

A Computational Approach to Relative Image Aesthetics

by

Jaya Vijetha Gattupalli

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved July 2016 by the
Graduate Supervisory Committee:

Baoxin Li, Chair
Jianming Liang
Hasan Davulcu

ARIZONA STATE UNIVERSITY

August 2016

ABSTRACT

Computational visual aesthetics has recently become an active research area. Existing state-of-art methods formulate this as a binary classification task where a given image is predicted to be beautiful or not. In many applications such as image retrieval and enhancement, it is more important to rank images based on their aesthetic quality instead of binary-categorizing them. Furthermore, in such applications, it may be possible that all images belong to the same category. Hence determining the aesthetic ranking of the images is more appropriate. To this end, a novel problem of ranking images with respect to their aesthetic quality is formulated in this work. A new data-set of image pairs with relative labels is constructed by carefully selecting images from the popular AVA data-set. Unlike in aesthetics classification, there is no single threshold which would determine the ranking order of the images across the entire data-set.

This problem is attempted using a deep neural network based approach that is trained on image pairs by incorporating principles from relative learning. Results show that such relative training procedure allows the network to rank the images with a higher accuracy than a state-of-art network trained on the same set of images using binary labels. Further analyzing the results show that training a model using the image pairs learnt better aesthetic features than training on same number of individual binary labelled images.

Additionally, an attempt is made at enhancing the performance of the system by incorporating saliency related information. Given an image, humans might fixate their vision on particular parts of the image, which they might be subconsciously intrigued to. I therefore tried to utilize the saliency information both stand-alone as well as in combination with the global and local aesthetic features by performing two separate sets of experiments.

In both the cases, a standard saliency model is chosen and the generated saliency maps are convoluted with the images prior to passing them to the network, thus giving higher importance to the salient regions as compared to the remaining. Thus generated saliency-images are either used independently or along with the global and the local features to train the network. Empirical results show that the saliency related aesthetic features might already be learnt by the network as a sub-set of the global features from automatic feature extraction, thus proving the redundancy of the additional saliency module.

*To my parents,
Prabhavathy and Konda Reddy,
for crossing all the social barriers and supporting me throughout.*

ACKNOWLEDGMENTS

I would like to express my profound gratitude to Dr. Baoxin Li for his immense support and encouragement throughout this project. I specially thank him for his patience in understanding and trying to solve some problems during the period when I couldn't move further with my experiments. I further thank him for training me extensively in the areas of Deep Neural Networks and Computer Vision with his rich technical expertise.

I would also like to express my gratitude to my committee members Dr. Hasan Davulcu and Dr. Jianming Liang for taking time out and agreeing to be a part of my defense committee.

I sincerely thank my lab mate and senior Ph.D. student, Mr. Parag S. Chandakkar for helping me resolve some experimental issues that I faced during the course of this project. I also thank him for teaching me some Machine Learning related tricks which eventually helped in speeding-up the training process.

I appreciate the support from all my lab-mates for keeping the work place comfortable and making the lab servers available during the period when I had to finish the experiments on time.

TABLE OF CONTENTS

	Page
LIST OF TABLES	viii
LIST OF FIGURES	ix
CHAPTER	
1 INTRODUCTION	1
1.1 Motivation.....	1
1.2 Problem Statement.....	2
1.3 Related Work	3
1.3.1 Hand Crafted Aesthetic Features.....	3
1.3.2 Deep Learning Based Models.....	4
1.3.3 Relative Training.....	5
1.4 Contributions	5
1.4.1 Relative Aesthetics.....	5
1.4.2 Incorporating Saliency.....	6
1.4.3 Specific Contributions.....	6
2 ESTABLISHING BASELINE	8
2.1 Binary Classification: Data-Set.....	8
2.1.1 Various Aesthetic Data-sets.....	8
2.1.1.1 Photo.net.....	9
2.1.1.2 CUHK.....	9
2.1.2 The AVA Data-Sets.....	10
2.1.2.1 Advantages of using AVA data-set.....	10

CHAPTER	Page
2.1.2.2 Forming the Data-Set.....	11
2.1.2.3 Description of the Meta-Data.....	14
2.1.3 Forming a sub-data-set from AVA	15
2.1.3.1 Analyses on the Data-Set.....	15
2.1.3.2 Forming a Custom Data-Set for Binary Classification...	18
2.2 Binary Classification: BaseLine.....	18
2.2.1 Deep Learning for Aesthetics.....	18
2.2.2 RAPID.....	19
2.2.2.1 Two Versions of Input.....	20
2.2.2.2 Network Architecture.....	21
2.2.2.3 Training the Network.....	22
2.2.3 Inference For Ranking.....	22
3 RELATIVE LEARNING OF AESTHETICS	23
3.1 Building a New Data-Set.....	23
3.1.1 Constraints Imparted.....	24
3.1.2 Building the New Data-Set.....	24
3.2 Proposed Approach Using Relative Learning	25
3.2.1 Network Architecture	26
3.2.1.1 Double Channel Architecture.....	26
3.2.1.2 Siamese Characteristics.....	26
3.2.1.3 Layer Specifications.....	28

CHAPTER	Page
3.2.2 Ranking Loss Layer	29
3.3 Experiments and Results	30
3.3.1 Training the Relative Architecture.....	30
3.3.2 Testing the Model.....	30
3.3.3 Determining the Ranking Order using Binary Classification.....	31
3.3.4 Analyses on the Results.....	32
3.3.5 Analyzing the kernels and the feature maps.....	35
4 INCORPORATING SALIENCY	35
4.1 An Existing Model.....	35
4.1.1 Brief Description.....	35
4.1.2 Drawbacks.....	35
4.2 Forming Saliency Images	36
4.3 Experiments and Results.....	40
4.3.1 A Stand-alone Saliency Based Module.....	40
4.3.2 Incorporating Saliency into the Existing Model.....	42
5 CONCLUSION	44
REFERENCES.....	46

LIST OF TABLES

Table		Page
1.	Architecture of a Column in the Base-Line	21
2.	Results for Ranking and Binary Classification	32
3.	Performance of the System for Various Values of α	41

LIST OF FIGURES

Figure		Page
1.	Figure Showing Random Images Drawn from the Data-Set.....	13
2.	A Screen-Shot of the AVA Data-Set Given by the Owners of the Data-Set....	14
3.	Histogram of Mean Ratings of the Images from AVA Data-Set.....	15
4.	Histogram of the Variances of the Image Ratings in the Data-Set.....	16
5.	Figure Showing the Histogram of Images in Various Categories.....	17
6.	Picture Depicting the Architecture of the Base-Line.....	19
7.	Architecture of the Proposed Network.....	27
8.	Visual Results Produced by Our Network.....	34
9.	Visualization of the Weights of the First Convolutional Layer for Binary Classification Network.....	35
10.	Visualization of the Weights of the First Convolutional Layer for Ranking Network.....	36
11.	Input Image Passed into the Networks in Order to Analyze the Feature Maps.....	38
12.	Feature Maps Obtained from the Binary Classification Network.....	38
13.	Feature Maps Obtained from the Ranking Network.....	39
14.	Variations of Saliency-Images for Different Saliency Operations.....	43
15.	Modified Saliency Map Pixel Values for Various Values of α	45

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

Automatic assessment of image aesthetics is an active area of research due to its widespread applications. Most of the existing state-of-art methods treat this as a classification problem where an image is categorized as either beautiful (having high aestheticism) or non-beautiful (having low aestheticism). In [1], [2], this problem has been formulated as a classification/regression problem by mapping an image to a rating value. Various approaches such as [1], [2], [3], [4], [5], [6], [7], [8], [9], [10] have been proposed which either use photographic rules or hand-crafted features to assess the aesthetics of an image. Due to the recent success of deep convolutional networks, approaches such as [11], [12] claim to have learned the feature representations necessary to categorize the given image as either beautiful or non-beautiful.

The approaches based on photographic rules have certain limitations such as the implementations of these rules may be an approximation, thus affecting the accuracy of aesthetic assessment. The rules may not be enough to govern the process of how we decide the aesthetic value of an image. It is possible that some of the important rules have been left out or some erroneous ones have been included. These rules are mostly accompanied by generic image descriptors or task-specific hand-crafted features. Such approaches suffer from the disadvantages of generic/hand-crafted features that they may not be suited for a special task such as aesthetic assessment or the feature space does not fully represent the key characteristics which make an image aesthetic. The deep neural network based

approaches overcome these disadvantages by learning the feature representations from the data.

