The Effect of Explanations and Algorithmic Accuracy on Visual Recommender Systems of Artistic Images

Vicente Dominguez
Pontificia Universidad Católica de Chile & IMFD
Santiago, Chile
vidominguez@uc.cl

Pablo Messina
Pontificia Universidad Católica de Chile & IMFD
Santiago, Chile
pamessina@uc.cl

Ivania Donoso-Guzmán
Pontificia Universidad Católica de Chile & Conversica
Santiago, Chile
indonoso@uc.cl

Denis Parra
Pontificia Universidad Católica de Chile & IMFD
Santiago, Chile
dparra@ing.puc.cl

ABSTRACT

There are very few works about explaining content-based recommendations of images in the artistic domain. Current 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, and trust. In this paper, 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 and relevance. Furthermore, our results show that the observed effects are also dependent on the underlying recommendation algorithm used. We tested two algorithms: Deep Neural Networks (DNN), which has high accuracy, and Attractiveness Visual Features (AVF) with high transparency but lower accuracy. Our results indicate that algorithms should not be studied in isolation, but rather in conjunction with interfaces, since both 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 which synthesizes the effects between different variables involved in the user experience with explainable visual recommender systems of artistic images.

IUI '19, March 17–20, 2019, Marina del Ray, CA, USA

ACM ISBN 978-1-4503-6272-6/19/03…$15.00
https://doi.org/10.1145/3301275.3302274

CCS CONCEPTS

• **Information systems** → **Recommender systems**; *Personalization*; • **Human-centered computing** → *User studies*; • **Computing methodologies** → Neural networks.

KEYWORDS

Visual recommender systems, explainable AI, art

ACM Reference Format:
Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and Denis Parra. 2019. The Effect of Explanations and Algorithmic Accuracy on Visual Recommender Systems of Artistic Images . In *24th International Conference on Intelligent User Interfaces (IUI '19), March 17–20, 2019, Marina del Ray, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3301275.3302274

1 INTRODUCTION

Online artwork recommendation has received little attention compared to other areas such as movies [1, 17] and music [6, 29]. Most research on artwork recommendation deals with studies on museum data [3, 4, 37, 41], but there is little work with datasets of online artwork e-commerce systems [20, 31]. 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[1].

The first works in the area of artwork recommendation date from 2006-2007 such as the CHIP [3] 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. [41], 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 [9, 20], which is rather expected considering that

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

DNNs are the top performing models for obtaining visual features for several tasks, such as image classification [28], and scene identification [38]. More recently, Messina et al. [31] 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 [25, 30], and also domain knowledge is an important variable to explain the user experience with a recommender system [2, 24, 35].

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 [34]. In contrast, features of visual attractiveness, despite being less accurate to predict user preference [31], could be easily explained, based on color, brightness or contrast [36]. Explanations in recommender systems have been shown to have a significant effect on user satisfaction [39], 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.

Objective. In this paper, we research the effect of explaining artistic image suggestions. In particular, we conduct a user study on Amazon Mechanical Turk (N=121) under three different interfaces and two different algorithms. The three interfaces are: i) no explanations, ii) explanations based 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* (http://www.UGallery.com/). Finally, we contribute with a Structural Equation Model based on the framework by Knijnenburg et al. [24] in order to fully explain the user experience with a explainable VBCR of artistis images.

Research Questions. To drive our research, the following three questions were defined:

- **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?

Outline. Our work is structured as follows: In Section 2 we survey relevant related work 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 5 that presents the global SEM which connects all the studied variables, and finally section 6 concludes the paper and provides an outlook for future work.

2 RELATED WORK

Relevant related research is collated in two sub-sections: first, we review research on recommending artistic images to people. Second we summarize studies on explaining recommender systems. Both are important to our problem at hand. The final paragraph in this section highlights the differences to previous work and our contributions to the existing literature in the area.

Recommendations of Artistic Images. The works of Aroyo et al. [3] with the CHIP project and Semeraro et al. [37] 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. [4], 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 [20], and the work of Dominguez et al. [9] and Messina et al. [31] compared neural against traditional visual features. None of the aforementioned works performed a user study under explanation interfaces to generalize their results.

