



PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE  
ESCUELA DE INGENIERÍA

# **STUDYING THE EFFECTS OF EXPLAINING RECOMMENDATIONS OF ARTISTIC IMAGES**

**VICENTE DOMÍNGUEZ MANQUENAHUEL**

Thesis submitted to the Office of Research and Graduate Studies  
in partial fulfillment of the requirements for the degree of  
Master of Science in Engineering

Advisor:

DENIS PARRA

Santiago de Chile, January 2019

© MMXIX, VICENTE DOMÍNGUEZ MANQUENAHUEL



PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE  
ESCUELA DE INGENIERÍA

# **STUDYING THE EFFECTS OF EXPLAINING RECOMMENDATIONS OF ARTISTIC IMAGES**

**VICENTE DOMÍNGUEZ MANQUENAHUEL**

Members of the Committee:

DENIS PARRA

HANS LÖBEL

EDUARDO GRAELLS

HERNÁN DE SOLMINIHAC

Thesis submitted to the Office of Research and Graduate Studies  
in partial fulfillment of the requirements for the degree of  
Master of Science in Engineering

Santiago de Chile, January 2019

© MMXIX, VICENTE DOMÍNGUEZ MANQUENAHUEL

*Gratefully to my family, my  
girlfriend and my pets.*

## ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Denis Parra of the Department of Computer Science at Pontificia Universidad Católica de Chile. He was always available to answer my questions or to guide me whenever I had a problem. He consistently steered me in the right the direction whenever he thought I needed it. He helped me to go two conferences during my master and an autumn school.

I would like to thanks my parents, specially to my mother. She was always supporting me and encourage me to do all the things that I wanted to do in my life. I will always be grateful that she was always there for me.

Finally, I would like to thank to Pablo Messina for coworking with me the last years. I would like to thank too Ivania Donoso for helping me in the very last moments. And also, thanks to Millennium Institute for Foundational Research on Data (IMFD) for founding this research.

## TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
LIST OF FIGURES	vii
LIST OF TABLES	ix
ABSTRACT	x
RESUMEN	xi
1. INTRODUCTION	1
1.1. Motivation . . . . .	1
1.2. Artwork Recommendations . . . . .	1
1.3. Amazon Mechanical Turk . . . . .	3
1.4. Objective and Contribution . . . . .	4
1.5. Research Questions . . . . .	5
1.6. Outline . . . . .	6
2. RELATED WORK	7
2.1. Recommendations of Artistic Images . . . . .	7
2.1.1. Explainability and transparency in Recommender Systems . . . . .	7
2.2. User-centric evaluation of recommender systems . . . . .	9
2.3. Differences to Previous Research & Contributions. . . . .	9
3. METHODOLOGY	11
3.1. Materials . . . . .	11
3.2. Visual Recommendation Approaches . . . . .	12
3.2.1. DNN Visual Feature (DNN) Algorithm . . . . .	12
3.2.2. Attractiveness Visual Features (AVF) Algorithm . . . . .	13
3.2.3. Computing Recommendations . . . . .	14
3.3. The Explainable Recommender Interfaces . . . . .	15

3.4. User Study Procedure . . . . .	17
3.5. System Architecture . . . . .	20
4. RESULTS	23
4.1. Demographic Results . . . . .	23
4.2. A Model of the UX with an Art Recommender . . . . .	25
4.2.1. Confirmatory Factor Analysis . . . . .	25
4.2.2. Structural Equation Model . . . . .	25
5. FUTURE WORK	28
6. CONCLUSIONS	29
REFERENCES	30

## LIST OF FIGURES

1.1	HIT published on Amazon Mechanical Turk. . . . .	5
3.1	Screenshot of the search interface of <i>UGallery</i> . Users can filter by different facets on the left side. . . . .	11
3.2	Model architecture of the AlexNet Convolutional Deep Neural Network used to extract visual features from images. . . . .	13
3.3	Interface 1: Baseline recommendation interface without explanations. . . . .	15
3.4	Design choices for explainable recommender interfaces, based on Friedrich and Zanker (Friedrich & Zanker, 2011). In (a) we explain the recommendation based on transparent visual features, while in (b) we explain based on item similarity, without details of the features used. . . . .	17
3.5	Interface 2: Explainable recommendation interface with textual explanations and top-3 similar images. . . . .	18
3.6	Interface 3: Explainable and transparent recommendation interface with features' bar chart and top-1 similar image. . . . .	18
3.7	Study procedure. After the pre-study survey and the preference elicitation, users were assigned to one of three possible interfaces. In each interface they evaluated recommendations of two algorithms: DNN and AVF. . . . .	19
3.8	System architecture. . . . .	21
4.1	The structural equation model for the data of the experiment using Knijnenburg's evaluation framework for recommender systems. Significance levels: $***p < .001$ , $**p < .01$ , $*p < 0.05$ . $R^2$ is the proportion of variance explained by the model. Numbers on the arrows (and their thickness) represent	

the  $\beta$  coefficients (and standard error) of the effect. Factors are scaled to have  
an  $SD$  of 1. . . . . 26



## LIST OF TABLES

3.1	Results of users' perception over several evaluation dimensions, defined in Section 3.4. Scale 1-100 (higher is better), except for Average rating (scale 1-5). DNN: Deep Neural Network, and AVF: Attractiveness visual features. The symbol $\uparrow^1$ indicates interface-wise significant difference (differences between interfaces using the same algorithms). The * symbol denotes algorithm-wise statistical difference (comparing a dimension between algorithms, using the same interface). . . . .	22
3.2	NASA TLX Results. . . . .	22
4.1	CFA output . . . . .	24

## ABSTRACT

Explaining suggestions from recommender systems is an important research area since it has shown a significant effect upon several dimensions of the user experience. However, there are few works about explaining content-based recommendations of images in the artwork domain. Even more, these works do not provide a perspective of the many variables involved in the user perception of several aspects of the system such as domain knowledge, relevance, explainability, diversity and trust. In this work, we aim to fill this gap by studying three interfaces, with different levels of explainability, for artistic image recommendation. Our experiments with  $N=121$  users confirm that explanations of recommendations in the image domain are useful and increase user satisfaction, perception of explainability, relevance, and diversity. Furthermore, our results show that the observed effects are also dependent on the underlying recommendation algorithm used. We tested the interfaces with two algorithms: Deep Neural Networks (DNN), which has high accuracy, and another method with high transparency but lower accuracy, Attractiveness Visual Features (AVF). Notably, the explainable visual features of the AVF method increased the perception of explainability but did not increase the perception of trust, unlike DNN, which improved both dimensions. These results indicate that algorithms in conjunction with interfaces play a significant role in the perception of explainability and trust for image recommendation. Finally, using the framework by Knijnenburg et al., we provide a comprehensive model and analysis which synthesize the relations and effects between different variables involved in the user experience with explainable visual recommender systems of artistic images.

**Keywords:** recommender systems, image recommendation, explainable interfaces, artwork recommendation, machine learning.

## RESUMEN

Explicar las sugerencias de los sistemas de recomendación es un área importante de investigación, ya que ha demostrado un efecto significativo en varias dimensiones de la experiencia del usuario. Sin embargo, hay muy pocos trabajos sobre la explicación de recomendaciones basadas en contenido de imágenes en el dominio de obras de arte. Más aún, estos trabajos no proporcionan una perspectiva de las muchas variables involucradas en la percepción del usuario en diversos aspectos del sistema, como el dominio del tema, la relevancia, la explicación, la diversidad y la confianza. En este trabajo, nuestro objetivo es llenar este vacío estudiando tres interfaces, con diferentes niveles de explicabilidad, para recomendaciones de imágenes artísticas. Nuestros experimentos con  $N = 121$  usuarios confirman que las explicaciones de recomendaciones en el dominio de la imagen son útiles y aumentan la satisfacción del usuario, la percepción de la explicación, la relevancia y la diversidad. Además, nuestros resultados muestran que los efectos observados también son dependientes del algoritmo de recomendación subyacente utilizado. Probamos las interfaces con dos algoritmos: Deep Neural Networks (DNN), que tiene una alta precisión, y otro método con alta transparencia pero menor precisión, Attractiveness Visual Features (AVF). En particular, las características visuales explicables del método AVF aumentaron la percepción de explicabilidad, pero no aumentaron la percepción de confianza, a diferencia de DNN, que mejoró ambas dimensiones. Estos resultados indican que los algoritmos en conjunto con las interfaces juegan un papel importante en la percepción de la explicación y la confianza de la recomendación de imágenes. Finalmente, utilizando el sistema de Knijnenburg et al., proporcionamos un modelo comprehensivo que sintetiza las relaciones y los efectos entre diferentes variables involucradas en la experiencia del usuario con sistemas recomendadores visuales explicables de imágenes artísticas.

**Palabras Claves:** sistemas recomendadores, recomendación de imágenes, interfaces explicables, inteligencia de máquina.

# 1. INTRODUCTION

## 1.1. Motivation

In the latest decade, online artwork sales are booming due to the influence of social media and new consumption behavior of millennials, and at the current growth rate, they are expected to reach \$9.58 billion by 2020<sup>1</sup>. The artist are moving their artworks to the online market, but this market is different to art exhibits in two main features:

- **Recommendations:** In an art exhibit the artist has the possibility to recommend some of her works. If the customer explain to the artist which kind of art she prefer most, the artist could make suggestions or recommendations of piece of artworks to purchase.
- **Explanations:** In addition to the last point, in an art exhibit the artist has the option to explain why is she recommending a piece of artwork.