While deep learning approaches have significantly advanced the state-of-art for this task, I observe that classifying a given image as beautiful or non-beautiful may not always be the natural choice for certain applications. It may also be more intuitive for humans to compare two images rather than giving an absolute rating to an image based on its aesthetic value. Moreover, all images in a set could belong to the beautiful or non-beautiful category according to a classification model. In such cases, it is mandatory that these images are arranged according to their aesthetic value. For example, a machine-learned enhancement system [14] has to provide an enhanced version of the query image to the user. To do so, it needs to compare two images with respect to their aesthetics to determine which set of enhancements results into a more beautiful image. In an image retrieval engine, it would be desirable to have an option to retrieve images having low/similar/high aesthetic value as compared to the query image. Thus, there is a necessity for developing algorithms which rank images based on their aesthetic value.

1.2 PROBLEM STATEMENT

Motivated by these issues, the novel problem of picking a more beautiful image from a pair is introduced in this work. This problem is termed as “Relative Aesthetics”. A new data-set of image pairs is created to attempt this problem. The pairs are carefully chosen from the popular AVA data-set [15] to satisfy certain constraints. For example, it is observed that comparing images from unrelated categories (for example, a close-up of a car and a wedding scene) does not make sense and hence such pairs are avoided.

Additionally, there exists no single threshold which can rank the pairs correctly across the entire data-set. In other words, if images were categorized into beautiful and non-beautiful, then some of the pairs in this newly formed data-set could contain both beautiful or both non-beautiful images. The details of data-set creation and its statistical analysis are provided in Chapters II and III. The data-set and the model would be made public for analysis and further advancement of this field.

This problem draws certain parallels with “relative attributes” [16]. The authors of [16] observe that training on a relatively labeled data allows them to build models which capture more general semantic relationships. They also mention that by using attributes as a semantic bridge, their model can relate to an unseen object category quite well. On the other hand, the current problem presents different challenges. In [16], they compare two images with respect to attributes (for example, more natural, furrrier, narrower etc.) which are better defined than the aesthetics of two images. Thus even though it is trivial to use models trained on categorical data to solve these ranking tasks, I found that using relative learning principles allows the proposed approach to outperform previous state-of-art classification models by gaining a more general and a semantic-level understanding of the proposed problem.

1.3 RELATED WORK

1.3.1 Hand-Crafting of Aesthetic Features:

Computational aesthetics research in the earlier years was focused on employing photographic rules, hand-crafted features or generic image descriptors. Intuitive and common properties such as color [1], [7], [8], texture [1], [2], content [6], [5], combination

of photographic rules, picture composition and hand-crafted features [5], [4], [6] have been used. The most commonly used photographic rules include Rule of Thirds used in [5], [4], [1]. Other compositional rules include low depth of field, opposing colors etc. [5]. Common color features such as lightness, color harmony and distribution, colorfulness have been quantified for aesthetics assessment purposes by computational models [1], [7], [8]. Texture features based on wavelets edge distribution, low depth of field, amount of blur have also been used [2], [5]. Approaches specifically trying to model content in the image by detecting people [6], [5], [4], generic image descriptors such as SIFT [17] have been proposed in [5].

1.3.2 Deep Learning Based Models:

Inspired by the then success of deep neural network on various tasks such as image classification [18], [19], object segmentation [20], facial point detection [21], Decaf features [22] for style classification [23] etc., [11] proposed a deep learning based aesthetics assessment approach. This approach classifies given image as beautiful or non-beautiful depending on the entire image as well as its local patches. Another such approach was presented in [12] where the authors aggregate the information from multiple patches in the multiple-instance-learning manner to improve the aesthetics assessment results further. However, most of these approaches treat the aesthetics assessment task as a binary classification problem, which may not always be the best choice for certain applications, as discussed before.

1.3.3 Relative Learning:

The concept of training on relatively-labeled data to improve model performance and provide it with certain semantic understanding of the problem is well-explored. The work on relative attributes [16] model predicts the relative strength of individual property in images. It allows for comparison with an unseen object category in the attribute space. Model learned in such a way enables richer text descriptions of images. Relative attribute feedback was used in conjunction with semantic language queries to improve the image search capability in [24]. There are many such applications where relative learning has explored a new dimension of the problem and improved the overall understanding of the model of a given task.

1.4 CONTRIBUTIONS

1.4.1 Relative Aesthetics:

In this work, I propose to employ the relative learning principles for the task of image aesthetics assessment. This task is extremely subjective and have vaguely defined properties than even the attributes such as bigger, higher, more natural etc. To allow for learning using hand-crafted features, various data-sets have been proposed such as Photo.net, CUHK, AVA data-set. The first two data-sets contain a few thousand images whereas the AVA data-set [15] contains 250,000 images. Thus AVA data-set is used to form image pairs which in turn facilitates the learning of the proposed approach. A Siamese deep neural network architecture [25] is proposed with a relative ranking loss, which takes an image pair as input and ranks them with respect to their aesthetic value. The back-propagation happens with the loss obtained from the ranking function, which helps the

network explore the attributes of certain images which makes them more aesthetic than the other images.

1.4.2 Incorporating Saliency:

As an extension to the above work, the role of saliency in determining the aesthetic value of an image is analyzed. A previous work in this area [30], presented a saliency-based model to enhance the classification performance of professional and non-professional images. Their model exhibited significant gain in classification accuracy as compared to the previous state-of-art. However, the authors hand-crafted the required global aesthetic-features and further performed the experiments on an extremely small and easy data-set. I therefore try to experiment on a much larger data-set using automatic feature extraction thus throwing more insight into the role of saliency in image aesthetics.

1.4.3 Specific Contributions:

My contributions through this work therefore are as follows:

1. A novel problem termed as “relative aesthetics” is formulated, which involves picking a more beautiful image from a given pair of images. A new data-set is created which has such relative labels formed from the popular AVA data-set by careful and constrained selection of image pairs.
2. The relative learning paradigm is incorporated into the proposed deep network and train it end-to-end. To the best of my knowledge, there is no prior work on studying aesthetics in a relative manner using deep neural networks.

3. It is shown that our model trained on relatively-labeled data is able to outperform a recent state-of-art method [11] trained on a similar sized, categorically labeled data-set for the proposed task.

4. The role of saliency in determining the aesthetic quality of an image is analyzed and some related experiments are performed. This is the first work to incorporate saliency into aesthetics using automatic feature extraction using deep-learning.

The rest of the document is organized as follows.

Chapter II explains the available data-sets in this area along with discussion on establishing a base-line data-set and model. All the experimental results in this work are compared against the results of the base-line. Chapter III describes the proposed relative deep neural network based approach, the experimental setup and results and analyses. Chapter IV throws some insight into the role of saliency in determining the aestheticism of an image.

I finally conclude in Chapter V.

CHAPTER 2

ESTABLISHING BASELINE

As mentioned in the previous chapter, the problem of *ranking* images based on their aesthetics value is completely novel and has never been attempted previously. Therefore, there exists no previous models that can be used as reference to compare the results of the proposed approach. Henceforth, a state-of-art aesthetic binary-categorization model is considered in order to establish a base-line. i.e a standard model is trained for the task of binary classification but altered at the final layer to be able to rank a pair of images based on their aesthetic value. The results achieved by this base-line are compared against the proposed method.

In order to achieve reliable ranking results from the binary classification model, we need to first establish a stable base-line. Therefore, in this chapter, I discuss the data-set used, the network architecture of the base-line and the training procedure. At the end of this chapter, the classification results achieved by this implementation are compared to the ones mentioned in the paper [11]. The results suggest that this implementation is in close agreement with the model proposed in the paper [11] and therefore can be used as a base-line in this work.

2.1 BINARY CLASSIFICATION: DATA-SET

2.1.1 Various Aesthetic Data-Sets:

In order to facilitate computational assessment of aesthetic quality, various data-sets are built and made publicly available. Photo.net, CUHK and AVA data-set are some

of the well-known data-sets in this area. Each sample in these data-sets consists of an image, their corresponding aesthetic label and some other meta-data related to the photographic style, semantic content etc. The labels take different forms in different data-sets like binary valued labels, real valued ones etc. In either case, the intention of the label is to represent how aesthetic the given image is.

2.1.1.1 Photo.net:

Photo.net contains 3,581 images which are collected from an online community and contains two scores per image [15]. Both the scores are given by users of the online community and are in the range of 1 to 7. The first score corresponds to the aesthetic value of a picture whereas the second one represents how original the picture is. However, because of the less volume of images and existence of some known biases (like the one discussed in [29]), this dataset is not used for our experiments. [29] shows that images receiving high aesthetic scores have outer frames externally added by the owners of the images to enhance visual appearance.

