4.1 Dataset and features

To recommend new movies to users, we use two types of datasets described in the next subsections: a dataset with visual features extracted from the movies, and a dataset of user reviews, used to validate the results.

4.1.1 Visual feature dataset

According to Deldjoo et al. [19], for the construction of a content-based recommendation system, visual features extracted from movie trailers are an alternative to extracting visual features for full length movies. Therefore, to carry out our work we selected 77 Disney movies, where 56 are animated movies and 21 are live-action movies, and their trailers were downloaded from YouTube for analysis. We chose those movies because they have available trailers and are present in the user rating dataset, presented in Section 4.1.2.

The categorization of movies between animation and live-action was done manually in order to provide a proof-of-concept on this matter.

The extraction of visual features was performed based on the trailers for each movie. After downloading the trailers, we extract all frames from all videos to extract the visual features of each frame.

The entire process of extracting the features was carried out with R software [63] version 3.6.3, using various computer vision algorithms as detailed below:

1. **Color saturation**: Mean of saturation of the HSV (Hue-Saturation-Value) color space. With the aid of the *colordistance* [84] package, we extract the mean of the saturation channel of all the pixels of each frame and calculate the mean of all the frames of the video.

2. **Color value**: Mean of brightness, the value of the HSV color space. Extracted in the same way as color saturation.

3. **Faces per frame**: Mean of number of faces per frame. For each frame, faces were detected using the *opencv* [60] package, as shown in Figure 4.1, and the average faces per frame for all video frames was calculated.

4. **Frames with text**: Proportion of frames with text, such as the movie name lettering and phrases displayed (selected trailers have no caption). Each frame with some text detected was considered a frame with text, and then the proportion of these frames over the total number of frames in the video was calculated. The texts were identified using Tesseract OCR [70] with the aid of the *tesseract* [61] package.

5. **Shot cuts**: Count of shot cuts. The shot cuts were identified by the absolute difference between each frame and its successor, both on grayscale. When a certain percentage of pixels in a frame differed between the successor frame, the successor frame was considered a new cut of the scene. The absolute difference was calculated using the *Rvision* [35] package.

6. **Clutter metric**: Mean of the proportion of edge pixels of frames. This metric quantifies the clutter present in the video, as proposed in [78]. It was identified how many border pixels each frame has using Canny Edges [14] detector with the aid of the *imager* [7] package, as shown in Figure 4.2, and the proportion of these pixels concerning the frame size was calculated. The average of the clutter metric of all frames of the video was considered.

Therefore, all visual features were normalized according to the Equation 4.1:

$$z_i = \frac{x_i - \mu}{\sigma}, \tag{4.1}$$

where μ is the mean and σ is the standard deviation of all observations of the feature, x_i is the original value of observation i of the feature and z_i is the normalized value of observation i. So that they were in the same unit of measurement.

Figure 4.1: Trailer frame for the movie *"Moana"* (Walt Disney Pictures, 2016) where the face was identified using the *opencv* [60] package.

In this way, our final visual features dataset is composed as follows: one observation (row) per movie and 7 columns, one of them being the movie identification, five being the visual features described above, all at the same scale, and a seventh with the movie category information, where animation is represented by 1 and live-action is represented by 0.

Figure 4.2: Trailer frame for the movie *"Moana"* (Walt Disney Pictures, 2016) where the border pixels were identified through the *imager* [7] package.

4.1.2 Rating dataset

To perform the experiment validation, we used the MovieLens 25M Datasets [41], a popular dataset that contains more than 25,000,000 ratings applied to 62,000 movies by 162,000 users. Only the rating data of the 77 selected Disney movies, as mentioned in the last section, were used. Also, all these 77 movies had more than 100 ratings (total is

410,354 rates) and maintained users who rated at least 5 of the movies (34,495 users). These dataset numbers are specified in Table 4.1.

Commonly, ratings are used in two ways to compare recommendation systems performance: *explicitly* and *implicitly* [74]. When explicit rating was used, the rating provided by the user was considered as it is, i.e., values between 0 and 5. As an implicit rating, the rating was binarized: 1 when the rating is 4 or higher and 0 otherwise. In this work, we use and compare both methods.

Table 4.1: Number of items at each dataset.

Characteristic	Quantity	Dataset
Animation movies	56	Visual Features
Live-action movies	21	Visual Features
Ratings	410,354	MovieLens
Users	34,495	MovieLens

4.2 Methodology

In this section, the methodology used to calculate the similarity, the proposed scores for the recommendation, and the validation method are presented.

4.2.1 Similarity

Recommendations are made through the similarity between the movie trailers, recommending movies visually similar to the movies that the user has already rated with high scores [19, 20, 30]. In this work, we propose two calculations for the similarity: using *just visual features* and using *visual features with movie category*. The similarity between two movies is calculated using cosine similarity according to Equation 4.2:

$$s_{xy} = \frac{\mathbf{f}_x^T \mathbf{f}_y}{||\mathbf{f}_x||\,||\mathbf{f}_y||}, \tag{4.2}$$

where $\mathbf{f}_x$ and $\mathbf{f}_y$ are the feature vectors of the movies x and y, respectively, as defined in Section 4.1.1.

The difference between the two calculations proposed in this work is that the feature vectors in the first case are composed only of visual features, while in the second, the category information is added. The similarity values are computed using the cosine similarity function that has been normalized to vary between 1 (most similar) and 0 (less similar).

4.2.2 Scores

As a way of defining whether a movie is more recommended for a certain user than another, each movie is classified using a score. We tested two forms of calculation: using a *Rating Similarity* and using *Just similarity*.

Similar to Deldjoo et al. [19], for each m movie present in a probe set P, rated by user u with a score of 4 or more, a list L of possible recommendations were generated, consisting of the movie m and more all other movies not rated by user u.

To calculate the *Rating Similarity*, the score considers the rating values provided by user u concerning a movie i from the list L, and calculated according to Equation 4.3:

$$r_{ui} = \frac{\sum_{j \in T} r_{uj} s_{ij}}{\sum_{j \in T} s_{ij}}, \tag{4.3}$$

where j is a movie rated by the user u present in a train set T, r_{uj} is the rating assigned by the user u to the movie j and s_{ij} is the similarity between movies i and j.

When we use *Just similarity*, the calculation considers only the similarity between the rated movies. Then for each movie, both scores are calculated so that the higher the score, the more recommended the movie is for that user.

4.2.3 Combination of features

In order to achieve better performance, we tested 25 different combinations of visual features as defined in Table 4.2. Starting with pairs of visual features, all results were calculated using the methodology proposed in this section. From the pair with the best result, each of the other features was added, forming triplets, and so on until obtaining the result of the method using all available visual features.

4.2.4 Validation

To evaluate our method, we performed 10-fold cross-validation. At each iteration, one fold was reserved to be the P probe set, and the remaining 9-folds make up the T train set. In P, we consider only movies rated 4 or higher.

After calculating each score for each movie in the L list, the recommendation list is generated through top-N movies with the highest score. The idea is that the movie m is

Table 4.2: Combinations of features, as defined in Section 4.1.1.

Id	Combination of Visual Features
1	Color Saturation + Color Value
2	Color Saturation + Faces per Frame
3	Color Saturation + Text per Frame
4	Color Saturation + Shot Cuts
5	Color Saturation + Clutter metric
6	Color Value + Faces per Frame
7	Color Value + Text per Frame
8	Color Value + Shot Cuts
9	Color Value + Clutter metric
10	Faces per Frame + Text per Frame
11	Faces per Frame + Shot Cuts
12	Faces per Frame + Clutter metric
13	Text per Frame + Shot Cuts
14	Text per Frame + Clutter metric
15	Shot Cuts + Clutter metric
16	Text + Clutter + C. Saturation
17	Text + Clutter + C. Value
18	Text + Clutter + Faces
19	Text + Clutter + Shot Cuts
20	Text + Clutter + S. Cuts + C. Saturation
21	Text + Clutter + S. Cuts + C. Value
22	Text + Clutter + S. Cuts + Faces
23	Text + Clutter + S. Cuts + C. Value + C. Saturation
24	Text + Clutter + S. Cuts + C. Value + Faces
25	Text + Clutter + S. Cuts + C. Value + Faces + C. Saturation

among the recommended N movies. The validation metric used is the Recall [17], implemented according to Equation 4.4:

$$Recall(N) = \frac{h}{|P|}, \tag{4.4}$$

where $|P|$ is the size of the probe set and h the number of hits, that is, the number of times that movie m was present among the recommended N movies in each $|P|$ recommendation. The experiments evaluated the results of the top-N for $N = 3, 5, 7,$ and 10.

Since list L is made up of all the non-relevant items in our dataset for each user plus the relevant movie, list L is variable length, averaging 64 movies. Consequently, to be able to compare our results with other work, even if indirectly, we also generate a random recommendation, where the top-N are selected from the L list at random, with equal probability. Thus, it is evaluated how much our results are better than the random recommendations and compared how much the results of another work are better than if the random recommendations had been made in their settings.

4.3 Results

We performed experiments with different proposed metrics. The movies were recommended both with the score that uses the rating assigned to the movie, *Rating Similarity*, and with the score that considers only similarity, called *Just Similarity*. Both scores being calculated with the similarity composed only of visual features and similarity that includes movie category information (identified with *w/ category*). Besides, the experiments were run with both the explicit dataset and the implicit dataset when possible.

For each of these combinations of configurations, the results were calculated using each of the 25 different combinations of visual features, where in each experiment 197,678 recommendations were made. The results in terms of Recall@10 are shown in Table 4.6, where the best result is highlighted in bold. The best result in terms of feature combination was obtained with the **text per frame**, **clutter metric**, **shot cuts** and **color value** visual features, with *Recall*@10 = 34.80%, when the calculation is done using Just Similarity w/ category. In fact, it is possible to see that the best results are found in this calculation configuration, regardless of the combination of features used.

Table 4.3: Recalls of the experiments executed with the **explicit dataset** using the visual features text per frame, clutter metric, shot cuts and color value, in addition to the results of the experiment of random recommendations.

	Recall@3	Recall@5	Recall@7	Recall@10
Rating Similarity	6.60%	10.76%	15.43%	22.16%
Just Similarity	10.96%	17.84%	24.13%	32.57%
Rating Similarity w/ category	5.68%	9.76%	14.06%	21.03%
Just Similarity w/ category	12.69%	19.30%	25.76%	34.73%
Random	4.99%	7.93%	11.59%	16.59%

We evaluated more in detail the results of the calculations using the best combination of features in Table 4.3 (performed using the *explicit dataset*) and Table 4.4 (performed using the *implicit dataset*). Experiments were also performed using random recommendations using both datasets to have comparable results in terms of dataset composition and sample size, and all experiments performed significantly better. We performed the independent t-test for difference of means, all experiments performed significantly better, α = 5%, than the random result, obtaining $p < 0.001$.

Comparing the results between *explicit* and *implicit datasets*, with respect to the *Just Similarity* approach, performing the independent t-test for difference of means with α = 5%, the results of the recommendations made using the *explicit dataset* were significantly better in the comparison for Recall@10, with p = 0.04. In the other results there was no significant difference. When comparing the *Just Similarity w/ category* approach, the *explicit*

Table 4.4: Recalls of the experiments executed with the **implicit dataset** using the visual features text per frame, clutter metric, shot cuts and color value, in addition to the results of the experiment of random recommendations.

	Recall@3	Recall@5	Recall@7	Recall@10
Just Similarity	10.89%	17.69%	23.98%	32.25%
Just Similarity w/ category	12.30%	18.74%	25.01%	33.99%
Random	4.81%	7.84%	11.67%	16.83%

dataset performed significantly better, obtaining p = 0.002 for Recall@3, p = 0.003 for Recall@5, $p < 0.001$ for Recall@7 and Recall@10.

The use of the category (animation or live-action) in combination with visual features generated significantly better results whenever the recommendations were made using *Just Similarity*. When *Rating Similarity* is used, the results are inverted, and using the movie category generated significantly lower results. Therefore, on the scoring method used to make the top-N recommendations, the use of *Just Similarity* generated significantly better results in all experiments. In addition, the advantage of using *Just Similarity* is being able to make recommendations also in cases where data on interactions with movies are implicit.