Nowadays, online artwork markets lack of this two features that are important to the artist and the customers. The objective of this thesis is to fill this lack of features of the online markets. To do that, we tested different algorithms with real users to make recommendation of artworks, and ways to explain these recommendations to the user.

## 1.2. Artwork Recommendations

Online artwork recommendation has received little attention compared to other areas such as movies (Amatriain, 2013; Gomez-Urbe & Hunt, 2016), music (Maes et al., 1994; Celma, 2010) or points-of-interest (Ye, Yin, Lee, & Lee, 2011; Yuan, Cong, Ma, Sun, & Thalmann, 2013; Trattner et al., 2016). Most research on artwork recommendation deals with studies on museum data (Aroyo et al., 2007; van den Broek, Kok, Schouten, & Hoenkamp, 2006; Semeraro, Lops, De Gemmis, Musto, & Narducci, 2012; Benouaret &

---

<sup>1</sup><https://www.forbes.com/sites/deborahweinswig/2016/05/13/art-market-cooling-but-online-sales-booming/>

Lenne, 2015), but there is little work with datasets of online artwork e-commerce systems (He, Fang, Wang, & McAuley, 2016; Messina, Dominguez, Parra, Trattner, & Soto, 2018).

The first works in the area of artwork recommendation date from 2006-2007 such as the CHIP (Aroyo et al., 2007) project, which implemented traditional techniques such as content-based and collaborative filtering for artwork recommendation at the Rijksmuseum, and the *m4art* system by Van den Broek et al. (van den Broek et al., 2006), which used histograms of color to retrieve similar artworks where the input query was a painting image. More recently, deep neural networks (DNN) have been used for artwork recommendation and are the current state-of-the-art model (He et al., 2016; Dominguez et al., 2017), which is rather expected considering that DNNs are the top performing models for obtaining visual features for several tasks, such as image classification (Krizhevsky, Sutskever, & Hinton, 2012), and scene identification (Sharif Razavian, Azizpour, Sullivan, & Carlsson, 2014). More recently, Messina et al. (Messina et al., 2018) compared the performance of visual features extracted with DNNs versus traditional visual features (brightness, contrast, LBP, etc.), finding that DNN visual features had better predictive accuracy. Moreover, they conducted a pilot study with a small group of art experts to generalize their results, but they did not conduct a user study with a larger sample of experts and non-experts art users. This aspect is important since past works have shown that off-line results might not always replicate when tested with actual users (McNee, Kapoor, & Konstan, 2006; Konstan & Riedl, 2012), and also domain knowledge is an important variable to explain the user experience with a recommender system (Knijnenburg, Willemssen, Gantner, Soncu, & Newell, 2012a; Parra & Brusilovsky, 2015; Andjelkovic, Parra, & ODonovan, 2018).

The aforementioned works miss one important aspect of the user experience with recommender systems: explainability. Artwork recommendations based on visual features obtained from DNNs, although accurate, are difficult to explain to users, despite current efforts to make the complex mechanism of neural networks more transparent to users

(Olah, Mordvintsev, & Schubert, 2017). In contrast, features of visual attractiveness, despite being less accurate to predict user preference (Messina et al., 2018), could be easily explained, based on color, brightness or contrast (San Pedro & Siersdorfer, 2009). Explanations in recommender systems have been shown to have a significant effect on user satisfaction (Tintarev & Masthoff, 2015), and no previous work has shown how to explain recommendations of images based on visual features. Hence, there is no study of the effect on users when explaining images recommended by a Visual Content-based Recommender (Hereinafter, VCBR). To the best of our knowledge, there is neither a research which fully combines in a single model different independent variables such as interface, explanation, algorithms, and domain knowledge, in order to explain several dimensions of the user experience with a VCBR such as perception of relevance, diversity, explainability and trust.

### **1.3. Amazon Mechanical Turk**

An important difficulty when conducting a user study is getting the study subjects, i.e., the users. Another difficult aspect is collecting the minimum amount of users needed to detect a significant statistical difference between the conditions of the experiment, in case this difference exist. It becomes even harder if you need subjects with certain traits or skills and you need to get these users within a short period of time. We had all these requirements to conduct our user study, and that is why we tried to recruit them from a crowdsourcing platform.

Crowdsourcing is the act of outsourcing some products or services to the crowd (Schenk & Guittard, 2009). Several tasks that in the past were made in a laboratory nowadays are made by this practise. Image tagging and classification, sentiment analysis, audio transcription, document classification and user studies, are some of the tasks that are needed for the scientist to do their researches. Thanks to the crowdsourcing platforms they are possible to be completed in short period of time.

In this context, there are many crowdsourcing platforms that can be used to do the tasks listed before such as MicroWorkers<sup>2</sup>, ClickWorker<sup>3</sup> or Amazon Mechanical Turk<sup>4</sup> (AMT). One of the most popular platforms to do scientific crowdsourcing is AMT. We chose AMT over the other alternatives because it is a very well documented platform and it has a large amount of working user subscribers, also known as *AMT workers*. AMT had over 500,000 registered workers by 2011<sup>5</sup>. Using this platform, a group of researchers were able to create Imagenet (Deng et al., 2009a), one of the biggest dataset of classified images and one of the main reasons on the development of the deep neural networks models such as Alexnet (Krizhevsky et al., 2012).

AMT is also known for being a good website to make user studies in an simply and fast way (Kittur, Chi, & Suh, 2008). It has filters that can be applied to show your HIT to specific types of users, like users that has over than 90% of approval on their tasks completed. It provides two types of accounts: requester and worker. Worker is the user that receives money for doing a Human Intelligence Tasks (HIT). Requester is the user that publish a HIT and pays to the workers that completed the task.

For all the reasons explained before we decided to use this platform to conduct our user study. Using this platform we were able to collect the opinions of the users about their experience of getting recommendations artistic images with different types of interfaces and levels of explanations.

#### **1.4. Objective and Contribution**

In this work, we analyze the effect of explaining artistic image suggestions. In particular, we conduct a user study on AMT (N=121) under three different interfaces and two different algorithms. The three interfaces are: i) no explanations, ii) explanations based

---

<sup>2</sup> <https://www.microworkers.com/>

<sup>3</sup> <https://www.clickworker.com/>

<sup>4</sup> <https://www.mturk.com/>

<sup>5</sup> Amazon Web Services Developer Forum

Task Link Instructions (Click to collapse)

We are conducting an academic study about artwork recommendation.

- In this study, you will be asked to select your favorite paintings from a list and then to evaluate recommendations. You will also fill 3 short surveys.
- **Inattentive answers really hurt our results, please don't do this if you can't give it your full attention.**
- The estimated length of this task is 10 minutes.

You will Select the link below to complete the survey. At the end of the survey, you will receive a code to paste into the box below to receive credit for taking our survey.

**Make sure to leave this window open as you complete the task.** When you are finished, you will return to this page to paste the code into the box.

Task link:

[http://niebla.ing.puc.cl/iui\\_user\\_study](http://niebla.ing.puc.cl/iui_user_study)

Provide the survey code here:

e.g. 123456

Submit

Figure 1.1. HIT published on Amazon Mechanical Turk.

on similar images, and iii) explanations based on visual features. Moreover, the two algorithms are: Deep Neural Networks (DNN) and Attractiveness Visual Features (AVF). In our study, we used images provided by the online store *UGallery*<sup>6</sup>. Finally, we contribute with a Structural Equation Model based on the framework by Knijnenburg et al. (Knijnenburg, Willemsen, et al., 2012a) in order to fully explain the user experience with a explainable VBCR of artists images.

### 1.5. Research Questions

To drive our research, the following two questions were defined:

<sup>6</sup> <http://www.UGallery.com/>



- **RQ1.** Given three different types of interfaces, one baseline interface without explanations and two with explanations but different levels of transparency, which one is perceived as most useful?
- **RQ2.** Furthermore, based on the visual content-based recommender algorithm chosen (DNN or AVF), are there observable differences in how the three interfaces are perceived?
- **RQ3.** How do independent variables such as algorithm, explainable interface and domain knowledge interact in order to explain the user experience with the recommender systems in terms of perception of relevance, diversity, explainability and trust?

## 1.6. Outline

Our work is structured as follows: In Section 2 we survey relevant related work in the area and explain how our work differs from previous work in the area. Section 3 introduces the explainable interface recommendation approaches and the algorithms, and discusses the study procedure to evaluate these. Then, in Section 4 we present the results, including the subsection 4.2 that presents the global SEM which connects all the studied variables, and finally sections 5 and 6 concludes the thesis and provides an outlook for future work in the area.

## **2. RELATED WORK**

We organized the relevant related research in three sub-sections: first, we review research on recommending artistic images. Second we summarize related work on explainability and transparency in recommender systems. Both are important to our problem at hand. Then, the next sub-section is about user-centric evaluation of recommender systems. The final paragraph in this section highlights the differences to previous work and our contributions to the existing literature in the area.

### **2.1. Recommendations of Artistic Images**

The works of Aroyo et al. (Aroyo et al., 2007) with the CHIP project and Semeraro et al. (Semeraro et al., 2012) with FIRSt (Folksonomy-based Item Recommender syStem) made early contributions to this area using traditional techniques. More complex methods were implemented recently by Benouaret et al. (Benouaret & Lenne, 2015), using context obtained through a mobile application, that makes a museum tour recommendation. Finally, the work of He et al. addresses digital artwork recommendations based on pre-trained deep neural visual features (He et al., 2016), and the work of Dominguez et al. (Dominguez et al., 2017) and Messina et al. (Messina et al., 2018) compared neural against traditional visual features. None of the aforementioned works performed a user study under explanation interfaces to generalize their results.