Explaining Recommender Systems. There are some related works in the general area of explanations for recommender systems [22, 39]. Though a good amount of research has been published in the area about making explanations using tags [43], social connections [40], linked-open data [33], methods with soft-probabilistic logic [26] as well as visually-enhanced recommendation interfaces [5, 11, 19, 27, 35, 42],

to the best of our knowledge, no previous research has conducted a user study to understand the effect of explaining recommendation of artistic images based on different visual features.

The closest works in this aspect are researches oriented to automatically add caption to images [13, 32] or to explain image classifications [21], but they are not directly related to personalized recommender systems.

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. In a previous study, we conducted a preliminary analysis of these effects [10], but here we contribute with a full model. Our contributions are then threefold: 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. [24], in order to characterize all the variables involved in the user experience with a explainable VCBR of art images.

3 METHODS

In the following section we describe in detail our study methods. First, we introduce the dataset chosen for the purpose of our study. Second, 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.

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 [44]. They support emergent artists by helping them sell their artwork online. 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 .

Visual Recommendation Approaches

As mentioned earlier in this paper, 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 do not apply.

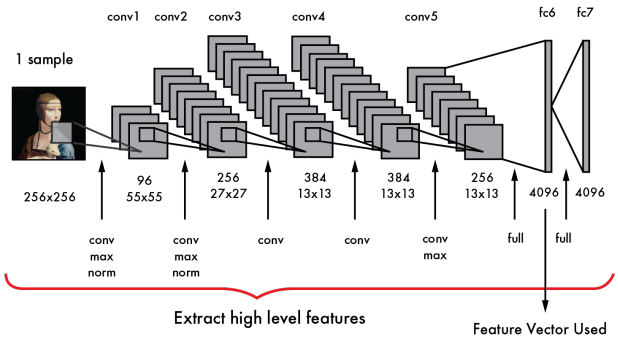

Figure 1: Model architecture of the AlexNet Convolutional Deep Neural Network used to extract visual features from images.

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 [28], which architecture is shown in Figure 1. In particular, we use an AlexNet model pre-trained with the ImageNet dataset [8]. Using the pre-trained weights, for every image a vector of 4,096 dimensions was generated with the Caffe[2] framework. We resized every image to a $227x227$ image. This is the standard pre-processing needed to use the AlexNet.

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 in [36] proposed several explainable visual features that to a great extent, can capture the attractiveness of an image posted on Flickr. 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[3]: 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 how detailed is the image.

[2]http://caffe.berkeleyvision.org/
[3]http://opencv.org/

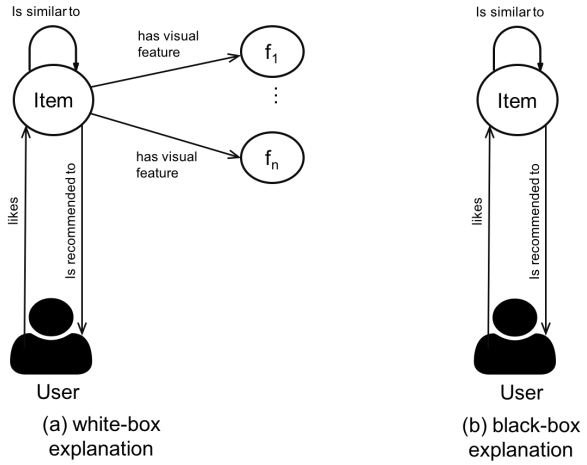

Figure 2: Design choices for explainable recommender interfaces, based on Friedrich and Zanker [14]. In (a) we explain the recommendation based on transparent visual features, while in (b) we explain based on item similarity, without details of the features used.

