Algorithmic and HCI aspects for explaining recommendations of artistic images

VICENTE DOMINGUEZ, Pontificia Universidad Católica de Chile & IMFD, Chile
IVANIA DONOSO-GUZMÁN, Pontificia Universidad Católica de Chile, Chile
PABLO MESSINA, Pontificia Universidad Católica de Chile & IMFD, Chile
DENIS PARRA, Pontificia Universidad Católica de Chile & IMFD, Chile

Pre-Print, final version to be PUBLISHED in ACM TiiS https://doi.org/10.1145/3369396.

Explaining suggestions made by recommendation systems is key to make users trust and accept these systems. This is specially critical in areas such as art image recommendation.

Traditionally, art works are sold in art galleries where people can see them physically, and artists have the chance to persuade the people into buying them. On the other side, online art stores only offer the user the action of navigating through the catalog, but nobody plays the persuading role of the artist. Moreover, few works in recommendation systems 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 several aspects of the user experience with a recommender system of artistic images, from algorithmic and HCI perspectives. We conducted two user studies in Amazon Mechanical Turk to evaluate different levels of explainability, combined with different algorithms. While in study 1 we focus only on a desktop interface, in study 2 we attempt to understand the effect of explanations in mobile devices.

In general, our experiments confirm that explanations of recommendations in the image domain are useful and increase user satisfaction, perception of explainability and relevance. In the first study, our results show that the observed effects are dependent on the underlying recommendation algorithm used. In the second study, our results show that these effects are also dependent of the device used in the study, but with a smaller effect.

Finally, using the framework by Knijnenburg et al., we provide a comprehensive model, for each study, which synthesizes the effects between different variables involved in the user experience with explainable visual recommender systems of artistic images.

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

Additional Key Words and Phrases: Visual Recommender Systems, Explainable AI, Art

ACM Reference Format:
Vicente Dominguez, Ivania Donoso-Guzmán, Pablo Messina, and Denis Parra. 2020. Algorithmic and HCI aspects for explaining recommendations of artistic images . 1, 1 (July 2020), 32 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors' addresses: Vicente Dominguez, Pontificia Universidad Católica de Chile & IMFD, Santiago, Chile, vidominguez@uc.cl; Ivania Donoso-Guzmán, Pontificia Universidad Católica de Chile, Santiago, Chile, indonoso@uc.cl; Pablo Messina, Pontificia Universidad Católica de Chile & IMFD, Santiago, Chile, pamessina@uc.cl; Denis Parra, Pontificia Universidad Católica de Chile & IMFD, Santiago, Chile, dparra@ing.puc.cl.

XXXX-XXXX/2020/7-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Online artwork recommendation has received little attention compared to other areas such as movies [2, 21] and music [9, 41]. Most research on artwork recommendation deals with studies on museum data [4, 6, 56, 62], but there is little work with datasets of online artwork e-commerce systems [26, 45]. Traditionally artwork sales happened in art galleries where people can see different physical artworks, and artists had the chance to encourage people to buying their work. In the last decade, online artwork sales have boomed with the influence of social media and new consumption behavior by millennials, at the current growth rate, they are expected to reach $9.58 billion by 2020[1]. With this type of e-commerce the user can navigate through the catalog, but nobody plays the key role of the artist: encourage people buying their artworks.

Studying the recommendation of images is very different from recommending music or movies, making it an attractive area for research. In order to validate whether a user likes a song or movie we need to make sure they watched the movie or listened to the music [13]. In the case of visual art, while there might be contextual factors involved, just by looking at a painting or photo, a user can tell her positive or negative preference. It is also different from a user deciding if she likes fashion simply from looking at a picture of the clothes; as size weight and fitting come into play later. Nevertheless, fashion is certainly the topic dominating visual recommendation systems [27, 43].

The first works in the area of artwork recommendation date from 2006-2007 such as the CHIP [4] 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. [62], 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 [16, 26], which is rather expected considering that DNNs are the top performing models for obtaining visual features for several tasks, such as image classification [38], and scene identification [57]. In our recent work, Messina et al. [45] 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 expert and non-expert art users. This aspect is important since past works have shown that off-line results might not always replicate when tested with actual users [37, 44], and also domain knowledge is an important variable to explain the user experience with a recommender system [3, 36, 49].

The aforementioned works miss one important aspect of the user experience with recommender systems: how explainability combined with the recommendation algorithm affect the user experience with a visual art recommender system. There is one influential work in this area by Cramer et al. [13], which studied the effect of transparency in the explanation of a content-based art recommender. There are two implications of this work: (i) explaining to the user why a recommendation was made increased acceptance of the recommendations, but (ii) trust in the system itself was not improved by increased transparency in the explanation. These results are indeed important, but the study had two limitations: (1) they tested only a single recommendation algorithm, and without a baseline we cannot tell if the algorithm accuracy had an influence on the results, and (2) the transparent explanation was always optional, since the users had to click a button "Why" in order to receive the explanations. Another aspect missed is how the type of device used can influence the user experience with a visual art recommender system. The study by Han et al. [23] for instance, showed that it is possible to capture user preferences in a recommender system using a model of actions available only on mobile multitouch devices (such as double-pinch or zooming with two

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

fingers). Moreover, Amatriain [2] shows important differences in the Netflix interface between desktop and mobile versions, since the type of users' preference signal can be of different type.

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 [48].

In contrast, features of visual attractiveness, despite being less accurate to predict user preference [45], could be easily explained, based on color, brightness or contrast [55]. Explanations in recommender systems have shown to have a significant effect on user satisfaction [61], but there are more variables that affect the user experience, not only the explanation or the type of explanation. We believe that variables like the device of the study and the option of explanation, also affect the user experience in a significant way.

To the best of our knowledge there are no studies on the effect combined of the user device and the optino of explanation. Hence, explaining recommendations by a Visual Content-based Recommender (VCBR) [15] never has been tested in different types of devices in a single study. There is also no research proposing a model that fully combines all the variables involved in the user experience, using different types of devices, to explain several dimensions of the user experience with a VCBR such as perception of relevance, diversity, satisfaction and trust.

Objective. In this paper, we study the effect of explaining artistic image suggestions. In particular, we conduct two user studies on Amazon Mechanical Turk. The first one was under three different interfaces and two different algorithms. The three interfaces were: i) no explanations, ii) explanations based on similar images, and iii) explanations based on visual features. The two recommendation algorithms were: Deep Neural Networks (DNN) and Attractiveness Visual Features (AVF). The first algorithm represents items with latent features (a vector of real numbers) from a pre-trained neural network so it can be considered accurate but a non-transparent algorithm, while the second method uses explicit visual features such as image brightness, sharpness, etc. Therefore, we consider it less accurate, but a more transparent algorithm. In the second study we used one algorithm (the best from study 1) and two interfaces, designed for two types of devices, desktop and mobile. In each interface there were two conditions of explanation: optional and mandatory. The algorithm in this second study was based only on visual features from deep neural networks. In both studies, we used images provided by the online store *UGallery* (http://www.UGallery.com/). Finally, we contribute with two Structural Equation Models based on the framework by Knijnenburg et al [36] in order to fully explain the user experience with an explainable VBCR of artistic images.

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

- **RQ1**. Given a recommender interface with different levels of explanations (none, similarity-based, feature-based), which level is perceived as most useful?
- **RQ2**. Considering the visual recommender algorithm chosen, one accurate, but non-transparent and another less accurate, but more transparent, are there observable differences in how the levels of explainability in the interfaces are perceived?
- **RQ3**. How do independent variables such as algorithm, explainable interface and domain knowledge interact in order to explain the user experience (UX) with the recommender system in terms of several measures?
- **RQ4**. When using one algorithm with high accuracy, are there significant differences between types of devices used (mobile vs. desktop)?
- **RQ5** Does it have an effect on UX if the recommendation explanations are left optional rather than always visible?

Using Amazon's Mechanical Turk. One important decision we made in order to collect user feedback for our experiments was by using Amazon's Mechanical Turk (AMT). While some authors raise important risks on conducting studies in AMT which can hinder ecological validity [24], other studies show that AMT can be a safe source to conduct user studies [5], while others reach similar conclusions, but under certain constraints to prevent misuse [32]. Several studies have used AMT in the past for validating hypotheses with respect to recommendation systems [3, 7, 35]. Based on these previous attempts we followed certain criteria to decrease the risk of making our evaluation invalid. Among them, using questions to check for users' attention in several steps of the study, as well as making sure that the answers had sufficient intra-user variability in pre- and post-study surveys.

Outline. Our work is structured as follows: In Section 2 we survey relevant related work and explain how our research differs from previous work in the area. The following three sections are about the study 1. In the first section, we describe the methods and materials. We introduce the explainable interface recommendation approach, the algorithms, and discuss the study procedure to evaluate them. In the second section we present the results of the study. In the last of these three sections, we present and discuss the study SEM, which connects all the studied variables. The following three sections have the same structure of the three mentioned before, but are about the study 2. To finish, section 11 concludes the paper and provides an outlook for future work.

2 RELATED WORK

In this section we present the previous work that motivates our research. Since the areas described might have been explored in different research fields, we focus mainly on summarizing the work directly related to personalization and recommender systems. Hence, we classify the most relevant related work in three areas: a) Artwork Recommender Systems, b) Explainability and transparency, c) User-centric evaluation of recommender systems and d) Multi-touch devices.

2.1 Artwork Recommender Systems

The works of Aroyo et al. [4] with the CHIP project and Semeraro et al. [56] 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. [6]. They used context information from 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 [26], and the work of Dominguez et al. [16] and Messina et al. [45] compared neural against traditional visual features. None of the aforementioned works performed a user study under explanation interfaces to generalize their results.

2.2 Explainability and transparency in Recommender Systems

Herlocker et al. [28] introduced the idea of explaining recommendations as a way of making 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, [59] 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 [60].

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 [29] 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 an explanation were affected most by the transparency of the system. The role of transparency in recommender systems is discussed in [58], showing that ”people feel more confident with recommendations that they perceive more transparent.”

Cramer et al in their work [13] studied topics very similar to those addressed in our own work. They analyzed the effect of transparency and trust of an artwork recommender system. The difference with respect to 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.