#### **2.1.1. Explainability and transparency in Recommender Systems**

Herlocker et al. (Herlocker, Konstan, & Riedl, 2000) introduced the idea of explaining recommendations as a way to make the system more transparent to users' decisions and to improve users' acceptance of recommender systems. Based on successful previous results from expert systems, they expected that interfaces of collaborative filtering recommenders would benefit from explanations as well. They studied different ways to explain recommendations and rated histograms as "the most compelling way to explain

the data behind the prediction.” A study with 210 users of MovieLens, a well-known movie recommender system, showed that users value explanations and would like to add them to the recommender interface (86% of the respondents of a survey). The authors also think that explanation facilities can increase the filtering performance of recommender systems, though they could not find explicit evidence to support it and called for further well-controlled studies in this area. Furthermore, (Tintarev & Masthoff, 2007) noticed that explanations might have different objectives, and identified seven different aims for explanations: transparency, scrutability, trustworthiness, effectiveness, persuasiveness, efficiency and satisfaction. More recently, in the handbook of recommender systems there is a whole chapter that addresses the design and evaluation of explanations in recommender systems (Tintarev & Masthoff, 2011).

One of the main effects of the explainability of intelligent systems is viewed in the user’s perception of trust. A recent study in intelligent systems (Holliday, Wilson, & Stumpf, 2016) showed that the effect of the explanation in the user’s trust changes over time. This study showed that people who were exposed to the intelligent system without explanation decreased their trust over time, but the ones who received explanations increased their trust. This happened because users who do not receive explanation were affected most by the transparency of the system. The role of transparency in recommender systems is discussed in (Sinha & Swearingen, 2002), showing that ”people feel more confident with recommendations that they perceive more transparent.”

Cramer et al in their work (Cramer et al., 2008) studied topics very similar to ones addressed by our work. They analyzed the effect of transparency and trust of an artwork recommender system. The difference with respect our study is that they did not consider the effect of different interfaces in different devices. Also, we are using newer algorithm than the one used in their work.

## **2.2. User-centric evaluation of recommender systems**

Traditionally, evaluation of recommender systems has relied mainly on prediction accuracy, but over the years researchers and professionals implementing recommender systems have reached consensus that this evaluation must consider additional measures such as diversity, novelty, and coverage. Beyond these metrics, recent research has increasingly considered user-centric evaluation measures such as perceived diversity, controllability and explainability. For instance, Ziegler et al. (Ziegler, McNee, Konstan, & Lausen, 2005) studied the effect of diversification in lists of recommended items, Tintarev and Masthoff (Tintarev & Masthoff, 2007) investigated on recommender systems' transparency, Cramer et al. (Cramer et al., 2008) studied explainability in recommender systems, and Knijnenburg et al. (Knijnenburg, Bostandjiev, O'Donovan, & Kobsa, 2012) tried to explain the effects of user-controllability on the user experience in a recommender system. Nevertheless, as a result of a lack of a unified framework, comparing the results of different studies or replicating them is not a simple task. Two recent user-centric evaluation frameworks addressed this issue. On one side, Pu et al. (Pu, Chen, & Hu, 2011) proposed ResQue, identifying four main dimensions (perceived quality, user beliefs, user attitudes and behavioral intentions) and a set of constructs to evaluate each one. On the other side, Knijnenburg et al. (Knijnenburg, Willemsen, Gantner, Soncu, & Newell, 2012b) defined dimensions and relations between them (objective systems aspects, subjective system aspects, experience, interaction, situational characteristics and personal characteristics), but encouraged the users of this framework to choose their own constructs based on some specified guidelines. We finally decided to use this last framework to evaluate our experiments from a user centric view, because it gives an holistic view of the results involved in the user study.

## **2.3. Differences to Previous Research & Contributions.**

To the best of our knowledge this is the first work which studies the effect of explaining recommendations of images based on visual features. Our contributions are three-fold: i)

we analyze and report the effect of explaining artistic recommendations especially for VCBR, ii) with a user study we validate off-line results stating the superiority of neural visual features compared to attractiveness visual features over several dimensions, such as users' perception of explainability, relevance, and trust, and iii) we present a structural equation model, based on the framework by Knijnenburg et al. (Knijnenburg, Willemsen, et al., 2012a), in order to characterize all the variables involved in the user experience with a explainable VCBR of art images.

### 3. METHODOLOGY

In the following section we describe in detail our study methods. First, we introduce the dataset chosen for the purpose of our study. Second, we introduce the two algorithms chosen for our study are revealed. Third, we explain the design choices for the three different explainable visual interfaces implemented which we evaluate. Finally, the user study procedure is explained.

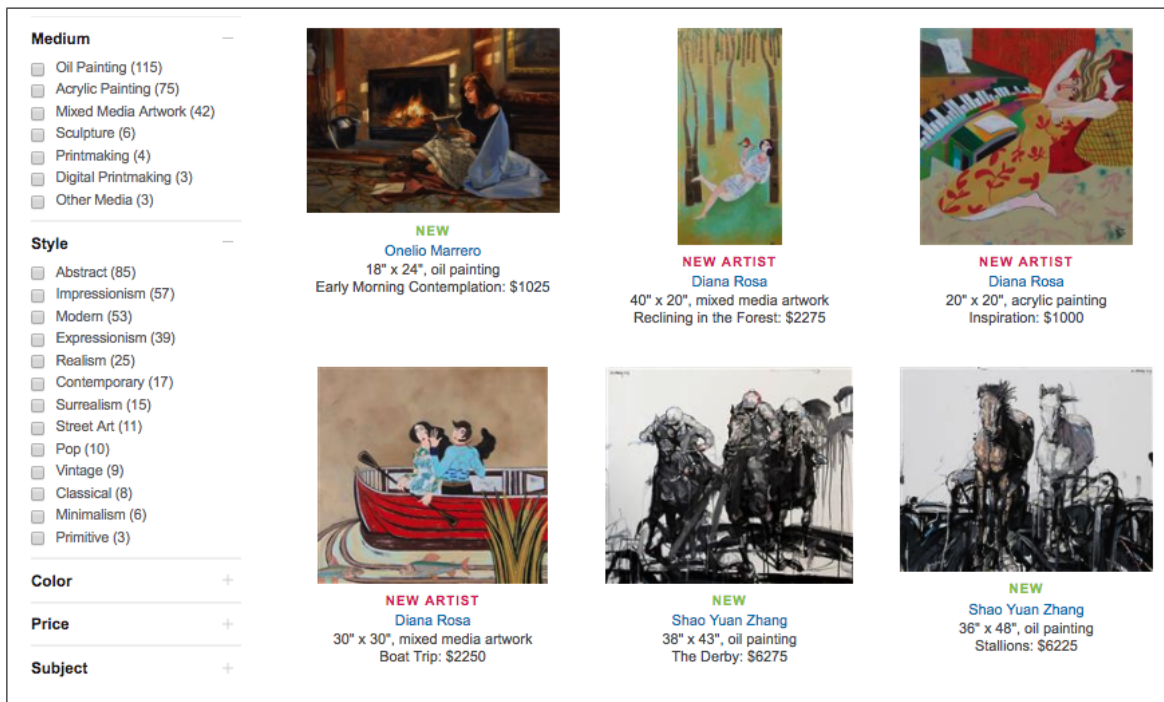


Figure 3.1. Screenshot of the search interface of *UGallery*. Users can filter by different facets on the left side.

#### 3.1. Materials

For the purpose of our study we rely on a dataset provided by the online web store *UGallery*, which has been selling artwork for more than 10 years (Weinswig, 2016). They support emergent artists by helping them sell their artwork online. The *UGallery* website allows users (customers) to search for items and to browse the catalog based on different

attributes with a predefined order: orientation, size, medium, style and others, as seen on the left side of Figure 3.1.

For our research, UGallery provided us with an anonymized dataset of 1,371 users, 3,490 items and 2,846 purchases (transactions) of artistic artifacts, where all users have made at least one transaction. On average, each user bought 2-3 items over recent years .

### **3.2. Visual Recommendation Approaches**

As mentioned earlier in this work, we make use of two different content-based visual recommender approaches in our work. The reason for choosing content-based methods over collaborative filtering-based methods is grounded in the fact that once an item is sold via the UGallery store, it is not available anymore (every item is unique) and hence traditional collaborative filtering approaches based on co-occurrence do not apply.

#### **3.2.1. DNN Visual Feature (DNN) Algorithm**

The first algorithmic approach we employed was based on image similarity, itself based on features extracted with a deep neural network. The output vector representing the image is usually called an image’s visual embedding. The visual embedding in our experiment was a vector of features obtained from an AlexNet, a convolutional deep neural network developed to classify images (Krizhevsky et al., 2012), which architecture is shown in Figure 3.2. In particular, we use an AlexNet model pre-trained with the ImageNet dataset (Deng et al., 2009b). Using the pre-trained weights, for every image a vector of 4,096 dimensions was generated with the Caffe (<http://caffe.berkeleyvision.org/>) framework. We resized every image to a  $227 \times 227$  image. This is the standard pre-processing needed to use the AlexNet.

