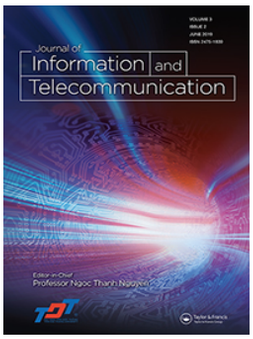


Taylor & Francis
Taylor & Francis Group

Journal of Information and Telecommunication

ISSN: 2475-1839 (Print) 2475-1847 (Online) Journal homepage: www.tandfonline.com/journals/tjit20

An algorithmic approach to estimate cognitive aesthetics of images relative to ground truth of human psychology through a large user study

Tousif Osman, Shahreen Shahjahan Psyche, Tonmoay Deb, Adnan Firoze & Rashedur M. Rahman

To cite this article: Tousif Osman, Shahreen Shahjahan Psyche, Tonmoay Deb, Adnan Firoze & Rashedur M. Rahman (2019) An algorithmic approach to estimate cognitive aesthetics of images relative to ground truth of human psychology through a large user study, Journal of Information and Telecommunication, 3:2, 156-179, DOI: 10.1080/24751839.2018.1542574

To link to this article: https://doi.org/10.1080/24751839.2018.1542574

Published online: 11 Nov 2018.

JOURNAL OF INFORMATION AND TELECOMMUNICATION
2019, VOL. 3, NO. 2, 156–179
https://doi.org/10.1080/24751839.2018.1542574

An algorithmic approach to estimate cognitive aesthetics of images relative to ground truth of human psychology through a large user study

Tousif Osman, Shahreen Shahjahan Psyche, Tonmoay Deb, Adnan Firoze and Rashedur M. Rahman

Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh

ABSTRACT

This research introduces a learning model that estimates the cognitive perception of aesthetics. Taking psychology into account, this bridges the gap between human and machine. The goal is to build a machine-learning model that can estimate beauty in images perceived by human eyes. We have summand our research [Firoze, A., Osman, T., Psyche, S. S., & Rahman, R. M. (2018). Scoring photographic rule of thirds in a large MIRFLICKR dataset: A showdown between machine perception and human perception of image aesthetics. Asian Conference on Intelligent Information and Database Systems (pp. 466–475), Springer; Osman, T., Psyche, S. S., Deb, T., Firoze, A., & Rahman, R. M. (2018). Differential color harmony: A robust approach for extracting Harmonic Color features and perceive aesthetics in a large dataset. International Conference on Big Data and Cloud Computing, Springer] together with the idea of humans' personal preferences and achieved higher than state of the art performances. An extensive user study (374 participants) has been conducted to support claims. Several photographical compositional metrics have been used. Colour gradient, rule of thirds and human subject's psychology has been picked as features. The consideration of user's perspective or psychology is one of the key contributions of this research.

ARTICLE HISTORY
Received 28 June 2018
Accepted 28 October 2018

KEYWORDS
Visual aesthetics; visual perception; cognitive machine-learning; colour analysis; rule of thirds; computer vision; image processing; image composition; Flickr

1. Introduction

In this present era, A.I. and Machine Learning are one of the main streams of computational developments. This approach of computing allows us to solve problems that were computationally unsolvable in the previous epoch of computing. However, sometimes the cutting-edge technologies fail to comprehend human cognition. One such cognitive behaviour is perceiving aesthetics. This research introduces and devises an artificial system that can comprehend the aesthetical attractiveness of an image. The building blocks of the research are: a significantly large user study and three computational models based on photographic metrics and psychology. Firstly, few metrics have been defined that come unconsciously

CONTACT Rashedur M. Rahman rashedur.rahman@northsouth.edu

into the consideration of the human mind when it perceives beauty. Next, we have developed algorithms to extract features from images for each of the metrics. Finally, we have combined these features with the techniques of Machine Learning and mimicked human's cognitive ability to perceive visual aesthetics. This research work is primarily a continuation of our previous work (Firoze, Osman, Psyche, & Rahman, 2018), where we have used the photographic rule called 'Rule of Thirds' (ROT), and perceived appeal with a soft computational approach. We have also incorporated another of our work (Osman, Psyche, Deb, Firoze, & Rahman, 2018) with this research, where we have introduced a technique to analyse colours considering a small fragment of human's perception of colours and predicted beauty with that metric. In this research, we have considered and introduced another metric – *Human Context* – while perceiving beauty. This contextual analysis has allowed us to gain a huge improvement in the results. Adding user's context in the machine learning model allowed the synthetic system to produce results more close to human perception. We have analysed a large user study to understand human's choice of preference and devised a system that can perceive aesthetics considering Human Context combined with other defined features. Finally, we have compared machine perceived aesthetics with human perceived aesthetics through a rating system and confronted this system's applicability in the applied world.

2. Background study

Few computer vision researchers have attempted to automate the semantics of aesthetics of images before us. We did not reinvent the wheel completely and used some of the existing techniques along with our novel concept of user context. Recently, Kong, Shen, Lin, Mech, and Fowlkes (2016), employed a deep neural network for aesthetics perception. After assembling training and performance evaluation on aesthetics and attributes database (AADB), they developed a Convolutional Neural Network (CNN) based architecture that fuses both attributes and contents of a photo to rank an image.

Datta and Wang (2010) introduced ACQUINE for real-time aesthetics classification. It is based on a support vector machine (SVM) classifier and involves fast feature extraction and classification. The focus of the paper was to move toward understanding human emotional preferences on the image. In another work, Lu, Lin, Jin, Yang, and Wang (2015) proposed a novel deep neural network architecture focusing on both feature learning and training classifiers. Apart from hand-crafted features, it computes features itself. For achieving image aesthetics based on content, they developed network adaptation technology. Further, they boosted performance upon experiments on AVA dataset by employing image style along with semantic attributes. However, Datta, Joshi, Li, and Wang (2006) approached studying different visual properties of an image in terms of aesthetics evaluation. The measurement of aesthetics was binary: high and low. According to the experiment, they claimed that a model with only 15 features with an SVM classifier was sufficient for achieving satisfactory accuracy.

ROT method was leveraged by Mai, Le, Niu, and Liu (2011) to quantify aesthetics of an image. Initially, they extracted the salient regions from the images using Graph Based Visual Saliency (GBVS), Fourier Transform (FT), Global Contrast (GC) and other 'objectness' methods. Additionally, authors used generic objectness analysis as a proxy for saliency detection. Then they moved on to measure aesthetics using machine learning techniques.

Amirshahi, Hayn-Leichsenring, Denzler, and Redies (2014) have rated 30 paintings by humans and the same ones by algorithms and tried to find a correlation between the two in terms of beauty or aesthetics. In their findings they have found that computationally generated ROT metrics only partially mimics the real humans perception of beauty. Maleš, Heđi, and Grgić (2012) detected the ROT using saliency algorithms. Context Aware and Global Contrast based salient region detector were the two algorithms used to find saliency. After finding the salient regions, the used PCA. Next, for classification, they took the advantages of Linear Discriminant Analysis, Mahalanobis Linear Discriminant Analysis, Quadratic Discriminant Analysis and Support Vector Machines.

Lu, Peng, Zhu, and Li (2016) focused on learning multimodal features from images and proposed labelled-latent Dirichlet allocation (EL-LDA) Model. Later, unsupervised Gaussian mixture models (GMM) is learned for 7435 high aesthetics and low aesthetics images. The similar experiment was done for Supervised LASSO regression approach with a complete dataset. Their framework can discover harmonious colours from a natural image. Later, as an extension of that paper (Lu, Peng, Yuan, Li, & Wang, 2016), they focused on neighbouring spatial regions for gaining colour harmonic information. This approach appeared more robust and outperformed previous colour-harmony based models. Phan, Fu, and Chan (2017) worked on palette data, where they designed ordering palettes. Further, they used dimensionality reduction for palette colour reordering, which implied applying robust interpolation techniques on data, and concluded research achieving a state-of-the-art result on summarization.

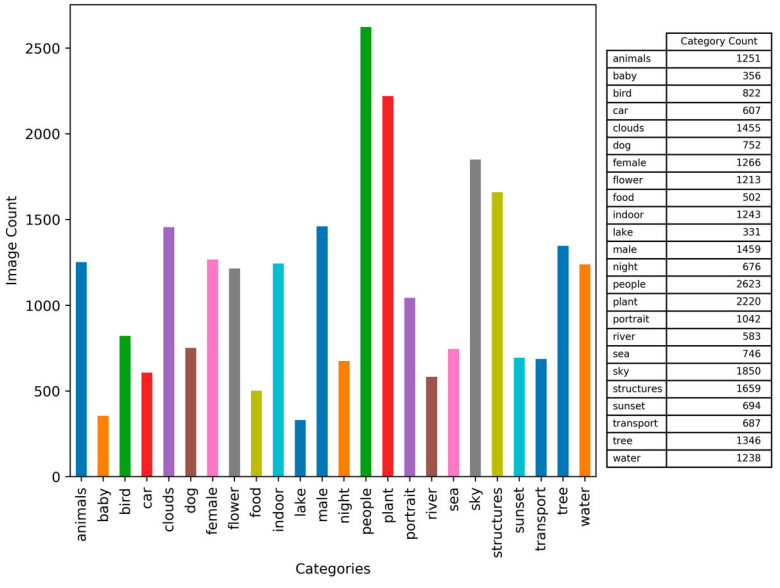
All of the mentioned work made an outstanding amount of contribution in the domain of machines' perception of aesthetes. However, these researches ignored one key factor of humans' perception of aesthetics and that is human psychology. In our research, we have considered this ignored factor and showed that consideration of human psychology can make a significant improvement in the final results. Furthermore, in contrast to researches in progress and that has been done in this domain, we have combined different feature extraction methods and built one novel system that can perceive aesthetics with a significantly improved accuracy.