2.1.1.2 CUHK:

CUHK is a data-set that contains 12,000 images half of which are of high quality and the other half are low quality. i.e the data-set contains labels which are binary (1/0), '1' for high quality images and '0' for low quality ones. This data-set is obtained from the well-known photography challenge website called *dpchallenge.com* where images are posted online and for each image, a score of 1-10 is given by various users. The CUHK data-set considered the mean of the scores given by all users for a given image and picked

the pictures which fall in the top 10% and the bottom 10% of all the image ratings. In other words, they considered pictures which are either voted as extremely beautiful (labelled as ‘1’) or extremely non-beautiful (labelled as ‘0’). However, this data-set can be considered immensely easy and some machine learning models achieved classification accuracies higher than 90%. Therefore, this dataset is not used for experiments in this work. Additionally, the proposed model requires data-set that has real-valued labels and a very high number of samples which is another reason for not choosing this data-set.

2.1.2 AVA-DATASET:

2.1.2.1 Advantages of using AVA-Dataset:

AVA dataset [15] incorporates the desired characteristics from the above two data-sets and ignores the undesired ones, forming a super-set of the above mentioned data-sets. It consists of a total of 250,000 images extracted from the same photography challenge website from which CUHK extracted its images, i.e *dpchallenge.com*. Each image has a rating given by various users on a scale of 1-10. AVA-dataset however, doesn’t summarize the ratings by considering the mean unlike the CUHK data-set. This data-set instead provides the distribution of ratings given by users for a particular image, thus facilitating future researchers to analyze the consensus or diversity among ratings given to various images. This data-set therefore can be used for binary classification task as well as ranking or regression tasks as follows. If a threshold is set on the mean ratings of the images, a binary labeled sub-data-set can be formed by considering images falling above the threshold as beautiful and below as non-beautiful. This sub-data-set can be used for the

task of binary classification. On the other hand, this data-set can be used for ranking/regression tasks by considering the mean values of the ratings.

Thus, this data-set is used for all the experiments conducted in this work. Since the number of images present in this data-set is humongous and belong to a wide variety of semantic categories, we believe the experimental results on this data-set generalizes well to other previously mentioned data-sets. Another important reason for choosing this data-set is that it allows for automatic feature extraction using deep-learning because of its huge volume of images, which may not be possible if other data-sets are used. Additionally, as mentioned in the previous section, hand-crafting of features for aesthetic quality analysis may not be the best method for feature extraction as quite often we (humans) may be ignorant of what kind of features make an image aesthetic or may consider some erroneous rules which do not contribute to aestheticism. This data-set is therefore chosen to train a deep-learning based model to compare aesthetic quality of a given pair of images.

2.1.2.2 Formation of the Data-Set:

As mentioned previously, this data-set is built by extracting images from the famous photography challenge website called *dpchallenge.com*. A detailed description about the protocol of posting pictures on the website and extracting them to form the data-set is given below.

The website organizers often create challenges and users are free to upload their images. Some examples of these challenges are Abandoned Buildings, Dichotomy, Fashion etc. Once an image is uploaded by a user, other users will be able to give a rating on a scale of 1-10 based on his/her own judgement of its aesthetic value. Such ratings

collected over time for each image are compared against each other and a winner image is chosen for each challenge. There are thousands of such challenges posted by the organizers till date and each challenge has thousands of images posted by the users.

This data-set is formed by considering a total of 255,530 images from 1398 challenges, whose descriptions are given in Appendix A. Each of these images are then put under one/two of the pre-decided 65 semantic categories by the owners of the AVA-dataset [15].

Some examples of the categories are Abstract, Cityscape, Nature etc. Complete list of all the 65 categories is given in appendix B of this document. Additionally, the owners considered creating a category called “no-category” to bag images that do not significantly belong to any category. To summarize, the images of this data-set has wide variety of content as can be shown in figure 1. The figure shows some sample images randomly selected from the data-set to allow the reader to obtain a coarse idea about the data-set. The images shown in the above figure also justify the reason behind computing automatic aesthetic-features instead of hand-crafting them. As can be observed from the above figure, determining rules that evaluate the aestheticism of an image is quite difficult or sometimes impossible owing to the possibility of wide variety of semantic content.

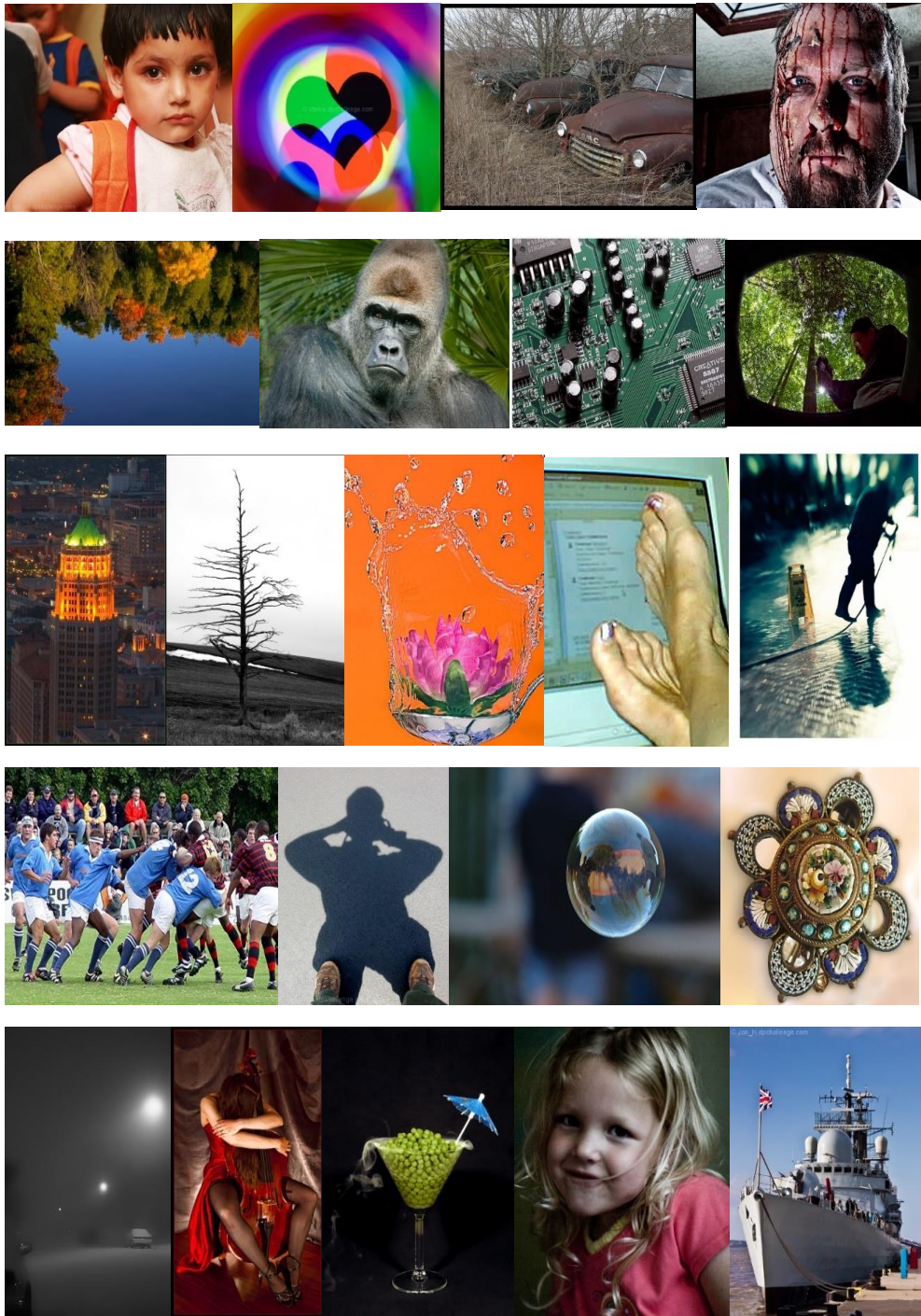
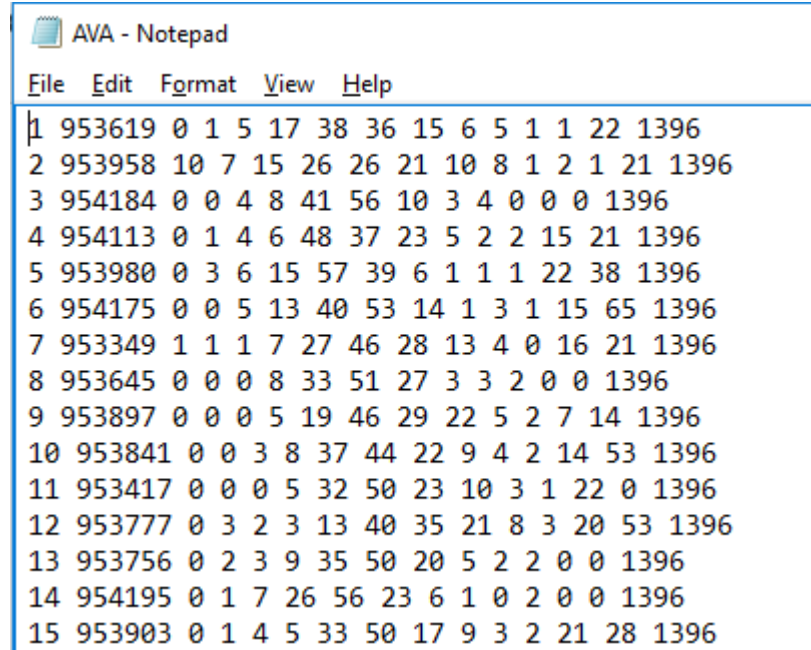


Figure 1: Figure showing random images that drawn from the data-set; It can be observed that the semantic content is drastically different between the images in the data-set.

