Image Aesthetics Assessment using Deep Learning (based on High Level Image Attributes)

Madhura Phatak
Research Scholar
Department of CSEGH Raisoni College of Engineering
Nagpur, India.

Dr. Prashant Borkar
Department of CSE
GH Raisoni College of Engineering
Nagpur , India.

***Abstract*— Aesthetic assessment of images has been getting a lot of attention for the past decade in the field of computer vision. Large amounts of social media and advertising data in the form of images is continuously analyzed to assign it an aesthetic quality value to improve businesses as well as for gaining more popularity across the web. Visual perception by humans cannot be fully replicated by a machine and continuously more work is being published on aesthetic classification of images. In this paper, we have presented a convolutional neural network model which automatically extracts high level features and distinguishes a set of images into pleasing and non-pleasing categories. Our dataset has been compiled from a variety of sources on the web to make it as diverse as possible. Compared to the traditional handcrafted methods and other machine learning models, our CNN model has provided a better classification accuracy of 68% on our dataset.**

Keywords*— *Machine learning, CNN, deep learning, image aesthetic classification, high level attributes

I. INTRODUCTION

In recent years, the assessment of image aesthetics has received a great deal of attention. Evaluation methods of image aesthetics are largely dependent on successful aesthetic features. The conventional approach uses hand-crafted features to test aesthetics. Nowadays, though, research focuses more on the Convolutional Neural Networks for more precise visual evaluation of aesthetics assessment [1]. Image Aesthetic Assessment refers to the evaluation of images on the basis of high and low quality of aesthetic value of the picture. It plays an important role in the field of computer vision which deals with automatic extraction and analysis of useful information from a picture or sequence of pictures. What makes this task a little challenging is that aesthetic value of an image may be differently perceived by each individual and thus automation of this process may not be very adept at accurately judging the aesthetic quality of an image.

An artistic picture can depend on the scenes as well as the objects depicted. Several pieces of research and studies based on image aesthetics estimation have been found in the past. The goal is to automatically segregate images into appealing and unattractive pictures [2]. The very first approach is to identify a collection of image features that they deem would affect the photographs ' aesthetic quality, and then develop some mathematical models to extract these features. However, this hand-crafted function is not only difficult but also inadequate to take into account the complete and fused essence of the aesthetics of photographs. This could therefore lead to incorrect evaluations [3].

Another important approach is the modelling of image features, using deep learning. In particular, compared to the conventional approach, using the Convolutional Neural Networks (CNN) that have been trained in a large-scale image database can be used to obtain more specific photo quality assessments. These techniques based on self-operating feature modelling generally perform well in terms of concluding whether the given image is aesthetically pleasing or not [4]. Different image properties or attributes affect the image, such as a low-level attributes, middle-level attributes, and high-level attributes [5].

Many visual features play a part in the perception of an image by a human. There are two basic categories of such features among others i.e. Low Level and High-Level features or attributes.

A. Low Level Attributes

Low-level features include spatial characteristics such as edge density, straight-edge density and entropy, and color characteristics such as hue, saturation, brightness and texture [6].

B. Middle Level Attributes

These are the objects in an image. Presence of Salient Object is an important object that includes a large object well separated from the background.

C. High Level Attributes

High level attributes are built on top of low-level attributes to detect larger shapes and objects in an image.

- *Rule of Thirds:* The main subject lies on the intersection or on the lines of a 3x3grid.
- *Depth of Field:* The region of interest is in focus and background is blurred, often to emphasize on the object of interest.
- *Opposing Colors:* Color pairs of opposing hues are prominent in the image[6].

II. SYSTEM ARCHITECTURE

As shown in Figure 1, we shall see the system architecture step by step starting from creating an image repository, feature extraction, the CNN framework, classification process that classifies image as appealing and

non-appealing.