3. User study and data source

One of the major keystones of this research is our user study. This study enabled us to train our model, evaluate our system and the most crucial work of this research: user context analysis. We used MIR-Flickr (Huiskes & Lew, 2008) as our image data source. The full collection has 25,000 images, from which we randomly picked 5000 from the Flickr website under creative commons license. MirFlickr has been regarded as the gold standard of datasets in visual retrieval systems. Images of this dataset may have more than one category tag and the entire dataset has 24 different categories. We have randomly selected 5000 images from this dataset and created our image dataset for this research. We have reduced the dataset so that real users can feasibly rate each of the images. Figure 1 shows the dataset's categorical distribution.

Although we have selected 5000 images, category count is lot more than that, because one image can have more than one category tags.

A user study on this data was conducted where 374 participants participated. Participants were mix-gendered; about 37% of the participants ware female and rest ware



	Category Count
animals	1251
baby	356
bird	822
car	607
clouds	1455
dog	752
female	1266
flower	1213
food	502
indoor	1243
lake	331
male	1459
night	676
people	2623
plant	2220
portrait	1042
river	583
sea	746
sky	1850
structures	1659
sunset	694
transport	687
tree	1346
water	1238

Figure 1. Categorical data distribution of images.

male. Participants ware aged between 19 and 29 years where the median age was 22 years. Majority of the participants were students from North South University, Bangladesh and about 5% of them were members of the photography club. We have developed a survey system to ease the process of the user study. This system eliminated bias from the user study by randomly selecting 35 images of different categories from our image dataset. This system made sure no image is selected twice in a cycle of selecting 5000 images. An image is selected second time after the system selected all images and completed a cycle. This allowed us to have an evenly distributed user response. We collected 12,748 users' responses on 5000 images, which means for 2252 images we have 3 user responses per image and 2748 images has 2 users' feedbacks per image. System asked the participants to answer the following questionnaire for each selected image:

(1) Give a numerical rating to the picture you are seeing: 5 being best, 1 being worst
(2) Click on the regions of the picture that attracts your attention.
(3) Name the object(s) you see in the picture at glance.

Question 1 tells us about the user's perception of aesthetics. This is the most valuable piece of information we have acquired from our user study. This is the user given score we used throughout our research to train models, measure accuracy etc. Question 2 and 3 allowed us to set a baseline for different aspects of our study. We have plotted a histogram of the user given scores in Figure 2. Figure 2(a) is the histogram of all user given score and Figure 2(b) shows the histogram of the average score of each image.

4. System design and workflow

The underlying principle of this research is to train a machine-learning model and perceive aesthetics. We have defined three features sets – Rule of Thirds (ROT), Colour Harmony,

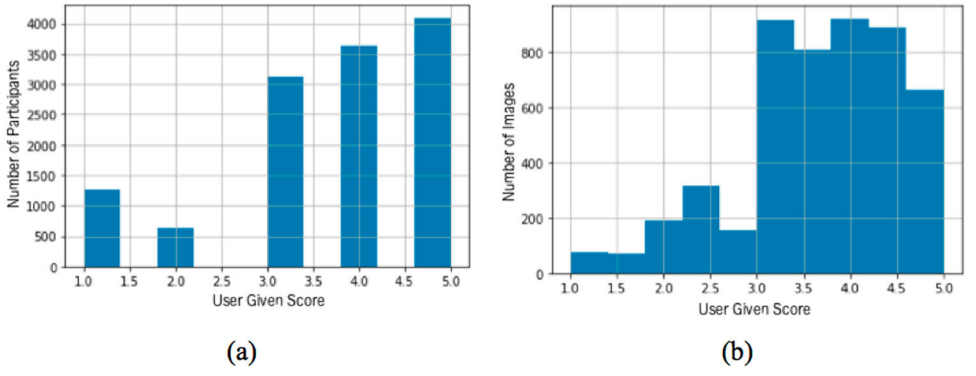


Figure 2. Histogram of user response. (a) Histogram of all user score. (b) Histogram of average score.

and User Context – to train our model. Extracting these features sets are the heart of our system. The simplified architecture of this system is shown in Figure 3. Initially, the system receives input images from the image dataset. Next, images are fed into a feature extraction module. This module extracts three feature sets and creates sample dataset for further processing. Afterwards, the sample dataset is split into 8:2 ratios for training and testing respectively. After that, training data is passed into the model generation module. Next, the generated model is applied on the testing data and the machine perceived aesthetics score is calculated. Details of the individual components of our system have been explained in the following subsections.

4.1. ROT feature

We have used Rule of Thirds (ROT) metric in this system to calculate a feature set. In our prior work (Firoze et al., 2018), we considered ROT as the scale of beauty measurement and geometrically calculated visual appeal in a soft-computational approach. In contrast to our prior work, we used a similar principal but calculated feature-set which is not a beauty measurement. Rather this feature-set holds information related to aesthetics.

ROT us defined as a compositional rule which essentially states that in an image, humans find objects to be more appealing when they are along the gridlines (if we draw a 3 by 3 grid) on the frame (Peterson, 2003). The famous Golden Ratio is the source of this heuristics (Weisstein, 2002) and the proportions were discoveries of the

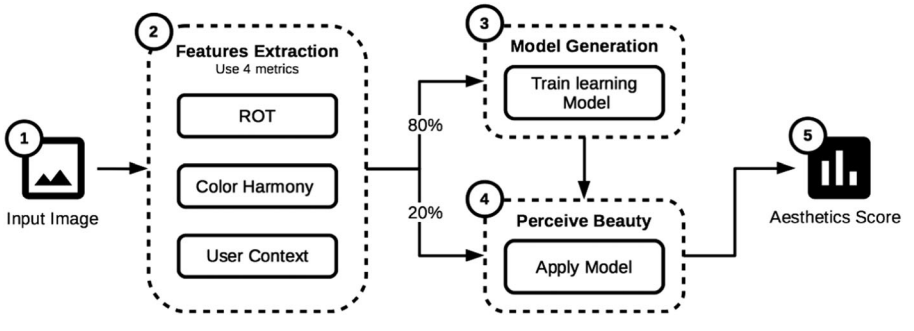


Figure 3. Simplified workflow of the system.

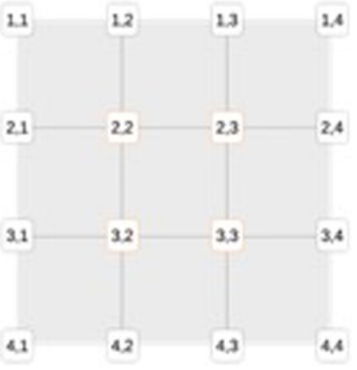


Figure 4. Corner points of gridlines.

ancient Greeks (Peterson, 2003). The ROT was first documented in 1797 in the book – 'Remarks on Rural Scenery' by J. T. Smith (Caplin, 2008).

To simplify the idea of ROT, if we draw a 3 by 3 evenly spaced grid on top of the frame of an image as in Figure 4, then each of these lines divides the image according to the golden ratio. Figure 4 simplifies the understanding of the subdivision.

If an image has its salient regions near or overlapping any of these gridlines, then they tend to be visually more attractive – is the core concept of ROT. Furthermore, salient regions closer to the intersecting points of the gridlines – points (2, 2), (2, 3), (3, 2), and (3, 3) in Figure 4 – will be of utmost beauty as it infers that the subject is along the golden ratio both vertically and horizontally.

The task of extracting ROT features was done in several stages. In our research, we have developed a module called 'ROT Engine' that extracts ROT features from an input image. Figure 5 represents the simplified workflow of the ROT engine. First, the ROT engine is fed an image as input. Then, the image is passed to the pre-processor (the next stage of our ROT engine). To apply the compositional metric, we must first compute the salient regions of the image. Hence, the next stage has been developed using GBVS (Harel, Koch, & Perona, 2007) saliency algorithm to detect salient regions. In our research (Firoze et al., 2018) we have demonstrated that GBVS performs better compared to other saliency algorithms, e.g. Itti-Koch saliency algorithm (Itti & Koch, 2000).

An image can have multiple salient regions but, in this research, we are considering at most two regions. We have come to this conclusion by analysing the response to question 3 of our user study. Looking into users' responses, we have observed the majority of the users named one or two objects. Observing this, we concluded that in general at most two salient object has prominence in regular human eyes.

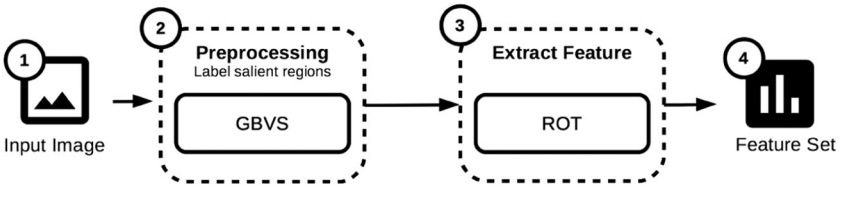


Figure 5. Simple workflow of the ROT engine.