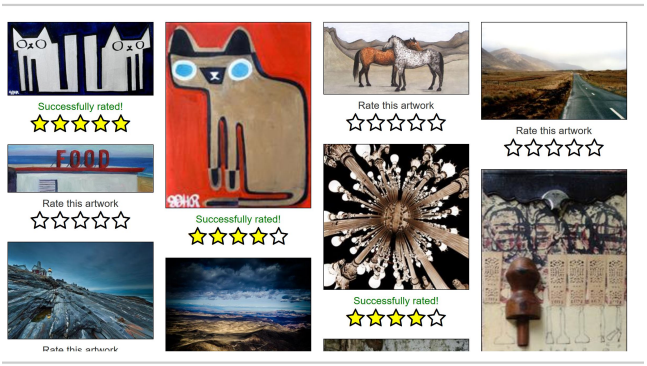

Figure 3: Interface 1: Baseline recommendation interface without explanations.

- *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 [36].
- *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 [12], 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 [36], as well as in Messina et al. [31].

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:

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

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. [31], for different users, the 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:

$$sim(V_i, V_j) = cos(V_i, V_j) = \frac{V_i \cdot V_j}{\|V_i\|\|V_j\|} \quad (2)$$

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 [28], and in the case of AVF, the features were extracted based on San Pedro et al. [36].

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 [14]. 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

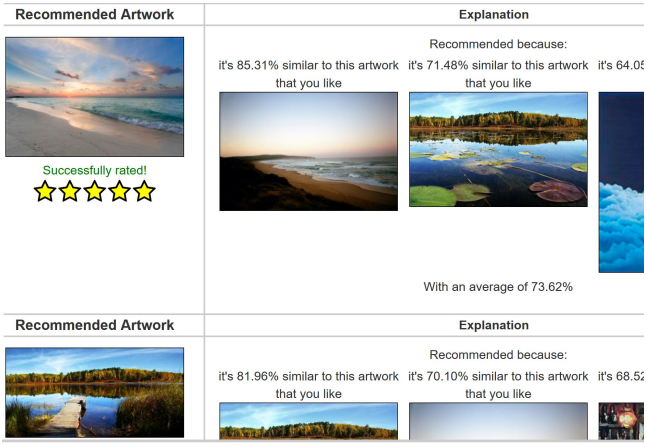

Figure 4: Interface 2: Explainable recommendation interface with textual explanations and top-3 similar images.

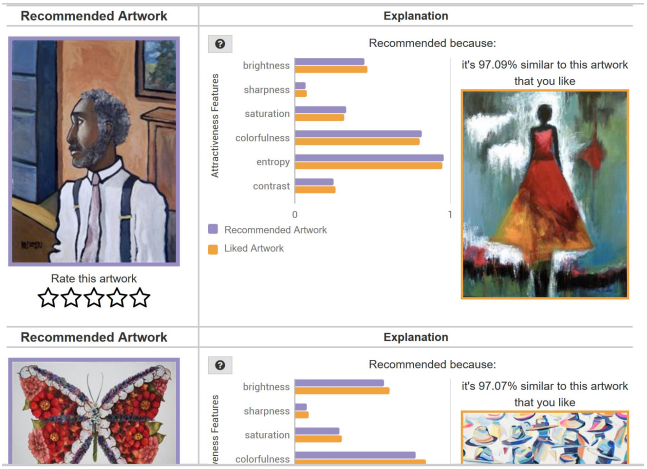

Figure 5: Interface 3: Explainable and transparent recommendation interface with features' bar chart and top-1 similar image.

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 unitelligible [28], while the representation obtained with attractiveness visual features [36] such as brightness, colorfullness, or luminance is comprehensible for humans.

Combining these options, we use explanations based on the content-based paradigm as presented by Friedrich and Zanker [14], where the attractiveness visual features are used to explain the recommendations in a white-box fashion, Figure 2 (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 2 (b).

Then, our study contains interface conditions depending on how recommendations are displayed: i) no explanations, as shown in Figure 3, ii) black-box explanations based on the top-3 most similar images a user liked in the past, as shown in Figure 4, 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 5. In all three cases the interfaces are vertically scrollable. While Interface 1 (baseline) is able 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.