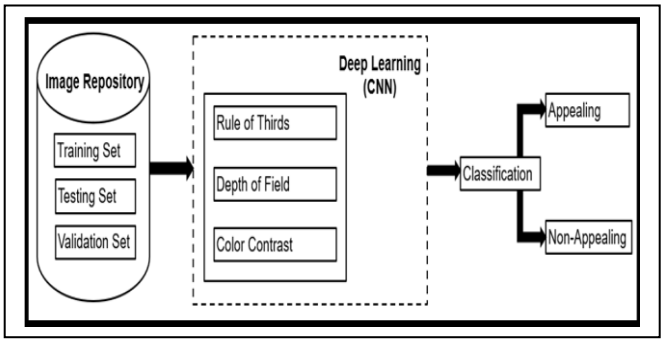

Fig. 1. System Architecture

III. APPROACH

We have considered a unique combination of high-level attributes, that is, the Rule of Thirds, Depth of Field and Color Contrast to be the key factors in evaluating our model.

A. *Handcrafted Methods*

a) Color Contrast: The abstract mathematical model is a color model. This color model that explains how colors can be interpreted as numbers tuples, usually three or four values or color components[7]. The perception of an image being pleasing or not pleasing depends on the colors or a combination of colors used in that image. Contrast is the difference between the color and brightness of the object and color and brightness of the background. This color difference makes the focused object distinguishable [8][9]. We have extracted the foreground and background of an image using the Grabcut algorithm. Further, we are calculating the RGB mean values of the extracted foreground and background images and taking their difference. The difference value of RGB mean values is compared with a given threshold and segregated into high color contrast images and low color contrast images.

High Contrast images are usually the images that have two contrasting colors for object and the background. These images will contain dark shadows and bright highlights. The figure below is a high contrast image.

Low Contrast images are do not have a big difference between their shadows and highlights. These images use supplementary colors on the color wheel. The lack of brightness may result in a flat image as the object and background may not be distinctly visible. Fig 2, depicts the background and foreground separation in an image.

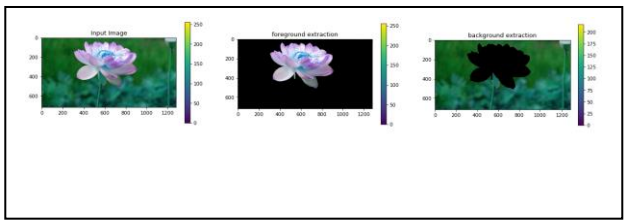

Fig. 2. Snapshot for foreground and background extraction using GrabCut Algorithm

b) Rule of Thirds[ROT]: Image composition is an important aspect in the photo Aesthetics approach. Professional photographer divides an image using horizontal lines and vertical lines. Photographers are forced to place the object at one of the intersections. This visual element is known as Rule of Thirds(RoT)[10]. Under this Rule of Thirds, if the pictures follows above mentioned rules, these would appeal to the eyes[11]. It is the positioning of the important elements in your scene along those lines, or at the points where they meet for better visual results[6]. An off-centre composition is more pleasing to the eye and looks more natural than one where the subject is placed right in the middle of the frame. It also encourages you to make creative use of negative space, the empty areas around your subject[12]. We have extracted the foreground and back-ground of an image using the grabcut algorithm. For Rule of Thirds, we use the foreground extracted image. On this image, we first calculated the centroid of the object and then the intersection points of the grid. Next, we checked the distance of the centroid from the intersection points. Based on a specific given threshold, the image is classified as either it follows RoT or it doesn't follow RoT. We apply following steps to find Rule of Thirds (RoT)as depeicted in Figure 3.

Step 1: Extract Foreground of an image.
Step 2: Convert the image first into greyscale and find its Contour.
Step 3: Find the Centroid of the Contour.
Step 4: Create 3X3 grid and find its intersection points.
Step 5: Find the distance between Center and its intersection point of the grid.
Step 6: If the minimum distance is less than the given threshold, then image follows ROT else it doesn't follow ROT.