The two extracted salient regions have been labelled as Primary Salient Region (PSR) and Secondary Salient Region (SSR). Next, the ROT engine proceeds to the next stage where the engine measures whether the image is maintaining ROT. First, we have calculated the centroid $C(x, y)$ of the salient areas to produce ROT features. The centroids are calculated as follows:

$$C_x = \frac{X_1 + X_2 + X_3 + \cdots + X_k}{S} \tag{1}$$

$$C_y = \frac{Y_1 + Y_2 + Y_3 + \cdots + Y_k}{S} \tag{2}$$

Here, X_1 through X_k are the X coordinates, Y_1 through Y_k are the y coordinates, and S is the summation of pixels in a salient blob. We have constructed devized functions (Equations (3) and (4)) that calculate orthogonal distances from each vertical or horizontal gridline to the centroids of the salient regions.

$$D_{\text{horizontal}}(C, i)|_{i=1,2} = \sqrt{\left(C_x - \frac{i*w}{3}\right)^2} \tag{3}$$

$$D_{\text{vertical}}(C, j)|_{j=1,2} = \sqrt{\left(C_y - \frac{j*h}{3}\right)^2} \tag{4}$$

Here, in Equations (3) and (4), h is the height of the image and w is the width of the image. As we have 2 salient regions and 4 gridlines, we are getting 8 orthogonal distances. In the equations, C represents one of the centroids – PSR or SSR; C_x and C_y represent the X and Y component of the centroids. Lastly, the i and j represent the number of horizontal and vertical gridline respectively. When i is 1, Equation (3) calculates the horizontal orthogonal distance $D_{\text{horizontal}}$ from the centroid C to left most vertical grid and when it is 2, Equation (3) calculates the orthogonal distance $D_{\text{horizontal}}$ from C to right most vertical grid line. When j is 1, Equation (4) calculates vertical orthogonal distance D_{vertical} from centroid C to top horizontal grid and when j is 2, Equation (4) calculates the vertical orthogonal distance D_{vertical} from C to bottom horizontal grid line.

The distance is measured in in relative measure that is a ratio of pixels with full image dimensions (and it is important to note that the images are of variable sizes). Considering this, we have come up with the following equations to normalize distance scores and generate scores which we can consider as ROT feature.

$$N_{\text{horizontal}} = \frac{D_{\text{horizental}}}{2h/3} \tag{5}$$

$$N_{\text{vertical}} = \frac{D_{\text{vertical}}}{2w/3} \tag{6}$$

We have normalized the scores by dividing the distances by 2/3 of the width and height of the image for horizontal and vertical distance respectively. The reason being: the centroid can have a maximum distance of 2/3 * width (maximum horizontal distance centroid can be from the grids) or 2/3 * height (maximum vertical distance centroid can be from the grids). In Equations (5) and (6), w = width and h = height of the image.

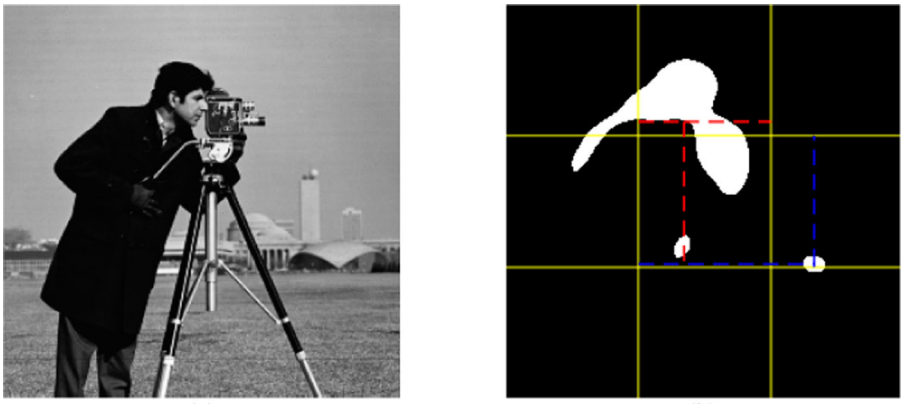


Figure 6. Sample Image and extracted ROT features (a) Sample Image 'cameraman.tif" and (b) Extracted features of ROT with respect to PSR and SSR. The darker and lighter dashed lines respectively in Figure 6(b) are the orthogonal distances between the centroids and gridlines for PSR and SSR respectively.

Figure 6 shows an example of ROT feature extraction. Figure 6(a) is the famous cameraman.tif sample input image. Red and blue dashed lines in Figure 6(b) are the orthogonal distances between the centroids and gridlines for PSR and SSR respectively. Our 8 ROT feature values for Figure 6 are: 0.17, 0.33, 0.05, 0.55, 0.66, 0.16, 0.49, and 0.01. Unlike our previous work (Firoze et al., 2018), these values do not tell anything specific about aesthetics but have some underlying correlation with the aesthetics.

4.2. Harmonic colour feature

In this model, we have explained the way to extract the Harmonic Colour Features (HCF). This module is based on our previous work (Osman et al., 2018). The basic idea is: when a colour changes its base colour – maintaining a pattern, the human eyes find it appealing (Stone, Adams, & Morioka, 2008). Figure 7 is an example of colour change where the shade changes gradually in Figure 7(a), the base and shade changes gradually Figure 7(b), but in Figure 7(c) the colour changes randomly. Even a non-artistic person will find Figure 7(a,b) more attractive than Figure 7(c).

We have demonstrated a simplified workflow of the HCF extraction module in Figure 8. First we receive the input image. Secondly, we change the colour space from RGB (Red, Green, Blue) to HSV (Hue, Saturation, and Value). Stone et al. (2008) state that a small change in HSV can produce a harmonic colour with respect to the original one. Therefore, we transform the colour space from RGB to HSV in this module. In the RGB scale, colours

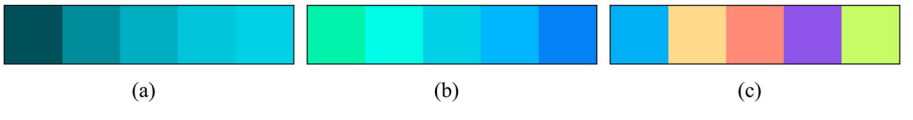


Figure 7. Three colour palates having different colour harmony. (a) Colour combination having a harmony of shade, (b) colour combination having harmony of Hue, (c) random colour combination.

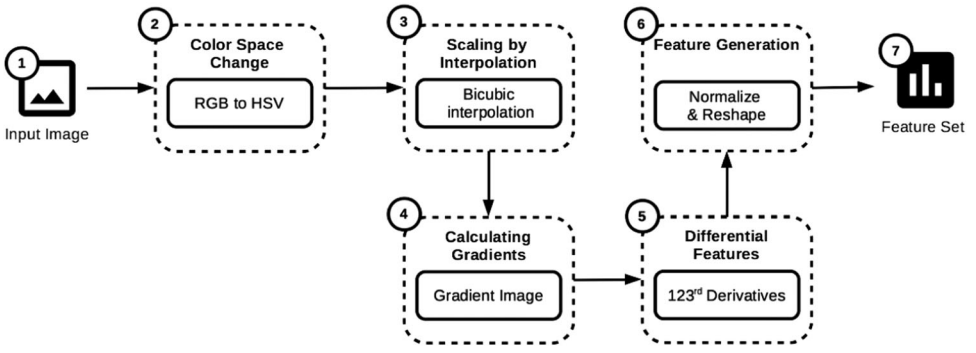


Figure 8. Simple workflow of the HCF module.

are represented using 3 root colours: Red, Green, and Blue. On the other hand, in the HSV scale, Hue is the primary component because it represents the base; saturation holds the degree of whiteness; finally, value holds the degree of blackness.

Figure 9(a) is a sample image and the three following Figure 9(b–d) – represent the RGB component of that sample image accordingly. Figure 9(e–g) represent the HSV/HSL component of the sample image consequently. As the images in our dataset are of variable sizes, we used bicubic interpolation (Keys, 1981) and created a 128 by 128 square image for each image which enabled producing fixed-length feature vector. Figure 10 shows the transformed square image of the given sample image in Figure 9(a).

Finally, we have calculated the rate of change of the colour components. Here, we have calculated a gradient image by averaging the distance of each pixel of the colour components with its surrounding 8 adjacent pixels. We have used Equation (7) on each pixel for each component to create the gradient image.

$$g(x, y) = \sum_{i,j=-1,-1}^{2,2\ i,j\neq 0,0} \frac{cmp(x, y) \ - \ cmp(x+i, \ \ y+j)}{8} \tag{7}$$

In Equation (7), *cmp* is the component variable while *x* and *y* are the iterators of the three components. The following pseudocode is the process of formulating a gradient image.

```
For p_ij in comp:
  For i from −1 to 2
    For j from −1 to 2
      If i, j is not (0,0)
        avg_dist := avg_dist + (comp[p_ij[ − comp]p_i + i, p_j + j])/8
      End if
    End for
  End for
  gra_comp[p_ij] := avg_dist
End for
```

Here, p_{ij} is the iterating point of the colour components and (*i, j*) represents the column and row iterators respectively. *gra_comp* holds the final gradient component of the particular component. Figure 11 shows the gradients, where Figure 11(a) shows the combined

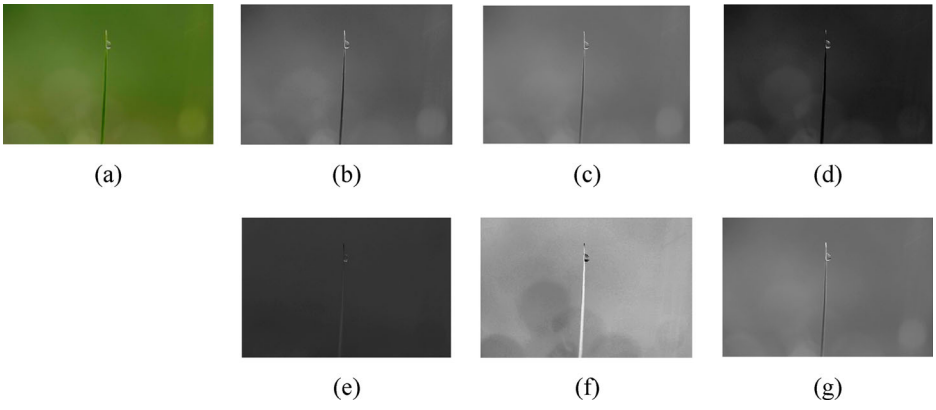


Figure 9. Sample image and colour components. (a) Sample image, (b) red component of RGB, (c) green component of RGB, (d) blue component of RGB, (e) hue component of HSV, (f) saturation component of HSV and (g) value component of HSV.

gradient of three components; Figure 11(b–d) shows the individual gradient components respectively.