2.1.2.3 Description of the Meta-Data:



```
AVA - Notepad
File Edit Format View Help
1 953619 0 1 5 17 38 36 15 6 5 1 1 22 1396
2 953958 10 7 15 26 26 21 10 8 1 2 1 21 1396
3 954184 0 0 4 8 41 56 10 3 4 0 0 0 1396
4 954113 0 1 4 6 48 37 23 5 2 2 15 21 1396
5 953980 0 3 6 15 57 39 6 1 1 1 22 38 1396
6 954175 0 0 5 13 40 53 14 1 3 1 15 65 1396
7 953349 1 1 1 7 27 46 28 13 4 0 16 21 1396
8 953645 0 0 0 8 33 51 27 3 3 2 0 0 1396
9 953897 0 0 0 5 19 46 29 22 5 2 7 14 1396
10 953841 0 0 3 8 37 44 22 9 4 2 14 53 1396
11 953417 0 0 0 5 32 50 23 10 3 1 22 0 1396
12 953777 0 3 2 3 13 40 35 21 8 3 20 53 1396
13 953756 0 2 3 9 35 50 20 5 2 2 0 0 1396
14 954195 0 1 7 26 56 23 6 1 0 2 0 0 1396
15 953903 0 1 4 5 33 50 17 9 3 2 21 28 1396
```

Figure 2: A screen-shot of the AVA data-set as given by the owners of the data-set [15].

Furthermore, each row in the data-set represents a sample which consists of a total of 15 columns. Figure 2 is a screen-shot of 15 such sample from the data-set. The description of each column is given below.

Col 1: Column # 1 is the serial number of the image in the data-set, it runs from 1-2555330.

Col 2: Column # 2 is the image id with which each picture can be identified from the *dpchallenge.com* website. It is unique for all the 2555330 images in the dataset.

Col 3 – 12: Each column determines how many users rated the given image with each of the ratings 1-10. i.e the value of column 3 determines the number of users rated the image with 1-rating, column 4 determines the number of users voted for 2-rating and so on up to column 12 which determines the number of users voted for a 10-rating. For each image in the data-set an average of 210 ratings are collected.

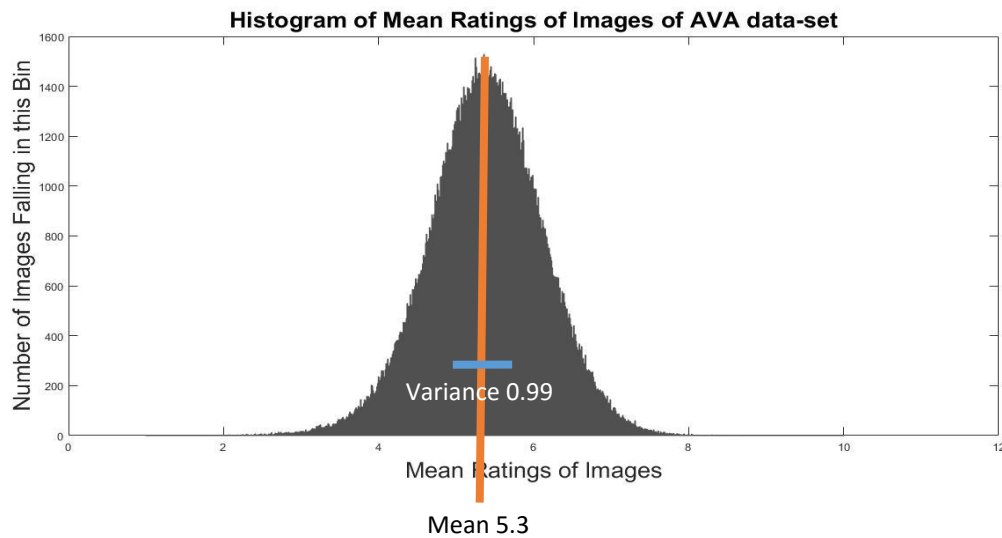


Figure 3: Histogram of Mean Ratings of the Images from AVA data-set. A gaussian is fit to this data and the mean and variance are found out to be 5.33 and 0.99 respectively.

Col 13-14: These two columns determine which semantic category the picture belongs to.

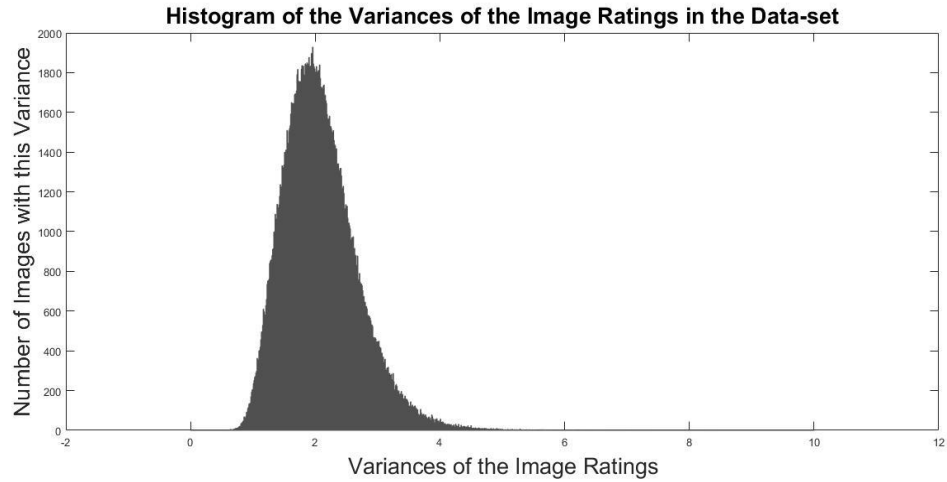
Col 15: This column determines which challenge the image belongs to.

2.1.3 Forming a sub-data-set from AVA:

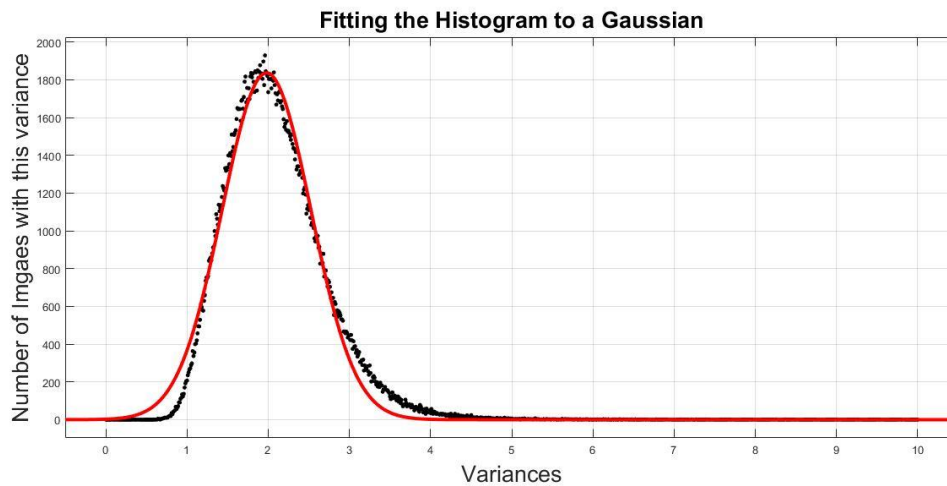
2.1.3.1 Analyses on the Data-Set:

In order to form a sub-data-set for training the base-line for binary classification and later form pairs out of this sub-data-set to train the ranking model, some analyses are performed on the AVA-data-set.

1. The distribution of the ratings of the images is analyzed by calculating the mean ratings of all the image and plotting them on a histogram (Fig 3). It can be inferred,



(A)



(B)

Figure 4: Figures showing the histogram and the curve fitting for the variances of the image ratings in the data-set.

the distribution of the ratings in the data-set is an approximate Gaussian distribution with the mean at around 5.39 and a variance of 0.99. This implies that there are very high number of images in the data-set that have an average rating of 5-6 than images that are either beautiful (i.e rating greater than 7) or non-beautiful (rating lower than 4).