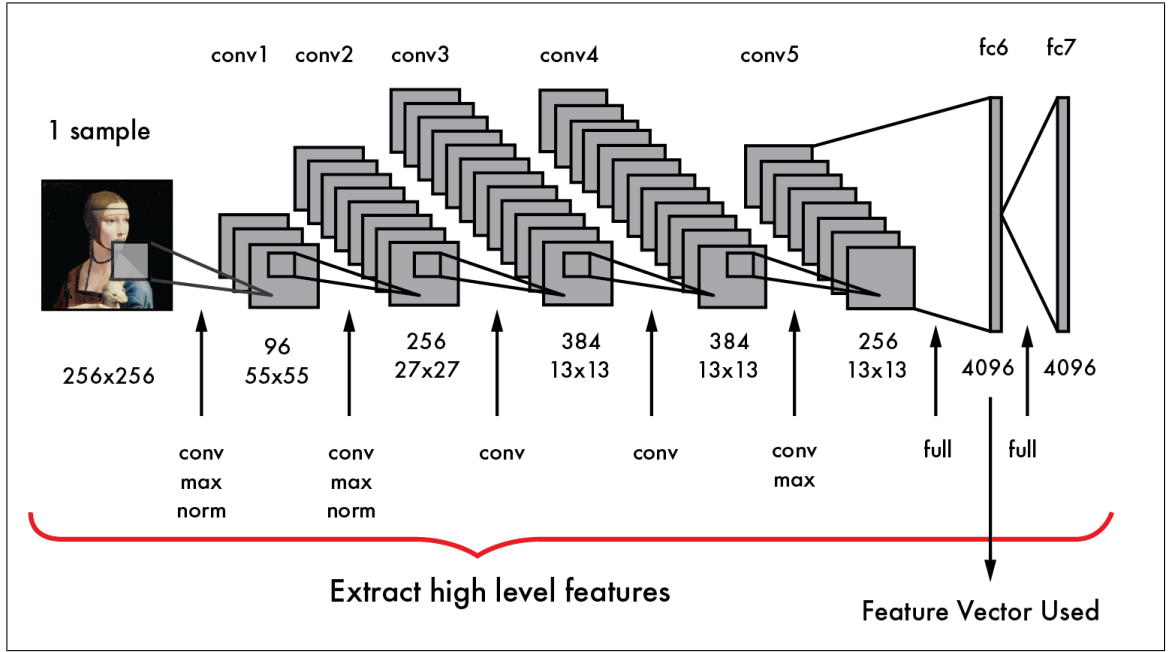


Figure 3.2. Model architecture of the AlexNet Convolutional Deep Neural Network used to extract visual features from images.

### 3.2.2. Attractiveness Visual Features (AVF) Algorithm

The second content-based algorithmic recommender approach employed was a method based on visual attractiveness features. San Pedro and Siersdorfer proposed several explainable visual features that to a great extent, can capture the attractiveness of an image posted on Flickr (San Pedro & Siersdorfer, 2009). Following their procedure, for every image in our *UGallery* dataset we obtain a vector of explicit visual features of attractiveness, using the OpenCV software library<sup>1</sup>: brightness, saturation, sharpness, colorfulness, naturalness, entropy, and RGB-contrast. A more detailed description of these features:

- *Brightness*: It measures the level of luminescence of an image. For images in the *YUV* color space, we obtain the average of the luminescence component *Y*.
- *Saturation*: It measures the vividness of a picture. For images in the *HSV* or *HSL* color space, we obtain the average of the saturation component *S*.
- *Sharpness*: It measures the how detailed is the image.

<sup>1</sup><http://opencv.org/>



- *Colorfulness*: It measures how distance are the colors from the gray color.
- *Naturalness*: It measures how natural is the picture, grouping the pixels in Sky, Grass and Skins pixels and applying the formula in (San Pedro & Siersdorfer, 2009).
- *RGB-contrast*: Measures the variance of luminescence in the RGB color space.
- *Entropy*: Shannon’s entropy is calculated, applied to the histogram of values of every pixel in grayscale used as a vector. The histogram is used as the distribution to calculate the entropy.

These metrics have also been used in another study (Elsweiler, Trattner, & Harvey, 2017), where authors show how to nudge people with attractive images to take up more healthy recipe recommendations. To compute these features, we used the original size of the images and did not pre-process them. More details on how to calculate these visual features can be found in the articles of San Pedro and Siersdorfer (San Pedro & Siersdorfer, 2009), as well as in Messina et al. (Messina et al., 2018).

### 3.2.3. Computing Recommendations

Given a user  $u$  who has consumed a set of artworks  $P_u$ , a constrained profile size  $K$ , and an arbitrary artwork  $i$  from the inventory, the score of this item  $i$  to be recommended to  $u$  is:

$$score(u, i)_X = \frac{\sum_{r=1}^{\min\{K, |P_u|\}} \max_{j \in P_u}^{(r)} \{sim(V_i^X, V_j^X)\}}{\min\{K, |P_u|\}}, \quad (3.1)$$

where  $V_z^X$  is a feature vector of item  $z$  obtained with method  $X$ , where  $X$  can be either a pre-trained AlexNet (DNN) or attractiveness visual features (AVF).  $\max^{(r)}$  denotes the  $r$ -th maximum value, e.g., if  $r = 1$  it is the overall maximum, if  $r = 2$  it is the second maximum, and so on. We compute the average similarity of the top- $K$  most similar images because as shown in Messina et al. (Messina et al., 2018), for different users, the

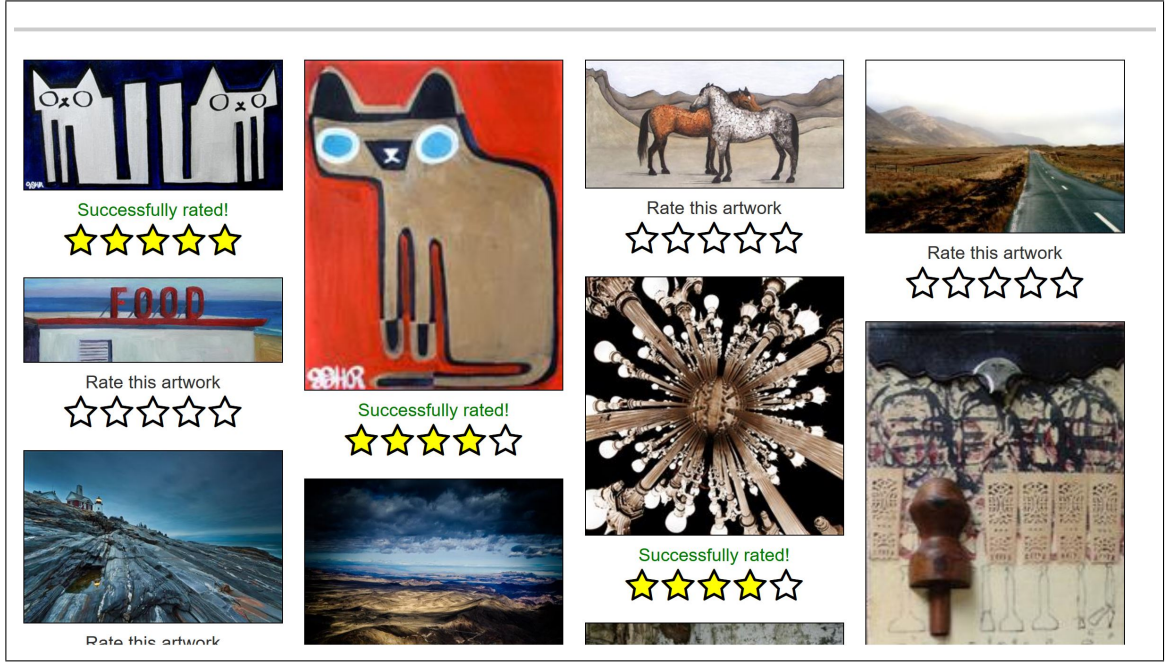


Figure 3.3. Interface 1: Baseline recommendation interface without explanations.

recommendations match better using smaller subsets of the entire user profile. Users do not always look to buy a painting similar to one they bought before, but they look for one that resembles a set of artworks that they liked.  $sim(V_i, V_j)$  denotes a similarity function between vectors  $V_i$  and  $V_j$ . In this particular case, the similarity function used was cosine similarity:

$$sim(V_i, V_j) = cos(V_i, V_j) = \frac{V_i \cdot V_j}{\|V_i\| \|V_j\|} \quad (3.2)$$

Both methods use the same formula to calculate the recommendations. The difference is in the origin of the visual features. For the DNN method, the features were extracted with the AlexNet (Krizhevsky et al., 2012), and in the case of AVF, the features were extracted based on San Pedro et al. (San Pedro & Siersdorfer, 2009).

### 3.3. The Explainable Recommender Interfaces

In our study we explore the effect of explanations in visual content-based artwork recommender systems. In order to guide our design of explanation interfaces, we used

the taxonomy introduced by Friedrich and Zanker (Friedrich & Zanker, 2011). Based on this taxonomy, three dimensions characterize explanations: (i) the recommendation paradigm (collaborative filtering, content-based filtering, knowledge-based, etc.), (ii) reasoning model (white-box or black-box explanation), and (iii) the exploited information categories (user model, recommended item, alternatives). In our case, the dimensions (i) recommendation paradigm and (iii) information categories are set, since we are using a content-based filtering approach (CBVR) and the information used to make explanation is directly obtained from the item, visual features of images. Then, our alternatives for designing explainable interfaces in this research are in the reasoning model: white-box (transparent) or a black-box (opaque) explanation.