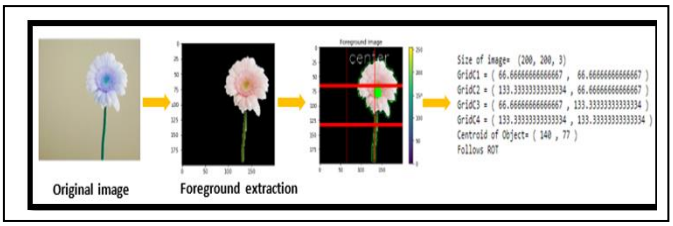

Fig. 3.1. Snapshot of an image that follows ROT

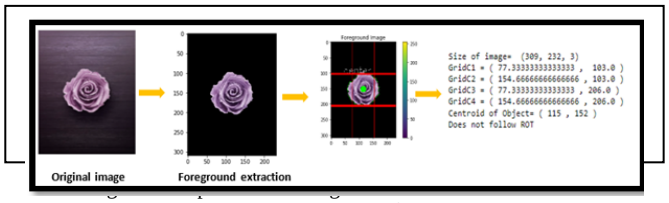

Fig. 3. Snapshot of RoT

b) Depth of Field[DOF]: Depth of Field grabs the attention of Viewer[13]. In depth of field more focus is on subject area and background is blur. as photographer can take decision about which scene elements should be in sharp focus and which should be out of focus. Using this artistic tool called Depth of Field, Photographers can grab attention of Viewer and also can affect the mood of photograph. In traditional cameras, decision such as which region of interest

should in sharp focus and which should be out of focus, are made by controlling the focus distance as well as lens' aperture[14].

The region of interest is in sharp focus and the background is blurred in a depth of Field. It may be more effective, emphasizing the subject while de-emphasizing the background[6][15]. Three main factors that will affect how you control the depth of field of your images are: aperture (f-stop), distance from the subject to the camera, and focal length of the lens on your camera[16].

In every picture there is a certain area of your image in front of, and behind the subject that will appear in focus[17]. We apply following steps,

Step 1: Extract background of an image using Grab cut algorithm.

Step 2: Convert the image into Greyscale.

Step 3: Use Laplacian function on the Greyscale image to find blur value.

Step 4: If the blur value is less than the given threshold value then image follows DoF else it doesn't.

Here, threshold value is set on the basis of trial and error.

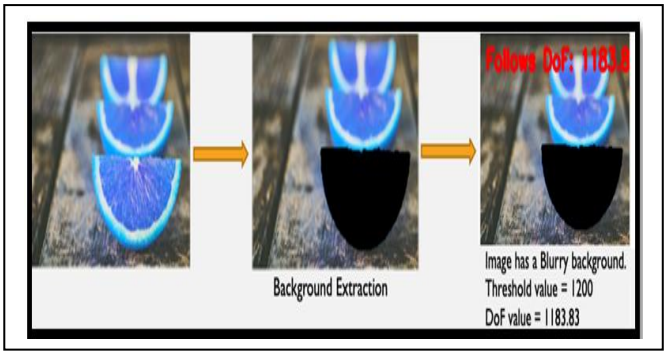

Fig. 4.1. Snapshot of an image that follows DoF

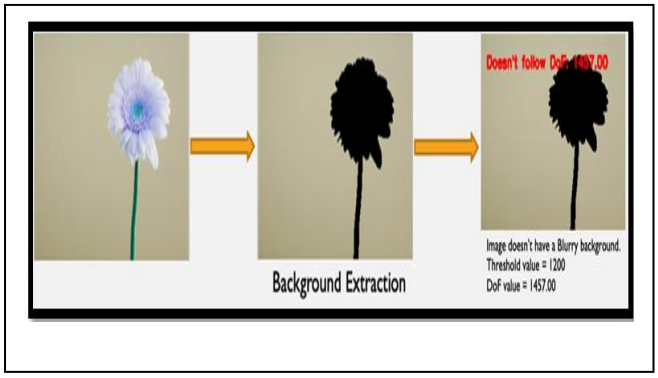

Fig. 4.2. Snapshot of an image that does not follows DoF

Fig. 4. Snapshot of Depth of Field[DoF]

B. *CNN [Convolutional Neural Networks]Approach*

The number of images has increased explosively over the last few years through the social network. Recently, aesthetic evaluation of both images has attracted the attention of many researchers. Image aesthetic assessment can help people choose or filter beautiful images from the crowd, or say appealing images [18]. Aesthetic assessment is an abstract field. Some visual features include the aesthetic evaluation of images. Visual characteristics are those that directly affect an individual's visual perception. This visual feature includes color, texture, shape, edge detection, pixel level control [19].