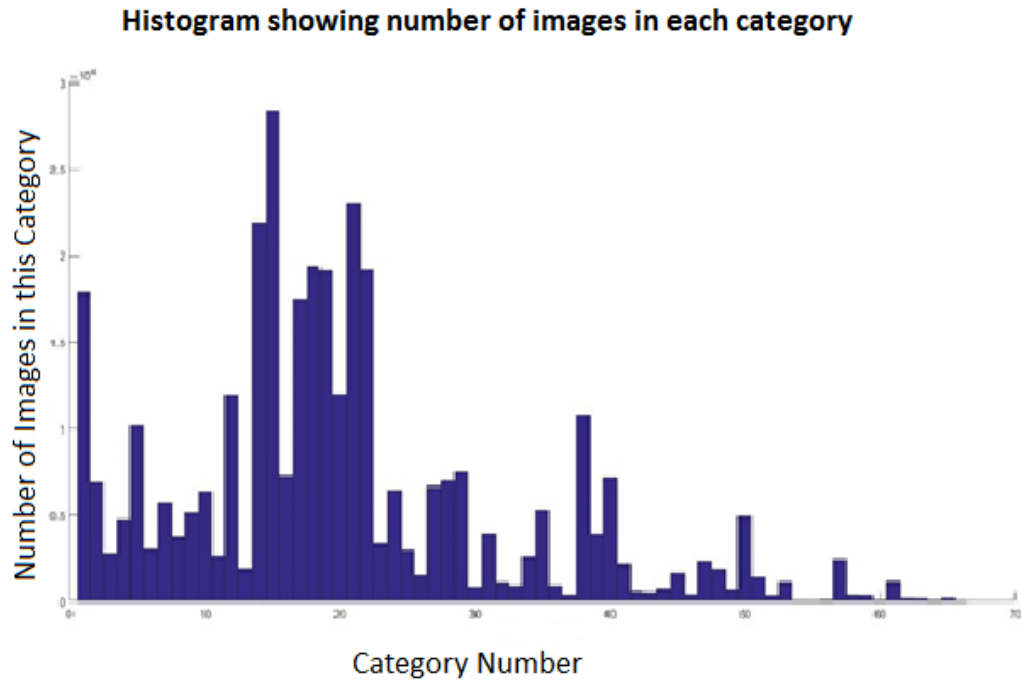


Figure 5: Figure showing the histogram of images in various categories; It can be observed that very few categories cover majority of the images in the data-set.

2. To analyze the distribution of difference/agreement in opinion for the images in the data-set, a histogram of the variances of image-ratings is plotted and is shown in figure 4(a). It can be noticed that the distribution follows an approximate Gaussian with mean at 1.97 and approximate variance of 0.77; figure 4(b). A higher value of variance for a particular image implies that the voters have varied opinions about the image, while a lower value of variance shows good agreement between the opinions of different voters.
3. Another analysis that is performed on the data-set is about the distribution of images in various categories. As already mentioned, each picture is put under 1-2 categories, where the category can be 'no-category' as well. A total of 196,961

images are obtained by excluding the ones that do not belong to any category. A histogram showing the number of samples in each category is given below. It shows that some categories like 1-Abstract, 15-Nature, 21-Black and White cover major percentage of the entire data.

2.1.3.2 Forming a Custom Data-Set for Binary Classification:

A sub data-set is formed from the above described AVA dataset [15] by considering pictures which have variance less than 2.6 and belonging to all categories and challenges. This step ensures that there is less disagreement between voters for a given image thus forcing the neural network to learn features that are relevant for aesthetic analysis in general. A random 80,000 images are chosen from this pool which are used for the task of binary classification. To convert the real valued average ratings into binary labels, the threshold is set as 5.5. i.e images that have mean ratings greater than 5.5 are given the binary label '1' and the ones with mean ratings less than 5.5 are given a label '0'. An equal split of positive and negative samples is present in the dataset. Of the 80,000 images, 40,000 are used for training, 3,000 for validation and the remaining are used for testing.

2.2 BINARY CLASSIFICATION: BASE-LINE

2.2.1 Deep Learning for Aesthetics:

A recent deep neural network based aesthetic quality assessment model [11] is chosen as a base-line in this work. This [11] is the first work in this area using deep learning which achieved a significant jump in classification accuracy than the previous state-of-art

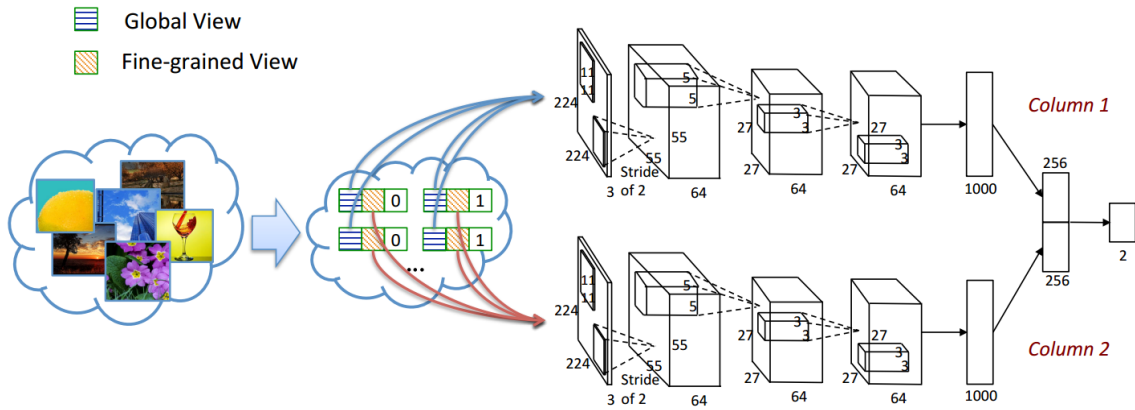


Figure 6: Picture depicting the architecture of the base-line. Two versions of input, Global View and Local View are passed to the two columns of network. Ref: [11]

models. While all the previous models [1], [2], [3], [4], [5], [6], [7], [8], [9], [10] in the area relied on hand-crafting of features using some pre-determined photographic rules, this [11] is the first work which relied on automatic feature extraction using deep learning. Results reported by [11] indicate that deep learning might be a promising direction to deal with the problem of computational analysis of aesthetics. This can be attributed to the difficulty (or sometimes impossibility) in hand-crafting the photography/aestheticism related rules, which can be easily learnt with deep learning based models.

2.2.2 RAPID:

The base-line considered in this work is called RAPID [11] which is an abbreviation for “RAting Pictorial aesthetics using Deep learning”. The central idea in this model is to input two variations of input image (a global view and a local view) to two deep convolutional columns which are concatenated in the end. This concatenated feature vector is passed through a set of fully connected layers to finally give out a probability value

$p(y=1/X)$. This value is thresholded at 0.5 to determine if the predicted label is 1/0. The below picture shows the architecture of the network along with the variations of input. More details about the architecture are given in the below paragraphs.

2.2.2.1 Two Variations of Input:

The model consists of two identical columns each of which is a deep convolutional neural network. The two columns only differ in the kind of input given to it. A global view of the image is given to the first column whereas a random local patch is given to the second column. The reason for inputting two versions of the same input image is as follows. The model is expected to learn some global aesthetic image features like the rule of thirds, the golden ratio etc. from the given global view of the image. Similarly, local features like the resolution are learnt from the local view. [11] conducted experiments on (i) two views of the image separately using two separate single column models and (ii) both the view together using the double-column model and the results are compared. It is shown in [11] that the double column (DCNN: Double Column Neural Network) model outperformed the single column ones in the aesthetic quality categorization experiments. DCNN showed an accuracy of 73.25% as compared to 71.20% when only local view is considered and 67.79% when only global view is considered [11]. I therefore consider using both the views of the image for all the experiments in this work.

Input	Convolution	Max-pooling	Convolution	Max-pooling
3 X 224 X 224	2, 64, 11, 2	2 X 2	1, 64, 5, 1	2 X 2

Convolution	Convolution	Dropout	Dense	Dropout	Dense
1, 64, 3, 1	1, 64, 3, 1	0.5	1000	0.5	256

Table 1: Architecture of a column in the base-line. Convolution is represented as (padding, # filters, receptive field, stride)

2.2.2.2 Network Architecture:

Diving into the details of the model, both the columns in the DCNN model [11] are identical and each has the following architecture. The input is a 224 X 224 image patch which is passed through a convolutional layer with kernel size 11 X 11, a stride of 2, containing 64 filters. The resulting tensor is then passed through a maxpooling layer of window size 2 X 2. Another set of convolutional and max pooling layers is applied with kernel size 5 X 5, 64 filters and a pooling window of 2 X 2. Two convolutional layers both with 3 X 3 kernel size and 64 filters are then applied followed by two fully connected layers with 1000 and 256 hidden units respectively. A dropout ratio of 0.5 is applied on the dense layers. The two 256 vectors obtained from the two columns are concatenated in the end which is then connected to a single neuron which outputs the probability of the image being aesthetic, i.e $p(y=I/X)$. The architecture details are shown in Table 1 for the readers' reference.