Unfortunately, the gradient image itself fails to be a feature itself as it has too large dimensionality. There are 5000 samples and the feature vector length is 3×128^2. Regular learning algorithm will fail to map feature set this large and it will take an unrealistic amount of time to process. Applying differentiation and min–max, it will extract a robust feature-set that contains the information about the change of colour in an image with minor loss and can be used in any algorithm straightforwardly in contrast to the more popular Principal Component Analysis (PCA). However, by applying

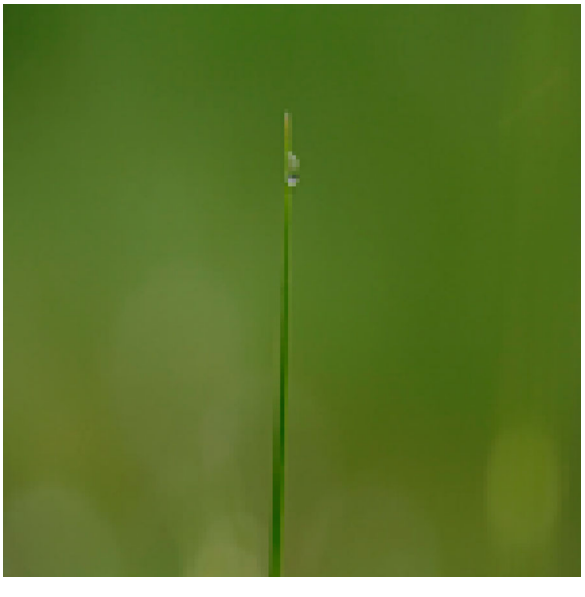


Figure 10. Transformed 128×128 image.

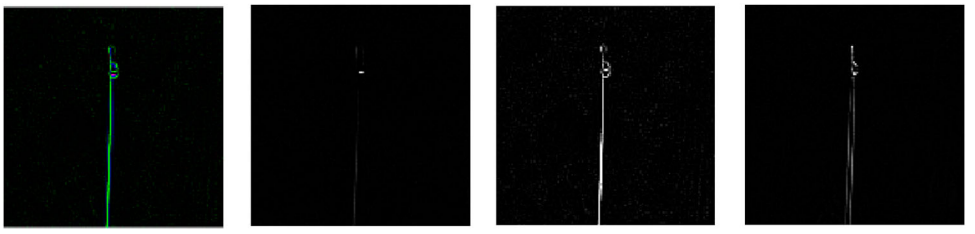


Figure 11. Sample image and colour components. (a) Combined gradient image, (b) gradient image with respect to hue, (c) gradient image with respect saturation, (d) gradient image with respect to lightness.

differentiation, we lose some detail information. Nonetheless, we are only interested in patterns of colour change rate and continuous differentiation preserves this changing information. We have validated in our previous work that, a lower dimension gradient can hold the colour correlation for the entire image. We have used Equation (8) to do this operation.

$$\Delta f \quad = \quad [X_2 - X_1, \quad X_3 - X_2 \ldots, \quad X_m - X_{m-1}] \tag{8}$$

In Equation (8), Δf is the first derivative, X_1 to X_m are the discrete series values of which we do differentiation. In this step, to create a lower dimension gradient image of size 5 by 5, first we have taken the 123rd derivatives and then again 123rd derivatives of the transposed matrix of that calculated matrix. Next, we have taken 123rd discrete differentials first on the transpose of the original image and then again in the transpose of the calculated matrix. Now, we have two 5×5 matrices which we have combined by taking the average. We can choose the order of derivatives based on our requirement of feature length.

Figure 12(a–c) shows the differential matrices that have been derived from each colour component respectively. In the sixth step of extracting harmonic colour feature, we have transformed the 2D matrix of size 5 by 5 to 1D matrix of size 1 by 25. These formulated matrices are our harmonic colour features.

In Figure 13, we can see Harmonic Colour vectors before normalization. In Figure 9 (a), we observe the base colour is green and the most part of the image is constituted with different shades of green. Thus, the rate of change of Hue is less while the other

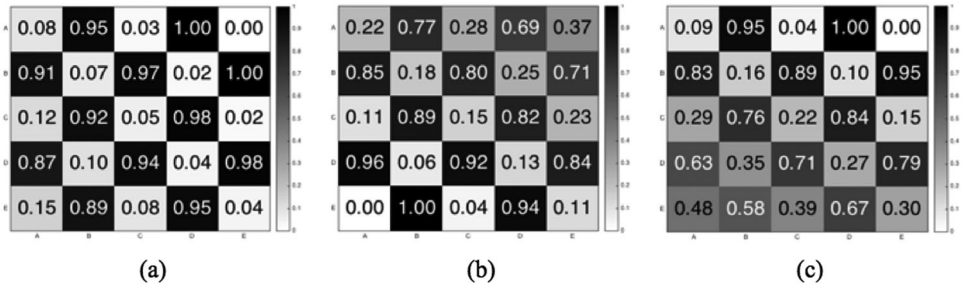


Figure 12. Differential Feature matrices. (a) Hue component, (b) saturation component and, (c) value component.

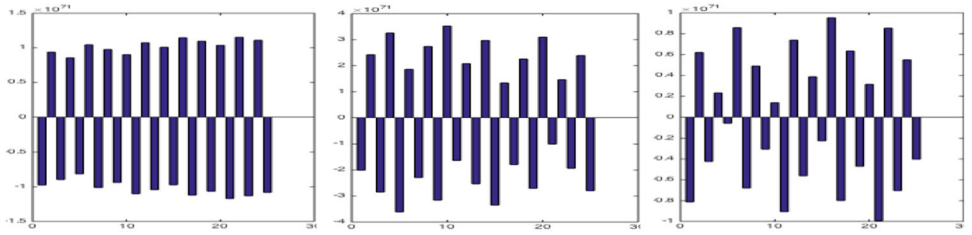


Figure 13. HCF plot. (a) Hue HCF, (b) saturation HCF, (c) value HCF.

two components, Saturation and Lightness, change rapidly which also our Figure 13 shows.

Colour changes rapidly in Figure 14 as well. Figure 14(a) represents the sample image and from Figure 14(b) to 14.d represents the HCF of that sample image. Here we can observe that the rate of change of hue is also rapid.

$$z_i = \frac{x_{i-\min(x)}}{\max(x) \quad - \quad \min(x)} \tag{9}$$

In this step, we have also used the min–max normalization method using Equation (9) to normalize the feature set. After normalization, in the seventh and final step we produced feature set of the score between 0 and 1. At the end of this feature extraction method, we produced 25 HCF for each component and altogether, 75 HCF for all three-colour components. After that, we have pre-processed these features sets using PCA. We have taken 4 components that hold 95% of the variance.

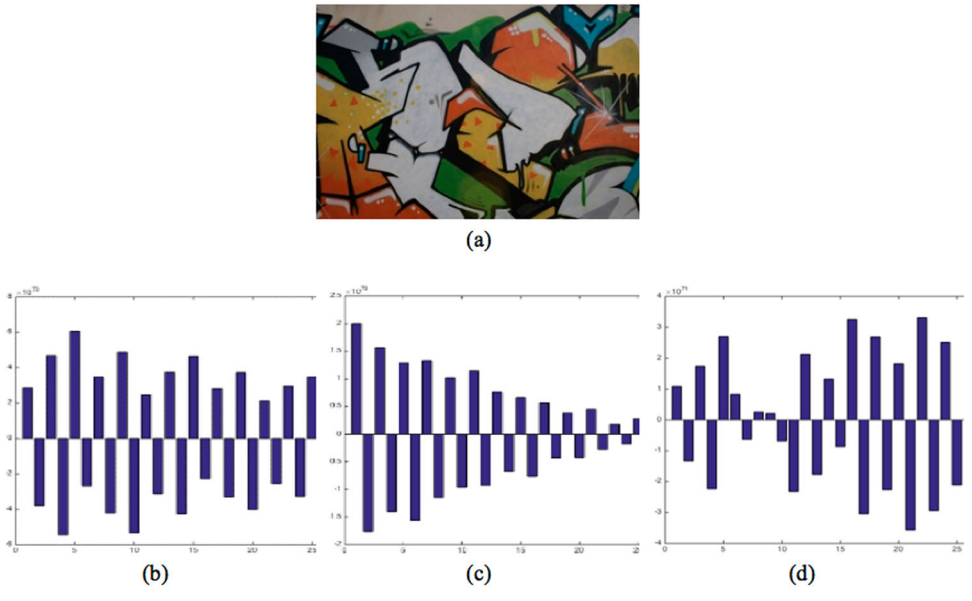


Figure 14. HCF plot. (a) Hue HCF, (b) saturation HCF, (c) value HCF.

4.3. User context feature

The novel work of this research and the part that made a significant improvement in the results of our system is the User Context (UC). The general idea of UC can be explained as follows: suppose a person likes cars and when he/she is shown a picture of car, due to the person's personal preferences, he/she is likely to find the picture appealing although it might not have aesthetics at all. Our research findings show that UC plays a significant role in human perception of aesthetics, and machines ability to perceive aesthetics can be improved greatly by considering UC. Incorporating psychological aspects in analysing aesthetic is a relatively new approach and in this section, we have described the workflow of the UC module of our research.

The simplified idea of UC is: we want to find a pattern in peoples' choices. We have used the category tags of our dataset to understand UC. In our user study (Section 3), we randomly presented every participant 35 pictures where each picture has multiple category tags. The major focus of our user study was distributing the images evenly. As the category count was uneven but significantly greater than the image count, we statistically concluded most participants would receive at least one image from each category even if we select the images randomly. Therefore, each participant is expected to be exposed and scored most categories. Hence, for every participant, we can create a 24-length vector having scores in the range of 1–5 for all 24 categories. These category scores for a user are the average of scores given to different categories. We have produced 374 such vectors for each participant and created the sample dataset for UC analysis. In spite of our statistical assumption, there exist few participants who were not given image from some categories. We imputed the missing category values with the mean of available scores for the respective user. We are considering the participants of the user study as the 'user' of this section of our work.