A recent work in the area of explainable recommendations is the work of Zhang et al. [69]. This work is a survey of the recent works in the topic of explainable recommendations. They provide examples of many types of explanation, but nothing about the art domain. Also, the topic of explainable deep learning for recommendations has a short review and it is a sub-section of the chapter ”Open Directions and New Perspectives”. So the problems addressed in our paper are something that is still open.

Scrutable models is another topic related to explainability and trust. There are several works on this topic, one of the first papers in this area is the work by Cook and Kay [11]. They provided the users the possibility of accessing the system's model that represent the user. This paper showed that the participants of their study were very interested in knowing more about their users models and how it works. One more recent work is the paper by Wasinger et al. [65] where they present a scrutable user model in a mobile application of a personalized restaurant menu. Their findings show that most of the users find the application more useful than the traditional menu because of the scrutability of the software. Since the algorithms in our paper cannot be configured by the user, we rely on using explanation to do our algorithms somehow scrutable.

In terms of applications, Zhao et al. [70] presented Pharos - a content-centric system able to recommend items, people and communities. They try to tackle the cold-start problem and also explain the recommendations by visualizing a social map with terms organized in latent communities. A within-subjects study with ten users shows that Pharos helped the subjects to complete exploratory tasks faster and better than BlogCentral, an existent tool. Although user knowledge and tasks were considered and they did not have significant differences between Pharos and BlogCentral, the small amount of subjects calls for a larger user study to generalize these results. Another drawback of this system is the lack of personalization. The social map displayed the same communities and terms to every user, and users' feedback suggests adding this feature in a future version. Zhang et al. [68] went beyond textual explanation by presenting a visual interface for a critiquing-based RS. In an e-commerce system, they present various critiques by a set of meaningful icons, and their results show how the visual presentation and the aided interaction improve the shopping experience of the users. The difference with these works is that they do not work with visual art recommendations.

In [39] there is a discussion on how intelligent systems should explain themselves. To investigate, they consider two aspects of an explanation: soundness (how truthful the explanation is) and completeness (the extent to which the explanation describes the underlying system). The two main findings were that completeness is more important than soundness and that oversimplification can lead to users trusting the system less. More recently Cai et al [8] evaluated normative versus comparative explanations to explain the results of a deep neural net sketch-recognition algorithm. In our work we tested comparative explanations by saying "this painting is X % similar than this one"

A recent survey [1] indicates that the current XAI research is not well integrated with knowledge from visualization, cognitive psychology, perception and human-computer interaction (HCI). This is also argued in [46], where the researcher states that current XAI research relies on the researcher's intuitions of what a "good" explanation is instead of relying on frameworks from social sciences. Wang et al. [64] proposed a theory driven framework for designing XAI models. They propose this framework for classification and clustering applications, that in general are not personalized like recommender systems. Moreover, the current works in XAI usually do not deploy or evaluate explanations with interactive applications or by conducting studies with real users [1], although there are exceptions [8, 39, 68, 70].

2.3 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. [71] studied the effect of diversification in lists of recommended items, Tintarev and Masthoff [59] investigated on recommender systems' transparency, Cramer et al. [13] studied explainability in recommender systems, and Knijnenburg et al. [35] 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. [51] 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. [36] 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 a holistic view of the results involved in the user study.

2.4 Mobile devices

Recent research shows that the multi-touch interaction (MTI) of smartphones and tablets can help predict user preference beyond the actions available in the desktop environment. In this line of research, Guo et al. [22] developed a model that improved the ranking of search results by using interactions available on mobile multitouch devices compared to a the usual click-log interactions on a desktop environment. Similarly, Han et al. [23] were able to improve the prediction of users' favorite items on non-ranked lists of papers by considering multitouch interactions such as dragging direction, speed and position. These studies found that the same interaction can represent different user actions in mobile and desktop devices. These results encourage us to test whether providing explanations in different formats (in text, lists, graphs, etc.) and in different devices (desktop and multi-touch) will show important differences attributed to the interactions available under each technology.

3 STUDY 1: ALGORITHMS AND EXPLANATION STYLES

In the following subsections we describe in detail our study methods. We first 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. To finish, we explain the user study procedure.

3.1 Study 1: Hypotheses

Based on analyses of the results from previous literature, we set the following hypotheses for study 1:

- Hypothesis 1: A recommendation interface with explanations will be better perceived than an interface without explanations. Among the interfaces with explanation, the most transparent will be perceived as most useful.
- Hypothesis 2: The visual content-based recommender algorithm chosen (accurate, not transparent vs. transparent, but not accurate), will have an interaction on how the explainable recommmendation interfaces are perceived.
- Hypothesis 3: Several variables beyond algorithm and the recommendation interface explain the user experience (UX) with the recommender system, among them, the user experience with the domain.

3.2 Study 1: Design

This user study has a mixed design. The within-subjects variable is an algorithm, for which we consider two types based on the mechanism to extract visual features from the images: Deep Neural Networks (DNNs) and Attractiveness visual features (AVF). The between subjects variable corresponds to the type of explainable interface: (i) non-explainable, (ii) explainable, but not transparent, and (iii) explainable and transparent. The main difference between conditions (ii) and (iii) lies in the characteristics of the visual features obtained with the recommendation methods. While DNN outputs accurate visual features, they are not interpretable by humans; they are just latent vectors with numbers. On the other side, AVF outputs visual features interpretable by humans (contrast, brightness, naturalness, etc.), but it is less accurate than DNN in predicting visual user preferences.

Although the main contribution of this study is related to the effects of explainability on visual recommendation systems, we decided to set algorithm type as the within-subject factor due to results from our own previous studies. In [16, 45] we studied different features for visual art recommendations, which were evaluated off-line with a traditional train/validation/test protocol, and we found that features from DNNs resulted in better user preference rankings than AVF. However, in the latest paper by Messina we also conducted a small study with 8 expert curators, and although DNN was better than AVF, the difference in some metrics was not significant, such as **precision@10**. For this reason, we wanted to make sure that in this study we had enough statistical power to identify a difference in algorithms (DNN vs. AVF), as well as interactions with different types of explainable interfaces upon different variables measured in user-centric studies. For this reason, we chose algorithm as within-subjects factor, to confirm in this study what was implied in our previous work in Messina et al. (2018).

3.3 Apparatus & 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 [66]. They support emergent artists by helping them sell their artwork online. For our research, UGallery provided us with an anonymized dataset

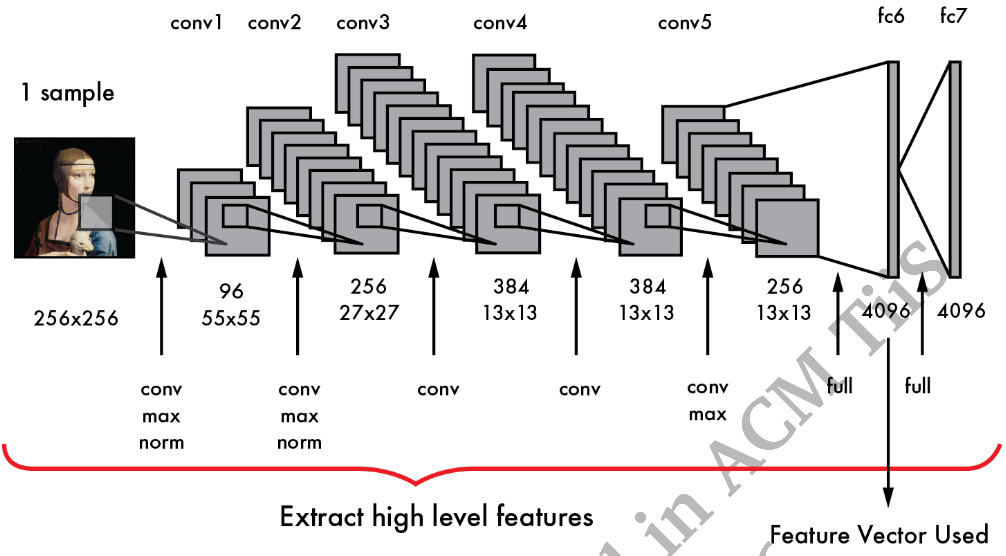


Fig. 1. Model architecture of the AlexNet Convolutional Deep Neural Network used to extract visual features from images.

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.

3.4 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.

3.4.1 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 [38], which architecture is shown in Figure 1. In particular, we use an AlexNet model pre-trained with the ImageNet dataset [14]. Using the pre-trained weights, for every image a vector of 4,096 dimensions was generated with the Caffe framework [31]. We resized every image to a 227x227 image. This is the standard pre-processing needed to use the AlexNet.