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 6 summarizes the conditions. The interface conditions were: *Condition 1*: interface 1 without explanations, as in Figure 3; *Condition 2*: using interface 2, each item recommendation is explained based on the top 3 most similar images in the user profile, as in Figure 4; and *Condition 3*: only for AVF algorithm, based on a bar chart of visual features, as in Figure 5, 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 6, 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. [7].

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

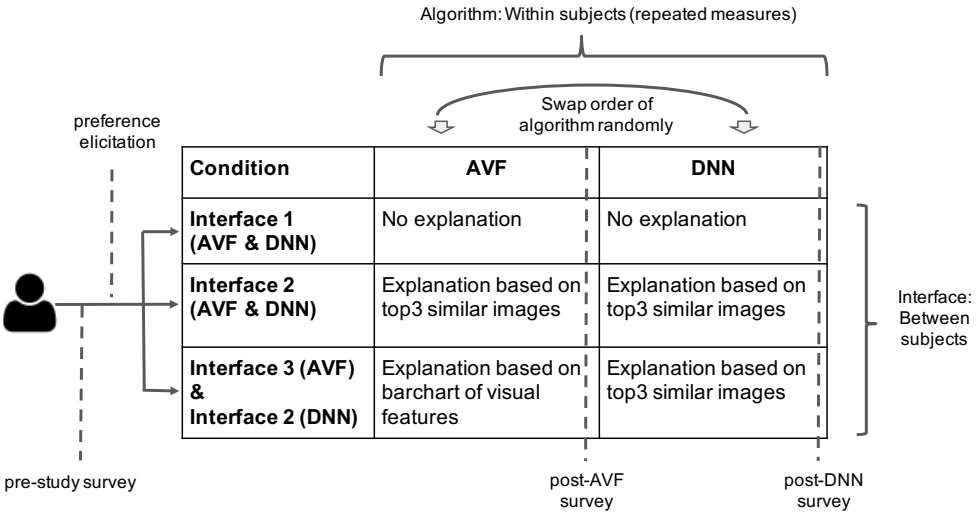

Figure 6: Study procedure. After the pre-study survey and the preference elicitation, users were assigned to one of three possible interfaces. In each interface they evaluated recommendations of two algorithms: DNN and AVF.

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.

Also, we measured the cognitive load perceived by the users during the experiment using the NASA TLX (task load index) workload assessment [18]. This evaluation was conducted in the post-algorithm survey. The results are presented in Table 3 and they were also integrated into the final model in Figure 7.

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.

4 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 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

Table 1: Results of users' perception over several evaluation dimensions, defined in Section 3. 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).

	Explainable		Relevance		Diverse		Interface Satisfaction		Use Again		Trust		Average Rating	
Condition	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF	DNN	AVF
Interface 1 (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
Interface 2 (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
Interface 3 (DNN: Top-3 similar, AVF: 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. sign. between interfaces by multiple t-tests, Bonferroni corr. $\alpha_{bonf} = \alpha/n = 0.05/3 = 0.0017$. Stat. sign. between algorithms using pairwise t-test, $\alpha = 0.05$.

Table 2: Results of the confirmatory factor analysis (CFA), indicating two constructs: *Effort* and *Satisfaction*.

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

Table 3: NASA TLX Results. The symbol $\uparrow^2$ indicates interface-wise significant difference (differences between interfaces using the same algorithms).

	Mental		Hurry		Insecure	
Condition	DNN	AVF	DNN	AVF	DNN	AVF
Interface 1 (No Explanations)	19.90	23.24	10.78	13.41	12.22	12.88
Interface 2 (DNN & AVF: Top3 images)	20.05	18.46	11.54	12.08	7.62	6.59
Interface 3 (DNN: Top3 imag., AVF: chart)	23.41	26.37	14.29	15.73	13.32	16.37$\uparrow^2$

are recommended (70.4), but since the relevance is rather insufficient (56.2), the perception of trust is reported as low (55.4).

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 ($AVF = 58.7$).

5 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. [24].

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. CFA is used to test whether the created factors are consistent with the hypothesized model.