Finally, the best result of the experiment was obtained when using the *explicit dataset*, with *Just Similarity* and using the movie category in combination with the **text per frame**, **clutter metric**, **shot cuts** and **color value** visual features, obtaining a *Recall*@10 = 34.73%, this means that the recommendations for this experiment were 109% **better** than the random recommendation. If we consider the result of *Recall*@3 = 12.69% the results were even better, 154% **better** than the random recommendations. The best result is generated by combining these visual features may indicate that, in general, individuals who watch Disney movies tend to enjoy movies with similar scene-cut dynamics and visual color styles that often represent the emotions portrayed. Additionally, it suggests that movies with similar image complexity, including abundant information and text, tend to attract similar audiences.

Table 4.5: Comparison with the competitive method.

	Recall@10	Random results	Improvement
Our result	34.73%	16.59%	109%
Deldjoo et al. [20]	2.16%	1.00%	116%

To compare our results, we chose the method based on deep learning recently proposed by Deldjoo et al. [20]. In their work, authors use the previous version of the dataset used in our work, MovieLens 20M, and similarly evaluate the results: they generate top-N recommendations based on a list of 1001 movies, being 1000 non-relevant movies and 1 relevant movie. Each time the relevant movie is among the N recommendations, it is considered a hit, and the recall is calculated in the same way as defined in our Equation 4.4 in Section 4.2.4. In our work, the top-N recommendations are based on lists containing all non-relevant items in our dataset for each user plus the relevant movie, as described in Section 4.2.2. Thus, the lists have a variable size, with an average of 64 movies. So, to make our methods comparable, we decided to compare both work with the results of random recommendations.

In the competitive method, the list that generates the top-N recommendation is composed of 1001 items, the probability that the relevant item will be randomly contained in the top-n recommendation where $N = 10$ is 1.00%. Having obtained an *Recall*@10 = 2.16%, we can say that they made recommendations 116% **better** than random recommendations. The comparison is in Table 4.5. Thus, although our work uses a more traditional approach, our results (109% for *Recall*@10 and 154% for *Recall*@3) are similar to deep learning methods.

For the area of aesthetics, the main benefit of using a more traditional approach, in addition to the low computational cost, is the interpretability guaranteed both by the feature extraction method and in the calculation and analysis of recommendations. Unlike deep learning features extraction methods, where the relations of the extracted information with our human vision system are not always interpretable, in traditional methods we are able to make that relationship directly and extend our understanding of user preferences beyond the system recommendation.

Table 4.6: Recall@10 results of all experiments and all combinations of visual features defined in Table 4.2, identified by id. The best feature combination result is highlighted in bold.

	Explicit dataset				*Implicit dataset*	
Id	Rating Similarity	Just Similarity	Rating Similarity w/ category	Just Similarity w/ category	Just Similarity	Just Similarity w/ category
1	19.36%	19.21%	18.59%	23.37%	18.95%	23.24%
2	21.60%	23.41%	20.46%	26.83%	23.69%	26.12%
3	22.32%	24.68%	21.17%	27.55%	25.22%	26.59%
4	22.46%	25.20%	20.34%	24.67%	24.23%	24.23%
5	20.72%	21.17%	19.28%	27.70%	21.69%	26.78%
6	20.39%	28.75%	19.32%	30.36%	28.75%	30.00%
7	21.26%	27.07%	19.73%	30.06%	27.26%	29.37%
8	21.35%	24.23%	19.31%	27.20%	23.72%	27.14%
9	20.06%	24.09%	18.48%	30.06%	24.56%	29.48%
10	22.14%	29.81%	19.42%	30.62%	28.73%	29.17%
11	22.52%	27.22%	21.45%	28.47%	26.87%	27.86%
12	21.89%	24.39%	19.34%	28.58%	24.19%	27.79%
13	23.59%	28.32%	21.98%	31.87%	27.83%	30.83%
14	21.50%	26.81%	18.40%	34.68%	26.44%	32.66%
15	20.21%	29.11%	18.99%	31.51%	29.24%	30.90%
16	21.65%	23.96%	19.80%	29.10%	24.26%	27.94%
17	21.21%	26.24%	19.92%	31.47%	26.60%	30.55%
18	22.06%	26.43%	19.72%	32.42%	25.66%	30.67%
19	22.83%	30.97%	20.97%	34.09%	30.59%	32.96%
20	22.75%	30.41%	21.46%	30.62%	29.91%	29.84%
21	**22.16%**	**32.57%**	**21.03%**	**34.73%**	**32.25%**	**33.99%**
22	23.24%	31.63%	21.71%	33.80%	31.29%	32.71%
23	21.68%	30.36%	20.87%	31.47%	30.09%	30.67%
24	22.60%	33.01%	21.43%	34.52%	32.69%	33.86%
25	22.05%	30.92%	21.19%	31.69%	30.56%	30.94%

5. INFLUENCE IN PAINTINGS

At its origin, studies of aesthetics within philosophy had art as one of the central objects of their questioning [47]. Therefore, it is natural that a series of work have been exploring the aesthetic characteristics of paintings and their styles. In recent years, as paintings are digitized in high quality, it becomes possible to study paintings computationally [73]. The main themes widely explored by computational aesthetics are mainly related to solving issues such as artist identification and style prediction, although other problems such as retrieval of similar paintings, painting dating, detection of forgery are also very popular [12].

However, the application of algorithms and large-scale automatic evaluations of works of art has generated discussions. Not only the development and application of these new technologies are mostly unknown by art scholars, but there is a lot of concern about their implementation. As analyzed by Spratt and Elgammal [72], a good part of the concerns on the part of art history researchers is precisely due to the lack of knowledge and disbelief in how computers could perform such subjective tasks performed by specialists. However, according to the authors, part of the responsibility for these concerns is how computer scientists disseminate their work, generalizing the power of computer analysis to global and complex problems, rather than seeking to collaborate with art historians to solve specific problems in the field.

Foka [31] emphasizes that art historians are not looking for systems that make interpretations automatically, as new methods applied recently have gone beyond analyzing the content of the images of artworks. Furthermore, Foka listed topics that deserve greater attention in computer science for facilitating the work of art historians, such as the creation of a painting recovery system, signature detection, ethnicity recognition, among others, and reinforces the importance of approach between these two areas in collaborative work.

Thus, the work of Foka [31] and Sprat and Elgammal [72] converge in the thought that computational advances applied in art history have to be questioned not because they threaten to replace art historians in their tasks, but on the contrary, because they may not have a practical utility as the computer scientists may expect.

Keeping this discussion and questions in mind, in the present work, we explore the problem of identifying influence among artists, as further detailed in next sections. It is necessary to emphasize that we do not intend to solve the question of influences through our methods. Instead, we intend to provide new evidence of possible relationships between artists and identify in which characteristics these relationships can be noticed, in order to be input for specialists who are investigating these relationships. In addition, we had the intention of bringing our work closer to researchers in the arts area as a way of opening a channel of mutual communication and collaboration, as highlighted in the previously mentioned work.

5.1 Influence analysis on paintings

Influence in an artist's work is a topic that often sparks discussions among art historians because it is a complex matter and involves objective and subjective tasks [12]. According to Hermerén [42], there are basic conditions for art scholars to assess influence in the arts: temporal and contact conditions, which refers to the contact between the influenced artist and the influencer's work, change conditions, which says that some characteristic of the influenced artist's work has changed after the contact with the influencer's art, and similarity conditions, which refer to remarkably similar visual characteristics.

From a computational point of view, these influence relationships can be identified through the similarity between the artists' works of art, with the aid of visual features. Bressan et al. [13] extracted two types of low-level local feature vectors, one composed of SIFT features and the second using image color statistics, to compose a similarity score based on Fisher kernel [44]. Shamir and Tarakhovsky [69] used the WND-CHARM scheme, commonly used for biomedical image analysis, to extract 4027 features from 994 paintings by 34 artists, and calculated a matrix of similarities that can be visualized through a phylogeny, tree-shaped diagram commonly used in biology to visualize the relationship between species [77].

Saleh et al. [65] also addressed the issue of identifying influence among artists using visual features extracted from their artworks. The authors' idea was to create an influence suggestion system using semantic visual features, inspired by style classification work, to suggest influences due to the similarity of the artworks. For this, the authors used a dataset composed of 1,710 images of paintings by 66 artists and containing 13 painting styles. To make an overall assessment of the results, they collected a set of 76 pairs of positive influences, claimed by art historians, to compose the ground truth. To calculate the similarity between the artists, the authors used high-level semantic features extracting the class feature vector, GIST descriptors, and HOG descriptors and calculated the Hausdorff distance between the artists, treating each artist as a set of points composed of each of your artworks. The evaluation of the results is done by calculating the recall, which is defined as the ratio between the number of correct influences detected and the total of known influences on the ground truth. Saleh et al. [65] best result was top-5 recall 34.21% using GIST features.

5.2 Face in paintings

An important human skill is to recognize the other. Throughout evolution, we have been honing our abilities to process unknown faces and familiar faces. In a recent study, researchers estimated that people know about 5,000 different faces on average [45]. As a result, faces catch our attention. For those looking at a painting, for example, faces are the main points of attention, when present. In an important study by Yarbus [85] on eye movement, in 1965, it was possible to notice that the fixations of the observers' eyes were particularly directed to the faces of the individuals in the painted scene already in the first moments of the observations [75].

In art, the perception of differences and similarities is a fundamental skill through which art historians analyze paintings. According to Schenk and Stumpel [67], although faces in painting have been studied from many angles and art historians make use of facial features in their analysis, they rarely cite face comparison as a method. For Schenk and Stumpel, art historians do not reflect on the fact that they apply facial recognition and memory skills, perhaps because recognizing faces is a very universal and everyday skill to be considered specific in the field of the study of art history.

In an experiment carried out with 96 lay participants in art, results showed that laypeople categorize faces in the same way as art specialists, with regard to their region or painting school [67]. The authors concluded that artists from the regions and schools involved in the tests used and reused recognizable facial types and that art scholars can make use of this phenomenon to make attribution of works of art. Furthermore, the authors stress that there is a need to study issues like this using a multidisciplinary approach, combining theories of art history, perception, and computerized facial recognition.

Some work have already explored computer analysis of faces in paintings. Sablatnig et al. [64] proposed a method to analyze the authorship of mini portraits of the Austrian royal family by evaluating the shape of faces and brush strokes. Gupta et al. [40] used a deep learning-based facial recognizer to verify the identity of renaissance-era portrait models, seeking to find which different paintings portrayed the same person.

Thinking about perceptual aspects related to facial aesthetics, with aesthetics being the study of the perception of what is beautiful, facial attractiveness can be considered as a result of aesthetic perception. Schmid et al. [68] carried out work with the objective of developing an attractiveness metric based on various face measurements, including symmetry, Golden ratio metrics, as well as metrics used by Renaissance artists as guides to paint beautiful faces.

Our work meets the analysis of influence between artists, assuming that the representation of faces is an important part of the artworks, where the authors use inspiration and dedicate a lot of work. Seeking to collaborate with insights both for the history of art and for research in computational aesthetics, we seek to point out in what kind of characteristics these faces are most similar, taking into account the composition of the painting, the proportions used in the construction of the faces, their position and the presence and intensity of facial expressions. In the next sections, we will present the dataset used in the present work and the extracted visual features (Section 5.3), the methodology used (Section 5.4), and, finally, the results (Section 5.5).

This is the third experiment of this work, and we focus again on showing that visual features can be used to respond to perceptual questions involving computational aesthetics. In this specific experiment, we focus on visual features extracted from the faces in the works-of-art.