3.4.2 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 in [55] 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 feature vector that represents, using the OpenCV software

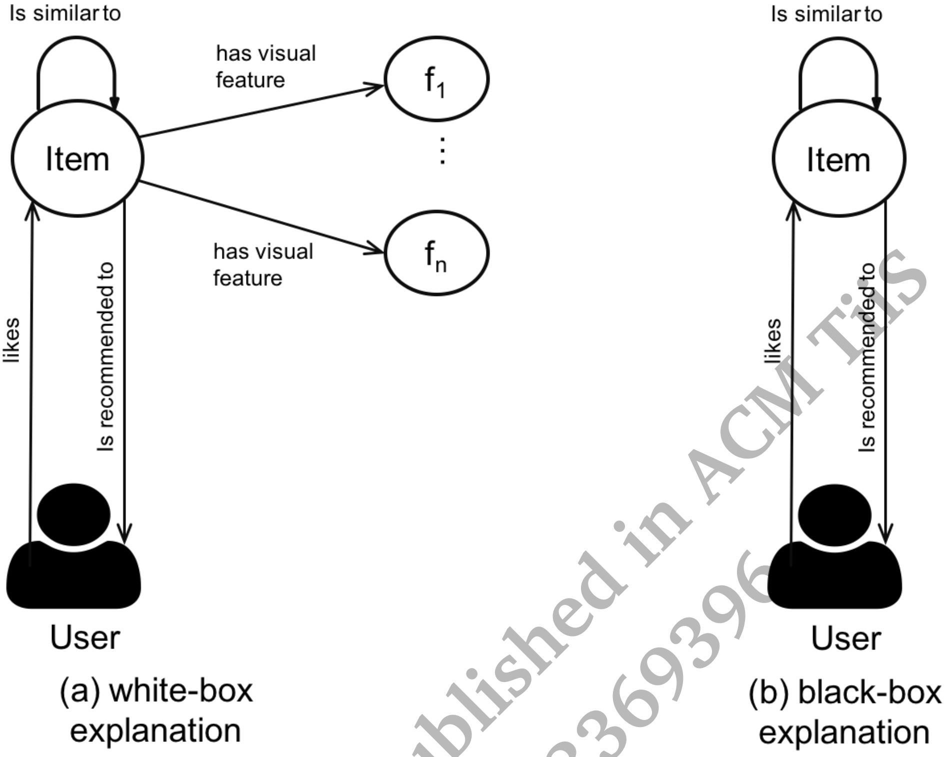


Fig. 2. Design choices for explainable recommender interfaces, based on Friedrich and Zanker [18]. In (a) we explain the recommendation based on transparent visual features, while in (b) we explain based on item similarity, without details of the features used.

library[2]: brightness, saturation, sharpness, colorfulness, naturalness, entropy, and RGB-contrast. The following offers a more detailed description of these features:

- *Brightness*: 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*: 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*: Measures how detailed the image is.
- *Colorfulness*: Measures how distant the colors are from the gray color.
- *Naturalness*: Measures how natural the picture is, grouping the pixels in Sky, Grass and Skins pixels and applying the formula in [55].
- *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 [17], 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

[2] http://opencv.org/

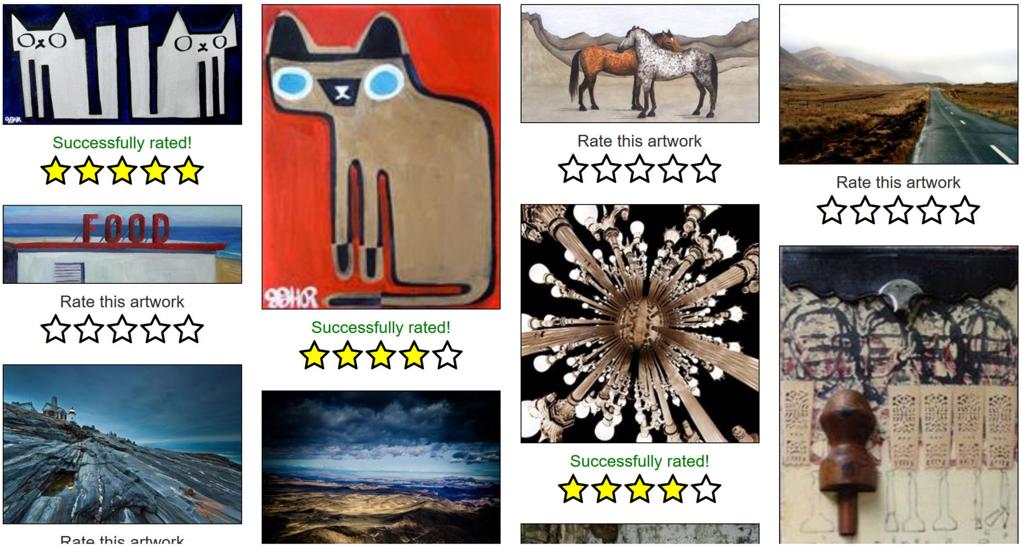


Fig. 3. Interface 1: Baseline recommendation interface without explanations.

how to calculate these visual features can be found in the articles of San Pedro and Siersdorfer [55], as well as in Messina et al. [45].

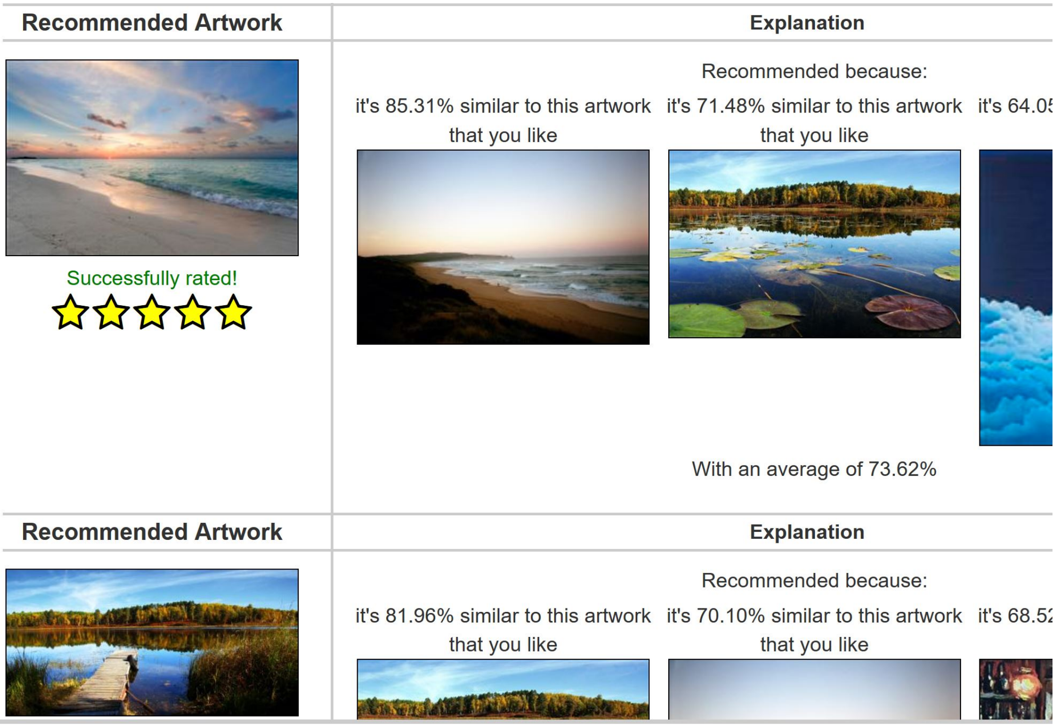
3.4.3 Computing Recommendations. Given a user u who has consumed a set of artworks P_u, a constrained profile size K, and an arbitrary artwork i from the inventory, the score of this item i to be recommended to u is:

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

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. [45], 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:

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

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 [38], and in the case of AVF, the features were extracted based on San Pedro et al. [55].



Recommended Artwork

Explanation

Successfully rated!

Recommended because:

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

With an average of 73.62%

Recommended Artwork

Explanation

Recommended because:

it's 81.96% similar to this artwork that you like

it's 70.10% similar to this artwork that you like

it's 68.52

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

3.5 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 [18]. 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 an explanation is directly obtained from the item and visual features of images. Then, our alternatives for designing explainable interfaces in this study 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 [38], while the representation obtained with attractiveness visual features [55] 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 [18], 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

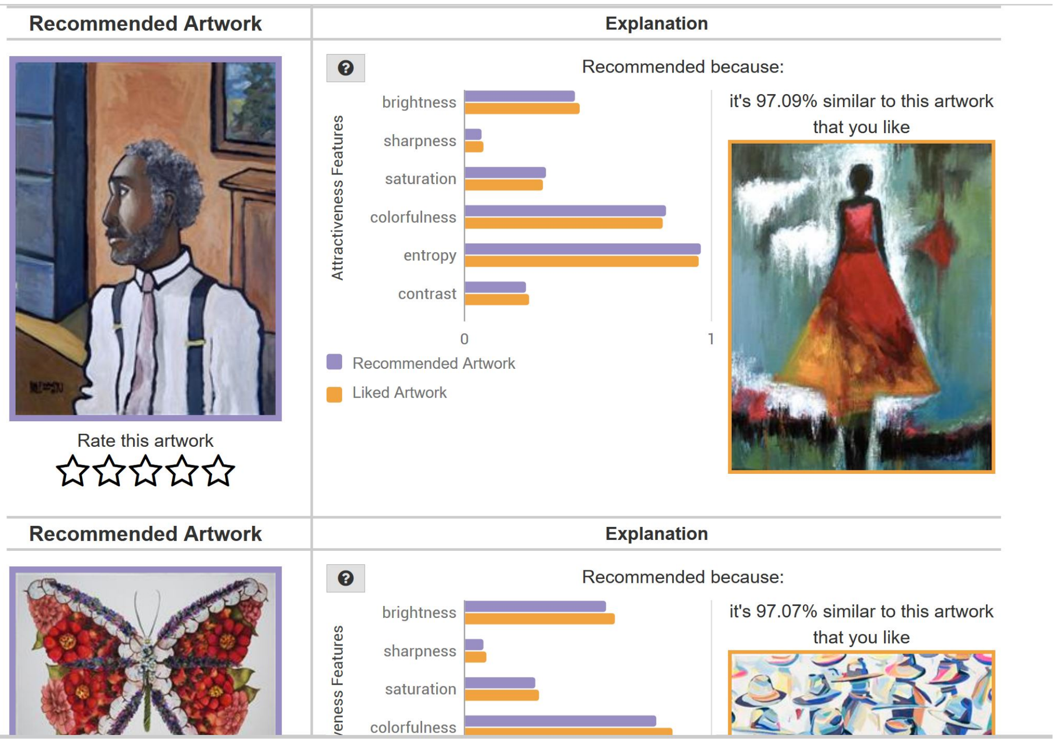


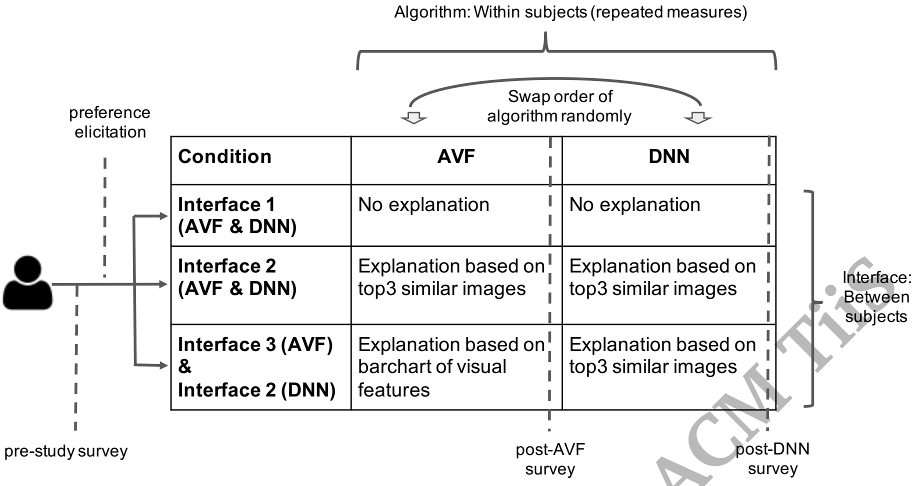
Fig. 5. Interface 3: Explainable and transparent recommendation interface with features' bar chart and top-1 similar image.

black-box fashion, just by indicating which similar items in the user preference list produced the recommendation, as in Figure 2 (b).