Confirmatory Factor Analysis. We conducted a CFA and examined the validity and reliability scores of the constructs measured in our study. The results are summarized in Table 2. 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

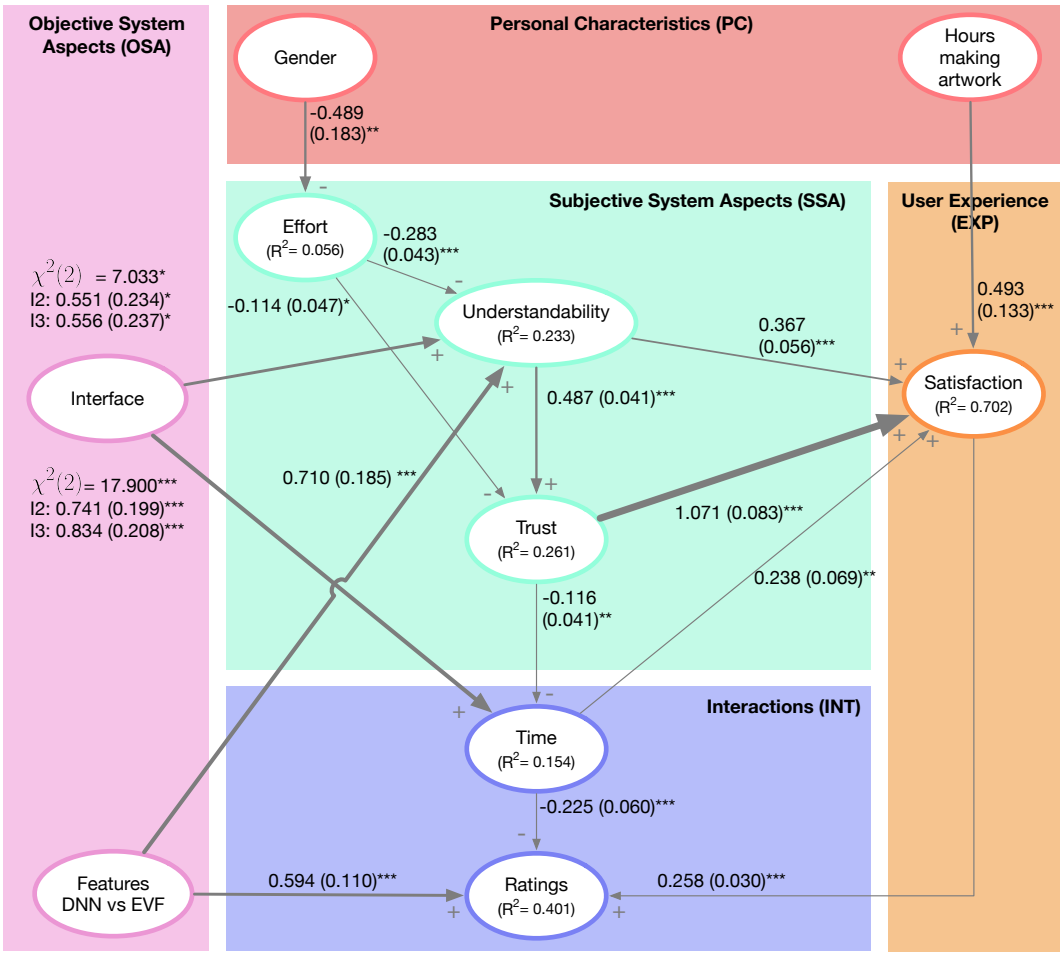

Figure 7: The structural equation model for the data of the experiment using Knijnenburg's evaluation framework for recommender systems. Significance levels: $***p < .001$, $**p < .01$, $*p < 0.05$. R^2 is the proportion of variance explained by the model. Numbers on the arrows (and their thickness) represent the β coefficients (and standard error) of the effect. Factors are scaled to have an SD of 1.

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.