2.2.2.3 Training the Network:

This network is trained on the 40,000 training that were chosen using the criteria described in the above section. A learning rate of 0.001 and L2 regularizer of 0.05 are applied during training. The learning is dropped by 25% after each 10 epochs and the network is trained until no significant improvement in the validation accuracy is observed. This model achieved a classification test accuracy of 59.92 % on our test set of 20,000 samples and 69.18 % on the standard test set provided by the AVA-dataset owners [15]. The significant difference in performance can be contributed to the huge number of categories our test set contains as opposed to only eight categories contained in the standard test set. Additionally, the ~4% drop in accuracy of this implementation as compared to 73.25% mentioned in the reference paper [11] can be attributed to the comparatively fewer number of images that I considered. Only 40,000 images are considered in this implementation as compared to 230,000 images used in [11]. A stable base-line is thus established and used in this work.

2.2.3 Inference for Ranking:

The model thus trained is used to calculate the ranking performance as previously mentioned by considering the soft probability values of the two images in a pair. The two images are passed into the model one after the other and the soft probability values omitted by the network are saved. During inference, the first image is predicted as more beautiful if the value of $p(y = 1|\mathbf{X}_1)$ is greater than $p(y = 1|\mathbf{X}_2)$ and vice versa (\mathbf{X}_1 and \mathbf{X}_2 represent the first and the second image in the pair respectively). These results are then compared with the ranking results of the proposed model which will be discussed in chapter III.

CHAPTER 3

RELATIVE LEARNING OF AESTHETICS

This chapter discusses the proposed approach where relative learning techniques are used to learn aesthetic related features and thus attempt the problem of ranking of images based on aesthetics. That means given a pair of images, this model predicts a binary label '1' if the first image is more beautiful than the second and '-1' otherwise. However, as mentioned in the previous chapter, there exists no data-sets with relative label information for the task of visual aesthetics. This chapter therefore, deals with building a data-set with relative labels and using it for the relative learning task. The proposed approach to attempt the task of ranking images based on their aesthetic value is discussed in the second part of this chapter followed by experiments, results and analyses.

3.1 BUILDING A NEW DATA-SET

Our task is to determine the more beautiful image in a pair. To the best of our knowledge, there exists no such dataset containing relatively-labeled pairs with respect to their aesthetic rating. A data-set containing 40,000 image pairs is created and used in this work. The individual images in these pairs belong to the AVA data-set [15]. Half the data-set is used for training and the other half for testing. I now describe the protocol used to form the pairs out of the images from the AVA data-set.

3.1.1 Constraints Imparted:

The protocol can be defined by the three constraints as follows:

1. The difference between the average ratings of the images in a pair should be ≥ 1 . This constraint on the rating difference ensures that the network emphasizes on the characteristics differences defining the aestheticism of both images.
2. Each image in the AVA data-set has 210 ratings on an average. As mentioned in Chapter II, the variance of all the ratings for each image is computed and plotted on a histogram. The mean of the fitted Gaussian is at 1.97 and the variance is 0.77. As mentioned in [15], the high variances among the image ratings is a result of the collective disagreement between the raters, which suggests that such images may have certain abstract/novel content or photographic style, preferred only by certain group of people. The images which cause such significant disagreements among the raters are avoided by only considering the images having rating-variances less than 2.6.
3. Additionally, pairs formed from images belonging to different categories are avoided since the characteristics which make an image aesthetic may vary with the category. For example, a beautiful picture of a car may have bright colors whereas a beautiful picture of a human face may have low-depth of field, lighter colors etc. Additionally, since the ratings in the AVA data-set are crowdsourced ratings; the opinions may exhibit a preference towards some category. The effect of these two factors can be mitigated by using pictures from the same category to form pairs.

3.1.2 Building the new data-set:

Relative labels are formed after such careful selection of pairs. A pair is labelled as 1 if the average rating of the first image is greater than that of the second image and -1

otherwise. The majority of the pairs in this data-set have the rating-difference ≈ 1 . To quantify, the rating-difference for about 85% of the training and test data is between 1 and 1.5. As the rating difference between the images of a pair decreases, choosing the more beautiful image in that pair gets difficult. To ensure that the proposed network is not biased towards our data-set, the experiments are replicated on a standard test set formed as follows. The creators of the AVA data-set [15] provide 20,000 images to evaluate different approaches. I use the aforementioned protocol and form 7,670 pairs and use it as a standard data-set.

3.2 PROPOSED APPROACH USING RELATIVE LEARNING

The comparison of the aesthetics of two images is dependent on several factors and people's visual preferences. Some of the factors include color harmony [7], colorfulness [1], inclusion of opposing colors [5], composition [26], visual balance [27] etc. They are also affected by the content in the picture [4], [6]. Though determination of aesthetics is a subjective process, there are some well-established rules in the photography community such as low depth-of-field, rule of thirds, golden ratio [28]. However, making hand-crafted features for such rules is difficult and often will lead to approximation or misrepresentation of those rules. Therefore, a deep neural network based approach [11] is taken and the relative ranking rules are incorporated into it by designing a suitable loss function. Most of the rules or aesthetic criteria can be defined using either an entire image or a part of it. Therefore, for each image in the pair, this network is trained on two views of an image as done in the previous chapter. First view being the entire image and the second one being a local patch. This enables the network to see different aspects of the input as discussed in

chapter II. The network architecture and its training procedure are discussed in detail in the subsequent sub-sections.

3.2.1 Network Architecture:

3.2.1.1 Double Channel Architecture:

The proposed deep convolutional neural network takes an image pair as input. For each image in the pair, it takes as input that image itself and its local patch. Since all images have to be of the same size, they are warped to be $224 \times 224 \times 3$. A same size local patch is also cropped from the original resolution image. The image is warped based on the findings in [11], which shows that local patches along with warped image gives the best result. The proposed network has two “channels” as shown in Fig. 7, corresponding to the input pair of images. A channel is defined as the part of our CNN which takes an image along with its local patch as input. Each channel has two “columns”. One column takes the warped image and the other one takes its local patch as input.

3.2.1.2 Siamese Characteristics:

The architecture is a Siamese network where each channel shares weights in a certain way, which is shown in Fig. 7 by means of color coding. The columns with the same color (i.e. either red or green) share the weights. This is because the ranking produced by the network should be invariant to the order of the images in the pair. Both channels have exactly the same architecture until they are merged at the final but one dense layer of

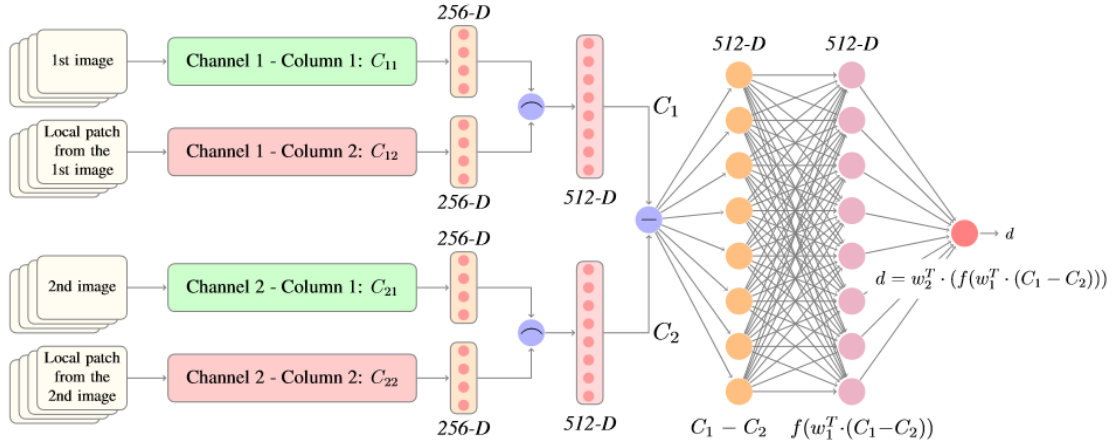


Figure 7: Architecture of the proposed network; Weights are shared between the columns C_{11} and C_{21} (shown in green), C_{12} and C_{22} (shown in red); The features obtained from C_{11} and C_{12} are concatenated (represented by $_$ symbol) to get C_1 and C_{21} and C_{22} are concatenated to get C_2 ; The vector $C_1 - C_2$ is passed through two dense layers to obtain a score d comparing the aesthetics of two images. $f(\cdot)$ denotes an ReLU non-linearity. Please refer to the text for further details.