Our study then 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.

3.6 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 the 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.


Algorithm: Within subjects (repeated measures)
Swap order of algorithm randomly
preference elicitation
pre-study survey
Interface: Between subjects
post-AVF survey
post-DNN survey

Condition	AVF	DNN
Interface 1 (AVF & DNN)	No explanation	No explanation
Interface 2 (AVF & DNN)	Explanation based on top3 similar images	Explanation based on top3 similar images
Interface 3 (AVF) & Interface 2 (DNN)	Explanation based on barchart of visual features	Explanation based on top3 similar images

Fig. 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.

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.

With respect to the complete study workflow, as shown in Figure 6, participants accepted the study on Mechanical Turk (https://www.mturk.com) and were redirected to a web application. After accepting a consent form, they were 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. [10].

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 a 1-5 scale to each image) to top ten recommendations of images employing either the DNN or the AVF algorithm (also assigned at random as previously mentioned). Finally, the participants were asked to answer a post-algorithm survey (1). 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. These questions are adapted from dimensions presented in the ResQue framework [51] and from previous user studies on recommendation systems [34, 49].

We also measured the cognitive load perceived by the users during the experiment using the NASA TLX (task load index) workload assessment [25]. This evaluation was conducted in the post-algorithm survey.

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.

3.6.1 Checking valid subjects. Since we ran this study on AMT, we had to ensure that our users were valid subjects, i.e., they were seriously paying attention to the study. With this goal, we set two validation questions: one in each study survey.

Table 1. Post study survey questions. The last three questions correspond to NASA TLX. We also asked the open question *If explanations did not make sense to you, explain why?* and asked on any comments they had on the study.

Representative name	Question
Understand	I understood why the art images were recommended to me.
Relevance	The art images recommended matched my interests.
Diversity	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.
Recommend	I would suggest that my colleagues use this recommender system when they want to find art images in the future.
Trust	I trusted the recommendations made.
Explainable	The explanations over the items recommended made sense to me.
Mental demand	How mentally demanding was the task?
Rush	How hurried or rushed was the pace of the task?
Insecure	How insecure, discouraged, irritated, stressed, and annoyed were you?

- **Pre-Study Survey:** On average, North Pole is warmer than the Sahara desert. Answers were *Yes* or *No*, to be chosen with radio buttons.
- **Post-Algorithm Survey:** On average, an elephant is heavier than a mouse. Answer was a number between 0 (no agreement) to 100 (full agreement), chosen with a traditional slider control widget.

To decide whether a user was paying attention, we set two conditions: (a) The user answered Yes to the first statement, and (b) The user answered with a number between 95 and 100 to the second statement. If any of the two conditions previously defined was not met, the user was classified as an invalid subject and her records were discarded from the data analysis and final results reported in this article.

4 RESULTS STUDY 1

The study consisted of 200 users out of which 121 were able to answer our validation questions successfully and hence were included in the results. 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.

Demographics. Our subjects were between 18 to over 60 years old. Thirty-six percent were between 25 to 32 years old, and 29% between 32 to 40 years old. Females made up 55.4%. Twelve percent just finished high school, 31% had finished 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 of 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. Just 20% of our subjects 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 to 6 months.

5 A MODEL OF THE UX WITH AN ART RECOMMENDER

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. [36]. 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.

5.1 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 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.

Table 2. Results of the confirmatory factor analysis (CFA), indicating two constructs: *Effort* and *Satisfaction*. Questions in gray are the ones that were removed from the factor.

Construct	**Item**	**Loading**
Effort	How mentally demanding was the task?	0.750
$\alpha = 0.865$ $AVE = 0.6883$	How insecure, discouraged, irritated, stressed, and annoyed were you?	0.826
	How hurried or rushed was the pace of the task?	0.906
Satisfaction $\alpha = 0.955$	I would use this recommender system again for finding art images in the future	0.973
$AVE = 0.880$	Overall, I am satisfied with the recommender interface.	0.875
	I would suggest that my colleagues use this recommender systemwhen they want to find art imagesin the future.	0.963
	The art images recommended matched my interests.	
	I trusted the recommendations made.	

5.2 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 factors and other variables. The model has a good [3] fit: $\chi^2(72) = 103.935$, $p = .008$; $RMSEA = 0.043$, $90\%CI : [0.022, 0.060]$,

[3] A model should not have a non-significant χ^2 (p > .05), but this statistic is often regarded as too sensitive. Hu and Bentler [30] 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.

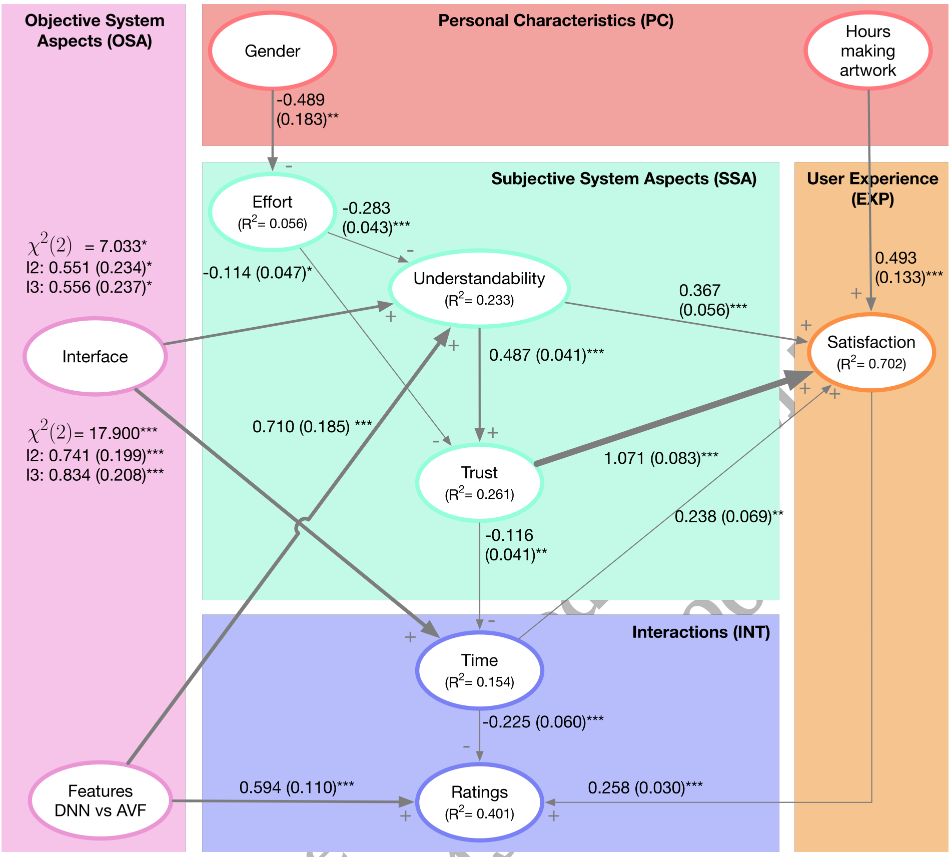


Fig. 7. The structural equation model for the data of the experiment using Knijnenburg's evaluation framework for recommender systems. Significance levels: $***p < .001$, $**p < .01$, $*p < 0.05$. R^2 is the proportion of variance explained by the model. Numbers on the arrows (and their thickness) represent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

$CFI = 0.997$, $TLI = 0.996$. We did not find any significant interaction effects between the the experimental manipulation.

To rule out a potential learning effect of users in the within-subjects condition (algorithm), in the analysis we also added a variable that represents whether the experiment is the first or second session for the same user. This variable was found not significant in all of the regressions analyzed, thus providing evidence of no asymmetric transfer effects between consecutive sessions.

Effect of algorithm: the algorithm used to create the features has a positive effect on understandability. When using DNN features users tend to understand better as compared to making content-based recommendations using AVF. 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 [20], 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 [20], also found 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 user satisfaction strongly depends on how much users 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 exert too much effort when interacting with the system, they also perceive less understanding of the recommendations.

5.3 Discussion study 1

With respect to Hypothesis 1 we found that users were more satisfied with interfaces that provided explanations. These kind of interfaces increased understandability, which increased trust and satisfaction with the system. It also increased the time users spent on the system, which increased satisfaction. These results show that the hypothesis was right. The user felt more comfortable with a system that is more transparent. Just by adding a simple explanation, made the users increase their satisfaction. The users spent more time in the system, probably because they enjoyed the experience. All of this made us rethink the importance of the user centric desing, and its role in the design of entire systems. If you gain the trust of the user, you gain her atention, and the systems need the user atention and participation to live.

With Hypothesis 2 we found that a more accurate algorithm, even if it is not transparent, increased understandability, which increased satisfaction. A better algorithm also increased the average ranking users provided to the images. There were no significant differences in how users perceived transparency, but they were more satisfied with the overall system when they were exposed to DNN. This showed us the importance the accuracy of an algorithm in a system. Independent of the transparency of the system, if you are not using an algorithm that has a decent performance, the understability and satisfaction of the user is affected. The design of the interface of the system is important, but it cannot be isolated from the algorithm. This is a core factor in terms of the user experience, even if the user does not know anything on how it works.