5.3 Dataset and Features

We build our dataset based on the ground truth presented by Saleh el al. [65], briefly described in Section 5.1. Firstly, it is important to mention that Saleh et al. did not make their dataset public available, so we need to create a dataset from scratch. Fortunately, they presented in their paper the ground truth of influences, composed of 66 artists. We searched the 66 artists on the WikiArt[1] website and found 62 of them. Through scraping, we downloaded 17,904 images of paintings from the 62 artists found. Since the objective of this work is to evaluate the influence based on the faces of the artworks, we firstly detected and cropped the faces using the OpenFace 2.0 software [5]. We chose this software to perform this task as it provides other information related to the cropped face that is useful for analysis, e.g. landmarks, pose, and gaze, detailed later. In this process, 8,435 faces were detected in 4,437 works of art performed by 56 artists. All faces are cropped, aligned vertically with nose to center, and a mask is applied to remove the background from the image. An example of face detection and cropping is shown in Figure 5.2, where we can see the detected face landmarks, the gaze and head orientation detection, and the cropped face output. Some artworks have more than one face detected, as in the example of Figure 5.3. In these cases, we consider only the largest face of each painting, so that a work is not represented more than once in the dataset. Therefore, the final dataset, which we call the *complete dataset*, is composed of 4,437 faces from 56 artists, as shown in Table 5.1, and the 56 influence relationships between the artists of the dataset that make up the ground truth are shown in Table 5.5. Note that the difference between Saleh's work (66) and our work (56) is due to some artists we discarded because artworks do not have faces. Figure 5.1 illustrates such process to build the dataset.

Furthermore, we test to compute the relationships only for a certain period. As the 20th century went through transformations in the paradigm of how art itself and style are seen within art [34], we made a cut based on the period of life of the artists, which we call the *temporal subset*. In this dataset, we kept all the artists who lived until 1900, i.e., 27 artists, identified in Table 5.1, and the ground truth is composed of 26 relations, identified in Table 5.5.

In addition to detecting and cropping faces, OpenFace provides a series of face information, which we use in our work, as follows:

[1] https://www.wikiart.org/

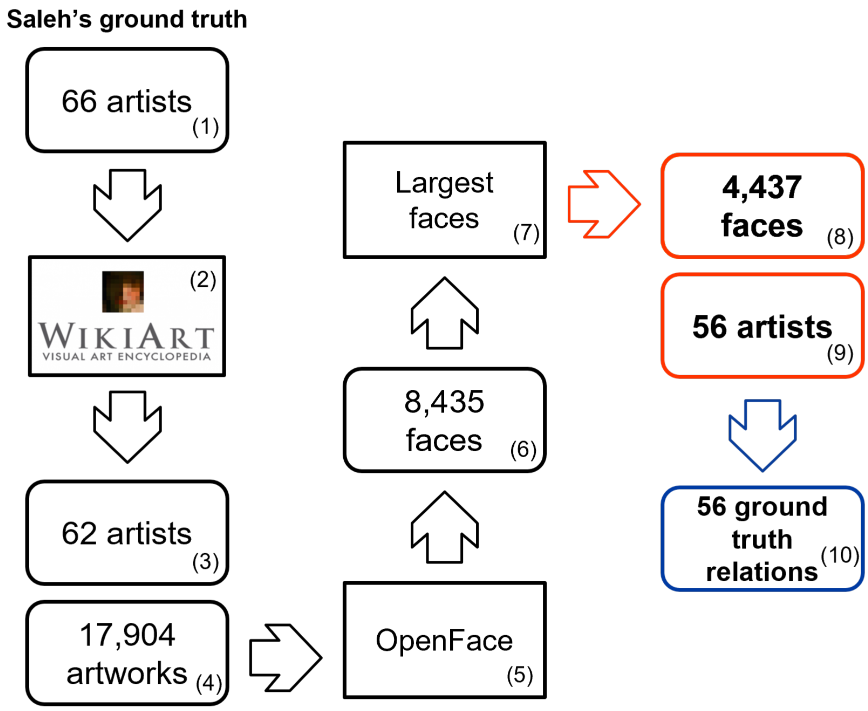


Figure 5.1: Dataset and ground truth construction process. Initially, we identified the 66 artists indicated by the ground truth of Saleh's [65] work (1), we searched for them on Wikiart (2) where we found 62 artists (3). Of these 62 artists, we downloaded 17,904 artworks (4) and through the Openface software (5) we detected 8,435 faces (6). Of the artworks with more than one face, we kept only the largest face (7), thus leaving 4,437 faces (8) from 56 different artists (9). These artists have 56 different relationships of influence among themselves, thus forming the ground truth used in this work (10).

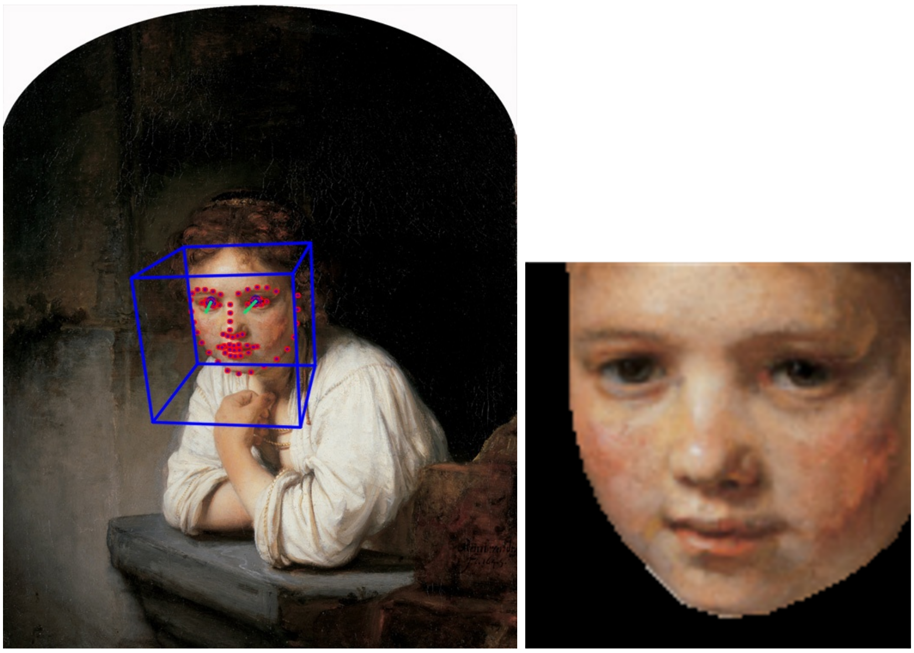
1. **Eye Gaze**: Two gaze direction vectors in world coordinates, one for the left eye and one for the right eye, and the horizontal and vertical angle of gaze direction for both eyes, illustrated in Figure 5.2a;

2. **Pose**: Vector of the location of the head relative to the camera in millimeters, and vector of the rotation in radians, in world coordinates with the camera being the origin, around the X, Y, Z axes illustrated in Figure 5.2a;

3. **Rigid shape**: Parameters of a Point Distribution Model (PDM), a linear model used to parameterize the shape of a face using a set of parameters, used in the landmark detection process, where the rigid shape parameters describe the position of the face in the image (scaling, rotation and translation) [5];

4. **Action Units (AUs)**: Intensity information (0 to 5) of 17 AUs and presence (0 absent, 1 present) of 18 AUs, used as a way to describe human facial expression, as proposed by Ekman and Friesen [23].

The aesthetic perception of faces is related to attractiveness. According to Graf and Landwehr [38], the literature on aesthetic preferences treats aesthetic taste and attrac-



(a) Face detected in paint. (b) Cropped face.

Figure 5.2: *Girl at a Window*, (Rembrandt, 1645). In (a) the detected face is shown with landmarks (red), gaze (green) and head orientation (blue). In (b) the face of (a) cut out and with the mask removing the background.

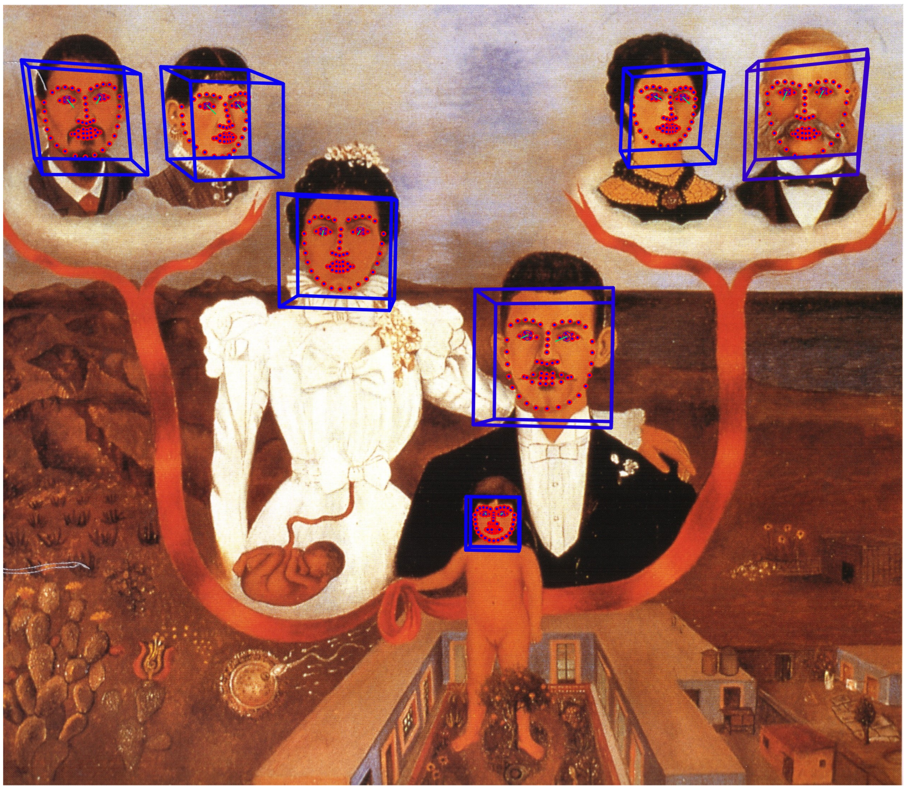


Figure 5.3: *Mis Abuelos, Mis Padres y Yo*, (Frida Kahlo, 1936). Example of painting with multiple faces detected. The face representing this painting in the dataset is the face of the most centered man in a white tie (the father).

tiveness judgments as equivalent concepts. Thus, measuring the attractiveness of a face is also measuring its aesthetics. In their work, Schimid et al. [68] systematically investigated the relationship between certain measurements of a face and its attractiveness. Using the calculation proposed by Schmid et al., applied to landmarks extracted by OpenFace, we extracted the following information:

1. **Neoclassical canons**: Measures proposed by artists from the Renaissance period as guides for drawing beautiful faces, define 9 pairs or trios of face segments that, according to them, should have equal sizes. Based on landmarks, 6 of the Neoclassical Canons were calculated and the measure used was the coefficient of variation between these pairs and trios of segments, where the closer to zero, the closer they would be to the ideal measure;

2. **Symmetry**: Three different measures of symmetry between the left and right sides, based on the centerline of the face, for the top of the eyebrows, inner edge, outer edge and base of the eyes, width of the nose, top and side of the lips, and width of the face, totaling 21 measures. The measurements being the ratio of the distances, the natural log of the ratio of the distances, and the adjusted distance difference. For the fitted difference and natural log of the ratio, a value of zero implies symmetry and the farther from zero, the more asymmetric. For proportion, a value of 1 indicates symmetry and the farther from 1, the more asymmetric;

3. **Golden ratios**: Measure 17 different ratios between the size of pairs of face segments, vertically and horizontally, such as mideye distance to interocular distance, nose width to lip height, mouth width to nose width, and so on. The original idea of the work that proposed these measures was to identify whether those values approached the golden ratio, that is, 1.618. The closer the measurements are to 1.618, the more beautiful the face would be.

Table 5.1: Artists that make up the final dataset. Artists with * on their death day represent artists who lived until the 19th century and are considered in the temporal subset.