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 factors and other variables. The model has a good [4] fit: $\chi^2(72) = 103.935$, $p = .008$; $RMSEA = 0.043$, $90\%CI : [0.022, 0.060]$, $CFI = 0.997$, $TLI = 0.996$.

[4]A model should not have a non-significant χ^2 (p > .05), but this statistic is often regarded as too sensitive. Hu and Bentler [23] propose cut-off values for other fit indices to be: CFI > .96, TLI > .95, and RMSEA < .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 compared to making content-based recommendations using AVF. 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 then has a positive effect on satisfaction, on its own and mediated by trust. This result is consistent with the model found in [16], 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 and analyze explanations. Gedikli et al, in [16], 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: *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 recommendations.

6 CONCLUSIONS & FUTURE WORK

In this paper, 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 on users. 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 to explain the features learned by a DNN of images.

Regarding RQ2, we see that the algorithm plays an important role in conjunction with the interface. DNN is perceived better than AVF in most dimensions, showing that further research should focus on the interaction between algorithm and explainable interfaces. 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. [24], 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 system 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 [15, 34] and test them using this user-centric recommender evaluation framework.

7 ACKNOWLEDGEMENTS

The authors from PUC Chile were funded by Conicyt, Fondecyt grant 11150783, as well as by the Millennium Institute for Foundational Research on Data (IMFD).

REFERENCES

[1] Xavier Amatriain. 2013. Mining large streams of user data for personalized recommendations. *ACM SIGKDD Explorations Newsletter* 14, 2 (2013), 37–48.

[2] Ivana Andjelkovic, Denis Parra, and John O'Donovan. 2018. Moodplay: Interactive music recommendation based on Artists' mood similarity. *International Journal of Human-Computer Studies* (2018).

[3] LM Aroyo, Y Wang, R Brussee, Peter Gorgels, LW Rutledge, and N Stash. 2007. Personalized museum experience: The Rijksmuseum use case. In *Proceedings of Museums and the Web*.

[4] Idir Benouaret and Dominique Lenne. 2015. Personalizing the Museum Experience through Context-Aware Recommendations. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 743–748.

[5] Bruno Cardoso, Gayane Sedrakyan, Francisco Gutiérrez, Denis Parra, Peter Brusilovsky, and Katrien Verbert. 2019. IntersectionExplorer, a multi-perspective approach for exploring recommendations. *International Journal of Human-Computer Studies* 121 (2019), 73–92.

[6] Oscar Celma. 2010. Music recommendation. In *Music Recommendation and Discovery*. Springer, 43–85.

[7] Anjan Chatterjee, Page Widick, Rebecca Sternschein, William Smith II, and Bianca Bromberger. 2010. The Assessment of Art Attributes. 28 (07 2010), 207–222.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 248–255.

[9] Vicente Dominguez, Pablo Messina, Denis Parra, Domingo Mery, Christoph Trattner, and Alvaro Soto. 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*. DOI: http://dx.doi.org/10.1145/3125486.3125495

[10] Vicente Dominguez, Pablo Messina, Christoph Trattner, and Denis Parra. 2018. Towards Explanations for Visual Recommender Systems of Artistic Images. In *Proceedings of the 5th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*. 69–73.

[11] Ivania Donoso-Guzmán and Denis Parra. 2018. An Interactive Relevance Feedback Interface for Evidence-Based Health Care. In *23rd International Conference on Intelligent User Interfaces*. ACM, 103–114.

[12] David Elsweiler, Christoph Trattner, and Morgan Harvey. 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*. ACM, 575–584.

[13] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John Platt, and others. 2015. From captions to visual concepts and back. (2015).

[14] Gerhard Friedrich and Markus Zanker. 2011. A taxonomy for generating explanations in recommender systems. *AI Magazine* 32, 3 (2011), 90–98.

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2414–2423.

[16] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I

explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367 – 382. DOI: http://dx.doi.org/https://doi.org/10.1016/j.ijhcs.2013.12.007

[17] Carlos A Gomez-Uribe and Neil Hunt. 2016. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)* 6, 4 (2016), 13.