Finally, w.r.t. Hypothesis 3, we found that the user experience is affected and explained by more variables than only the interface and algorithm. Time, Trust, Understanability and Effort were affecting the user experience in conjunction with the Interface and the Algorithm. Also, the more users spend time making art, the more their satisfaction with the system.

6 STUDY 2: DEVICE TYPE AND EXPLANATION OPTIONALITY

6.1 Study 2: Hypotheses

By analyzing the results from previous literature, we set the following hypotheses for study 2:

- Hypothesis 4: If we test the same algorithm in different devices, there will be significant differences in the user perception.
- Hypothesis 5: If the user can choose whether or not to receive a recommendation's explanation; this will have a positive effect in the user experience.

6.2 Study 2: Design

The study design is a between subjects with two variables: *device* (mobile, Desktop) and *explanation obligatoriness* (Optional, Mandatory)

In this study we used many materials and methods from the first study. We briefly describe the elements that were reused from study 1 and we provide more details on the new aspects used in study 2.

6.3 Materials

In this study we also used the dataset provided by Ugallery, described in 3.3.

6.4 Visual Recommendation Approach

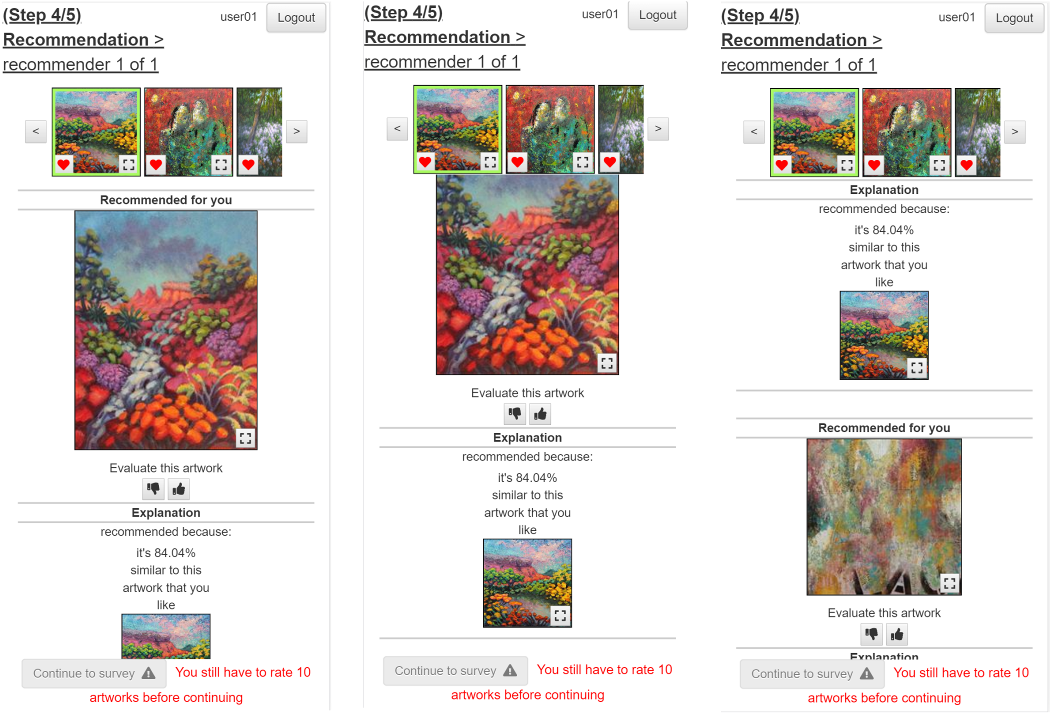
In this second study we were only concerned with studying the effect of explanations, so we kept a single algorithm throughout all the user study. Concretely, we implemented a hybrid content-collaborative recommendation algorithm inspired by both Youtube's recommender system [12] and VBPR [27]. Like the traditional Matrix Factorization approach, we model user-item preferences $\hat{x}_{u,i}$ as the dot product between a user vector and an item vector, i.e. $\hat{x}_{u,i} = \vec{u} \cdot \vec{i}$. However, instead of learning specific latent vectors for each user and item, we trained a single deep neural network with a fixed number of parameters capable of embedding any user and item into a latent vector space based on visual content. For items, we use ResNet50 to extract an $\mathbb{R}^{2048}$ vector that goes through 2 additional layers with SELU[4] activations to produce a final $\mathbb{R}^{128}$ item vector $\vec{i}$. For users, inspired by Covington et al. [12] we compute the latent vector of each item in the user's profile, then compute their average and finally apply 3 additional layers with SELU activations to generate a final $\mathbb{R}^{128}$ user vector $\vec{u}$. This design allowed us to quickly generate user vectors in real-time based on the artworks liked during the preference elicitation stage of the study, without having to train a new set of variables for each new user that enters the system. To train this network, we followed the approach of VBPR [27], namely, training the network for ranking with the BPR pairwise optimization framework [52], using the purchase transactions provided by *UGallery* as ground truth.

6.4.1 Computing Recommendations. Recommendations are generated in two stages: a filtering stage followed by a re-ranking stage. In the filtering stage we use our proposed neural network to compute a user vector from the items the user liked during the preference elicitation. The dot product between this user vector and each candidate item vector gives us the user's preference for each candidate item. We keep the top 100 candidate items from this stage and ignore the rest. Next, in the re-ranking stage we re-rank these 100 images so as to generate a top-10 recommendation which is both visually similar and diverse as follows: for each item we retrieve a $\mathbb{R}^{200}$ vector[5]. Then, for each candidate item we find 3 things: 1) the user's liked item most similar to it, 2) the similarity score with this liked item and 3) the ID of the visual cluster of this liked item[6]. We compute these similarities using cosine similarity. We then group candidate items by cluster ID (obtained from the most similar liked item in each case as explained before) and rank candidate items within each group by their similarity score. Finally, we proceed to fill a top-10 recommendation by rounds. In each round, we sort the groups based on the similarity score of the item at the top of the group. For each group we then pick the top item, add it to the recommendation list and remove it from the group. We keep repeating rounds like this until the top-10 recommendation is complete. The rationale

[4] SELU stands for Scaled Exponential Linear Unit, a new activation function with self-normalizing properties [33]

[5] We used ResNet50 pre-trained on ImageNet to extract $\mathbb{R}^{2048}$ vectors for all images. We then applied PCA over a set of 13,297 images crawled from *Ugallery* to reduce their dimensionality to $\mathbb{R}^{200}$. All this is precomputed beforehand.

[6] We used K-Means clustering to fit 100 clusters to the $\mathbb{R}^{200}$ item vectors obtained with ResNet50 + PCA.



recommended because:
it's 84.04% similar to this artwork that you like

You still have to rate 10 artworks before continuing

recommended because:
it's 84.04% similar to this artwork that you like

You still have to rate 10 artworks before continuing

recommended because:
it's 84.04% similar to this artwork that you like

You still have to rate 10 artworks before continuing

Fig. 8. Mobile interface with mandatory explanations. In the top of the image you can see the user profile. The highlighted image image in the user profile is the explanation of the current recommendation.

behind this round-based re-ranking stage is that it generates a more diverse recommendation, more evenly distributed among the visual clusters the user showed preference for.

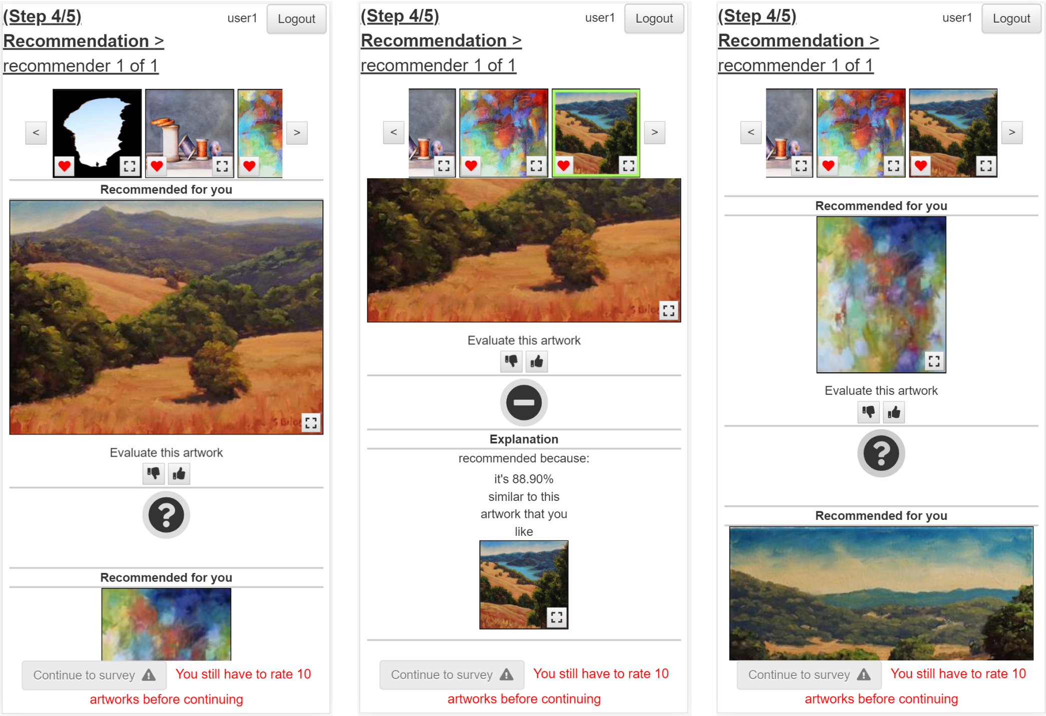
6.5 The Explainable Recommender Interfaces

In this study we studied the effect of different devices and the control of the user over the explanation of recommendations. Since in this study we used a DNN based algorithm, all the explanations were in a black-box fashion, showing the most similar image of the user's profile. The similarity was calculated using cosine similarity between the two image vectors.