Artist	Birth Day	Death Day	Detected Faces	Biggest Face per Painting
Frederic Bazille	1841-12-06	1870-10-28*	34	20
Giovanni Bellini	1430-01-01	1516-11-29*	137	65
William Blake	1757-11-28	1827-08-12*	42	28
Sandro Botticelli	1445-01-01	1510-05-17*	205	71
Francis Bacon	1909-10-28	1992-04-28	57	43
Max Beckmann	1884-02-12	1950-12-28	52	34
Gustave Caillebotte	1848-08-19	1894-02-21*	49	38
Robert Campin	1375-01-01	1444-04-26	73	29

Caravaggio	1571-09-29	1610-07-18*	118	60
Paul Cezanne	1839-01-19	1906-10-22	124	110
Marc Chagall	1887-07-07	1985-03-28	1287	68
Eugene Delacroix	1798-04-26	1863-08-13*	100	56
Albrecht Durer	1471-05-21	1528-04-06*	302	167
El Greco	1541-01-01	1614-04-07*	192	119
Théodore Géricault	1791-09-26	1824-01-26*	41	38
Lorenzo Ghiberti	1378-01-01	1455-12-01*	2	2
Francisco Goya	1746-03-30	1828-04-16*	244	162
Juan Gris	1887-03-23	1927-05-11	3	3
David Hockney	1937-07-09		71	38
Jean Auguste D. Ingres	1780-08-29	1867-01-14*	262	189
Jasper Johns	1930-05-15		3	2
Frida Kahlo	1907-07-06	1954-07-13	108	72
Wassily Kandinsky	1866-12-16	1944-12-13	7	7
Ernst Ludwig Kirchner	1880-05-06	1938-06-15	61	48
Gustav Klimt	1862-07-14	1918-02-06	62	49
Paul Klee	1879-12-18	1940-06-29	4	4
Leonardo da Vinci	1452-04-15	1519-05-02*	52	33
Roy Lichtenstein	1923-10-27	1997-09-29	4	4
August Macke	1887-01-03	1914-09-26	25	20
Kazimir Malevich	1879-02-23	1935-05-15	33	29
Edouard Manet	1832-01-23	1883-04-30*	122	106
Andrea Mantegna	1431-01-01	1506-09-13*	390	91
Franz Marc	1880-02-08	1916-03-04	2	2
Michelangelo	1475-03-06	1564-02-18*	175	66
Piet Mondrian	1872-03-07	1944-02-01	6	6
Berthe Morisot	1841-01-14	1895-03-02*	91	86
Robert Motherwell	1915-01-24	1991-07-16	1	1
Edvard Munch	1863-12-12	1944-01-23	64	55
Georgia O'Keeffe	1887-11-15	1986-03-06	1	1
Camille Pissarro	1830-07-10	1903-11-13	53	52
Pablo Picasso	1881-10-25	1973-04-08	185	157
Raphael	1483-01-01	1520-01-01*	344	124
Rembrandt	1606-07-15	1669-10-04*	438	347
Pierre-Auguste Renoir	1841-02-25	1919-12-03	736	616
Gerhard Richter	1932-02-09		9	8
Auguste Rodin	1840-11-12	1917-11-17	28	25
Henri Rousseau	1844-05-21	1910-09-02	57	31

Peter Paul Rubens	1577-06-28	1640-05-30*	696	284
Alfred Sisley	1839-10-30	1899-01-29*	4	4
Titian	1488-01-01	1576-08-27*	425	212
Jan van Eyck	1395-01-01	1441-07-09*	106	54
Vincent van Gogh	1853-03-30	1890-07-29*	152	145
Diego Velazquez	1599-06-06	1660-08-06*	147	113
Johannes Vermeer	1632-10-31	1675-12-15*	43	34
Andy Warhol	1928-08-06	1987-02-22	125	68
Norman Rockwell	1894-02-03	1978-11-08	281	141

Finally, using the images of the cropped faces, we extract color and clutter information, and using the landmarks, other proportion features in addition to the ones used to study the attractiveness:

1. **Colors**: Mean and standard deviation of each of the three color channels in the HSV space;
2. **Clutter**: Ratio of edge pixels compared to the number of pixels in the image;
3. **Proportions**: Difference between eye sizes, ratio of eye size to face size, ratio of center of mouth size to whole mouth size, ratio of mouth size to face size, and ratio of face size compared to the size of the entire painting.

In addition to mentioned individual features, we propose to explore the following feature groups:

- *Composition*: Color and clutter features;
- *Proportion*: Proportion features, Neoclassical Canons, Symmetry, and Golden ratio;
- *Position*: Features of gaze, pose, and rigid shape;
- *Expression*: Features of the intensity of AUs and amount of active AUs.

In the case of images (faces) where some feature cannot be extracted, we input the missing values with the median of the feature of the artist's paintings. This was the case for the attractiveness features (Neoclassical canons, Symmetry and Golden ratio), where 2,538 images had the face rotated, preventing measurements from being calculated on both sides of the face. We could use other strategies, e.g., look for another face in the same painting instead to input with median values the missing parameters, or even treat the 3D information to generate data about the face without rotation. For the moment, we use only the imputation schema, for simplicity, but other strategies can be tested in future work. For visual features extracted with world coordinates, the values were normalized, according to the Equation 5.1:

$$z_i = \frac{x_i - \mu}{\sigma}, \tag{5.1}$$

where μ is the mean and σ is the standard deviation of all observations of the feature, x_i is the original value of observation i of the feature and z_i is the normalized value of observation i. The normalization is important once the Hausdorff distance is used.

In the next section, we detail the methodology used to analyze the problem through the calculation of similarity and the form of evaluation of the results found.

5.4 Methodology

To identify possible relationships between artists based on the faces of their pieces-of-art, it is necessary to measure which artists are most similar to each other. For this, using the visual features previously mentioned in Section 5.3, we present the methodology for calculating the similarities, and the method used of validate and compare of the obtained results.

5.4.1 Similarity

For an artist j to have been influenced by an artist i, artist i must have been born before the artist j, or at least have been contemporaries. To ensure that the influence relationships follow this logic, we only consider relationships in which the artist influencer i was born before the death of the influenced artist j.

As each artist has painted pieces-of-art (each one with a face), to calculate the similarity between two artists we consider each artist as a set composed of the faces of his/her artworks, where artist i has the set of faces P^i and artist j has the set of faces P^j. From there, we calculated the similarity between artists using the asymmetric distance $D_{q\%}$ based on the Hausdorff distance proposed in [65] and defined by:

$$D_{q\%}(P^i, P^j) = \overset{q\%}{\underset{k}{\max}}\, d(p^i_k, P^j), \tag{5.2}$$

where we consider the distance measure $D_{q\%}(P^i, P^j)$ between influenced artist i of artist influencer j as the Euclidean distance q percentile between each painted face $p^i_k \in P^i$ of artist i for the set P^j of painted faces of artist j. We used q = 50%, which represents the median distance between the face p_k^i and the set P^j, for comparison purposes with the results of the work by Saleh et al. [65].

5.4.2 Evaluation

After calculating the distance between each artist and their possible influencers, we generate a list of the top-5 closest artists, in terms of distance $D_{q\%}$, for each of the artists in the dataset and compare it to the ground truth. As discussed in Section 5.3, the ground truth we used, provided by Saleh et al. [65], was constructed only with consensual influence relationships among art historians. Based on the 56 artists present in our dataset, the ground truth is composed of 56 influence relationships between pairs of these artists, thus being a sparse dataset, where most artists have a number of influencers less than 5 or even there is none.

As this is a sparse dataset, compared to our list of top-5 computed influence relationships, metrics such as accuracy are not the best assessment option. A good evaluation metric, which even allows us to make comparisons with Saleh's work, is to identify how many of the influence relationships calculated by our method are in accordance with the ground truth which represents the true influence relationships. Therefore, the metric used in the evaluation of the work is Recall, as defined below:

$$Recall = \frac{|h|}{N}, \tag{5.3}$$

where $|h|$ is the number of ground truth influence relationships found among the top-5 computed influence relationships, and N is the total amount of ground truth influence relationships.

As detailed in Section 5.3, we created 4 different groups of visual features: Composition, Proportion, Position, and Expression. Using each group separately, we generated the top 5 influence relationships for each artist, based on Equation 5.2, and ratings in terms of recall.

We also evaluate the results by combining the result of feature groups, which we call the *feature combination*. For this, from the set of top-5 influences computed by each feature group, we selected only the influence relationships that had the smallest distance $D_{q\%}$ between the artists, based on the median, that is, we kept only half of the influence relationships with the smallest distance of each feature group. We then combined influences from all groups, excluding repeated influence relationships, and then evaluated the results in terms of recall.

Finally, we also evaluate influence relationships that we call *second-degree influences*. We consider second-degree influences when the methodology suggests that a certain artist j influenced artist i, but in fact the j artist influenced an m artist who ultimately influenced the i artist. We hypothesize, once it is not reported in the ground truth, that characteristics of certain painters may have passed through generations and our methodology may help to identify these influences. Figure 5.4 provides an example of second-degree influences. Thus, we calculated the results through Equation 5.3 considering the second-degree influences on $|h|$, both for the complete dataset and the temporal subset, for each feature group and for the feature combination.

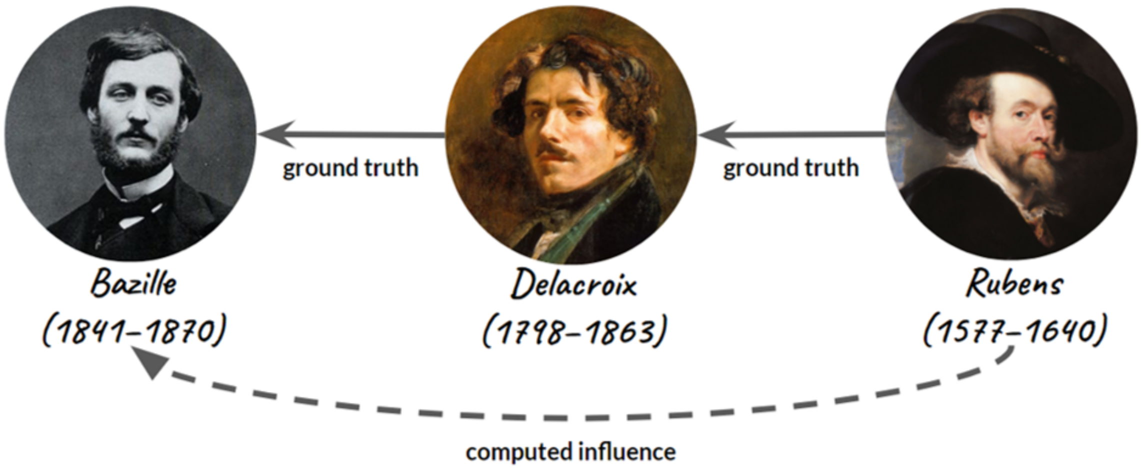


Figure 5.4: second-degree influence example. In this case, the ground truth presents Peter Paul Rubens as an influencer on Eugene Delacroix, and Eugene Delacroix as an influencer on Frederic Bazille, while in our computed relations Peter Paul Rubens appears as an influencer on Frederic Bazille. When considering the second-degree influences, we assume that the method identified this indirect relationship correctly.

5.5 Results

This section presents the results of our work applying the methodology described in our dataset. First, we present how we made the comparisons with the state-of-art, then we present the results by feature group, the results of the combination of features, and, finally, we talk about the visualization built to analyze the relationships.

5.5.1 Comparisons

To compare the results in a fair way, we recalculated the results of the work by Saleh et al. [65], using the same artists we have in our dataset, i.e., 56 artists and not 66. For the recalculation, we will use the top-5 relations of each artist calculated by Saleh et al. [65], where the authors originally reached Recall = 29% using Classeme, Euclidean

Distance and q = 50% features, and compared with ground truth composed by the 56 artists of our dataset, as shown in Table 5.5. We recalculate the results for our complete dataset, for the temporal subset, and considering the second-degree influences as detailed in this section earlier, and we present the results of the recalculation along with our results in next sections.

5.5.2 Feature group

As detailed in Section 5.4, we calculate the $D_{q\%}$ distance between artists considering each feature group separately and then compute the top-5 influence relationships for each artist. Each feature group generated 278 different influence relationships for the complete dataset and 138 influences for the temporal subset, which were evaluated in terms of recall, as described in Section 5.4.2. So, for comparison purposes, we also recalculated the result of the Saleh's work, composed of 290 influences considering the complete dataset and 124 influences for the temporal subset. The results can be seen in Table 5.2 and are discussed in the next sections.