Convolutional Networks are multi-stage, trainable architectures composed of multiple stages. The input and output of each stage are array sets, called function maps. For example, if we fed input a color image as input, then each feature map would be a 2D array. This 2D array contains a color channel of an inputted image. On the output side, each feature map represents an extracted feature at all input locations. Each stage consists of four layers: a convolution layer, ReLU, a pooling layer and a fully connected Layer. A typical Convolution layer is composed of total three-layer stages, which is followed by classification module [20].

a) Dataset Building: The dataset consists of 6004 images gathered from a wide range of heterogeneous sources. We have considered the following High-level attributes - Rule of Thirds, Depth of Field, and Color Contrast. We have 2000 images of each category with 4 extra images, following as well as not following each of the above rules. We have further divided the dataset into Appealing and Non-Appealing for the basis of training the model based on above attributes. The dataset is split randomly into training and testing sets. The Training set consists of 4202 images and the Testing set consists of 1802 images (0.3 of total number of images). For preprocessing, we have converted the images to grayscale, resized the images into size (128*128) and standardized the dataset by scaling.Figure 5 shows a snap-shot a dataset.

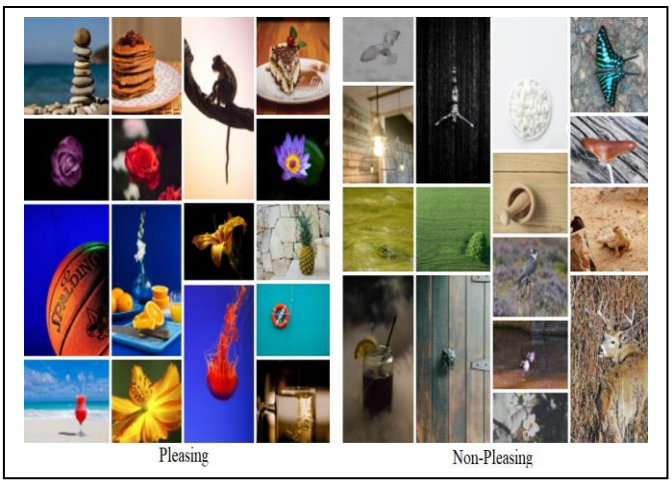

Fig. 5. The dataset consisting of Pleasing and Non-Pleasing images.

b) CNN Design: In this section, we will give the detailed explanation of the design and architecture of CNN model. The model receives black and white images of size 128*128 as input and has five convolution layers each followed by Max Pooling (of size 2*2). All these layers contribute in the automatic feature extraction for the three high level features considered. This is followed by 2 fully connected layers. All these layers use rectified linear activation function except the output layer. The output layer uses sigmoid activation for 2-class classification into Pleasing

and Non-Pleasing classes. The number of filters in the convolutional layers are 64,64,128,128 and 64 respectively. The kernel size of the convolutional layers is (7*7), (3*3), (3*3), (7*7) and (5*5) respectively.Figure 6 shows the CNN architecture.

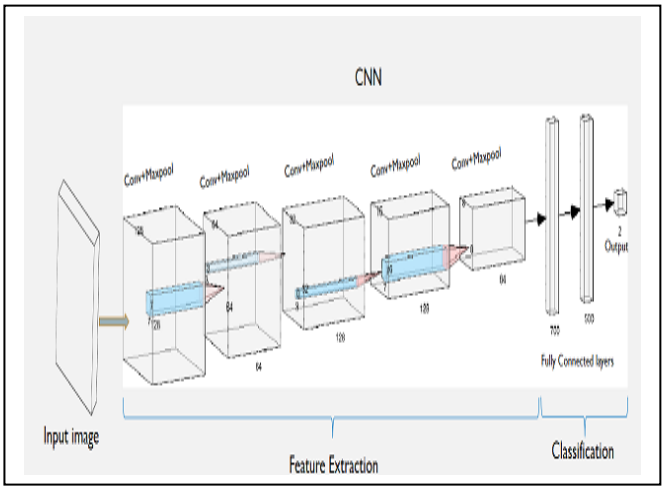

Fig. 6. CNN Architecture

IV. RESULTS

Our Deep Learning model provides a classification accuracy of 68% and gives a loss rate of 61%. The validation data has been tested with respect to the training dataset and graphs of accuracy and loss are plotted as shown in Figure 7.