512-D. The architecture of the upper channel (channel 1) is equivalent to the architecture of the base-line established. This channel has two columns which takes the image and its local patch as input. Since these two inputs are on a different spatial scale and trying to convey different aesthetic properties as discussed earlier, we do not set constraints on the weights of both the columns in a channel.

3.2.1.3 Layer Specifications:

The upper column in channel 1 (C_{11}) takes the entire image as input which is of size $224 \times 224 \times 3$. The column has five convolutional layers. The first convolutional layer has

64 filters each of size $11 \times 11 \times 3$ with stride 2. Second convolutional layer has 64 filters of size 5×5 with stride 1. Third and fourth layer have 64 filters of size 3×3 with stride 1. These are followed by two dense layers of size 1000 and 256 respectively. We apply 50% Dropout at these two dense layers. Max-pooling is applied after first two convolutional layers. Each max-pooling operation halves the input in both the directions. We use ReLU activation throughout. The inputs are appropriately zero-padded after each convolution layer so that the output is of the same size as its input.

The lower column of channel 1 (C_{12}) and both the columns of channel 2 (i.e. C_{21} and C_{22}) have the same architecture as C_{11} including dropout, maxpooling and zero-padding operations. The key thing to note here is that the weights are shared for: 1. the two columns which take the entire image as input i.e. C_{11} and C_{21} 2. the remaining two columns which take the local patches as input i.e. C_{12} and C_{22} . C_{11} and C_{21} each generate a 256-D representation (i.e. of the entire image). Similarly, C_{12} and C_{22} also generate 256-D features (i.e. of the local patch). The two 256-D representations from (C_{11} , C_{12}) are concatenated to form two 512-D representations. Similarly, the two 256-D representations of (C_{21} , C_{22}) are concatenated. The Fig. 7 shows this architecture and the sharing of weights.

3.2.2 RANKING LOSS LAYER

The proposed network should be able to rank two input images with respect to their aesthetics value. More formally, given two input images I_1 and I_2 , I_1 is predicted to be more beautiful than I_2 (also denoted as $I_1 > I_2$ here onward) if a positive value is obtained for

$d(I_1, I_2)$ and vice versa. In other words, $d(I_1, I_2)$ is a measure comparing aesthetics of two images.

$$d(I_1, I_2) = w^T \cdot (g(I_1) - g(I_2)) \dots \dots \dots (1)$$

Here, $g(I_1)$ and $g(I_2)$ are the CNN representations. In this network, $g(I_1)$ and $g(I_2)$ are represented by C_1 and C_2 respectively, as shown in Fig. 7. To increase the representational power, the vector $(C_1 - C_2)$ is passed through two dense layers separated by a ReLU non-linearity. Thus for our network, Equation 1 takes a slightly modified form as follows:

$$d(I_1, I_2) = w_2^T \cdot (f(w_1^T \cdot (C_1 - C_2))) \dots \dots \dots (2)$$

where $f(\cdot)$ denotes an ReLU non-linearity.

Keeping this in mind, we can now design our final loss function with the following properties:

1. It should propagate zero loss when all image pairs are ranked correctly
2. It should always produce a non-negative loss. The loss function is designed as follows.

$$L = \max(0, \delta - y \cdot d(I_1, I_2)) \dots \dots \dots (3)$$

here, y is a ground-truth label which takes value 1 if first image in the pair is more beautiful than the second one (i.e. $I_1 > I_2$) and it equals -1 if $(I_1 < I_2)$. The term $\max(0, \cdot)$ is necessary to ensure that only non-negative loss gets back propagated. The δ is a user-defined parameter which serves has two purposes. Firstly, it defines a required separation to declare $I_1 > I_2$ (or $I_1 < I_2$). That means if $y \cdot d(I_1, I_2) > \delta$, then no loss should be back-propagated for such pairs. Secondly, and more importantly, $\delta > 0$ avoids a trivial solution to our optimization objective. To clarify further, if $\delta = 0$, then for $y = 1$ and $y = -1$, a common trivial solution exists which makes either $w_1 = 0$ or $w_2 = 0$. δ is set to be equal to 3 in this

work as no performance boost is noticed by further increasing the separation between CNN feature representations of I_1 and I_2 .

In the further subsections, the training and testing procedures of our architecture are explained. Then I compare the aesthetic ranking results of our network against a state-of-art network that is trained on a categorical data.

3.3 EXPERIMENTS AND RESULTS

3.3.1 Training The Relative-Architecture

This architecture is trained using mini-batch SGD with a learning rate of 0.001, momentum = 0.9, weight decay of 10^{-6} and by employing Nesterov momentum. The learning rate is reduced by 15% after every 10 epochs. The batch size is set to 50. Apart from warping and cropping out the local patch, I only subtract the mean RGB value computed on the training set, from each pixel of the image. During training, when the network makes a wrong decision, it is forced to learn by exploiting the difference between some other characteristics of the image in the next iteration. I believe that over a number of epochs, it manages to discover the relevant image properties which better define image aesthetics. The training is stopped when the validation accuracy on a set of 3,000 images does not show significant improvement for 10 consecutive epochs.

20,000 image pairs are used for training containing all unique images i.e. total 40,000 images. Additionally, relative labels are used for training i.e. a pair is labeled as 1 if $r_1 - r_2 > I$, otherwise it is labeled as -1 . Here, r_i is the average rating of I_i in AVA data-set. More details about the data-set are discussed in the previous section of this chapter.

3.3.2 Testing The Model:

Given a new pair of images, I first subtract the mean of the training data from each pixel of both the images. It is to be noted that the test set does not share any pairs or any individual images with the training and validation set. Both the images and their patches are first passed into our network to get the value of $d(I_1, I_2)$ from Equation 2. I_1 is then predicted as a more beautiful image than I_2 if $d(I_1, I_2) > 0$ and vice versa. The test set contains 19,841 image pairs. The weights of the epoch at which the achieve highest ranking accuracy with the least amount of validation loss are used for testing.

3.3.3 Determining Ranking Order Using a Binary Classification Network:

A network is trained on categorically-labeled data using my own implementation of the RAPID approach [11] as discussed in chapter II. It is trained on the same set of 40,000 images that is used to train the relative network. However, in this case, these images have been categorized as either beautiful or non-beautiful depending on the average ratings obtained directly from the AVA data-set. The threshold that determines the class of an image is set to 5.5, since the ratings in the AVA data-set range from 1-10. The network omits a probability measure $p(y = I/I)$ which is probability of an image I belonging to the beautiful class

	Ranking on the custom test-set	Ranking on the pairs from standard test-set	Classification on the custom test-set	Classification on the standard test-set
Base-line	62.21	65.87	59.92	69.18
Proposed	70.51	76.77	59.41	71.60

Table 2: Results for ranking and binary classification

While testing for a pair of input images, the first image is passed through the network and the probability measure $p(y = I/I_1)$ is obtained. Passing the second image gives the value of $p(y = I/I_2)$. The first image is decided to be more beautiful than the second one if $p(y = I/I_1) > p(y = I/I_2)$. This test set contains 19,841 image pairs and is identical to the test set used for our approach as mentioned in Section 3.2.4. A significantly lower accuracy is obtained on this relative ranking problem using a similar-sized network, which suggests that a network trained on categorically-labeled data fails to learn the complex, relative ranking order in the data.

3.3.4 Analyses on the Results:

The experiments are run using the relative-network on the custom test set and the standard test set containing 20,000 and 7,670 image pairs respectively. A ranking accuracy of 70.51% and 76.77% are achieved on the custom test-set and on the standard test-set respectively. Here, ranking accuracy is defined as the fraction of pairs for which the model correctly picks the more beautiful image as per the ground-truth labels. We compare our

approach with the base-line described in chapter II. RAPID (base-line) produces a ranking accuracy of 62.21% and 65.87% on the custom and the standard test-set respectively.

Due to the relative-learning-based approach, I believe that the network has gained a semantic-level understanding of the properties which make an image highly aesthetic. To verify this, the task of binary classification is attempted on the custom data-set as well as the standard test-set. For this purpose, the top channel of our network i.e. C_{11} and C_{12} (see Fig. 6) is extracted and a fully connected layer along with a sigmoid is connected in the end to convert the values into decision values. Only the last dense layer is trained using the binary classification training data, that is described in the chapter II.

Results are computed by passing the input test image through the network to obtain the probability of that image being beautiful. Both the custom and the standard test set consisting of 10,000 and 20,000 images respectively are used to compute the results. The proposed approach obtains 59.41% classification accuracy on the custom test set as compared to 59.92% obtained by the base-line. An accuracy of 71.60% is achieved on the standard test set as compared to 69.18% obtained by the base-line. The proposed network outperforms RAPID on the ranking task and produces competitive performance on the classification task. The results of all the experiments are summarized in Table 2.