Table 5.2: Result of calculations using each group of visual features separately and comparison with the results of Saleh et al. [65]. The best results are highlighted in bold.

			Second-degree influences	
Feature group	Complete dataset	Temporal subset	Complete dataset	Temporal subset
Composition	25.00%	**50.00**%	29.85%	**53.85**%
Proportion	21.43%	38.46%	25.37%	50.00%
Position	23.21%	30.77%	26.87%	42.31%
Expression	21.43%	34.61%	29.85%	50.00%
Saleh et al. [65]	**37.50**%	46.15%	**38.81**%	46.42%

5.5.3 Feature combination

After computing the results for each feature group separately, we performed the feature combination as described in Section 5.4.2. Thus, for the complete dataset, we kept 139 of the 278 influence relationships for each feature group. Removing the repeated relationships, we kept a total of 410 influence relationships, reaching **Recall = 32.14%**. As for the temporal subset, we kept 69 of the 138 influence relationships of each feature group, and, removing the repeated relationships, we kept a total of 181 influence relationships, reaching **Recall = 65.38%**. For comparison purposes, we also recalculated the result of Saleh's work. The recalculated results for the complete dataset were Recall = 37.50%, based on 290 influence relations. For the temporal subset, the recalculation of results of Saleh's work reaches a Recall = 46.15%, based on 124 influence relations.

Finally, we assessed second-degree influences. For this, we recalculate the results of the feature combination for the complete dataset and the temporal subset considering as correct the second-degree influences as well. Thus, we arrive at a **Recall = 49.25%** with the complete dataset and **Recall = 82.14%** for the temporal subset. For comparison purposes, we also recalculated the results of Saleh's work considering second-degree influences, obtaining Recall = 38.81% for the complete dataset and Recall = 46.42% for the temporal subset. All such results of the feature combination presented can be consulted in Table 5.3, together with the comparison with the work by Saleh et al. [65].

Table 5.3: Results of calculations with feature combination and comparison with the results of Saleh et al. [65]. The best results are highlighted in bold.

			Second-degree influences	
Results	Complete dataset	Temporal subset	Complete dataset	Temporal subset
Our results	32.14%	**65.38**%	**49.25**%	**82.14**%
Saleh et al. [65]	**37.50**%	46.15%	38.81%	46.42%

It is interesting to note that the original result of Saleh's work was Recall = 29%. By keeping in the ground truth only the relations of artists that have faces, their results improved by 9%, reaching Recall = 37.50% in our complete dataset, and improved by over 17%, reaching Recall = 46.15% when we use the temporal subset. This seems to indicate that both our initial hypothesis that faces are important clues to identify influences between artists makes sense, as well as the temporal cut made in the temporal subset. This result improvement considering the temporal subset is also observed in our results, regardless of considering second-degree influence relations in the recall calculation or not.

The feature group that obtained the best results was the Composition group, which contains color and clutter information. This result indicates that these features are relevant to the stylistic representation of faces by artists. This includes variations in color usage and the level of detail and elements incorporated into the facial depictions. The results were even superior to the result of Saleh's work when compared to the temporal subset. The feature groups that obtained considerable improvement when considering second-degree influences were the Expression group, with an improvement of little more than 15%, and the Proportion and Position groups improved by more than 11%.

5.5.4 Visualization

To facilitate the exploration of the found results and obtain insights, we proposed an interactive web application[2], built in Shiny [15], where it is possible to visualize the influence relations proposed by our work in arc diagrams, as exemplified in Figure 5.5. In these diagrams, the artists are temporally ordered (from the left to de right) and when our methodology indicates some relationship of influence between them, they are connected by an arc, indicating that the more recent artist may have been influenced by the older artist. Furthermore, the size of each artist's node represents how many other artists he/she has influenced, according to our method, that is, the larger the node, the greater the number of those influenced by it, and the color of the node indicates what century the artist has lived up to. It is also possible to highlight in the diagram the influence relationships computed by our work that are in accordance with the ground truth, as well as the second-degree influence relationships considered, and which of our relationships that are not in the ground truth but were also identified by the work of Saleh et al. [65].

As detailed in Figure 5.6, the web application also allows viewing the relationships computed by each of the feature groups, considering the complete dataset or the temporal subset, and viewing all the influence relationships of a specific painter, either as influencer or influenced. Figure 5.7 presents artist Théodore Géricault's relationships as an example.

Since the artists are sorted in descending order by the date of their death, in general, the artists at the left end of the arch were influencing those at the right end of the arch. However, as there are relationships between contemporary artists, this is not always the case. To facilitate the identification of who is the influencer and who is influenced, it is also possible to see the list of relationships between the artists in a table format, as shown in Figure 5.8. In this table, there is the name of the artist, by whom he/she was influenced, and if the relationship is on the ground truth.

The web application also presents the top-10 artists who most influenced other artists, according to the influence relations computed by our work. The artists who most influenced other artists, according to the ground truth, can be found in Table 5.4. Based on the complete dataset, the influences computed using the Proportion feature group had 5 of the top-10 influences in agreement with those presented in Table 5.4: Titian, Peter Paul Rubens, Diego Velazquez, Pablo Picasso, and Giovanni Bellini. The other feature groups and the feature combination had between 3 and 4 artists in their top-10 according to the top-10 ground truth. Analyzing all feature groups and the feature combination, from the top-10 artist influencers according to the ground truth, Titian always appears as one of the top-10 artist influencers in our results, followed by Peter Paul Rubens and Giovanni Bellini.

[2]https://brunamdalmoro.shinyapps.io/influence_face_of_art/

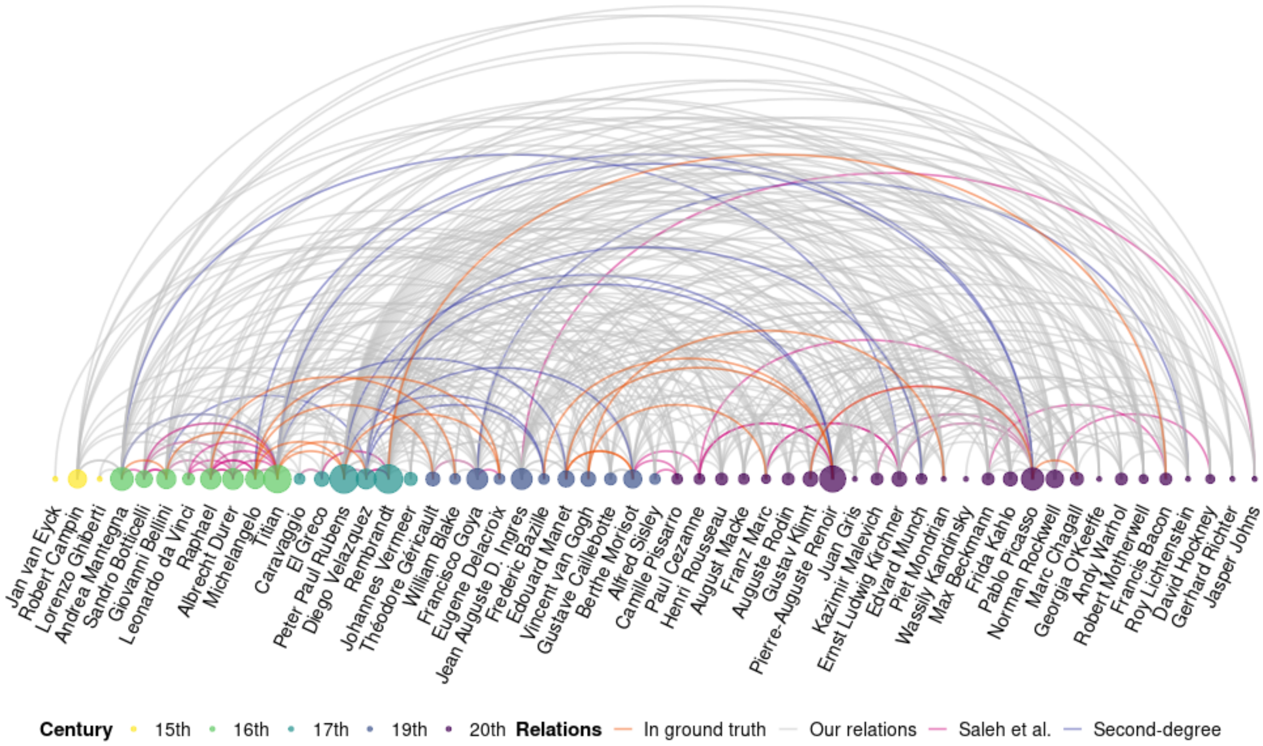


Figure 5.5: Arc diagram of the 410 influence relationships computed by our work using the complete dataset and feature combination. The larger the nodes, the more influential the artist was, according to our methodology. The colors of the nodes indicate which century each artist has lived through, and they are ordered in ascending order from left to right. The highlighted arcs refer to relationships that our methodology suggested and that are also in the ground truth (orange), they are second-degree influences present in the ground truth (blue), they were computed by Saleh et al. [65] (magenta).

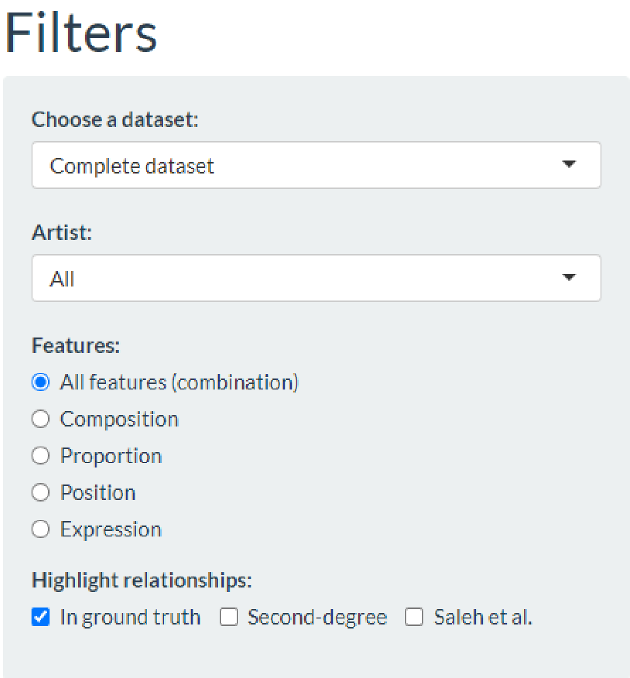


Figure 5.6: Filters present in the results visualization tool.

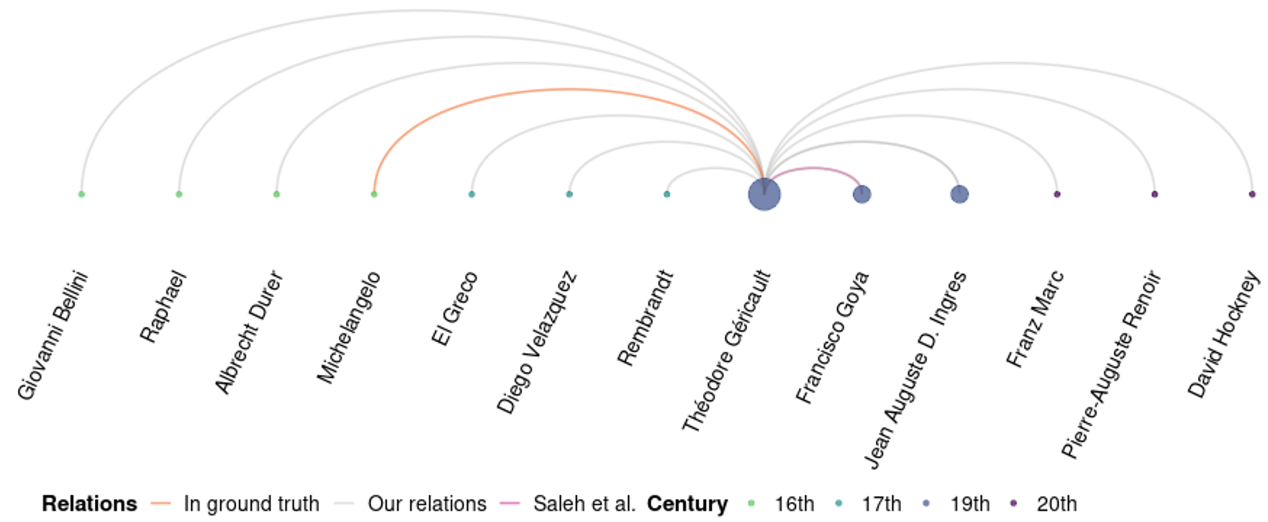


Figure 5.7: Arc diagram showing the influence relationships computed by our work for Géricault. In orange is highlighted Michelangelo's influence relationship on Théodore Géricault who is in the ground truth, and in pink is highlighted Théodore Géricault's influence relationship on Francisco Goya who is not in ground truth but also appears in Saleh et al. [65].