We have clustered the UC dataset and created user groups based on the choice of preference of each user. By analysing results, and trials and errors, we have selected Hierarchical Agglomerative Clustering (HAC) (Zepeda-Mendoza & Resendis-Antonio, 2013) algorithm to cluster the dataset. We have used Euclidean distance as a distance metric, average linkage as the linkage criteria and created 7 clusters of users and named them 'user groups'. We have plotted truncated dendrogram of the clustered points in Figure 15.

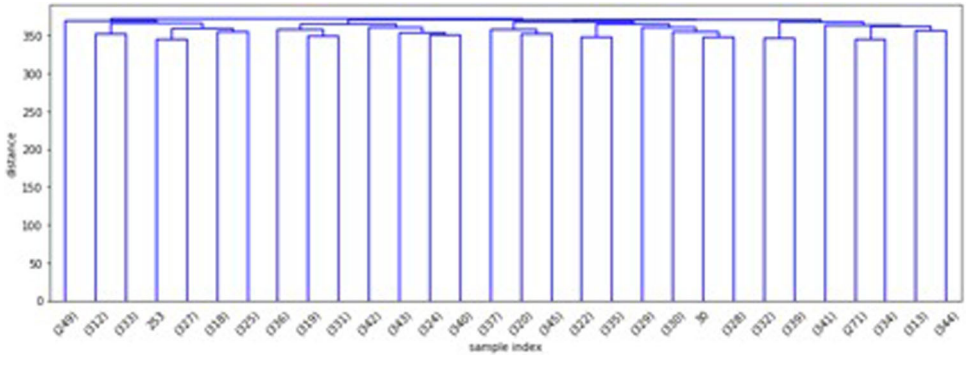


Figure 15. Context explanation and validity.

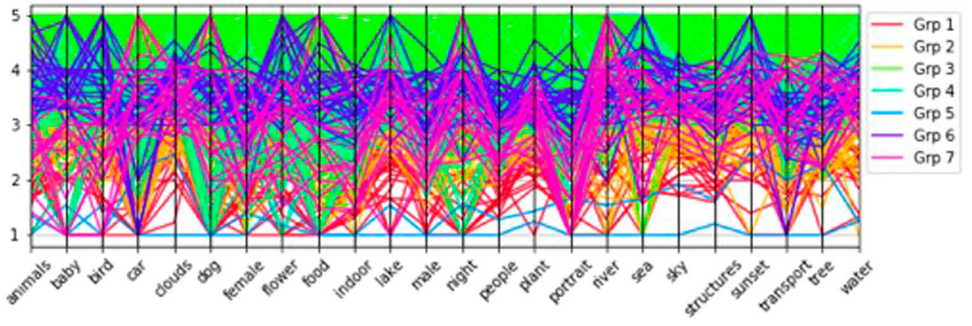


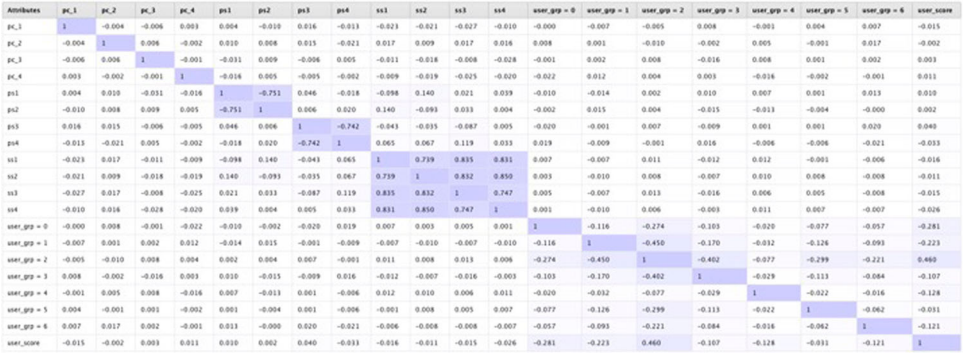
Figure 16. Context explanation and validity by groups and categories/tags.

The *x*-axis of Figure 15 is the sample index and the *y*-axis is the distance of the linkage points. The dendrogram has been truncated to display 30 non-singleton linkages and rest of the linkages have been merged into leaf nodes. We have plotted parallel coordinates plot of cluttered user groups in Figure 16. This plot enabled us to analyse and understand user's choice of preference.

In Figure 16, the vertical axis represents the scores (1–5) of images and the horizontal axis represents categorical divisions of the images. Each line of the figure represents the score given by a user to each category and the colour of the lines represents the user group it belongs. If we observe the figure, we can see some trends involving the groups. For instance, people from group 7 have scored the river category images with much higher score. In contrast, people from the same group have given low scores to those images, which fall into the portrait category. So, from here we can infer that for a group of people who like rivers are less likely to like portraits. Similarly, observing group 6 we can also infer that a group of people who like birds are less likely to like cars. From group 1 we can infer that people who are less fond of lakes are also less fond of nights. These are just a brief overview of this context data. This context data can be analysed further to get a deeper insight into the psychology behind human visual perception but that goes beyond the scope of this work. We have used the clustered group as the UC feature set while training our estimator and perceiving aesthetics.

4.4. Learning model selection and generation

After features extraction, we have created a machine-learning model to perceive aesthetics. We have used the values calculated in the previous sections as our feature set. We have produced set of 8 values (ps1, ps2, ps3, ps4, ss1, ss2, ss3, ss4) as ROT features and set of 4 values as Colour Harmony feature. We have encoded the 7 categorical UC groups to 7 columns (user_grp = 0, user_grp = 1, user_grp = 2, user_grp = 3, user_grp = 4, user_grp = 5, user_grp = 6) by assigning 1 if user belongs to the respective group or 0 otherwise. At the end, we produced our dataset composed of 19 features. Figure 17 shows the correlation between all the features and the user given score. We can observe in the figure, SSRs have a strong correlation between them. This is an unexpected phenomenon in the dataset. In addition, the PSR and SSR have a correlation between them and most importantly, the user group or the user context has a notable correlation with the user perceived aesthetics score.



Attributes	pc_1	pc_2	pc_3	pc_4	ps1	ps2	ps3	ps4	ss1	ss2	ss3	ss4	user_grp = 0	user_grp = 1	user_grp = 2	user_grp = 3	user_grp = 4	user_grp = 5	user_grp = 6	user_score
pc_1	1	-0.004	-0.006	0.003	0.004	-0.010	0.016	-0.013	-0.023	-0.021	-0.027	-0.010	-0.000	-0.007	-0.005	0.008	-0.001	0.004	0.007	-0.015
pc_2	-0.004	1	0.006	-0.002	0.010	0.008	0.015	-0.021	0.017	0.009	0.017	0.016	0.008	0.001	-0.010	-0.002	0.005	-0.001	0.017	-0.002
pc_3	-0.006	0.006	1	-0.001	-0.031	0.009	-0.006	0.005	-0.011	-0.018	-0.008	-0.028	-0.001	0.002	0.008	-0.016	0.008	0.001	0.002	0.003
pc_4	0.003	-0.002	-0.001	1	-0.016	0.005	-0.005	-0.002	-0.009	-0.019	-0.025	-0.020	-0.022	0.012	0.004	0.003	-0.016	-0.002	-0.001	0.011
ps1	0.004	0.010	-0.031	-0.016	1	-0.751	0.046	-0.018	-0.098	0.140	0.021	0.039	-0.010	-0.014	0.002	0.010	0.007	0.001	0.013	0.010
ps2	-0.010	0.008	0.009	0.005	-0.751	1	0.006	0.020	0.140	-0.093	0.033	0.004	-0.002	0.015	0.004	-0.015	-0.013	-0.004	-0.000	0.002
ps3	0.016	0.015	-0.006	-0.005	0.046	0.006	1	-0.742	-0.043	-0.035	-0.087	0.005	-0.020	-0.001	0.007	-0.009	0.001	0.001	0.020	0.040
ps4	-0.013	-0.021	0.005	-0.002	-0.018	0.020	-0.742	1	0.065	0.067	0.119	0.033	0.019	-0.009	-0.001	0.016	-0.006	-0.006	-0.021	-0.033
ss1	-0.023	0.017	-0.011	-0.009	-0.098	0.140	-0.043	0.065	1	0.739	0.835	0.831	0.007	-0.007	0.011	-0.012	0.012	-0.001	-0.006	-0.016
ss2	-0.021	0.009	-0.018	-0.019	0.140	-0.093	-0.035	0.067	0.739	1	0.832	0.850	0.003	-0.010	0.008	-0.007	0.010	0.008	-0.008	-0.011
ss3	-0.027	0.017	-0.008	-0.025	0.021	0.033	-0.087	0.119	0.835	0.832	1	0.747	0.005	-0.007	0.013	-0.016	0.006	0.005	-0.008	-0.015
ss4	-0.010	0.016	-0.028	-0.020	0.039	0.004	0.005	0.033	0.831	0.850	0.747	1	0.001	-0.010	0.006	-0.003	0.011	0.007	-0.007	-0.026
user_grp = 0	-0.000	0.008	-0.001	-0.022	-0.010	-0.002	-0.020	0.019	0.007	0.003	0.005	0.001	1	-0.116	-0.274	-0.103	-0.020	-0.077	-0.057	-0.281
user_grp = 1	-0.007	0.001	0.002	0.012	-0.014	0.015	-0.001	-0.009	-0.007	-0.010	-0.007	-0.010	-0.116	1	-0.450	-0.170	-0.032	-0.126	-0.093	-0.223
user_grp = 2	-0.005	-0.010	0.008	0.004	0.002	0.004	0.007	-0.001	0.011	0.008	0.013	0.006	-0.274	-0.450	1	-0.402	-0.077	-0.299	-0.221	0.460
user_grp = 3	0.008	-0.002	-0.016	0.003	0.010	-0.015	-0.009	0.016	-0.012	-0.007	-0.016	-0.003	-0.103	-0.170	-0.402	1	-0.029	-0.113	-0.084	-0.107
user_grp = 4	-0.001	0.005	0.008	-0.016	0.007	-0.013	0.001	-0.006	0.012	0.010	0.006	0.011	-0.020	-0.032	-0.077	-0.029	1	-0.022	-0.016	-0.128
user_grp = 5	0.004	-0.001	0.001	-0.002	0.001	-0.004	0.001	-0.006	-0.001	0.008	0.005	0.007	-0.077	-0.126	-0.299	-0.113	-0.022	1	-0.062	-0.031
user_grp = 6	0.007	0.017	0.002	-0.001	0.013	-0.000	0.020	-0.021	-0.006	-0.008	-0.008	-0.007	-0.057	-0.093	-0.221	-0.084	-0.016	-0.062	1	-0.121
user_score	-0.015	-0.002	0.003	0.011	0.010	0.002	0.040	-0.033	-0.016	-0.011	-0.015	-0.026	-0.281	-0.223	0.460	-0.107	-0.128	-0.031	-0.121	1

Figure 17. Correlation between all features and targeted score.