These alternatives depend on the type of visual features we use to represent the images. The vector representation of an image obtained from a Deep Convolutional Neural network is rather opaque since the features obtained are unintelligible (Krizhevsky et al., 2012), while the representation obtained with attractiveness visual features (San Pedro & Siersdorfer, 2009) such as brightness, colorfulness, or luminance is comprehensible for humans.

Combining these options, we use explanations based on the content-based paradigm as presented by Friedrich and Zanker (Friedrich & Zanker, 2011), where the attractiveness visual features are used to explain the recommendations in a white-box fashion, Figure 3.4 (a). Alternatively, we explain them in a black-box fashion, just by indicating which similar items in the user preference list produced the recommendation, as in Figure 3.4 (b).

Then, our study contains interface conditions depending on how recommendations are displayed: i) no explanations, as shown in Figure 3.3, ii) black-box explanations based on the top-3 most similar images a user liked in the past, as shown in Figure 3.5, and iii) transparent explanations employing a bar chart of attractiveness visual features, as well as showing the most similar image of the user's item profile, as presented in Figure 3.6. In all three cases the interfaces are vertically scrollable. While Interface 1 (baseline) is able

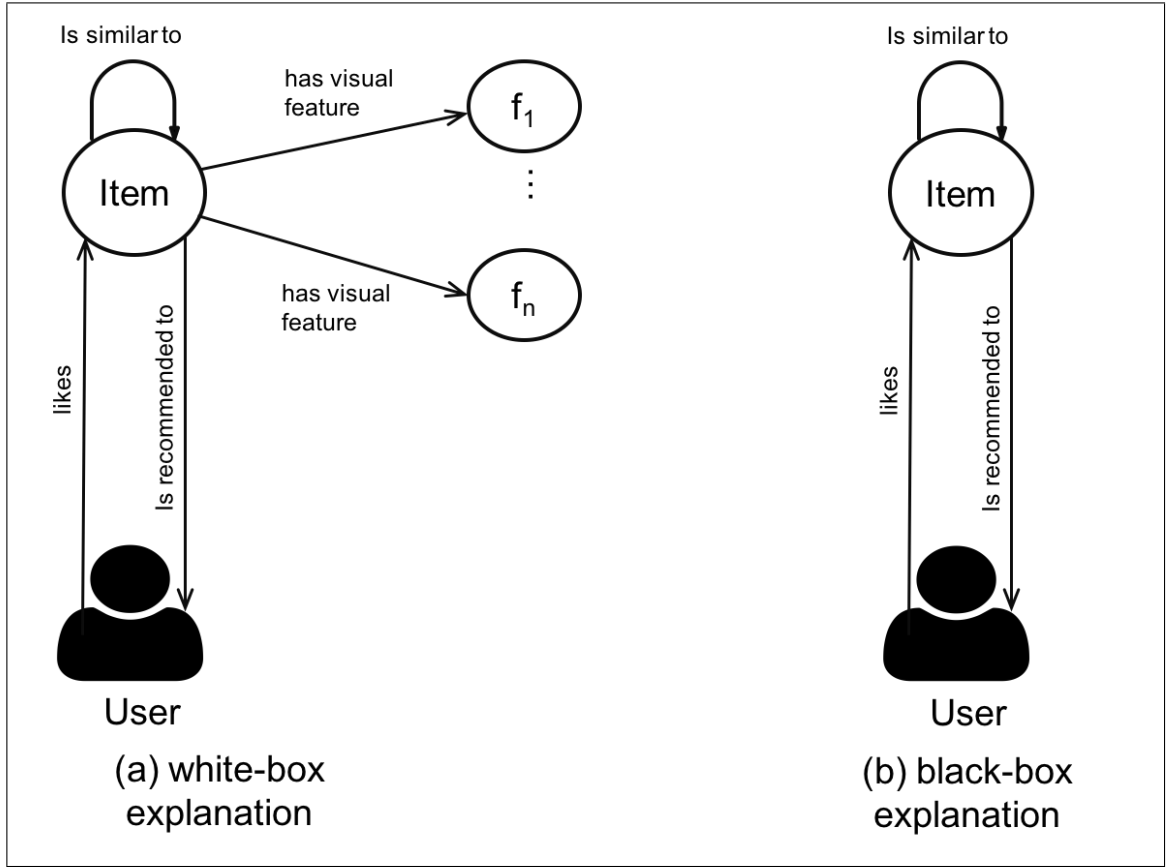


Figure 3.4. Design choices for explainable recommender interfaces, based on Friedrich and Zanker (Friedrich & Zanker, 2011). In (a) we explain the recommendation based on transparent visual features, while in (b) we explain based on item similarity, without details of the features used.

to show 5 images in a row at the same time, interfaces 2 and 3 are capable of showing one recommended image per row to the user.

### 3.4. User Study Procedure

To evaluate the performance of our explainable interfaces we conducted a user study in Amazon Mechanical Turk using a 3x2 mixed design: 3 interfaces (between-subjects) and 2 algorithms (within-subjects, DNN and AVF). The table within Figure 3.7 summarizes the conditions. The interface conditions were: *Condition 1*: interface 1 without explanations, as in Figure 3.3; *Condition 2*: using interface 2, each item recommendation is





Recommended Artwork	Explanation
 <p>Successfully rated!</p> <p>★★★★★</p>	<p>Recommended because:</p> <p>it's 85.31% similar to this artwork that you like    it's 71.48% similar to this artwork that you like    it's 64.05% similar to this artwork that you like</p> <div>    </div> <p>With an average of 73.62%</p>
Recommended Artwork	Explanation

Figure 3.5. Interface 2: Explainable recommendation interface with textual explanations and top-3 similar images.




Recommended Artwork	Explanation												
<div></div> <div>Rate this artwork</div> <div></div>	<div><div><div>?</div></div><div><div>Recommended because:</div><div><div>it's 97.09% similar to this artwork that you like</div><div></div></div></div></div> <div><div>Attractiveness Features</div><table><tr><td>brightness</td><td><div><div></div><div></div></div></td></tr><tr><td>sharpness</td><td><div><div></div><div></div></div></td></tr><tr><td>saturation</td><td><div><div></div><div></div></div></td></tr><tr><td>colorfulness</td><td><div><div></div><div></div></div></td></tr><tr><td>entropy</td><td><div><div></div><div></div></div></td></tr><tr><td>contrast</td><td><div><div></div><div></div></div></td></tr></table><div><div>0</div><div>1</div></div><div><div><div></div>Recommended Artwork</div><div><div></div>Liked Artwork</div></div></div>	brightness	<div><div></div><div></div></div>	sharpness	<div><div></div><div></div></div>	saturation	<div><div></div><div></div></div>	colorfulness	<div><div></div><div></div></div>	entropy	<div><div></div><div></div></div>	contrast	<div><div></div><div></div></div>
brightness	<div><div></div><div></div></div>												
sharpness	<div><div></div><div></div></div>												
saturation	<div><div></div><div></div></div>												
colorfulness	<div><div></div><div></div></div>												
entropy	<div><div></div><div></div></div>												
contrast	<div><div></div><div></div></div>												
Recommended Artwork	Explanation												

Figure 3.6. Interface 3: Explainable and transparent recommendation interface with features' bar chart and top-1 similar image.

explained based on the top 3 most similar images in the user profile, as in Figure 3.5; and *Condition 3*: only for AVF algorithm, based on a bar chart of visual features, as in Figure 3.6, but for DNN we used the explanation based on top 3 most similar images, because the

neural embedding of 4,096 dimensions has no transparent (*human-interpretable*) features to show in a bar chart.

To compute the recommendations for each of the three interface conditions two recommender algorithms were chosen: one based on DNN visual features, and the other based on attractiveness visual features (AVF). The order in which the algorithms were presented was chosen at random to diminish the chance of a learning effect.

With respect to the complete study workflow, as shown in Figure 3.7, participants accepted the study on Mechanical Turk (<https://www.mturk.com>) and they were redirected to a web application. After accepting a consent form, they are redirected to the pre-study survey, which collects demographic data (age, gender) and a subject's previous knowledge of art based on the test by Chatterjee et al. (Chatterjee, Widick, Sternschein, Smith II, & Bromberger, 2010).

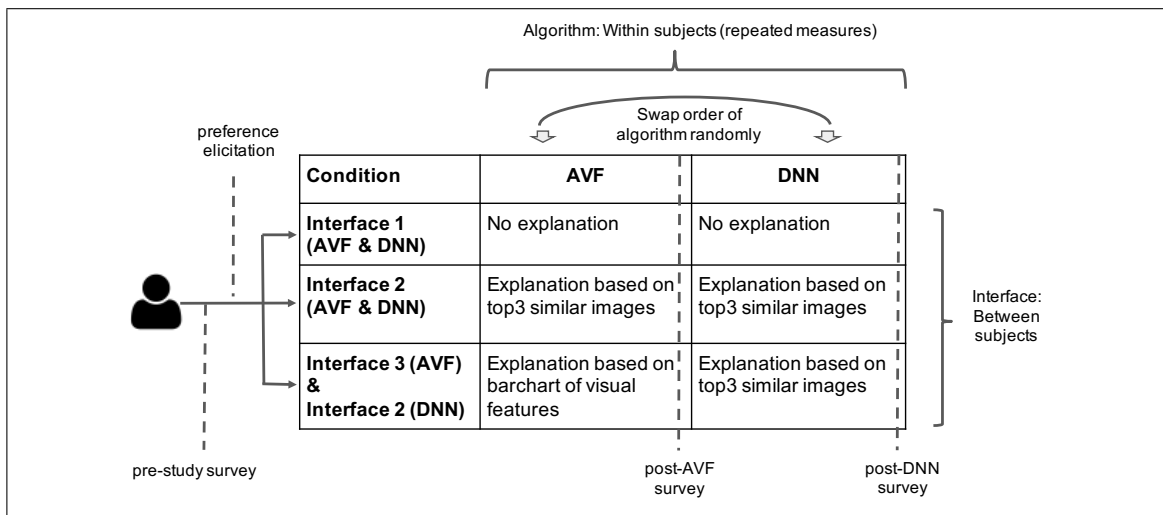


Figure 3.7. Study procedure. After the pre-study survey and the preference elicitation, users were assigned to one of three possible interfaces. In each interface they evaluated recommendations of two algorithms: DNN and AVF.