Show 10 entries Search:

	artist_name	influencer_name	ground_truth
391	Marc Chagall	Norman Rockwell	No
392	Max Beckmann	Norman Rockwell	No
393	Titian	Giovanni Bellini	Yes
394	Berthe Morisot	Edouard Manet	Yes
395	Pierre-Auguste Renoir	Edouard Manet	Yes
396	Giovanni Bellini	Andrea Mantegna	Yes
397	Eugene Delacroix	Michelangelo	Yes
398	Peter Paul Rubens	Michelangelo	Yes
399	Théodore Géricault	Michelangelo	Yes
400	Edouard Manet	Berthe Morisot	Yes

Showing 391 to 400 of 410 entries Previous 1 ... 37 38 39 40 41 Next

Figure 5.8: List of influence relationships shown in the arc diagram. The information presented are the name of the artist (artist_name), by whom he was influenced (influencer_name), and if the relationship is on the ground truth (ground_truth).

The other artists who are also among our top-10 artist influencers results are Michelangelo, Diego Velazquez, Pablo Picasso, and El Greco.

The small number of faces of some artists, which can be seen in Table 5.1, tends to affect the number of influencer relationships computed by our work. Artists like Georgia O'Keeffe, Juan Gris, and Lorenzo Ghiberti, for example, were not considered to be anyone's influencers. Therefore, since the intention in this work is also to suggest relations of influence

Table 5.4: According to the ground truth, list of the top-10 artist influencers and the number of artists influenced by them.

Artist influencer	Number of influenced artists
Michelangelo	6
Pablo Picasso	5
Edouard Manet	4
Titian	3
Vincent van Gogh	3
Paul Cezanne	3
Giovanni Bellini	2
El Greco	2
Peter Paul Rubens	2
Diego Velazquez	2
Eugene Delacroix	2
Edvard Munch	2

between artists regarding face painting, it makes sense that artists who rarely use faces in their work should not be an influence in this regard.

Several of the relationships computed by our methodology and by Saleh et al. [65], regardless of whether using the complete dataset or the temporal subset, are mostly concentrated among 16th-century artists, as can be seen in Figure 5.9, where the magenta arcs are concentrated on the light green node artists. The fact that both work - ours work and Saleh's work - have computed these influence relations may indicate that the artists of this period have much more influence relations than is indicated by the ground truth, or it may indicate that the artists who emerged in the periods subsequent ones began to produce artworks with different visual aspects.

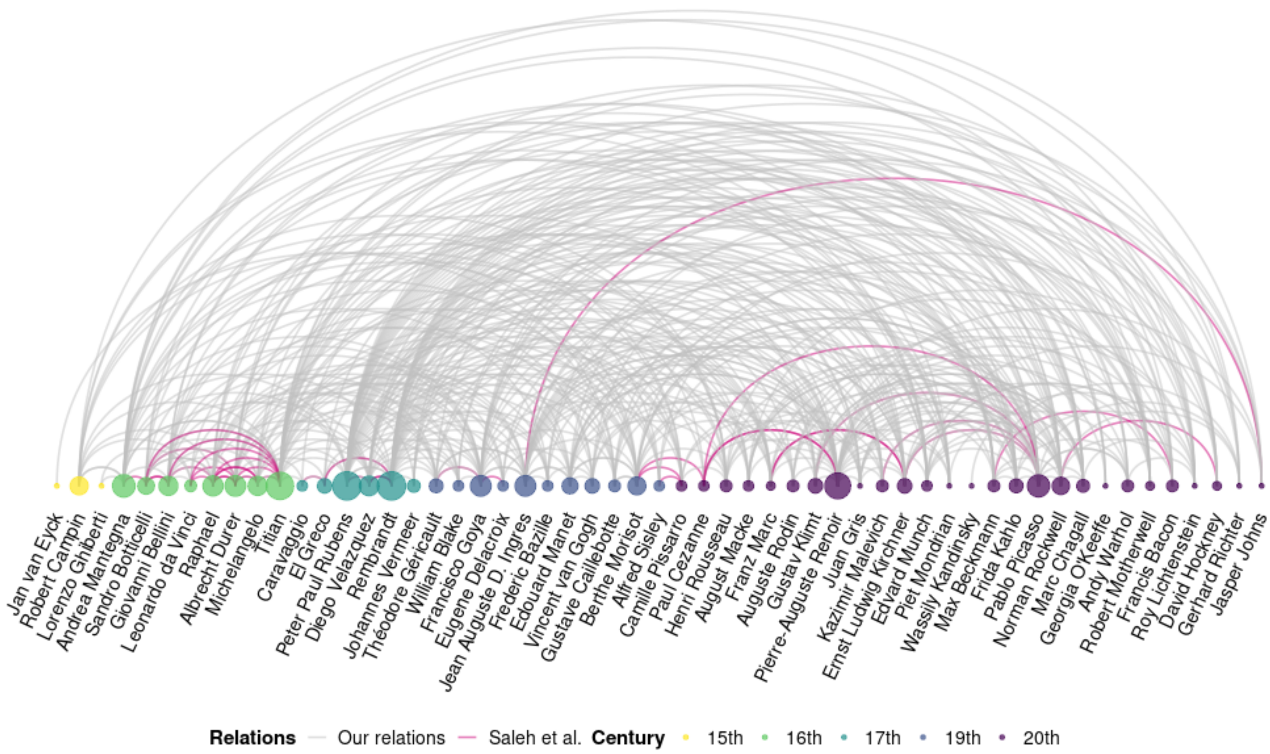


Figure 5.9: Arc diagram with the influence relationships computed by the complete dataset and Composition group. Highlighted in the round are the influence relations computed by our work and by Saleh's work, which are not in the ground truth.

Table 5.5: Relationship pairs that make up the ground truth of the 56 artists in our dataset. Artists with * represent relationships between artists who lived until the 19th century and are considered in the temporal subset.

Artist	Influenced by			
Frederic Bazille*	Alfred Sisley*	Edouard Manet*	Eugene Delacroix*	Pierre-Auguste Renoir
Giovanni Bellini*	Andrea Mantegna*			
William Blake*	Michelangelo*	Raphael*		
Francis Bacon	Diego Velazquez	Pablo Picasso	Rembrandt	Vincent van Gogh
Max Beckmann	Edvard Munch	Paul Cezanne		
Paul Cezanne	Camille Pissarro			
Marc Chagall	Pablo Picasso			
Eugene Delacroix*	El Greco*	Michelangelo*	Peter Paul Rubens*	
Albrecht Durer*	Andrea Mantegna*	Giovanni Bellini*		
El Greco*	Michelangelo*	Titian*		
Theodore Géricault*	Michelangelo*	Peter Paul Rubens*		
David Hockney	Pablo Picasso			
Jean Auguste D. Ingres*	Raphael*			
Frida Kahlo	Sandro Botticelli			
Wassily Kandinsky	Franz Marc	Paul Cezanne		
Gustav Klimt	Pablo Picasso			
Ernst Ludwig Kirchner	Albrecht Durer			
Roy Lichtenstein	Jasper Johns	Pablo Picasso		
August Macke	Edvard Munch			
Kazimir Malevich	Paul Cezanne			
Edouard Manet*	Berthe Morisot*	Diego Velazquez*		
Franz Marc	Vincent van Gogh	Wassily Kandinsky		
Michelangelo*	Ghiberti*			
Piet Mondrian	Vincent van Gogh			
Berthe Morisot*	Edouard Manet*			
Edvard Munch	Edouard Manet			
Pablo Picasso	El Greco	Francisco Goya		
Pierre-Auguste Renoir	Edouard Manet	Eugene Delacroix		
Auguste Rodin	Michelangelo			
Peter Paul Rubens*	Michelangelo*	Titian*		
Titian*	Giovanni Bellini*			
Vincent van Gogh*	Camille Pissarro			
Diego Velazquez*	Caravaggio*	Titian*		
Johannes Vermeer*	Caravaggio*			
Andy Warhol	Jasper Johns			

6. DISCUSSION

Computational aesthetics is a subfield of computer vision that seeks to understand the human aesthetic perception of images and image sequences, and create systems that make different aesthetic decisions trying to approach the judgment of a human being about the images [12] [47]. Visual features detection are the way used to extract information from images analyze them in different ways. In our work we explore three different problems related to the image through the aesthetic computation approach, testing and proposing visual features to analyze the images. We work with three different types of images: real images, computer-created images (animation), and paintings, in addition we explore also sequences of images (videos).

In the first work presented, we proposed a new way of approaching the problem of prediction of popularity of online videos proposed in [78], using the same technique and part of the same dataset, but treating the problem as a classification problem and using only visual features as predictors. Using Support Vector Machine with Gaussian Radial Basis Function we predict which of the 1,820 videos published on Facebook had more than a certain number of views seven days after their publication, based solely on visual features so that such analysis can be produced before publication. Our predictive model performed better when using the video characteristics and rigidity features, obtaining Kappa of 0.7324, sensitivity of 0.8930, and positive predictive value of 0.8930.

In the second work, we propose the construction of a content-based top-N recommendation system using the movie category (animation or live-action) explicitly with visual features to improve the recommendations. The proposed technique obtained the best result when using the *explicit dataset*, with *Just Similarity*, using the movie category with the text per frame, clutter metric, shot cuts, and color value visual features. Our method performed 109% better than the random recommendations and similarly to the work based on deep learning, where authors made recommendations 116% better than random recommendations.

Finally, in the third work, we assess the influence relationships between artists based on how they paint faces in their artworks. We use four different groups of visual features: Composition, Proportion, Position, and Expression. Testing the results separately by group, the group that obtained the best result, regardless of the test performed (complete dataset, temporal subset, second-degree influences), was composition, which includes color and clutter features, reaching a Recall = 53.85% in the temporal subset and considering second-degree influences. When evaluating the results by combining the closest influences computed from all features, we obtained even better results, reaching Recall = 82.14% in the temporal subset and considering second-degree influences. Our results surpassed Saleh's results, except for the result obtained with the complete dataset. The improvement in the result of Saleh et al. [65] when considered only artists who have faces reinforces the hypothesis that faces are elements that inspire influence among artists and that help to identify those relationships.

6.1 Revisiting the work goals

In this section, we retake our objectives and analyze them, as follows:

- **Objective 1:** *To explore different problems involving different types of images from the perspective of computational aesthetics.*

We work with real image and animation sequences, as well as paintings. Our results indicated that, regardless of the type of image or video, visual features are good information to use in predicting the popularity of videos posted online, to make better movie recommendations using in conjunction with the type of image (animation or live-action), and to explain the relationship between artists based on the painted faces of their artworks.

- **Objective 2:** *To use only visual features to solve problems as a way to test the power and usefulness of this information in different applications.*

The problems we explore in this work can be solved computationally in different ways, using metadata information, as well as textual information such as the name and description of the video, movies, or artwork, information about genre, style, among others. However, the information present in the images is relevant to the solution of all these problems.

In the present work, we explore the problems of popularity prediction, movie recommendation, and influence analysis on works of art using only visual features and we obtained good results, if compared with competitive methodologies. The advantage of using only visual features is that you can study a set of videos and images and make predictions and recommendations even when other information is not available, scarce, or unreliable.