We have randomly selected 80% of the dataset or 10,198 data points and created our training dataset, and created our testing dataset with rest of the 20% or 2550 data points. We have further created 3 folds of the training data while training our model and computed cross-validated performance results. A number of estimates have been tried out on the training dataset and based on the performance and overall results, we have selected the optimal estimator and trained our model. Our experiment results in Table 2 show that Gradient Boosted Trees (GBT) model gives us results with maximum accuracy for the given dataset. Therefore, we have used GBT to build our model.

We have used GBT estimator that uses H_2O 3.8.2.6 (Candel & LeDell, 2018) algorithm. We have configured our estimator with a learning rate of 0.0545, 50 trees, maximal depth of 3, and with 10 bins. We have programmatically assigned a combination of different values in the parameters and selected the optimal values of parameters that performs the best. We have plotted the performance improvement of the selected parameters in Figure 18.

Figure 18 illustrates the performance of the estimator for different values of the parameter. The horizontal axis of all four figures represents the error rate of the estimator and the vertical axis denotes the four parameters respectively. The dot on the plots shows the lowest point on the performance line and is the optimal parameter values. We have used these values and produced the model with the maximum performance. Figure 19 shows two boosted gradient trees as a sample from our trained model. These trees are the weak classifiers where leaf nodes are the weak classifier values and the links are the parameter value range.

4.5. Applying model and prediction

We applied our trained model to perceive aesthetics and calculated the accuracy of our estimator. In contrast, by achieving UC's improved accuracy, this feature reduces the generality of the model and concentrates on the applicable domain. The models we generated in our previous research (Firoze et al., 2018; Osman et al., 2018) could directly be applied on any image to perceive aesthetics. However, the model we have created in this research is a user biased model and the machine needs to learn the user first and then it can perceive aesthetics. Figure 20 shows the workflow of applying this model.

The system first needs to present users of the system images from all known categories. Next, the system needs to create a user context profile by finding which cluster group the

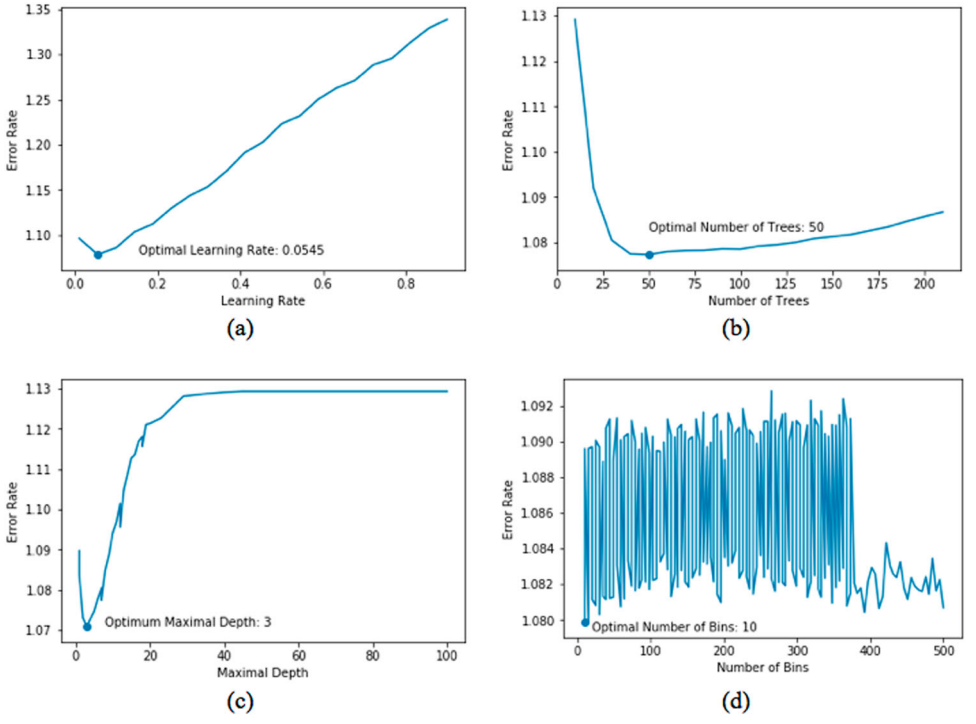


Figure 18. Performance of GBT parameters. (a) Optimization of learning rate, (b) optimal number of trees, (c) optimum maximal depth, and (d) optimal number of bins.

user belongs. After that, the system can perceive aesthetics biased to given user profile. When the system receives an input image, first it conducts feature extraction. Next, the system combines the features with the user profile and applies the trained model to perceive aesthetics.

5. Results analysis

We have calculated and cross-validated error rates of our model to show the performance of our system. Table 1 lists the detailed performance score of our system. In the table, we can see Root Mean Squared Error (RMSE) is 1.062, which means score predicted by our system may differ by one score on average from user given score. We have also

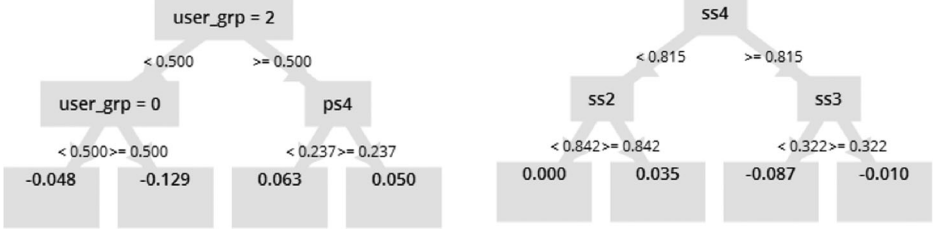


Figure 19. Two sample boosted tree.

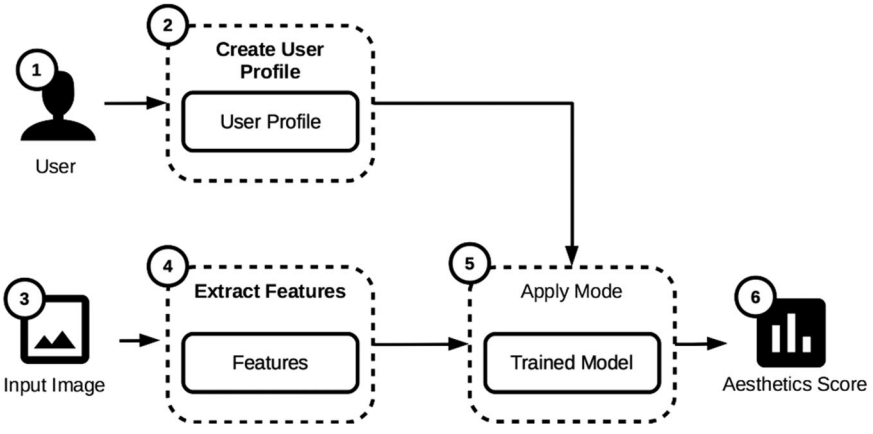


Figure 20. Workflow of perceiving aesthetics using our model.

experimented with other machine-learning models and in Table 2 we have listed the performance score of these models. Looking at the table, we can see that GBT has the lowest error rate of all the models we have experimented with. That is why we have selected GBT to train our system and because of space constraints on this research paper, we have only elaborated the building blocks of GBT.

We have also compared results generated by our model – UC Model – with No-UC Model – our prior work (Firoze et al., 2018; Osman et al., 2018) – and with CNN model presented in research work (Kong et al., 2016). We have randomly created equal sized training and equal sized testing data for the three models. The reason behind creating 3 different datasets is because UC Model acts on the 12,000+ dataset collected by user study whereas the other 2 models work 5000 data composed of the average score of each image given by users. This is because CNN and No-UC Model does not understand the concept of 'users'. For an instance, if we select a score of one image from dataset of CNN and No-UC model then we will have 2–3 scores given by users for that image in the dataset of UC model. If we select 1000 image scores from dataset of CNN and No-UC model, we will have 2000–

Table 1. Performance of our system

Performance Vector	Score
RMSE	1.062 +/− 0.000
Absolute Error	0.837 +/− 0.654
Relative Error Lenient	20.98% +/− 18.20%
Correlation	0.512
Prediction Average	3.702 +/− 1.236

Table 2. Performance of different Machine-learning Models

Model	RMSE	Relative Error Lenient
Gradient Boosted Trees	1.062	20.98%
Generalized Linear Model	1.08	21.30%
Deep Learning	1.097	21.40%
Decision Trees	1.109	21.50%
Random Forest	1.114	21.70%

Table 3. Performance of different Machine Learning Models

Model	RMSE	Absolute Error	AIC Base	AIC	BIC Base	BIC	R^2
User Context Model	1.062	0.837	4786.318	4353.172	4796.944	4369.111	0.2518093
No-user Context Model	1.094	0.858	4062.172	4062.042	4072.798	4077.982	0.001418694
CNN Model (Kong et al., 2016)	2.617	2.453	4107.048	4010.144	4117.674	4026.084	0.06380895

3000 scores in the dataset of UC Model for those exact images. Therefore, it is obvious these 3 models cannot work on the same dataset in our given scenario and the best possible solution is to randomly create 3 equal sized different datasets for the three models and compare based on that.