6.6 User Study Procedure

To evaluate the effect of the explanation of an algorithm in different device interfaces we conducted a user study on Amazon Mechanical Turk. We defined a user study with 2x2 between-subjects design. The variables were *device*, with possible values *Mobile* and *Desktop*, and *explanation obligatoriness*, with values *Mandatory* and *Optional*.

The full study procedure is shown in Figure 11. Participants accepted the study in one of the two HITs on Mechanical Turk (https://www.mturk.com) and were redirected to a web application. If a user accepted both HITs, we used just the data of the first HIT the user acomplished, we paid both HITs to the user anyway. 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. [10].



recommended because:
it's 88.90% similar to this artwork that you like

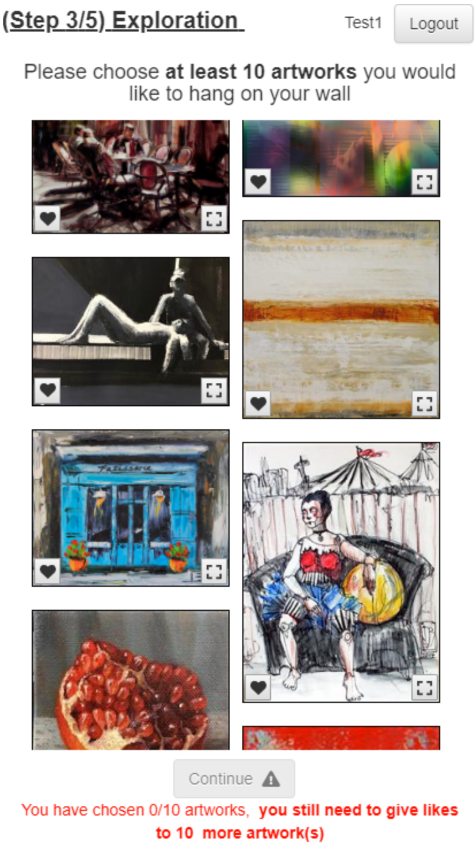
You still have to rate 10 artworks before continuing

Fig. 9. Mobile interface with optional explanations. The difference with the mandatory explanation interface is the question mark button. If the user touches it, the explanation of the recommendation is displayed. If the user touches the score mark button, the explanation is hid.

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 Explanation condition. In each condition, they again provided feedback (giving a "like" or "dislike") to top ten recommendations of images by employing our Youtube-like algorithm. Finally, the participants were asked next to answer a post-algorithm survey, whose questions are shown in Table 4. Similar to study 1, these questions are adapted from dimensions presented in the ResQue framework [51] and from previous user studies on recommendation systems [34, 49]. Once the participants finished answering the post study survey, they were redirected to the final view, where they received a survey code for later payment in Amazon Mechanical Turk.

7 RESULTS STUDY 2

The study entailed a total of 202 users out of which 177 were able to answer our validation questions successfully and hence were included in the results. We used the same questions and conditions of the study 1 to validate our users. Filtering out users not responding properly to these questions allowed us to include 89 users for the Desktop interface, where 45 users got Optional Explanations and 44 got Mandatory Explanations. For the Mobile interface we got 88 users, 44 in each explanation condition. In total, participants were paid an amount of $2.00 USD per study, which took them around 10 minutes to complete. We increased the payment in this second study compared to study 1. The reason is that in study 1 we just considered the reference of previous studies to calculate our budget [42, 54], while in this second study we considered AMT guidelines about wages as well.



(Step 3/5) Exploration

Test1 Logout

Please choose **at least 10 artworks** you would like to hang on your wall

Continue

You have chosen 0/10 artworks, **you still need to give likes to 10 more artwork(s)**

Fig. 10. Preference Elicitation interface, mobile device

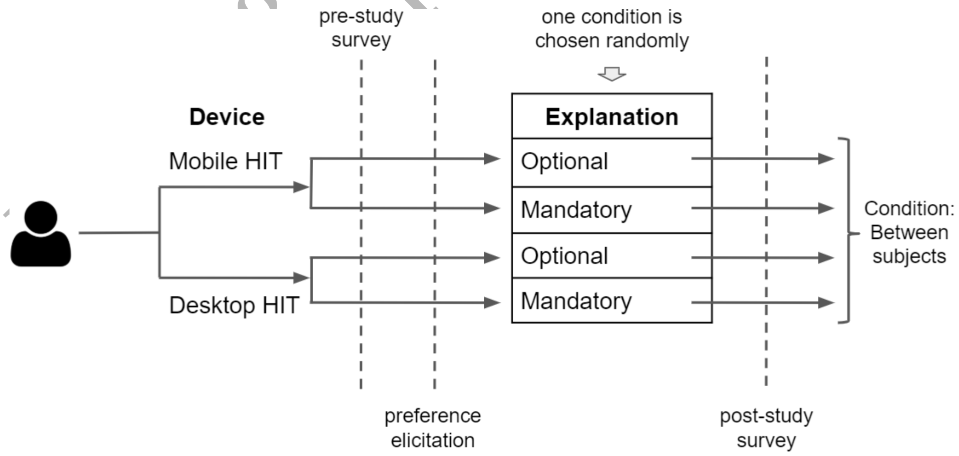


Fig. 11. Study procedure. After the pre-study survey and the preference elicitation, the users were randomly assigned to an explanation condition.

Table 4. Post study survey questions. The last three questiosn correspond to NASA TLX. We also asked the open question *If explanations did not make sense to you, why?*. We also asked about any comments they had regarding the study.

Representative name	Question
Understand	I understood why the art images were recommended to me.
Relevance	The art images recommended matched my interests.
Diversity	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.
Recommend	I would suggest to my colleagues this recommender system when they want to find art images in the future.
Trust	I trusted the recommendations made.
Use explanations	I used the explanations to decide about the recommendations.
Explainable	The explanations over the items recommended made sense to me.
Mental demand	How mentally demanding was the task?
Rush	How hurried or rushed was the pace of the task?
Insecure	How insecure, discouraged, irritated, stressed, and annoyed were you?

Our subjects ranged from 18 to over 60 years old. Thirty-seven percent were between 25 to 32 years old, and 28% between 32 to 40 years old. Females made up 48%. Those just out of high school came in at 16.9%, while 31% has completed some college, and 51.4% had a bachelor's, master's or Ph.D. degree. Only 2.2% reported some visual impairment. With respect to their understanding of art, 36% had zero experience, 46% had attended 1 or 2 lessons, and 18% reported to have attended 3 or more at high school level or above. Only 31% of our subjects reported that they had almost never visited a museum or an art gallery; 36% do this once a year; and 33% do this once every 1 to 6 months.

8 A MODEL OF THE UX WITH AN ART RECOMMENDER WITH DIFFERENT DEVICES

As in the first study we created a full model connecting the variables measures in the study based on Structural Equation Models (SEM). We grouped the variables in the same groups as we did in the first study: (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.

8.1 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 5. We constructed 3 factors: *Haste*, *Domain expertise* and *Satisfaction*. The items used share at least 50.3% of their variance with their designated 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.

In this study we attempted to create the construct *Effort* in the same fashion as study 1. However the question *How mentally demanding was the task?* from the NASA TLX survey did not load significantly well in the new confirmatory factor analysis (CFA). For this reason we decided to change the name of the construct to *Haste*, in such a way that it reflected the meaning of the remaining loading items.

Table 5. Results of the confirmatory factor analysis (CFA), indicating three constructs: *Haste*, *Satisfaction* and *Domain expertise*. Questions in gray are the ones that were removed from the factor.

Construct	**Item**	**Loading**
Haste $\alpha = 0.811$ $AVE = 0.703$	How mentally demanding was the task?	
	How insecure, discouraged, irritated, stressed, and annoyed were you?	0.730
	How hurried or rushed was the pace of the task?	0.935
Satisfaction $\alpha = 0.904$ $AVE = 0.773$	I would use this recommender system again for finding art images in the future.	0.920
	Overall, I am satisfied with the recommender interface.	0.750
	The art images recommended matched my interests.	
	I would suggest my colleagues to use this recommender system when they want to find art imagesin the future.	0.957
Domain expertise $\alpha = 0.619$ $AVE = 0.859$	In case you have taken art classes like: Studio Art Class, Art History Classes, Art Theory or Aesthetics Classes. How many times have you taken any of these classes at the high school level or above?	0.709
	In an average week how many hours do you spend making visual art?	0.825
	In the average week, how many hours do you spend looking at visual art or reading a publication that is related to visual art?	0.905
	On average, you visit art museums or art galleries about once every: 2 months, 6 months, a year or almost never.	0.689
	In case you have taken art classes like: Studio Art Class, Art History Classes, Art Theory or Aesthetics Classes. How many times have you taken of any of these classes at the high school level or above?	0.709

8.2 Structural Equation Model

We subjected the 3 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 mediocre fit [7] fit: $\chi^2(120) = 240.817, p = .000$; $RMSEA = 0.076$, $90\%CI : [0.062, 0.089]$, $CFI = 0.974$, $TLI = 0.968$. We did not find any significant interaction effects between the the experimental manipulation.

[7] A model should not have a non-significant χ^2 (p > .05), but this statistic is often regarded as too sensitive. Hu and Bentler [30] 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.