[18] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[19] Chen He, Denis Parra, and Katrien Verbert. 2016b. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications* 56 (2016), 9–27.

[20] Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. 2016a. Vista: A Visually, Socially, and Temporally-aware Model for Artistic Recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, USA, 309–316. DOI: http://dx.doi.org/10.1145/2959100.2959152

[21] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision*. Springer, 3–19.

[22] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. ACM, New York, NY, USA, 241–250. DOI: http://dx.doi.org/10.1145/358916.358995

[23] Litze Hu and Peter M. Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* 6, 1 (1999), 1–55. DOI: http://dx.doi.org/10.1080/10705519909540118

[24] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. *User Modeling and User-Adapted Interaction* (2012), 441–504. DOI: http://dx.doi.org/10.1007/s11257-011-9118-4

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

[26] Pigi Kouki, Shobeir Fakhraei, James Foulds, Magdalini Eirinaki, and Lise Getoor. 2015. Hyper: A flexible and extensible probabilistic framework for hybrid recommender systems. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 99–106.

[27] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2017. User preferences for hybrid explanations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 84–88.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[29] Pattie Maes and others. 1994. Agents that reduce work and information overload. *Commun. ACM* 37, 7 (1994), 30–40.

[30] Sean M McNee, Nishikant Kapoor, and Joseph A Konstan. 2006. Don't look stupid: avoiding pitfalls when recommending research papers. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. ACM, 171–180.

[31] Pablo Messina, Vicente Dominguez, Denis Parra, Christoph Trattner, and Alvaro Soto. 2018. Content-Based Artwork Recommendation: Integrating Painting Metadata with Neural and Manually-Engineered Visual Features. *User Modeling and User-Adapted Interaction* (2018). DOI: http://dx.doi.org/10.1007/s11257-018-9206-9

[32] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé, III. 2012. Midge: Generating Image Descriptions from Computer Vision Detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 747–756. http://dl.acm.org/citation.cfm?id=2380816.2380907

[33] Cataldo Musto, Fedelucio Narducci, Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2016. ExpLOD: A Framework for Explaining Recommendations Based on the Linked Open Data Cloud. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, USA, 151–154. DOI: http://dx.doi.org/10.1145/2959100.2959173

[34] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature Visualization. *Distill* (2017). DOI: http://dx.doi.org/10.23915/distill.00007 https://distill.pub/2017/feature-visualization.

[35] Denis Parra and Peter Brusilovsky. 2015. User-controllable personalization: A case study with SetFusion. *International Journal of Human-Computer Studies* 78 (2015), 43–67.

[36] Jose San Pedro and Stefan Siersdorfer. 2009. Ranking and Classifying Attractiveness of Photos in Folksonomies. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*. ACM, New York, NY, USA, 771–780. DOI: http://dx.doi.org/10.1145/1526709.1526813

[37] Giovanni Semeraro, Pasquale Lops, Marco De Gemmis, Cataldo Musto, and Fedelucio Narducci. 2012. A folksonomy-based recommender system for personalized access to digital artworks. *Journal on Computing and Cultural Heritage (JOCCH)* 5, 3 (2012), 11.

[38] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 806–813.

[39] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender Systems Handbook*. Springer, 353–382.

[40] Chun-Hua Tsai and Peter Brusilovsky. 2018. Explaining Social Recommendations to Casual Users: Design Principles and Opportunities. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion (IUI '18 Companion)*. ACM, New York, NY, USA, Article 59, 2 pages. DOI: http://dx.doi.org/10.1145/3180308.3180368

[41] Egon L van den Broek, Thijs Kok, Theo E Schouten, and Eduard Hoenkamp. 2006. Multimedia for art retrieval (m4art). In *Multimedia Content Analysis, Management, and Retrieval 2006*, Vol. 6073. International Society for Optics and Photonics, 60730Z.

[42] Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. 2013. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 351–362.

[43] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: Explaining Recommendations Using Tags. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09)*. ACM, New York, NY, USA, 47–56. DOI: http://dx.doi.org/10.1145/1502650.1502661

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