- **Objective 3:** *To use only interpretable visual features to generate insights for the area of aesthetics and related areas.*

One of the most relevant points in addressing problems from the perspective of computational aesthetics is to allow areas such as experimental aesthetics and psychology, for example, to take advantage of the results obtained to gain insights into how human perception works and what kind of aesthetic response each image information can generate. For this to be possible, the visual features used to generate the results must be interpretable.

Thus, generic information extracted through deep learning models that cannot be directly related to the perceptual characteristics of the images makes this type of collaboration between the areas of computational and experimental aesthetics difficult or impossible.

In the present work, we only use interpretable information extracted from the images, such as information on color, faces, symmetry, proportions, and facial expression. We explore the results according to feature groups to facilitate the extraction of insights and understand which visual features collaborated the most to achieve the best results. In the problem of influences of works of art, for instance, we developed a way to explore the relationships between artists based on the proposed visual features, as a way to suggest relationships to be evaluated by researchers in the history of art and to be another possible input to relate artists where the influence is not so accepted or has never been considered.

6.2 Used visual features

Table 6.1 presents the visual features used in each of the three problems presented in this work. We can see that the color information was present in the three problems, either through color classes or through the mean and standard deviation (s.d.) of the HSV channels. Another piece of information that was used in the three works and always obtained good results was clutter: in the three explored problems, it is in the set of features that obtained the best results (video rigidity in popularity problem, clutter metric in recommendation problem, and composition in arts problem). Faces were also information present in the three works, but in different ways, since in the first two methods the information used was to verify the presence of faces in the video frames, while in the third problem they were the focus of all analyses.

However, it is very difficult, if not reckless, to assume that certain visual features are always good information that should be used for any application, whether it was good for popularity prediction, movie recommendation, or art analysis. In the field of arts, for example, it is precisely the intention to solve global problems that make specialists in the field disbelieve in the potential of computer science and computer vision applied to the arts [72]. In our work, for example, when we restrict the scope a little more to a certain historical period, we get better results than when we try to generalize the problem over a large period. The same thing happens with the results of Saleh et al [65]. when we analyze the relationships of artists who portray faces in their paintings, their results improve. In problems related to the prediction of popularity, the type of video explored is very important. We explore videos from Facebook news pages, but in different contexts, such as video classes or lectures and the same attributes of the images do not get the same results. The same thing about movie recommendations, where we explore different types of images (live-action and animation), but from the same studio.

Table 6.1: All features used to explore the three problems in this work. In bold, features that were used in two or more problems.

Feature	Prediction popularity of online videos	Movie Recommendation	Influence in paintings
Video duration	x		
Frames per second	x		
Frame height	x		
Frame width	x		
Shot cuts	**x**	**x**	
Type of shot cuts	x		
Video speed	x		
Video rigidity	x		
Video category		x	
Clutter	**x**	**x**	**x**
Color classes	x		
Dominant color	x		
Color saturation		**x**	**x**
Color value		**x**	**x**
Color hue			x
Color saturation (s.d.)			x
Color value (s.d.)			x
Color hue (s.d.)			x
Frames with text	**x**	**x**	
Text size	x		
Frames with faces	x		
Faces per frame	**x**	**x**	
Face size	**x**		**x**
Eye gaze			x
Pose			x
Rigid shape			x
Action Units			x
Neoclassical canons			x
Symmetry			x
Ratio			x
Diff. size eyes			x
Eye size			x
Mouth ratio			x
Mouth size			x

Thus, in this work, we were able to assess the usefulness of visual features, which can be an important input in the development of new studies in the area, combined or not with other information. Therefore, we do not intend to reach general conclusions about which are the best features for each problem or still for all problems. We believe that many more experimentation and studies have to be made in order to investigate if there is some feature that is a common alternative for many problems in the context of computational aesthetic. We indicate that, in problems with images, using visual features and exploring the vast options, according to the objectives of the work, can generate good results and it seems to be a promising solution.

6.3 Limitations

Even though we have achieved the proposed objectives, our work has some limitations. In the popularity prediction work presented in Chapter 1, we used the dataset provided by Deldjoo et al. [19] with information but without the videos. If we had access to the videos, we could explore other information about the image that was not available, e.g. to test the grouping of the videos by subject, include the context of the page. It's possible that, with tests like these, we could get even better results. In the movie recommendation work, we tested a small subset of Disney movies. The main limitation was due to the cost of extracting visual features frame by frame, in a common computer. Finally, in the work of works of art, when combining visual features, even creating the threshold to reduce the relations of each group, we get a greater number of computed relations than the work by Saleh et al. [65]. So our results tend to be better. In the recommendations within the feature groups, this problem does not happen.

6.4 Future Work

Even though the objectives of this work have been achieved, we still intend to further improve our work in the future.

In the popularity prediction problem, we would like to test other analyzes by shifting the study scope to video popularity assessment classes available on YouTube, trying to identify which visual characteristics attract students more to watch a video of a certain subject than another. Having our own selection of videos we can have more freedom to explore different groupings of videos and assess whether each area of knowledge, for example, tends to be more successful using certain visuals than others.

In the movie recommendation problem, we want to expand the dataset and the number of features used. For that, instead of extracting the features of each frame of the video, it is possible to use the keyframes of each scene to do extraction, avoiding the limitation of video processing in terms of computational cost. It would also be interesting to compare the results of the recommendations comparing different movie and animation studios.

As for the artwork, we intend to improve the way we combine the results of feature groups, not only to adjust the number of computed influence relationships, but also to consider the feature groups proportionally according to their recall, or that is, groups of features that have better recall would contribute more to influence relationships than the group of features with less recall. Furthermore, as a way of getting closer to the problems in the history of art, we started to work together with Prof. Dr. Charles Monteiro, from the School of Humanities at PUCRS, who collaborated in the insights of this work.

In this way, we see this work as the beginning of our exploration into the questions of computational aesthetics, an area that has been defined for a few years and expanding how into different problems.

REFERENCES

[1] Abboushi, B.; Elzeyadi, I.; Taylor, R.; Sereno, M. “Fractals in architecture: The visual interest, preference, and mood response to projected fractal light patterns in interior spaces”, *Journal of Environmental Psychology*, vol. 61, Feb 2019, pp. 57–70.

[2] Aggarwal, C. C. “Content-based recommender systems”. Springer International Publishing, 2016, chap. 4, pp. 139–166.

[3] Aggarwal, C. C. “An Introduction to Recommender Systems”. Springer International Publishing, 2016, chap. 4, pp. 1–28.

[4] Araujo, V.; Dalmoro, B.; Musse, S. R. “Analysis of charisma, comfort and realism in cg characters from a gender perspective”, *The Visual Computer*, vol. 37, Jul 2021, pp. 2685–2698.

[5] Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; Morency, L. “Openface 2.0: Facial behavior analysis toolkit”. In: IEEE International Conference on Automatic Face & Gesture Recognition, 2018, pp. 59–66.

[6] Baluja, S.; Pomerleau, D.; Jochem, T. “Towards automated artificial evolution for computer-generated images”, *Connection Science*, vol. 6–2, Oct 1994, pp. 325–354.

[7] Barthelme, S. “imager: Image processing library based on ’cimg”’. Source: https://CRAN.R-project.org/package=imager, Jan 2021.

[8] Birkhoff, G. D. “Quelques éléments mathématiques de l’art”. In: Congresso Internazionale dei Matematici, 1929, pp. 315–333.

[9] Birkhoff, G. D. “Aesthetic Measure”. Harvard University Press, 1933, 226p.

[10] Bo, Y.; Yu, J.; Zhang, K. “Computational aesthetics and applications”, *Visual Computing for Industry, Biomedicine, and Art*, vol. 1–1, Sep 2018, pp. 6.

[11] Boser, B. E.; Guyon, I.; Vapnik, V. “A training algorithm for optimal margin classifiers”. In: ACM Conference on Computational Learning Theory, 1992, pp. 144–152.

[12] Brachmann, A.; Redies, C. “Computational and experimental approaches to visual aesthetics”, *Frontiers in Computational Neuroscience*, vol. 11, Nov 2017, pp. 102.

[13] Bressan, M.; Cifarelli, C.; Perronnin, F. “An analysis of the relationship between painters based on their work”. In: IEEE International Conference on Image Processing, 2008, pp. 113–116.

[14] Canny, J. F. “A computational approach to edge detection”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8–6, Nov 1986, pp. 679–698.

[15] Chang, W.; Cheng, J.; Allaire, J.; Xie, Y.; McPherson, J. “shiny: Web application framework for r”. Source: https://CRAN.R-project.org/package=shiny, Jan 2021.

[16] Chen, L.; Bampis, C. G.; Li, Z.; Sole, J.; Bovik, A. C. “Perceptual video quality prediction emphasizing chroma distortions”, *IEEE Transactions on Image Processing*, vol. 30, Dec 2021, pp. 1408–1422.

[17] Cremonesi, P.; Koren, Y.; Turrin, R. “Performance of recommender algorithms on top-n recommendation tasks”. In: ACM Conference on Recommender systems, 2010, pp. 39–46.

[18] Dal Molin, G. P.; Nomura, F. M.; Dalmoro, B. M.; de A. Araujo, V. F.; Musse, S. R. “Can we estimate the perceived comfort of virtual human faces using visual cues?” In: IEEE International Conference on Semantic Computing, 2021, pp. 366 – 369.

[19] Deldjoo, Y.; Elahi, M.; Cremonesi, P.; Garzotto, F.; Piazzolla, P.; Quadrana, M. “Content-based video recommendation system based on stylistic visual features”, *Journal on Data Semantics*, vol. 5–2, Feb 2016, pp. 99–113.

[20] Deldjoo, Y.; Elahi, M.; Quadrana, M.; Cremonesi, P. “Using visual features based on MPEG-7 and deep learning for movie recommendation”, *International Journal of Multimedia Information Retrieval*, vol. 7–4, Jun 2018, pp. 207–219.

[21] Deng, F.; Ren, P.; Qin, Z.; Huang, G.; Qin, Z. “Leveraging image visual features in content-based recommender system”, *Scientific Programming*, vol. 2018, Aug 2018, pp. 1–8.

[22] Douchová, V.; et al.. “Birkhoff’s aesthetic measure”, *Acta Universitatis Carolinae Philosophica et Historica*, vol. 21–1, Aug 2015, pp. 39–53.

[23] Ekman, P.; Friesen, E. “Facial action coding system”, *Environmental Psychology & Nonverbal Behavior*, vol. 3–2, Jan 1978, pp. 5.

[24] Elahi, M.; Deldjoo, Y.; Moghaddam, F. B.; Cella, L.; Cereda, S.; Cremonesi, P. “Exploring the semantic gap for movie recommendations”. In: ACM Conference on Recommender Systems, 2017, pp. 326–330.

[25] Facebook. “In-stream ads — facebook for business”. Source: https://www.facebook.com/help/publisher/943230709179046, Apr 2020.

[26] Fang, N.; Zhang, Y.; Wei, Q.; Guo, Y. “Composition-aware learning for aesthetic assessment of edited photos”. In: IEEE International Conference on Computer and Communications, 2020, pp. 2151–2155.

[27] Fechner, G. T. “Elemente der psychophysik”. Druck und Verlag von Breitkopf u. Härtel, 1860, 500p.

[28] Fechner, G. T. “Zur experimentalen Aesthetik”. S. Hirzel, 1871, 88p.

[29] Fechner, G. T. “Vorschule der Aesthetik”. Druck und Verlag von Breitkopf & Härtel, 1876, 583p.

[30] Filho, R. J. R.; Wehrmann, J.; Barros, R. C. “Leveraging deep visual features for content-based movie recommender systems”. In: IEEE International Joint Conference on Neural Networks, 2017, pp. 604–611.

[31] Foka, A. “Computer vision applications for art history: Reflections and paradigms for future research”. In: Electronic Visualization and the Arts, 2021, pp. 73–80.

[32] Forbes. “5 big numbers that show netflix’s massive growth continues during the coronavirus pandemic”. Source: https://www.forbes.com/sites/jonathanponciano/2020/10/19/netflix-earnings-5-numbers-growth-continues-during-the-coronavirus-pandemic/#6c2e7efd225e, Oct 2020.