Following this, they had to perform a preference elicitation task. In this step, the users had to “like” at least ten paintings, using a Pinterest-like interface. Next, they were

randomly assigned to one interface condition. In each condition, they again provided feedback (rating with 1-5 scale to each image) to top ten recommendations of images with employing either the DNN or the AVF algorithm (also assigned at random as discussed before). Finally, the participants were asked to next answer a post-algorithm survey. The dimensions evaluated in the post-algorithm survey are the same for DNN and AVF algorithms. They were presented in the form of statements where the user had to indicate their level of agreement in a 0 (totally disagree) to 100 (totally agree) scale:

- **Explainable:** I understood why the art images were recommended to me.
- **Relevance:** The art images recommended matched my interests.
- **Diverse:** The art images recommended were diverse.
- **Interface Satisfaction:** Overall, I am satisfied with the recommender interface.
- **Use Again:** I would use this recommender system again for finding art images in the future.
- **Trust:** I trusted the recommendations made.

This process is repeated for the second algorithm as well. Once the participants finished answering the second post study survey, they were redirected to the final view, where they received a survey code for later payment in Amazon Mechanical Turk.

### 3.5. System Architecture

The system built to run the experiment has four main components. The four components and their interactions can be seen in Figure 3.8. Details of these components are:

- The Frontend or Browser was built using React.js <sup>2</sup>, this is in charge of showing the recommendations and the items for the elicitation task. Is in charge of all the direct interactions between the user and the system.

---

<sup>2</sup> <https://reactjs.org/>

- The Recommender API was developed using Flask <sup>3</sup>. This part was in charge of computing the recommendations after receiving the user feedback from the preference elicitation task. This was running in different server with high computational power.
- The database management system used was PostgreSQL <sup>4</sup>. Is in charge of storing al the data collected in the study.
- The Server was developed using Express.js <sup>5</sup>. This was in charge of making the connections between all the components in the system. Also, is in charge of querying the Survey Monkey API to check if the users fulfilled the surveys.

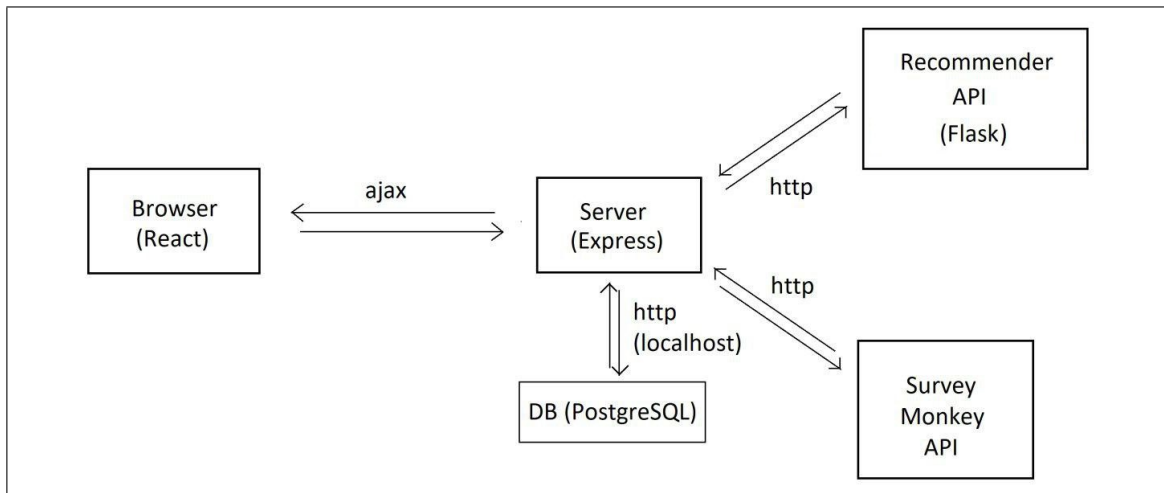


Figure 3.8. System architecture.

<sup>3</sup> <http://flask.pocoo.org/>

<sup>4</sup> <https://www.postgresql.org/>

<sup>5</sup> <https://expressjs.com/>



Table 3.1. Results of users’ perception over several evaluation dimensions, defined in Section 3.4. Scale 1-100 (higher is better), except for Average rating (scale 1-5). DNN: Deep Neural Network, and AVF: Attractiveness visual features. The symbol  $\uparrow^1$  indicates interface-wise significant difference (differences between interfaces using the same algorithms). The \* symbol denotes algorithm-wise statistical difference (comparing a dimension between algorithms, using the same interface).

Condition	Explainable		Relevance		Diverse		Interface Satisfaction		Use Again		Trust		Average Rating	
	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF
<b>Interface 1</b> (No Explanations)	66.2*	51.4	69.0*	53.6	46.1	69.4*	69.9	62.1	65.8	59.7	69.3	63.7	3.55*	3.23
<b>Interface 2</b> (DNN & AVF: Top-3 similar images)	83.5* $\uparrow^1$	74.0 $\uparrow^1$	80.0*	61.7	58.8	69.9*	76.6*	61.7	76.1*	65.9	75.9*	62.7	3.67*	3.00
<b>Interface 3</b> (DNN: Top-3 similar, AVF: feature bar chart)	84.2* $\uparrow^1$	70.4 $\uparrow^1$	82.3* $\uparrow^1$	56.2	65.3 $\uparrow^1$	71.2	69.9*	63.3	78.2*	58.7	77.7*	55.4	3.90*	2.99

Stat. significance between interfaces by multiple t-tests, Bonferroni corr.  $\alpha_{bonf} = \alpha/n = 0.05/3 = 0.0017$ . Stat. significance between algorithms using pairwise t-test,  $\alpha = 0.05$ .

Table 3.2. NASA TLX Results.

Condition	Mental		Hurry		Insecure	
	DNN	AVF	DNN	AVF	DNN	AVF
<b>Interface 1</b> (No Explanations)	19.90	23.24	10.78	13.41	12.22	12.88
<b>Interface 2</b> (DNN & AVF: Top-3 similar images)	20.05	18.46	11.54	12.08	7.62	6.59
<b>Interface 3</b> (DNN: Top-3 similar, AVF: feature bar chart)	23.41	26.37	14.29	15.73	13.32	16.37 $\uparrow^2$

## 4. RESULTS

### 4.1. Demographic Results

The study was finished by 200 users out of which 121 were able to answer our validation questions successfully and hence were included in the results. In total, we had two validation questions set to check for attention of our study participants. Filtering out users not responding properly to these questions allowed us to include 41 users for the Interface 1 condition, 41 users for Interface 2 condition and 39 users for Interface 3 condition. In total, participants were paid an amount of 0.40 USD per study, which took them around 10 minutes to complete.

Our subjects were between 18 to over 60 years old. 36% were between 25 to 32 years old, and 29% between 32 to 40 years old. Females made up 55.4%. 12% just finished high school, 31% had a some college degree, 57% had a bachelor's, master's or Ph.D. degree. Only 8% reported some visual impairment. With respect to their understanding about art, 20% did not have experience, 48% had attended 1 or 2 lessons, and 32% reported to have attended 3 or more lessons at high school level or above. 20% of our subjects also reported that they almost never visited a museum or an art gallery; 36% do this once a year; and 44% do this once every 1 or 6 months.

**Differences between Interfaces.** Table 3.1 summarizes the results of the user study. First we compared interface performance and then we looked at the algorithmic performance. The explainable interfaces (Interface 2 and 3) significantly improved the perception of explainability compared to Interface 1 under both algorithms. There is also a significant improvement over Interface 1 in terms of relevance and diversity, but this is only achieved by the DNN method when this is compared against the AVF method using the interface 3. Interestingly, this is the condition where the interface is more transparent, since it explains exactly what is used to recommend (brightness, saturation, sharpness, etc.). People report that they understand why the images are recommended (70.4), but

since the relevance is rather insufficient (56.2), the perception of trust is reported as low (55.4).

Table 4.1. CFA output

<b>Construct</b>	<b>Item</b>	<b>Loading</b>
Effort	Insecure	0.826
$\alpha = 0.865$	Rush	0.906
$AVE = 0.6883$	Mental Demand	0.750
Satisfaction	Satisfaction System	0.875
$\alpha = 0.955$	Use system	0.973
$AVE = 0.880$	Recommend friend	0.963

**Differences between Algorithms.** With the only exception of the dimension *Diverse* where AVF was significantly better, DNN was perceived more positively than AVF at large. In interfaces 2 and 3, the DNN method was perceived significantly better in 5 dimensions (explainability, relevance, interface satisfaction, interest for eventual use, and trust), as well as higher average rating.