Next, we have computed, compared, and presented the RMSE and Absolute Error of these three models in Table 3. This table at a glance shows us UC Model has the best performance and lowest error rate where CNN Model has the highest error rate in this problem domain. We have also calculated AIC, BIC, and R Squared (R^2) of a perceived score of test dataset of three models against user perceived score in Table 3. The baseline for AIC and BIC is slightly different as we have used three different datasets for evaluation of the performance of the three approaches. By definition, more the AIC and BIC drop from the baseline better influence the model has on the user given score. No-UC Model's drop of AIC is 0.13, CNN Model's drop AIC is 96.9 and UC Model's drop of AIC is 433.15 from the respective base AIC score. This certainly indicates the UC Model has a superior influence on the user perceived score. A similar outcome can be seen if we see the drops of BIC. Hence, observing the significant drop of UC Model, we can claim UC Model is the most influential model. Table 4 holds the statistical confidence interval difference of No-UC and CNN model (Kong et al., 2016) compared with the UC model when the confidence is 95%. As the intervals don't span over 0 or a significantly small number, we can statistically claim that the observed difference is significant.

We have further analysed the estimations of the three models using Figure 21(a–f). We plotted the histogram of UC Model in Figure 21(a), the histogram of No-UC Model in Figure 20(c) and CNN Model in Figure 21(e). The *x*-axis of the histograms represents score which represents both: the user and system given score. The *y*-axis represents the count of scores, and the colour of the bars represents the user given score and the system perceived score. The bars have slightly transparent colours; so, the colours are slightly different when the bars overlap. In Figure 21(a), we can see machine perceived score count is similar to human perceived score; for example, the users have given score 4 to around 800 images out of 2550, where the machine has given score 4 to around 600 images. Hence, we can abstractly conclude that UC Model's estimated score count is consistent with the user given score count to a high extent. In Figure 21(c) (model without UC), we can observe a very large number of images has been scored between 2 and 3.5 where the system has scored very few images in that range. Therefore,

Table 4. Confidence Interval compared with UC Model at 95% Confidence Level

Model	Confidence Interval
No-user Context Model	$0.0064 \pm 0.0001527461588$
CNN Model (Kong et al., 2016)	$0.311 \pm 0.00018352830376$

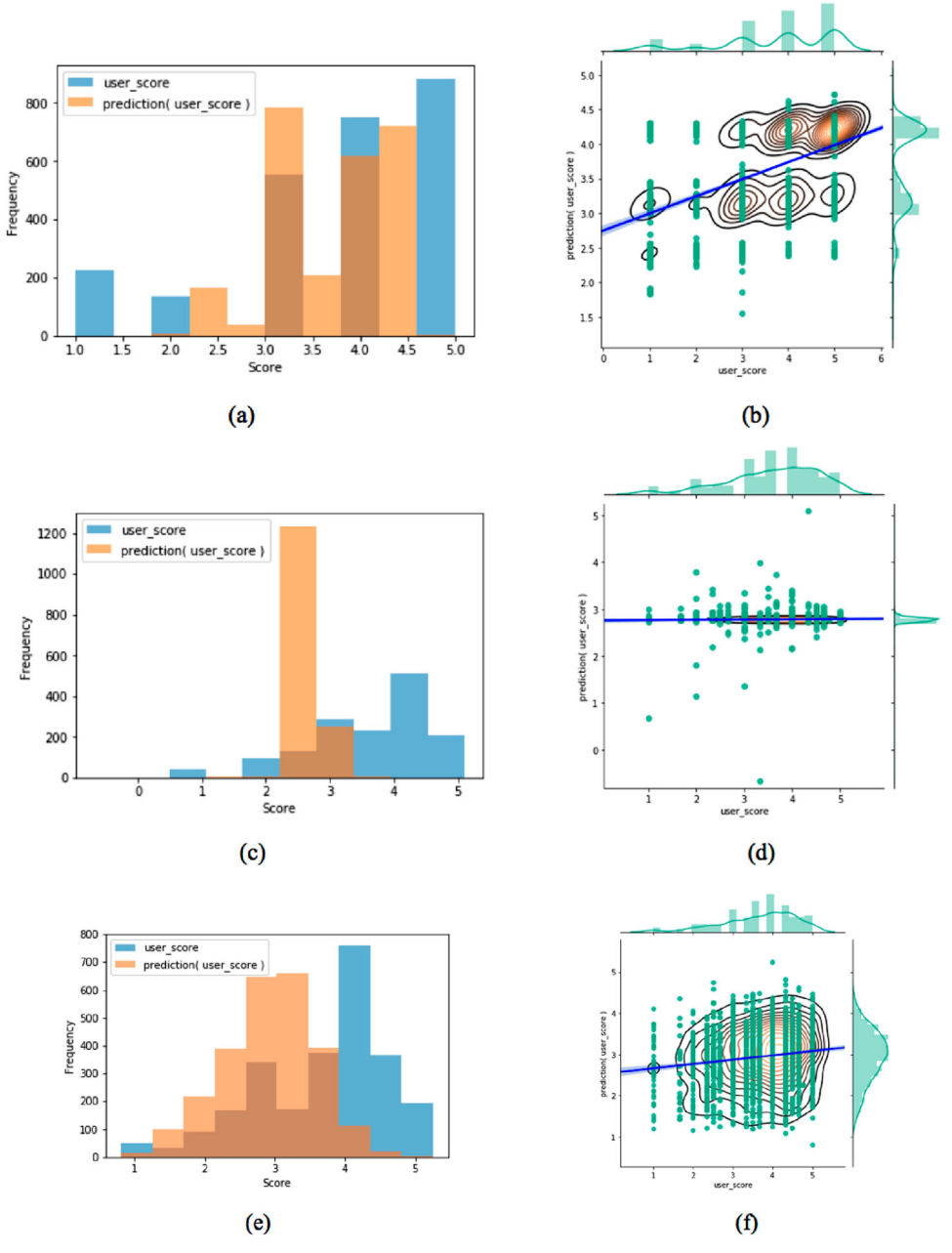


Figure 21. Analysis and comparison of different model. (a) Histogram of Model with User Context, (b) Scatter plot of Model with User Context, (c) Histogram of Model without User Context, (d) Scatter plot of Model without User Context, (e) Histogram of CNN Model, and (f) Scatter plot of CNN Model.

we conclude that the predicted scores of this model is inconsistent with the user given score and machine flatly scored most images. In the histogram of CNN Model, we can observe machine perceived score is normally distributed in a bell shape. It has scored

very few images in the between 4 and 5 where a large number of images has been scored in that range by users.

We have scatter plotted the perceived score against user given score to get a deeper understanding of the three predictions. Figure 21(b) represents scatter plot of UC Model, Figure 21(d) represents scatter plot of No-UC Model, and Figure 21(f) represents scatter plot of CNN Model. In these scatter plots, the user given score is represented in the *x*-axis, machine perceived score is represented with the *y*-axis. We have stacked the histogram (green bars) and contour (green curves) of user given score at the top edge of the plots and stacked the histogram and contour of machine perceived score at the right edge of the plots. Looking at the stacked histograms, we can have a rough idea about how many scores overlaps on a given data point. We have plotted 2D bivariate kernel density estimation (yellow counters) along with the scatter plot (green points). 2D Counters of this plot represent the density of the score count. Density is represented by brightness and number of loops. We have also plotted a regression line (the lightest column) on the scatter plot that goes through the average score count of scores. Density and stacked histogram of UC Model show that a large number of images has been scored 5 and our machine successfully maps majority of those images. This density also shows us that our machine scored a large number of images in between 3 and 4 where the user has scored those images 4, and the system scored images between 2.75 and 3.5 where users have scored those images 3. These estimations are very promising and consistent with user's perception. However, if we observe the score 1 and 2 we can see the predicted scores are denser quite above the user given score. We can conclude from this: a user belonging to a certain UC group can make the user like certain type of images; but the opposite scenario, a user belonging to a certain group does not necessarily make him not like certain type of images. This outcome supports our previous works where we have shown having a certain feature can increase the likelihood of having a good aesthetics in human eyes but not having a certain feature does not mean aesthetics is less in human eyes. Regression line on the plot shows us our system can successfully map scores above 2 with great accuracy.

If we observe the plot for No-user Context in Figure 21(d), we can see the model has scored most of the images in the range of 2.5–3.5. Although the RMSE score 1.094 of this model, the RMSE picture is considerably good, but the estimation is bad. As the users have scored many images in the range of 2.5–4.5, the error rate of this model came out good. The regression line has no slope at all and it shows us there is no overall predicted score change when user score changes. Regression line of CNN Model in Figure 21(f) has a slightly better rate of change; it changes from 2.5 to 3.5 as the user given score changes from 1 to 5. The density of this plot shows us that predicted scores are spread across the plot having a high density approximately centring at user given score 4 and system perceived score 3. We can tell from the plot that this model has an evenly large number of miss-estimation that counter the correct-estimations, which also explains the reason behind having a high RMSE score of 2.617.

We can conclude from these 3 scatter plots, the ideal regression line of estimation is a diagonal line having a 45°slope and the regression line of the UC Model is closest to the ideal line. Analysing these plots, we can conclude UC Model outperforms other two models with a great magnitude.