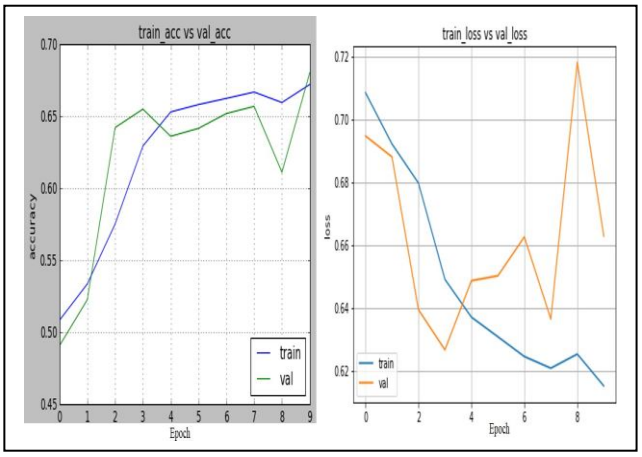

Fig. 7. Graphs showing the comparison of scores between Training and Validation sets in terms of loss rate and accuracy. The X axis represents the Number of Epoch and Y axis gives the corresponding accuracy/loss.

V. DISCUSSION AND CONCLUSION

The classification of an image based on its aesthetic appeal can be tricky and challenging since human perception is greatly subjective and unpredictable at times [4]. However, we have presented a CNN model which takes into consideration the three high level features - Rule of Thirds, Depth of field, and Color Contrast, and gives an accurate result for most of the images based on these features. As we increase the datasets, the accuracy increases further. Here we have computed the CNN accuracy and the accuracy of handcrafted mechanisms. We have considered comparisons for 1802 images.

The computation of accuracy can be done as follows: -

For CNN: -

True Positives (TP) = 643 False Positives (FP) = 319 False Negatives (FN) = 264 True Negatives (TN) = 576

We have achieved an accuracy of 67.64% in our deep learning model. This can be derived using the formula for Accuracy i.e.

$$\text{Accuracy} = \frac{\text{(True Positive + True Negative)}}{\text{Total Number of Images}}$$

$$\text{So, Accuracy} = \frac{(643 + 576)}{1802} = 67.64\%$$

For the Handcrafted or Machine Learning approach: -

True Positives (TP) = 446, False Positives (FP) = 498, False Negatives (FN) = 458, True Negatives (TN) = 400

$$\text{So, Accuracy} = \frac{(446 + 400)}{1802} = 47\%$$

Machine learning algorithms like Support Vector Machine provided an accuracy of 47%, hence CNN gives better performance in comparison. Figure 8 depicts the comparison between CNN and SVM module based on some outliers.

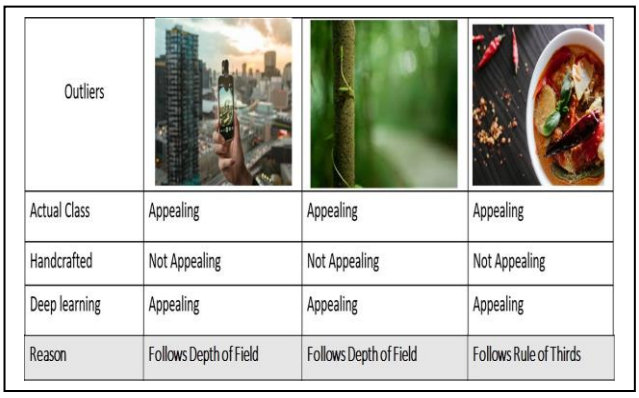

Fig. 8. A few examples of Comparison between Handcrafted module and CNN module based on outliers.