Overall, the results indicate that the explainable interface based on top 3 similar images works better than an interface without explanation. Moreover, this effect is enhanced by the accuracy of the algorithm, so even if the algorithm has no explainable features (DNN) it could induce more trust if the user perceives a larger predictive preference accuracy.

A very notable result is that the difference in Trust between the two algorithms is not significant under the non-explainable interface ( $DNN = 65.8$  vs.  $AVF = 59.7$ ), but this difference turns significant under the explainable interface conditions, either with non-transparent explanation ( $DNN = 76.1$  vs.  $AVF = 65.9$ ) or when comparing non-transparent ( $DNN = 78.2$ ) with transparent visual explanation ( $AVF = 58.7$ ).

## **4.2. A Model of the UX with an Art Recommender**

In order to provide a comprehensive and complete understanding of the dependent and independent variables involved in this study, as well as their relationships, we conducted an analysis based on Structural Equation Models (SEM). In order to reduce the number of variable combinations and to cluster the variables in cohesive groups, we followed the recommender systems evaluation framework by Knijnenburg et al. (Knijnenburg, Willemsen, et al., 2012a). In this way, we could group the variables in: (a) Personal Characteristics, (b) Objective System Aspects, (c) Subjective System Aspects, (d) Interactions, and (e) User Experience.

Prior to this analysis, we conducted a Confirmatory Factor Analysis (CFA) to reduce the number of variables and group them in more understandable constructs to be included in the SEM.

### **4.2.1. Confirmatory Factor Analysis**

We conducted a CFA and examined the validity and reliability scores of the constructs measured in our study. We constructed 2 factors: *Effort* and *Satisfaction*. The items used share at least 56.2% of their variance with their designated construct. To ensure the convergent validity of constructs, we examined the average variance extracted (AVE) of each construct. The AVEs were all higher than the recommended value of 0.50, indicating adequate convergent validity. To ensure discriminant validity, we ascertained that the square root of the AVE for each construct was higher than the correlations of the construct with other constructs.

### **4.2.2. Structural Equation Model**

We subjected the 2 factors we found in the CFA, all the items that could explain and mediate relations and the experimental conditions to structural equation modeling, which simultaneously fits the factor measurement model and the structural relations between

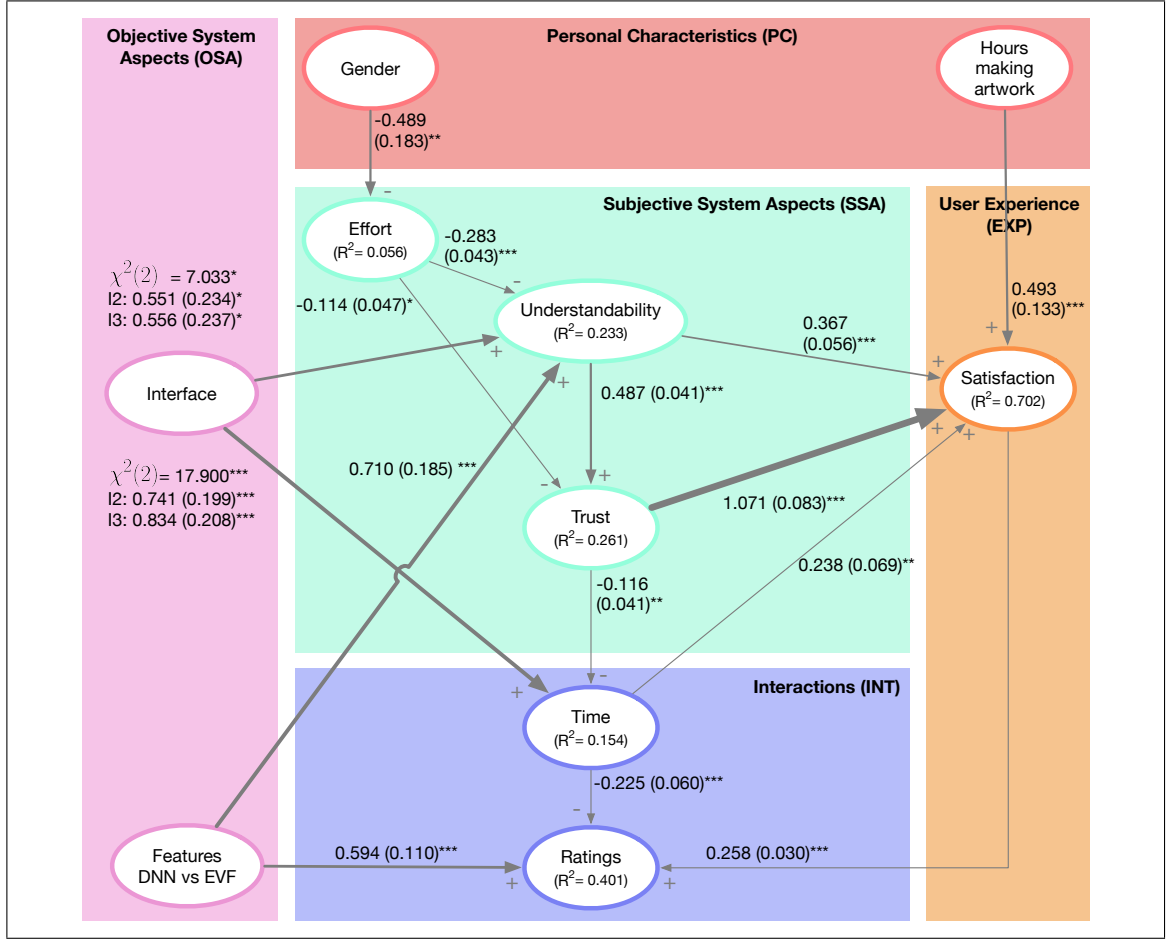


Figure 4.1. The structural equation model for the data of the experiment using Knijnenburg's evaluation framework for recommender systems. Significance levels: \*\*\* $p < .001$ , \*\* $p < .01$ , \* $p < 0.05$ .  $R^2$  is the proportion of variance explained by the model. Numbers on the arrows (and their thickness) represent the  $\beta$  coefficients (and standard error) of the effect. Factors are scaled to have an  $SD$  of 1.

factors and other variables. The model has a good <sup>1</sup> fit:  $\chi^2(72) = 103.935$ ,  $p = 0.008$ ;  $RMSEA = 0.043$ , 90%  $CI : [0.022, 0.060]$ ,  $CFI = 0.997$ ,  $TLI = 0.996$ .

<sup>1</sup>A model should not have a non-significant  $\chi^2$  ( $p > 0.05$ ), but this statistic is often regarded as too sensitive. Hu and Bentler (Hu & Bentler, 1999) propose cut-off values for other fit indices to be:  $CFI > 0.96$ ,  $TLI > 0.95$ , and  $RMSEA < 0.05$ , with the upper bound of its 90% CI below 0.10.

*Effect of algorithm:* The algorithm used to create the features has a positive an effect on understandability. When using DNN features users tend to understand better. As we saw in the last section, DNN also has a positive effect on the ratings.

*Effects of interface on understandability:* The model shows that the interfaces with explanations have a positive effect on understandability, which has a positive effect on satisfaction, on its own and mediated by trust. This result is consistent with the model found in (Gedikli, Jannach, & Ge, 2014), that indicates that users are "more satisfied with explanation facilities which provide justifications for the recommendations".

*Effects of interface on time:* Explainable interfaces also have a positive effect on time, that also has a positive effect on satisfaction. This suggests that users need to take time to understand analyze explanations. Gedikli et al, in (Gedikli et al., 2014), also found that this effect on their model.

*Effect of trust in satisfaction:* The effect that *trust* has upon satisfaction is almost 3 times larger than the effect of understandability. This highlights the fact that users' satisfaction strongly depend on how much they trust the system they are interacting with. It is interesting to notice that, based on our model, neither the interface nor the algorithm used to create the features has a direct effect on trust. Both effects are mediated by understandability, which could mean that users only trust something they understand.

*Effect of effort:* The *effort* has a negative effect on understandability and trust. When users have to make too much effort when interacting with the system, they also perceive a smaller understanding of the system and its recommendations.

## 5. FUTURE WORK

One of the aspects that were not evaluated in this study was the effect of different types of interfaces and multitouch systems. The device used to conduct the study is a relevant factor to consider, it could have a significant effect in the entire user experience. It is necessary to make a new user study with the interface or device as a variable to investigate.

Another factor that will be interesting to measure is the effect of making the explanations optional. If the conditions of the experiments are optional and mandatory explanation, probably the results will show different types of users, depending on their interest in the recommendations. Also, this will show if the users are using the explanations to make their evaluation over the items recommended.

In recent times, the traditional five-point star rating scale evaluation is disappearing. The online social networks are using more frequently the like/dislike scale because is something that gives to the user two options well defined, also reduces the variance in the results. Trying a different scale like the like/dislike and study its effect is something that must be done in the future.

The last idea for future work is to evaluate the performance online of a state of the art algorithm. The superiority of the deep neural networks was demonstrated in this work. Now that it is proved, we can use in an online environment the most recent algorithm. For example something like the Youtube algorithm (Covington, Adams, & Sargin, 2016) or VBPR (He & McAuley, 2016), or create a new model based on them.

## 6. CONCLUSIONS

In this work, we have studied the effect of explaining recommendation of images employing three different recommender interfaces, as well as interactions with two different visual content-based recommendation algorithms: one with high predictive accuracy but with unexplainable features (DNN), and another with lower accuracy but with higher potential for explainable features (AVF).