6. Discussion

We have successfully designed a learning model that can perceive aesthetics from user's perspective with great accuracy in this research work. Although the applicability of this model in the real-world application might be questioned, this is not the case in reality. Most recommendation system in the real world interacts with users discretely and it is a very common scenario this system already maintains user profiles for a better recommendation. Therefore, this system can be instigated with most real world applications and can be made operational in any required field of application. As an instance: Shutterstock®, 500px, GraphicRiver, etc. online graphics and photograph marketplace can easily integrate our system with their platform to provide precise recommendations. Furthermore, social medias involving images, i.e. Instagram, Facebook, Twitter, etc., can easily deploy our system to provide better content recommendations.

Although, many research has been done (Datta & Wang, 2010; Firoze et al., 2018; Kong et al., 2016; Osman et al., 2018, etc.) with the attempt to enable machines to perceive beauty but their applicability is limited to the research labs. Many of these experiments have promising results but these outcomes are not reliable enough for real-world applications. We have seen in seen in Section 5 although two of the selected approaches have acknowledgeable results, their graphical representation and statistical analysis do not agree accordingly. Considering these issues, we set out to find and develop a robust, practical, and real-word applications ready solution that can be adopted in any required applicable domain.

The outcome of this research is not only applicable to perceive aesthetics only but also other domain of applications that deals with human vision by incorporating the features extraction and procedures of this research. Along with learning and understanding cognition of aesthetics, this research introduces a new way of looking at the problem – Machine's perception of Aesthetics – by taking the study of psychology into account of Machine Learning.

7. Future work

We are currently working on two more image compositional metrics to improve machines perception of aesthetics. First of this metrics is repeated pattern, where the base idea states that an image is more appealing to human eyes if it has some repetition in patterns. We already made notable progress by analysing counter loops in images. Second compositional metric is symmetry. The fundamental idea of this metric is: an image creates a better visual ambiance if there exists symmetry in the image. We are analysing this phenomenon by considering the key points of images. We hope to design a learning model that outperforms the accuracy of our present work in the near future.

Acknowledgments

We would like to thank Professor John Kender (http://www.cs.columbia.edu/~jrk/), Professor of Computer Science at Columbia University for his insights.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was funded by North South University ICTRG 42nd grant (51st Syndicate) for the fiscal year 2016–17.

Notes on contributors

Tousif Osman is currently working as Software Engineer at Dutch-Bangla Bank Ltd. Though out his career and education he associated with number of researches and gained a profound expertise in learning systems, adaptive algorithms, data mining, visual data analysis and mining, computer vision etc. fields. He is very passionate about exploring and learning new domains. Currently his is planning to explore the domain of Information Security and its adaptive solutions. Tousif Osman received his bachelors on Computer Science and Engineering and is going to join his master's program. When he is not glued to his work and research he is a book worm, food lover, and enjoys traveler with his family and friends. If you want, you can reach him at tousif.osman@northsouth.edu.

Shahreen Shahjahan Psyche is currently pursuing her Master's in Computer Science at University of Texas at Dallas. Her primary research interests are in Machine Learning, Big Data, Reinforcement Learning, Fuzzy Systems, Computer Vision, Natural Language Processing and Blockchain. During and after her Bachelor's degree, she was directly involved in various relevant research. Shahreen is fascinated by the opportunities that can be pursued in this era of Artificial Intelligence and really eager to be a part of this movement which can very well shape the world in decades. Shahreen has completed her Bachelor's degree from North South University in 2016. After that she worked as a Lab Instructor and as a Research Assistant in North South University for 2 years. shahreen.psyche@northsouth.edu, shahreen.psyche@icloud.com.

Tonmoay Deb, currently a third-year undergraduate student majoring in Computer Science & Engineering at North South University. Apart from academic studies, he conducts active research in the field of Machine Learning, Computer Vision and Nature Language Processing domain. Recently, his independent paper has been accomplished a prestigious recognition at The Undergraduate Awards in addition to acceptance to IEEE ICMLA 2018 conferencein Orlando, Fl, USA. Previously, he worked as a Research Assistant on a funded project "Computational Aesthetics Perception" under the supervision of Mr. Adnan Firoze and Dr. Mohammad Rashedur Rahman. This passionate researcher has already published several papers with established publishers e.g., ACM, IGI-Global, Springer. His works can be found at https://sites.google.com/site/tonmoay.

Adnan Firoze is a Core Faculty Member at North South University, Bangladesh and formerly a Teaching Fellow at the Computer Science Department in Columbia University in the City of New York. He completed his Dual M.S. in computer science and journalism in 2016 with distinction. He graduated summa cum laude in B.S. in Computer Science from North South University, Dhaka, Bangladesh in 2012. After that he worked at Computer Vision and Cybernetics Group, Bangladesh (http://www.cvcrbd.org/researchers). His interdisciplinary research works are based on digital image processing, machine learning, fuzzy logic, neural networks and data mining. His previous research works have appeared in numerous prestigious conferences and journals, namely, IEEE's 2012 International Conference on Machine Learning and Cybernetics (ICMLC), ACM's 13th International Conference on Enterprise Information Systems (ICEIS), International Journal of Healthcare Information Systems and Informatics (IJHISI), to mention a few. In 2015, he co-authored a book chapter on hospital surveillance data analysis in Springer's 'Intelligent Information and Database Systems'.

Rashedur M. Rahman is working as a Professor in Electrical and Computer Engineering Department in North South University, Dhaka, Bangladesh. He received his PhD in Computer Science from University of Calgary, Canada and Masters from University of Manitoba, Canada in 2007 and 2003 respectively. He has authored more than 150 peer reviewed research papers in journals or conference proceedings in the area of parallel, distributed, grid and cloud computing, knowledge and data engineering. His current research interest is in data science, data replication on grid, cloud load characterization, optimization of cloud resource placements, computational finance, deep

learning, etc. He has been serving on the editorial board of a number of journals in the knowledge and data engineering filed. He also serves as a member of organizing committee of different international conferences.

ORCID

Tousif Osman http://orcid.org/0000-0003-3878-8385
Adnan Firoze http://orcid.org/0000-0002-2751-009X

References

Amirshahi, S. A., Hayn-Leichsenring, G. U., Denzler, J., & Redies, C. (2014). Evaluating the rule of thirds in photographs and paintings. *Art & Perception*, *2*(1-2), 163–182.

Candel, A., & LeDell, E. (2018). *Deep learning with H_2O* (6th ed.). Mountain View: H2O.ai.

Caplin, S. (2008). *Art and design in photoshop: How to simulate just about anything from great works of art to urban graffiti*. Hoboken: Taylor and Francis.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In *Computer vision – ECCV 2006* (pp. 288–301). Berlin: Springer.

Datta, R., & Wang, J. Z. (2010). *ACQUINE: Aesthetic quality inference engine – real-time automatic rating of photo aesthetics*. Proceedings of the International Conference on Multimedia Information Retrieval (pp. 421–424). New York, NY, ACM.

Firoze, A., Osman, T., Psyche, S. S., & Rahman, R. M. (2018). *Scoring photographic rule of thirds in a large MIRFLICKR dataset: A showdown between machine perception and human perception of image aesthetics*. Asian Conference on Intelligent Information and Database Systems (pp. 466–475), Cham, Springer.

Harel, J., Koch, C., & Perona, P. (2007). *Graph-based visual saliency*. Advances in neural information processing systems (pp. 545–552).

Huiskes, M. J., & Lew, M. S. (2008). *The MIR Flickr retrieval evaluation*. Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval (pp. 39–43), New York, NY, ACM.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10-12), 1489–1506.

Keys, R. (1981). Cubic convolution interpolation for digital image processing. *IEEE Transactions On Acoustics, Speech, And Signal Processing*, *29*(6), 1153–1160.

Kong, S., Shen, X., Lin, Z., Mech, R., & Fowlkes, C. (2016). Photo aesthetics ranking network with attributes and content adaptation. In *Computer vision – ECCV 2016* (pp. 662–679). Amsterdam: Springer.

Lu, X., Lin, Z., Jin, H., Yang, J., & Wang, J. Z. (2015). Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, *17*(11), 2021–2034.

Lu, P., Peng, X., Yuan, C., Li, R., & Wang, X. (2016). Image color harmony modeling through neighbored co-occurrence colors. *Neurocomputing*, *201*(Suppl. C), 82–91.

Lu, P., Peng, X., Zhu, X., & Li, R. (2016). An EL-LDA based general color harmony model for photo aesthetics assessment. *Signal Processing*, *120*(Suppl. C), 731–745.

Mai, L., Le, H., Niu, Y., & Liu, F. (2011). Rule of thirds detection from photograph. Multimedia (ISM), 2011 IEEE International Symposium (pp. 91–96). Warsaw, Poland.

Maleš, M., Heđi, A., & Grgić, M. (2012). Compositional rule of thirds detection. ELMAR, 2012 Proceedings (pp. 41–44). Zadar, Croatia.

Osman, T., Psyche, S. S., Deb, T., Firoze, A., & Rahman, R. M. (2018). Differential color harmony: A robust approach for extracting Harmonic Color features and perceive aesthetics in a large dataset. International Conference on Big Data and Cloud Computing (ICBDCC'18), Springer Karunya University, Coimbatore, India.

Peterson, B. (2003). *Learning to see creatively: Design, color, and composition in photography*. New York, NY: Amphoto Books.

Phan, H., Fu, H., & Chan, A. (2017). Color orchestra: Ordering color palettes for interpolation and prediction. *IEEE Transactions On Visualization And Computer Graphics*, *24*(6), 1942–1955.

Stone, T. L., Adams, S., & Morioka, N. (2008). *Color design workbook: A real world guide to using color in graphic design*. Kentucky : Rockport Pub.

Weisstein, E. W. (2002). *Golden ratio*. Retrieved from http://mathworld.wolfram.com/GoldenRatio.html

Zepeda-Mendoza, M. L., & Resendis-Antonio, O. (2013). Hierarchical agglomerative clustering. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), *Encyclopedia of systems biology* (pp. 886–887). New York, NY: Springer.