Also, the outliers that are wrongly classified using the traditional handcrafted modules, are correctly classified by the deep learning module. Thus, we can conclude that our Deep learning model provides better performance compared to the traditional algorithms available for aesthetic classification of images.

ACKNOWLEDGMENT

We would like to offer our gratitude towards Dr Patwardhan for valuable support. We would also like to offer

gratitude towards everyone that has contributed directly and indirectly towards this work.

REFERENCES

[1] Datta, Ritendra, Jia Li, and James Z. Wang. "Algorithmic inferencing of aesthetics and emotion in natural images: An exposition." 2008 15th IEEE International Conference on Image Processing. IEEE, 2008.

[2] Wang, Yanran, et al. "Beauty is here: evaluating aesthetics in videos using multimodal features and free training data." Proceedings of the 21st ACM international conference on Multimedia. 2013.

[3] Bianco, Simone, et al. "Predicting image aesthetics with deep learning." International Conference on advanced concepts for intelligent vision systems. Springer, Cham, 2016.

[4] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James. Z. Wang, "Rating Image Aesthetics Using Deep Learning", 2015 IEEE Transactions on Multimedia, Vol. 17, No. 11.

[5] Ke, Yan, Xiaoou Tang, and Feng Jing. "The design of high-level features for photo quality assessment." 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Vol. 1. IEEE, 2006.

[6] Sagnik Dhar, Vicente Ordonez, Tamara L Berg, "High Level Describable Attributes for Predicting Aesthetics and Interestingness," Colorado Springs: USA,2011.

[7] Falcoz, Paolo. "Simple Low Level Features for Image Analysis." Multimedia Techniques for Device and Ambient Intelligence. Springer, Boston, MA, 2009. 17-42.

[8] Steven W. Smith, Ph.D., The Scientist and Engineer's Guide to Digital Signal Processing. https://www.dspguide.com/ch23/5.htm

[9] Mojsilovic, Aleksandra, and Bernice Rogowitz. "Capturing image semantics with low-level descriptors." Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205). Vol. 1. IEEE, 2001.

[10] Mai, Long, et al. "Rule of thirds detection from photograph." 2011 IEEE International Symposium on Multimedia. IEEE, 2011.

[11] Zhang, Mingju, et al. "Auto cropping for digital photographs." 2005 IEEE International Conference on Multimedia and Expo. IEEE, 2005.

[12] Maleš, Matija, Adam Heđi, and Mislav Grgić. "Compositional rule of thirds detection." Proceedings ELMAR-2012. IEEE, 2012.

[13] Wang, James Ze, et al. "Unsupervised multiresolution segmentation for images with low depth of field." IEEE Transactions on Pattern Analysis and Machine Intelligence 23.1 (2001): 85-90.

[14] Jacobs, David E., Jongmin Baek, and Marc Levoy. "Focal stack compositing for depth of field control." Stanford Computer Graphics Laboratory Technical Report 1.1 (2012): 2012.

[15] Yiwen Luo and Xiaoou Tang, "Photo and Video Quality Evaluation: Focusing on the Subject," Department of Information Engineering the Chinese University of Hong Kong, Hong Kong.

[16] Liu, Zhi, et al. "Automatic segmentation of focused objects from images with low depth of field." Pattern Recognition Letters 31.7 (2010): 572-581.

[17] Li, Hongliang, and King N. Ngan. "Learning to extract focused objects from low dof images." IEEE transactions on circuits and systems for video technology 21.11 (2011): 1571-1580.

[18] Luo, Yiwen, and Xiaoou Tang. "Photo and video quality evaluation: Focusing on the subject." European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2008.

[19] Wang, Yanran, et al. "Beauty is here: evaluating aesthetics in videos using multimodal features and free training data." Proceedings of the 21st ACM international conference on Multimedia. 2013.

[20] LeCun, Yann, Koray Kavukcuoglu, and Clément Farabet. "Convolutional networks and applications in vision." Proceedings of 2010 IEEE international symposium on circuits and systems. IEEE, 2010.