[33] Galanter, P. “Xepa - autonomous intelligent light and sound sculptures that improvise group performances”. In: ACM Special Interest Group on Graphics and Interactive Techniques, 2014, pp. 386–393.

[34] Galenson, D. W. “Conceptual revolutions in 20th-century art”, *Historically Speaking*, vol. 10–5, Nov 2009, pp. 20–22.

[35] Garnier, S. “Rvision: Basic computer vision library”. Source: https://github.com/swarm-lab/Rvision, Jan 2021.

[36] Ghadiyaram, D.; Bovik, A. C. “Perceptual quality prediction on authentically distorted images using a bag of features approach”, *Journal of Vision*, vol. 17, Jan 2017, pp. 1–25.

[37] Goodall, C. R. “Computation using the QR decomposition”. Elsevier, 1993, vol. 9, chap. 13, pp. 467–508.

[38] Graf, L. K. M.; Landwehr, J. R. “Aesthetic pleasure versus aesthetic interest: The two routes to aesthetic liking”, *Frontiers in Psychology*, vol. 8, Jan 2017, pp. 15.

[39] Greenfield, G. “On the origins of the term "computational aesthetics"”. In: Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging, 2005, pp. 9–12.

[40] Gupta, A.; Mithun, N. C.; Rudolph, C.; Roy-Chowdhury, A. K. “Deep learning based identity verification in renaissance portraits”. In: IEEE International Conference on Multimedia and Expo, 2018, pp. 1–6.

[41] Harper, F. M.; Konstan, J. A. “The movielens datasets: History and context”, *ACM Transactions on Interactive Intelligent Systems*, vol. 5–4, December 2015, pp. 19.

[42] Hermerén, G. “Influence in art and literature”. Princeton University Press, 2015, 366p.

[43] Hoenig, F. “Defining computational aesthetics”. In: Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging, 2005, pp. 13–18.

[44] Jaakkola, T. S.; Haussler, D. “Exploiting generative models in discriminative classifiers”. In: International Conference on Neural Information Processing Systems, 1998, pp. 487–493.

[45] Jenkins, R.; Dowsett, A.; Burton, A. “How many faces do people know?”, *Royal Society of London. B, Biological Sciences*, vol. 285–1888, Oct 2018, pp. 20181319.

[46] Jiang, B.; Sui, D. Z. “A new kind of beauty out of the underlying scaling of geographic space”, *The Professional Geographer*, vol. 66–4, Dec 2014, pp. 676–686.

[47] Joshi, D.; Datta, R.; Fedorovskaya, E. A.; Luong, Q.; Wang, J. Z.; Li, J.; Luo, J. “Aesthetics and emotions in images”, *IEEE Signal Processing Magazine*, vol. 28–5, Sep 2011, pp. 94–115.

[48] Kalva, H.; Bovik, A.; Chen, H. H.; Egiazarian, K. O.; Wang, Z. “Introduction to the issue on perception inspired video processing”, *IEEE Journal of Selected Topics in Signal Processing*, vol. 8–3, May 2014, pp. 355–357.

[49] Khosla, A.; Sarma, A. D.; Hamid, R. “What makes an image popular?” In: International World Wide Web Conference, 2014, pp. 867–876.

[50] Kong, Q.; Rizoiu, M.; Wu, S.; Xie, L. “Will this video go viral: Explaining and predicting the popularity of youtube videos”. In: The Web Conference, 2018, pp. 175–178.

[51] Kong, S.; Shen, X.; Lin, Z. L.; Mech, R.; Fowlkes, C. C. “Photo aesthetics ranking network with attributes and content adaptation”. In: Computer Vision, 2016, pp. 662–679.

[52] Kuhn, M. “caret: Classification and regression training”. Source: https://CRAN.R-project.org/package=caret, Jan 2021.

[53] Kuhn, M.; Johnson, K. “Applied predictive modeling”. Springer, 2013, 600p.

[54] Landis, J. R.; Koch, G. G. “The measurement of observer agreement for categorical data”, *Biometrics*, vol. 33–1, Mar 1977, pp. 159–174.

[55] Levine, M. D. “Feature extraction: A survey”, *IEEE*, vol. 57–8, Aug 1969, pp. 1391–1407.

[56] Liu, Y.; Zhang, Q. “Interface design aesthetics of interaction design”. In: Design, User Experience, and Usability. Design Philosophy and Theory, 2019, pp. 279–290.

[57] McDonnell, R.; Breidt, M.; Bülthoff, H. H. “Render me real?: Investigating the effect of render style on the perception of animated virtual humans”, *ACM Transactions on Graphics*, vol. 31–4, Jul 2012, pp. 1–11.

[58] McWhinnie, H. J. “A review of research on aesthetic measure”, *Acta Psychologica*, vol. 28, Jun 1968, pp. 363–375.

[59] Neumann, L.; Sbert, M.; Gooch, B.; Purgathofer, W. “Computational aesthetics”, *Eurographics Workshop on Computational Aesthetics in Graphics, Visualization and Imaging*, vol. 25–1, Mar 2006, pp. 145–146.

[60] Ooms, J. “opencv: Bindings to ’opencv’ computer vision library”. Source: https://CRAN.R-project.org/package=opencv, Jan 2021.

[61] Ooms, J. “tesseract: Open source ocr engine”. Source: https://CRAN.R-project.org/package=tesseract, Jan 2021.

[62] Qu, W.; Song, K.; Zhang, Y.; Feng, S.; Wang, D.; Yu, G. “A novel approach based on multi-view content analysis and semi-supervised enrichment for movie recommendation”, *Journal of Computer Science and Technology*, vol. 28–5, Sep 2013, pp. 776–787.

[63] R Core Team. “R: A language and environment for statistical computing”. Source: https://www.R-project.org/, Jan 2021.

[64] Sablatnig, R.; Kammerer, P.; Zolda, E. “Hierarchical classification of paintings using face- and brush stroke models”. In: IEEE International Conference on Pattern Recognition, 1998, pp. 172–174.

[65] Saleh, B.; Abe, K.; Arora, R. S.; Elgammal, A. M. “Toward automated discovery of artistic influence”, *Multimedia Tools and Applications*, vol. 75–7, Aug 2016, pp. 3565–3591.

[66] Scha, R.; Bod, R. “Computationele esthetica”, *Informatie en Informatiebeleid*, vol. 11–1, Apr 1993, pp. 54–63.

[67] Schenk, A.; Stumpel, J. “Facial types in painting and recognition skills: Laymen as connoisseurs”, *Art & Perception*, vol. 5–2, Jun 2017, pp. 143 – 167.

[68] Schmid, K.; Marx, D.; Samal, A. “Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios”, *Pattern Recognition*, vol. 41–8, Aug 2008, pp. 2710–2717.

[69] Shamir, L.; Tarakhovsky, J. A. “Computer analysis of art”, *ACM Journal on Computing and Cultural Heritage*, vol. 5–2, Aug 2012, pp. 1–11.

[70] Smith, R. “An overview of the tesseract OCR engine”. In: IEEE International Conference on Document Analysis and Recognition, 2007, pp. 629–633.

[71] Spearman, C. “The proof and measurement of association between two things”, *The American Journal of Psychology*, vol. 15–1, Jan 1904, pp. 72–101.

[72] Spratt, E. L.; Elgammal, A. “Computational beauty: Aesthetic judgment at the intersection of art and science”. In: Computer Vision, 2015, pp. 35–53.

[73] Stork, D. G. “Computer vision and computer graphics analysis of paintings and drawings: An introduction to the literature”. In: Computer Analysis of Images and Patterns, 2009, pp. 9–24.

[74] Sun, Z.; Yu, D.; Fang, H.; Yang, J.; Qu, X.; Zhang, J.; Geng, C. “Are we evaluating rigorously? benchmarking recommendation for reproducible evaluation and fair comparison”. In: ACM Conference on Recommender Systems, 2020, pp. 23–32.

[75] Tatler, B. W.; Wade, N. J.; Kwan, H.; Findlay, J. M.; Velichkovsky, B. M. “Yarbus, eye movements, and vision”, *i-Perception*, vol. 1–1, Jan 2010, pp. 7–27.

[76] Thorat, P. B.; Goudar, R. M.; Barve, S. “Survey on collaborative filtering, content-based filtering and hybrid recommendation system”, *International Journal of Computer Applications*, vol. 110–4, January 2015, pp. 31–36.

[77] Trooskens, G.; De Beule, D.; Decouttere, F.; Van Criekinge, W. “Phylogenetic trees: visualizing, customizing and detecting incongruence”, *Bioinformatics*, vol. 21–19, 07 2005, pp. 3801–3802.

[78] Trzcinski, T.; Rokita, P. “Predicting popularity of online videos using support vector regression”, *IEEE Transactions on Multimedia*, vol. 19–11, Nov 2017, pp. 2561–2570.

[79] Tumblin, J.; Ferwerda, J. A. “Applied perception”, *IEEE Computer Graphics and Applications*, vol. 21–5, Sep 2001, pp. 20–21.

[80] van Noord, N.; Hendriks, E.; Postma, E. “Toward discovery of the artist’s style: Learning to recognize artists by their artworks”, *IEEE Signal Processing Magazine*, vol. 32–4, Jul 2015, pp. 46–54.

[81] Verge. “The entire world is streaming more than ever — and it’s straining the internet”. Source: https://www.theverge.com/2020/3/27/21195358/streaming-netflix-disney-hbo-now-youtube-twitch-amazon-prime-video-coronavirus-broadband-network, Oct 2020.

[82] Vieira, S. M.; Kaymak, U.; Sousa, J. M. C. “Cohen’s kappa coefficient as a performance measure for feature selection”. In: IEEE International Conference on Fuzzy Systems, 2010, pp. 1–8.

[83] Weinmann, M. “Visual Features - From Early Concepts to Modern Computer Vision”. Springer, 2013, chap. 1, pp. 1–34.

[84] Weller, H. “colordistance: Distance metrics for image color similarity”. Source: https://CRAN.R-project.org/package=colordistance, Jan 2021.

[85] Yarbus, A. L. “Eye Movements and Vision”. Springer, 1967, 222p.

[86] Youtube. “Press — youtube about”. Source: https://www.youtube.com/about/press/, Apr 2020.

[87] YouTube Support. “How to earn money on youtube”. Source: https://support.google.com/youtube/answer/72857?hl=EN, Apr 2020.

APPENDIX A – PAPERS WRITTEN DURING THE MASTER'S DEGREE

- Dal Molin, G. P.; Nomura, F. M.; Dalmoro, B. M.; de A. Araújo, V. F.; Musse, S. R. “Can we estimate the perceived comfort of virtual human faces using visual cues?" In: 2021 IEEE 15th International Conference on Semantic Computing (ICSC), 2021, pp. 366–369. Qualis A3. *(Published paper)*

- Araujo, V. F.; Melgare, J.; Dalmoro, B.; Musse, S. R. “Is the perceived comfort with CG characters increasing with their novelty", IEEE Computer Graphics and Applications, 2021, pp. 1–1. Qualis A3. *(Published paper)*

- Araujo, V. F.; Dalmoro, B. M.; Musse, S. R. “Analysis of charisma, comfort and realism in CG characters from a gender perspective", Springer The Visual Computer, Jul 2021. Qualis A3. *(Published paper)*

- Dalmoro, B. M.; Musse, S. R. “Predicting popularity of facebook videos through visual featuresusing support vector machine classifier". In: Anais do XLVIII Seminário Integrado de Software e Hardware, 2021, pp. 131–138. Qualis B1. *(Published paper)*

- Dalmoro, B.; Monteiro, C.; Musse, S. R. “Measuring the influence of painters through artwork facial features", 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2022, pp. 37-42. Qualis A3. *(Published paper)*

- Dalmoro, B.; Monteiro, C.; Musse, S. R. “Identifying influences between artists based on artwork faces and geographic proximity", SIBGRAPI 2022 Special Section of the Computers Graphics Journal. *(Accepted paper)*

MARISTA PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Pesquisa e Pós-Graduação
Av. Ipiranga, 6681 – Prédio 1 – Térreo
Porto Alegre – RS – Brasil
Fone: (51) 3320-3513
E-mail: propesq@pucrs.br
Site: www.pucrs.br