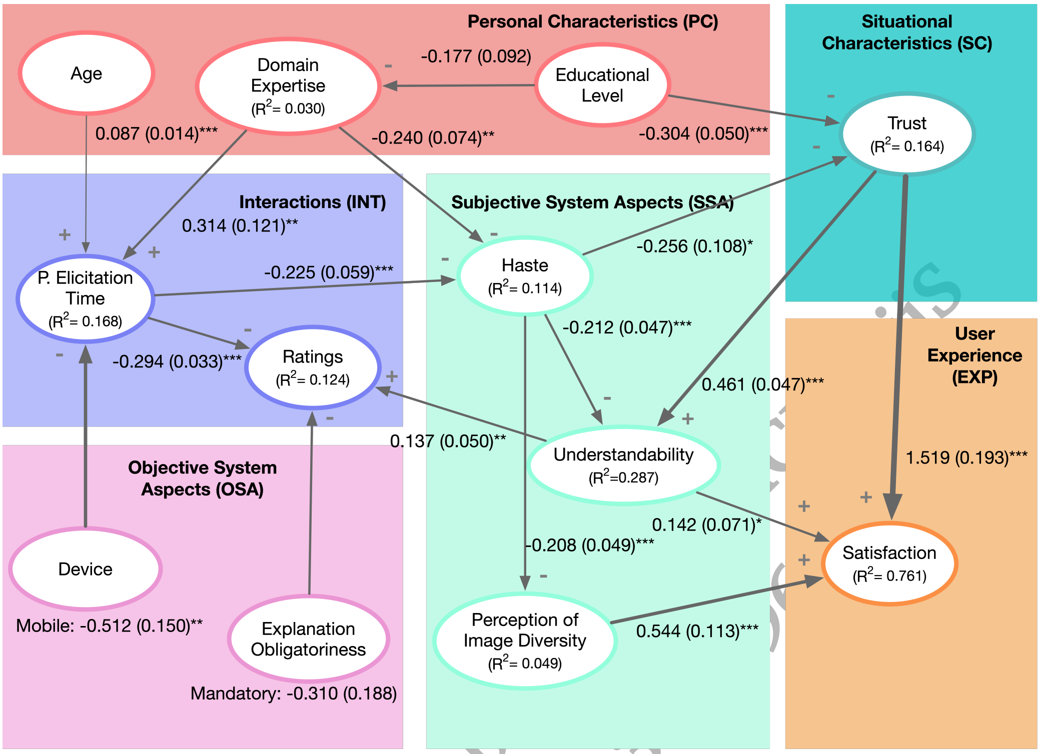


Fig. 12. The structural equation model for the data of the experiment using Knijnenburg's evaluation framework for recommender systems. Significance levels: $* * *p < .001$, $* * p < .01$, $*p < 0.05$. R^2 is the proportion of variance explained by the model. Numbers on the arrows (and their thickness) represent the β coefficients (and standard error) of the effect. Factors are scaled to have an *SD* of 1.

Effect of explanation obligatoriness: This effect was found not significant. Notice that in all conditions there were explanations available, but in one case the explanations were always displayed (mandatory), whereas in the other case the users could choose whether to display the explanations or rather hide them (optional). Then, we expected, as presented in hypothesis 5, that users who could control whether to see or not the explanations would have a better user experience due to the positive effect of user controllability observed in previous studies on recommender systems [7, 49, 63]. It was not the case.

Effects of device: The device directly affects the elicitation time, but indirectly affects the ratings through elicitation time. The effect is negative, which means that when using a mobile interface, users spent less time picking images they liked. This may happen because on a desktop laptop users can see almost all the shown images at once, which allows them to compare.

Effect of Pinterest Elicitation Time: the elicitation time negatively affects the ratings, which means that when users do the experiment on desktop they tend to spent more time picking images they like, which then translates into worse ratings and also in less stress during the task.

Effect of haste: *haste* has a negative effect on understandability and trust, just like in the previous experiment. It also has a negative effect on diversity perception, the more hurried the user feels, they tend to feel less diversity in the items recommended.

Effects on satisfaction: just like in the previous experiment, trust has a very large effect on satisfaction. In this experiment the effect is 10 times larger than the effect of understandability. The perception of image diversity positively affects satisfaction: the more diversity users perceive, the more satisfied they feel with the interface. The only Objective System Aspect (OSA) that affects this factor is the device, when using desktop the haste would be lower and then the perception of diversity would be higher.

9 DISCUSSION STUDY 2

This study was conducted with the aim of answering hypotheses 4 and 5. First, hypothesis 4 stated that there would be differences in user experience with the recommender systems between two devices: desktop and mobile. As we can see in the figure 12, the SEM of the study 2, the *device* variable had a significant effect. It affected directly the time in the preference elicitation phase, which is explained by the space and layout of the interfaces due to screen size. If the device was a mobile, users tend to spend less time selecting their preferences. We explain this considering that the larger screen of the desktop device promotes more comparisons among items shown in the screen, increasing exploratory behavior. In the mobile device, users focused on deciding whether a single image shown in the viewport was to be liked or not. Moreover, the SEM analysis shows an indirect relation between device and average rating: desktop device increases time spent by users, and the larger the time spent, the lower the ratings. This might be explained by the fact that in the mobile interface users evaluate each recommended image one at a time, while in the larger desktop interface users rate the paintings by watching also the context of the whole list. This result gives some hints about a variable not explicitly observed that can have a strong influence: evaluation of a list of recommendations versus evaluation of isolated recommended items.

With respect to hypothesis 5, we expected that increased user control upon showing or hiding the explanations would have a positive effect on subjects' experience. This was not the case, since we found no significant effects of providing explanation obligatoriness or optionality upon any variable, such as the perception of understandability. We rather found other aspects that had an effect. Interestingly, the perception of understandability had a positive significant effect on both satisfaction and ratings. Nevertheless, this perception of understandability was influenced by users' trust as a situational characteristic as well by users' feeling of haste during the study.

10 DISCUSSION & LIMITATIONS

10.1 General Discussion

(1) Apart from supporting with an online study where the visual recommender based on deep learning features performs better than the recommender based on explicitly-engineering features, one of the main take-aways of study 1 is that recommender interfaces with higher transparency are not always the best option compared to interfaces with simplified explanations or interpretability. In addition, the user satisfaction is positively affected by the users' trust on the system, as well by the perception of understability, time and experience with the art domain. These results are in line with other recommender systems studies [7, 49, 63]. In general, we suggest that developers and researchers to make sure that the algorithms used are accurate enough before attempting any design at the level of the interface for studying different types of explanations. Extending the findings by Cramer et al. [13], the users need to perceive the algorithm as competent in order to gain their trust. Design of the interface without making sure the algorithms are accurate enough will most likely fail.

(2) In the study 2 we only used the algorithm based on deep learning visual features with the explanations based on similar items, as the one by Cai et al [8]. It was interesting that in

the second study we again observed a strong effect of user trust as well as perception of understandability on the user satisfaction with the recommender system. The type of device had an incidence at the time of interacting with the interface, mostly based on the amount of item for which to compare with at the moment of preference elicitation. In a larger device users could compare many items, while in the mobile they could only see and judge one item at a time.

(3) Study 2 provides three important insights to the study of explainability on interactive image recommendation systems: **(I)** The first insight we highlight is the importance of the device (and its corresponding screen size) when designing the preference elicitation stage and not only the recommendation stage. When designing an interface for recommending images, the device and its corresponding screen size (small screen on mobile phone versus larger screen on a desktop device) might have a stronger impact on the preference elicitation stage that when users explore the actual recommendations. Preference elicitation plays an important role in collecting data for learning the user model, and thus we suggest to jointly study device, preference elicitation interface and recommendation interface in further studies. Focusing only on the interface that shows the recommended items might not give a full picture of the UX with the recommender. **(II)** Although in study 1 we found that providing explanations, as well as the type of explanation, had a significant effect on user experience compared to no explanations at all, in study 2 we found that providing users the chance to control whether they could see or hide the explanations had no effect compared to see them obligatorily. We thought that the limitations of the small mobile screen would produce an interaction between device and explanation obligatoriness, but it was not the case. Thus, providing explanations and different ways to show them seems to be a good avenue for future work, whereas providing users control over showing them of hiding them seems to have no much impact for user experience with the image recommender, even on screens with limited size like mobile devices. **(III)** Finally, our results provide additional evidence of important effects found on previous interactive recommendation studies [35, 49, 50], but in this case for recommendation of visual art. We report confirming evidence of the positive effect of users' inherent trust, as well as perceptions of understandability and diversity upon final satisfaction with the recommendation system.

(4) If we compare our results with other studies on explainable AI, either in terms of recommender systems [47] or in a traditional classifier [67], we can say that our results indicate that users evaluate positively explanations of predictions, but more transparency is not always the best option, even with a well-known accurate recommendation algorithm.

10.2 Limitations

One limitation in this work is that we built our recommendation models and hypotheses based on the off-line analysis of transactions from an online store, and in this article we show results of an online study where the signal of user preference are likes rather than sales. In a future study, we will ask users whether they would buy the paintings recommended so we can more precisely understand these two different preference signals.

Another important limitation in our experiments is that we conducted them in a single session, rather than following users' interest over a longer period of time. Holliday et al [29] showed that users' trust upon the system varies over time, so in the future we should conduct the experiments over a longer period of time and not just within a single session.

Another limitation to consider is that there were some changes between study 1 and study 2 that can hinder their comparison. The first one is that we changed the wage from 0.4 US dollars per minute in study 1, to 2 dollars per minute in study 2. We did this in order to adjust to ethical wage guidelines in AMT compared to what previous papers had reported [42, 54]. Another aspect making comparisons difficult among studies one and two is the number of subjects in each study. In the second study we were more careful in designing the full validation pipeline (not only the traditional validation questions) and for these reasons, in the second study we had 177 valid users (from 200 participating) versus only 121 valid users in study 1 (also with 200 subjects participating).

In the design of study 1, since we presented many conditions, combining different kinds of interfaces and algorithms, there are effects that cannot be explained just by one variable. One limitation is that there is a significant difference in the results between interface 3 and interface 2, but it cannot be attributed just to the change of interface and explanation. We also need to consider that in interface 2, recommendations were based on top 3 similar images, and in interface 3 the recommendations were based on top 1 similar image. This could affect the user experience in a way that cannot be ignored

We used visual features that capture the attractiveness of a picture to recommend art. There are more features that capture more aspects of art images, like the emotions and content. In [40] they present features inspired by psychology and art theory to make an affective classification of images. One limitation is that it is expensive to test several algorithms and features in a user study. As the amount of conditions increases, the amount of participants needed increases too. We decided to use the AVF over other sets of features because they were already tested in art recommendation task before [45] and performed well. We wanted to test more features inspired by emotion like the aforementioned, but the amount of participants needed to made the study is a limitation, and the performance of these features in an art recommendation task was not tested.

11 CONCLUSIONS & FUTURE WORK

This paper describes the effects of explaining recommendation of images by conducting two user studies. In the first study we employed 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). In the second study we studied the effects of the type of device (mobile or desktop), combined with the option of choosing whether to receive explanations in the recommendations. In the first study we answered research questions 1-3 and in the second we answered the last two.