The first result, which answers RQ1, shows that explaining the images recommended has a positive effect vs. no explanation. Moreover, the explanation based on top 3 similar images presents the best results, but we need to consider that the alternative method, explanations based on visual features, was only used with the AVF. This result should be further studied in other image dataset, and it opens a new branch of research in terms of new interfaces which could help to explain the features learned by a deep neural network of images.

Regarding RQ2, we see that the algorithm used plays an important role in conjunction with the interface. DNN is perceived better than AVF in most dimensions evaluated, showing that further research should focus on the interaction between algorithm and explainable interfaces. In the future we will expand this work to other datasets, beyond artistic images, to generalize our results.

Finally, with respect to RQ3, we have provided a holistic model, based on the framework by Knijnenburg et al. (Knijnenburg, Willemsen, et al., 2012a), which explains the relations among different independent variables (interface, algorithm, art domain expertise) and several metrics to measure the user experience with an explainable recommender systems of artistic images. In future work, we would like to use more advance models for explaining art recommendations based on recent models of neural style transfer (Gatys, Ecker, & Bethge, 2016; Olah et al., 2017) and test them using this user-centric recommender evaluation framework.



## REFERENCES

- Amatriain, X. (2013). Mining large streams of user data for personalized recommendations. *ACM SIGKDD Explorations Newsletter*, 14(2), 37–48.
- Andjelkovic, I., Parra, D., & ODonovan, J. (2018). Moodplay: Interactive music recommendation based on artists mood similarity. *International Journal of Human-Computer Studies*.
- Aroyo, L., Wang, Y., Brussee, R., Gorgels, P., Rutledge, L., & Stash, N. (2007). Personalized museum experience: The rijksmuseum use case. In *Proceedings of museums and the web*.
- Benouaret, I., & Lenne, D. (2015). Personalizing the museum experience through context-aware recommendations. In *Systems, man, and cybernetics (smc), 2015 ieee international conference on* (pp. 743–748).
- Celma, O. (2010). Music recommendation. In *Music recommendation and discovery* (pp. 43–85). Springer.
- Chatterjee, A., Widick, P., Sternschein, R., Smith II, W., & Bromberger, B. (2010, 07). The assessment of art attributes. , 28, 207-222.
- Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for youtube recommendations. In *Proceedings of the 10th acm conference on recommender systems* (pp. 191–198).
- Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., . . . Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5), 455.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009a). ImageNet: A Large-Scale Hierarchical Image Database. In *Cvpr09*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009b). Imagenet: A large-scale hierarchical image database. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp. 248–255).

Dominguez, V., Messina, P., Parra, D., Mery, D., Trattner, C., & Soto, A. (2017). Comparing neural and attractiveness-based visual features for artwork recommendation. In *Proceedings of the workshop on deep learning for recommender systems, co-located at recsys 2017*. Retrieved from <https://arxiv.org/pdf/1706.07515.pdf> doi: 10.1145/3125486.3125495

Elsweiler, D., Trattner, C., & Harvey, M. (2017). Exploiting food choice biases for healthier recipe recommendation. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval* (pp. 575–584).

Friedrich, G., & Zanker, M. (2011). A taxonomy for generating explanations in recommender systems. *AI Magazine*, 32(3), 90–98.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2414–2423).

Gedikli, F., Jannach, D., & Ge, M. (2014). How should i explain? a comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72(4), 367 - 382. Retrieved from <http://www.sciencedirect.com/science/article/pii/S1071581913002024> doi: <https://doi.org/10.1016/j.ijhcs.2013.12.007>

Gomez-Uribe, C. A., & Hunt, N. (2016). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems*

(*TMIS*), 6(4), 13.

He, R., Fang, C., Wang, Z., & McAuley, J. (2016). Vista: A visually, socially, and temporally-aware model for artistic recommendation. In *Proceedings of the 10th acm conference on recommender systems* (pp. 309–316). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/2959100.2959152> doi: 10.1145/2959100.2959152

He, R., & McAuley, J. (2016). Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the thirtieth aaai conference on artificial intelligence* (pp. 144–150).

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations [Conference Proceedings]. In *Proceedings of the 2000 acm conference on computer supported cooperative work* (p. 241-250). ACM. Retrieved from <http://doi.acm.org/10.1145/358916.358995>

Holliday, D., Wilson, S., & Stumpf, S. (2016). User trust in intelligent systems: A journey over time. In *Proceedings of the 21st international conference on intelligent user interfaces* (pp. 164–168).

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. Retrieved from <https://doi.org/10.1080/10705519909540118> doi: 10.1080/10705519909540118

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 453–456).

Knijnenburg, B. P., Bostandjiev, S., O'Donovan, J., & Kobsa, A. (2012). Inspectability and control in social recommenders [Conference Proceedings]. In *Proceedings of the sixth*

*acm conference on recommender systems* (p. 43-50). ACM. Retrieved from <http://doi.acm.org/10.1145/2365952.2365966>

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012a). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 441–504. doi: 10.1007/s11257-011-9118-4

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012b). Explaining the user experience of recommender systems [Journal Article]. *User Modeling and User-Adapted Interaction*, 22(4-5), 441-504. Retrieved from <http://dx.doi.org/10.1007/s11257-011-9118-4>

Konstan, J. A., & Riedl, J. (2012). Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2), 101–123.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Maes, P., et al. (1994). Agents that reduce work and information overload. *Communications of the ACM*, 37(7), 30–40.

McNee, S. M., Kapoor, N., & Konstan, J. A. (2006). Don't look stupid: avoiding pitfalls when recommending research papers. In *Proceedings of the 2006 20th anniversary conference on computer supported cooperative work* (pp. 171–180).

Messina, P., Dominguez, V., Parra, D., Trattner, C., & Soto, A. (2018). Content-based artwork recommendation: Integrating painting metadata with neural and manually-engineered visual features. *User Modeling and User-Adapted Interaction*. doi: 10.1007/s11257-018-9206-9

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*. (<https://distill.pub/2017/feature-visualization>) doi: 10.23915/distill.00007

Parra, D., & Brusilovsky, P. (2015). User-controllable personalization: A case study with setfusion. *International Journal of Human-Computer Studies*, 78, 43–67.

Pu, P., Chen, L., & Hu, R. (2011). A user-centric evaluation framework for recommender systems [Conference Proceedings]. In *Proceedings of the fifth acm conference on recommender systems* (p. 157-164). ACM. Retrieved from <http://doi.acm.org/10.1145/2043932.2043962>

San Pedro, J., & Siersdorfer, S. (2009). Ranking and classifying attractiveness of photos in folksonomies. In *Proceedings of the 18th international conference on world wide web* (pp. 771–780). New York, NY, USA: ACM. Retrieved from <http://doi.acm.org/10.1145/1526709.1526813> doi: 10.1145/1526709.1526813

Schenk, E., & Guittard, C. (2009, 01). Crowdsourcing: What can be outsourced to the crowd, and why ?

Semeraro, G., Lops, P., De Gemmis, M., Musto, C., & Narducci, F. (2012). A folksonomy-based recommender system for personalized access to digital artworks. *Journal on Computing and Cultural Heritage (JOCCH)*, 5(3), 11.

Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 806–813).

Sinha, R., & Swearingen, K. (2002). The role of transparency in recommender systems. In *Chi'02 extended abstracts on human factors in computing systems* (pp. 830–831).

Tintarev, N., & Masthoff, J. (2007). Effective explanations of recommendations: user-centered design [Conference Proceedings]. In *Proceedings of the 2007 acm conference on recommender systems* (p. 153-156). ACM. Retrieved from <http://doi.acm.org/10.1145/1297231.1297259>

Tintarev, N., & Masthoff, J. (2011). Designing and evaluating explanations for

recommender systems [Book Section]. In *Recommender systems handbook* (p. 479-510). Springer US. Retrieved from [http://dx.doi.org/10.1007/978-0-387-85820-3\\_15](http://dx.doi.org/10.1007/978-0-387-85820-3_15)

Tintarev, N., & Masthoff, J. (2015). Explaining recommendations: Design and evaluation. In *Recommender systems handbook* (pp. 353–382). Springer.

Trattner, C., Oberegger, A., Eberhard, L., Parra, D., Marinho, L., et al. (2016). Understanding the impact of weather for poi recommendations. *Proceedings of RecTour Workshop, co-located at ACM RecSys*.

van den Broek, E. L., Kok, T., Schouten, T. E., & Hoenkamp, E. (2006). Multimedia for art retrieval (m4art). In *Multimedia content analysis, management, and retrieval 2006* (Vol. 6073, p. 60730Z).

Weinswig, D. (2016). *Art Market Cooling, But Online Sales Booming*. <https://www.forbes.com/sites/deborahweinswig/2016/05/13/art-market-cooling-but-online-sales-booming/>. ([Online; accessed 21-March-2017])

Ye, M., Yin, P., Lee, W.-C., & Lee, D.-L. (2011). Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international acm sigir conference on research and development in information retrieval* (pp. 325–334).

Yuan, Q., Cong, G., Ma, Z., Sun, A., & Thalmann, N. M. (2013). Time-aware point-of-interest recommendation. In *Proceedings of the 36th international acm sigir conference on research and development in information retrieval* (pp. 363–372).

Ziegler, C.-N., McNee, S. M., Konstan, J. A., & Lausen, G. (2005). Improving recommendation lists through topic diversification [Conference Proceedings]. In *Proceedings of the 14th international conference on world wide web* (p. 22-32). ACM. Retrieved from <http://doi.acm.org/10.1145/1060745.1060754>