The first result, which addresses Hypothesis 1, 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 one type of visual feature, the AVF. This result should be further studied in other image datasets, and it opens a new branch of research in terms of new interfaces to explain the features learned by a Convolutional DNN of images.

Regarding Hypothesis 2, 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.

With respect to Hypothesis 3, we have provided a holistic model to each user study, based on the framework by Knijnenburg et al. [36], which explains the relations among different independent variables (interface, algorithm, device, art domain expertise) and several metrics to measure the user experience with an explainable recommender system of artistic images.

With the second study we were able to address the last two Hypotheses. We could not reject our Hypothesis 4, since mobile devices tend to decrease the stress of the users, which increases satisfaction. It is interesting to notice that the device has an effect that spreads to the whole model of user experience.

Finally for Hypothesis 5, we can see in figure 12, that the explanation obligatoriness has a direct but not significant effect on the ratings and only upon this variable. Although the effect is not significant it could not be ignored because the model has a relatively good fit and the effect is high (-0.310). This path states that when the users receive explanations in a mandatory fashion, they gave lower ratings than the people who have the option to decide whether to receive explanations or not. In this case, the explanations did not affect the whole user experience, just the evaluation of the images. We therefore realized in the first user study that the form of explanation has a major effect, but the explanation obligatoriness has, comparatively, no major effect.

In the future, we would like to use more advance techniques to explain on automatic decision models like LIME [53], but adapted to personalized recommender systems. We also would like to combine these types of explanation techniques with recent models of neural style transfer [19, 48] and test them using this user-centric recommender evaluation framework.

12 ACKNOWLEDGEMENTS

The authors from PUC Chile were funded by CONICYT, FONDECYT *Iniciacion* grant 11150783 as well as by FONDECYT *Regular* grant 1191791, and also by the Millennium Institute for Foundational Research on Data (IMFD).

13 ADDITIONAL MATERIAL

13.1 Study 1 general results

Table 6. Results of users' perception over several evaluation dimensions, defined in Section 3.6. Scale 1-100 (higher is better), except for Average rating (scale 1-5). DNN: Deep Neural Network, and AVF: Attractiveness visual features.

	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*	74.0	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*	70.4	82.3*	56.2	65.3	71.2	69.9*	63.3	78.2*	58.7	77.7*	55.4	3.90*	2.99

Table 7. NASA TLX Results.

	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↑[2]

Table 8. Results of users' perception over several evaluation dimensions, defined in Table 9 . Scale 1-100 (higher is better) Mob: Mobile Interface, and Desk: Desktop Interface.

	Understood		Relevance		Diversity		Interface Satisfaction		Use Again		Suggest		Trust		Use Explanations		Explainable	
Condition	Mob	Desk	Mob	Desk	Mob	Desk	Mob	Desk	Mob	Desk	Mob	Desk	Mob	Desk	Mob	Desk	Mob	Desk
Optional Explanation	79.41	82.18	79.02	77.20	64.64	69.13	78.95	81.04	76.84	81.56	76.09	80.89	80.27	80.13	47.05	48.47	71.23	72.04
Mandatory Explanation	82.70	84.48	75.14	71.50	78.23	71.18	81.07	82.20	76.57	81.36	73.55	76.77	77.34	78.05	54.75	52.64	83.73	80.77

13.2 Study 2 general results

Table 9. NASA TLX results, Prec@10 and Time tracked in elicitation and recommendation stage.

	Prec@10		Elicitation Time		Recommendation Time		Insecurity		Mental Demand		Hurry	
Condition	Mob	Desk	Mob	Desk	Mob	Desk	Mob	Desk	Mob	Desk	Mob	Desk
Optional Explanation	0.72	0.63	110.69	197.69	52.09	40.99	9.18	4.64	21.91	20.47	15.02	8.02
Mandatory Explanation	0.64	0.60	116.22	174.06	66.48	43.29	8.36	6.23	22.50	21.82	12.61	11.57

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 582.

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

[3] 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).

[4] 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.

[5] Christoph Bartneck, Andreas Duenser, Elena Moltchanova, and Karolina Zawieska. 2015. Comparing the similarity of responses received from studies in Amazon's Mechanical Turk to studies conducted online and with direct recruitment. PloS one 10, 4 (2015), e0121595.

[6] 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.

[7] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: a visual interactive hybrid recommender system. In Proceedings of the sixth ACM conference on Recommender systems. ACM, 35–42.

[8] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The Effects of Example-based Explanations in a Machine Learning Interface. In Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19). ACM, New York, NY, USA, 258–262. DOI: http://dx.doi.org/10.1145/3301275.3302289

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

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

[11] Ronny Cook and Judy Kay. 1994. The justified user model: a viewable, explained user model. In In Proceedings of the Fourth International Conference on User Modeling. Citeseer.

[12] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 191–198.

[13] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. User Modeling and User-Adapted Interaction 18, 5 (2008), 455.

[14] 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.

[15] 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 Proceedings of the 24th International Conference on Intelligent User Interfaces. ACM.

[16] 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

[17] 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.

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

[19] 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.

[20] 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

[21] 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.

[22] Qi Guo, Haojian Jin, Dmitry Lagun, Shuai Yuan, and Eugene Agichtein. 2013. Mining touch interaction data on mobile devices to predict web search result relevance. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 153–162.

[23] Shuguang Han, I-Han Hsiao, and Denis Parra. 2014. A study of mobile information exploration with multi-touch interactions. In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. Springer, 269–276.

[24] PD Harms and Justin A DeSimone. 2015. Caution! MTurk workers ahead—Fines doubled. Industrial and Organizational Psychology 8, 2 (2015), 183–190.

[25] 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.

[26] Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. 2016. 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

[27] Ruining He and Julian McAuley. 2016. VBPR: visual Bayesian Personalized Ranking from implicit feedback. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. AAAI Press, 144–150.

[28] 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, 241–250. http://doi.acm.org/10.1145/358916.358995

[29] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In Proceedings of the 21st International Conference on Intelligent User Interfaces. ACM, 164–168.

[30] 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

[31] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. arXiv preprint arXiv:1408.5093 (2014).

[32] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 453–456.

[33] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. Self-normalizing neural networks. In Advances in Neural Information Processing Systems. 971–980.

[34] BP Knijnenburg, N Rao, and A Kobsa. 2012b. Experimental materials used in the study on inspectability and control in social recommender systems. Technical Report. Institute for Software Research, University of California, Irvine.

[35] Bart P. Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012a. Inspectability and control in social recommenders. In Proceedings of the sixth ACM conference on Recommender systems (RecSys '12). ACM, 43–50. http://doi.acm.org/10.1145/2365952.2365966

[36] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012c. 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

[37] 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.

[38] 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.

[39] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In 2013 IEEE Symposium on Visual Languages and Human Centric Computing. IEEE, 3–10.

[40] Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In Proceedings of the 18th ACM international conference on Multimedia. ACM, 83–92.

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

[42] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. Behavior research methods 44, 1 (2012), 1–23.

[43] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 43–52.

[44] 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.

[45] 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

[46] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence (2018).

[47] Caio Nóbrega and Leandro Marinho. 2019. Towards explaining recommendations through local surrogate models. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. ACM, 1671–1678.

[48] 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.

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

[50] Denis Parra and Shaghayegh Sahebi. 2013. Recommender systems: Sources of knowledge and evaluation metrics. In Advanced Techniques in Web Intelligence-2. Springer, 149–175.

[51] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (RecSys '11). ACM, 157–164. http://doi.acm.org/10.1145/2043932.2043962

[52] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. AUAI Press, 452–461.

[53] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 1135–1144.

[54] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In CHI'10 extended abstracts on Human factors in computing systems. ACM, 2863–2872.

[55] 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

[56] 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.

[57] 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.

[58] Rashmi Sinha and Kirsten Swearingen. 2002. The role of transparency in recommender systems. In CHI'02 extended abstracts on Human factors in computing systems. ACM, 830–831.

[59] Nava Tintarev and Judith Masthoff. 2007. Effective explanations of recommendations: user-centered design. In Proceedings of the 2007 ACM conference on Recommender systems (RecSys '07). ACM, 153–156. http://doi.acm.org/10.1145/1297231.1297259

[60] Nava Tintarev and Judith Masthoff. 2011. Designing and Evaluating Explanations for Recommender Systems. Springer US, 479–510. http://dx.doi.org/10.1007/978-0-387-85820-3_15

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

[62] 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.

[63] 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.

[64] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI, Vol. 19.

[65] Rainer Wasinger, James Wallbank, Luiz Pizzato, Judy Kay, Bob Kummerfeld, Matthias Böhmer, and Antonio Krüger. 2013. Scrutable user models and personalised item recommendation in mobile lifestyle applications. In International Conference on User Modeling, Adaptation, and Personalization. Springer, 77–88.

[66] 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].

[67] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User trust dynamics: An investigation driven by differences in system performance. In Proceedings of the 22nd International Conference on Intelligent User Interfaces. ACM, 307–317.

[68] Jiyong Zhang, Nicolas Jones, and Pearl Pu. 2008. A Visual Interface for Critiquing-based Recommender Systems. In Proceedings of the 9th ACM Conference on Electronic Commerce (EC '08). ACM, New York, NY, USA, 230–239. DOI: http://dx.doi.org/10.1145/1386790.1386827

[69] Yongfeng Zhang and Xu Chen. 2018. Explainable recommendation: A survey and new perspectives. arXiv preprint arXiv:1804.11192 (2018).

[70] Shiwan Zhao, Michelle X. Zhou, Quan Yuan, Xiatian Zhang, Wentao Zheng, and Rongyao Fu. 2010. Who is talking about what: social map-based recommendation for content-centric social websites. In Proceedings of the fourth ACM conference on Recommender systems (RecSys '10). ACM, New York, NY, USA, 143–150. DOI: http://dx.doi.org/10.1145/1864708.1864737

[71] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In Proceedings of the 14th international conference on World Wide Web (WWW '05). ACM, 22–32. http://doi.acm.org/10.1145/1060745.1060754