We can therefore infer that a deep neural network trained with an appropriate loss function which accounts for such relatively labeled data, significantly outperforms a state-of-art network trained on same data with categorical labels. The proposed network is also able to achieve a competitive performance on an aesthetics classification problem with trivial modifications to its architecture and no fine-tuning at all. This shows that it has



Figure 8: Rankings produced by our network are shown above. Top and bottom rows show correct and wrong predictions respectively for a total of 4 pairs. Each of them are enclosed in either red/green boxes. For every pair, our network ranks the right image higher than the left image. Please view in color.

gained a certain semantic-level understanding of the factors involved in making an image aesthetic.

Fig. 8 illustrates some ranking results obtained by our network. The wrong predictions in the bottom row show that the network lacks semantic knowledge about objects and natural phenomena. For example, even though the picture containing two birds has better color harmony/contrast, the lightning phenomena is a rare capture, making it more picturesque.

3.3.5. Analyzing the Kernels and the Feature Maps:

Another analysis is performed by visualizing the weights of the kernels of both the binary classification network and the ranking network. For this task, the kernels of the first convolutional layer of the global column are extracted and printed in the form of images. This layer contains a total of 64 different kernels each of size 3 X 11 X 11. These kernels when displayed as images, result in RGB images as follows.



Figure 9: Visualization of the weights of the first convolutional layer for binary classification network.

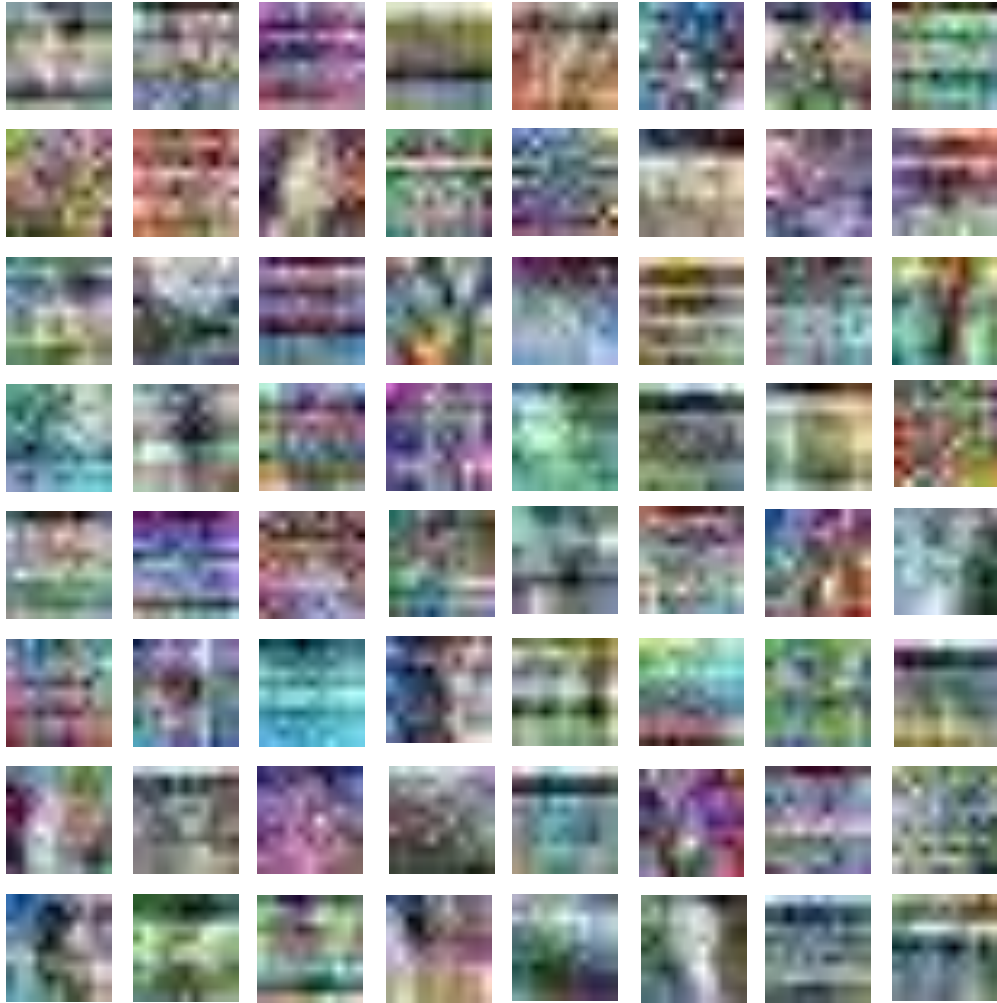


Figure 10: Visualization of the weights of the first convolutional layer for ranking network.

Figure 9 shows the kernels learnt by the binary classification network and figure 10 shows the kernels learnt by the ranking network. It can be noticed from the above figures that the ranking network learnt more defined kernels with some kind of patterns. On the other hand, the weights of the binary classification network seemed to have some level of randomness which may not be desired. This could be because of the fact that the binary classification network is missing higher amount of rating related information because of the set threshold

value of the binary classification data-set. However, ranking data-set retains this information to a better extent because of the relative labelling. As a result of this, the binary classification network would have learnt more vague and some level of generic features as compared to the ranking network which might have learnt sharp aesthetic related features. However, strong conclusions cannot be drawn by considering these kernel visualizations all alone.

Secondly, an image is passed into these two networks and the obtained feature maps from the first convolutional layer are visualized. A total of 64 feature maps are extracted and printed for both the networks one for each of the 64 kernels. The dimensions of the feature maps are 110 X 110 owing to the stride 2 of the previous convolutional layer.

Figure 11 shown the image that passed into the network and figures 12 and 13 are the feature maps obtained from the binary classification and the ranking networks respectively. It can be noticed from the below figures that the feature maps of the ranking network have high level of contrast between different regions of the images as compared to the binary classification's feature maps. This can be attributed to the properties of the specific image that is chosen, i.e. figure 11. Figure 11 is rich in colors and has the property of blurred back-ground. These properties seemed to be well captured by the feature maps of the ranking network as compared to the binary classification one. Also, if a different image is given to the network, then the networks might try to learn different kinds of properties instead of the contrast or blur. Therefore, more extensive analysis needs to be performed by passing various kinds of inputs into the network in order to draw a conclusion.



Figure 11: Input image passed into the networks in order to analyze the feature maps.



Figure 12: Feature maps obtained from the binary classification network.



Figure 13: Feature maps obtained from the ranking network.

Additionally, the task of analyzing what photography related properties the network learnt better requires extensive analyses of the data-set as well. Qualitative subjective tests are necessary in order to first build a sub-set of the data-set that satisfy some or all of the photographic properties and then utilize this to analyze the networks performance. This analysis is not performed as a part of this work and could be an interesting area to explore.

CHAPTER 4

INCORPORATING SALIENCY

4.1 AN EXISTING MODEL:

4.1.1 Description:

In this chapter, the role of visual attention in learning aesthetic features is explored. This idea is inspired from [30] where the authors believed that there exists a strong correlation between visual attention and visual aesthetics. They also assumed that the salient regions of a photograph contain the subject thus leading to a higher contribution in the aestheticism of the image. Further they proposed a method that incorporates the saliency information as a new set of features along with the pre-determined global features. These features are hand-crafted and a one-dimensional Support Vector Machine is used for the classification task in this work.

4.1.2 Drawbacks:

This model [30] exhibited significantly higher classification accuracy as compared to the then state-of-art approaches. However, the experiments presented in [30] are performed by extracting the top 10% and the bottom 10% images of the Photo.net data-set, which is an extremely small volume of images. This work therefore explores the role of saliency in determining the aesthetic features of an image on a large-scale data-set. Please note that the sub-set of AVA-dataset that is chosen in this work is not only huge in volume but also extremely difficult as it contains images with wide variety of semantic content. Additionally, the current data-set contains images with varied ratings, unlike the images in [30] which are either extremely beautiful (top rated images) or non-beautiful (bottom rated

images). Therefore, performing the experiments on this larger set of images may through more insight into the role of saliency in aesthetics.

Further, [30] may have some additional drawbacks arising from the fact that the representational power of their model is very limited. i.e they extracted the features using hand-crafted methods and used a linear SVM with no kernel for classification. Aesthetic related features in general may be quite complex and therefore may not be well-handled by a model with such low representational power. Therefore, the afore-mentioned deep-learning based model is utilized to enhance the performance of the system by incorporating saliency related information. Two separate sets of experiments are performed by inputting the saliency data independently as well as in combination with the global and local aesthetic features. The results are compared with the ones from chapter III.

4.1 FORMING THE SALIENCY-IMAGES

In order to facilitate learning of saliency based aesthetic features by the model, a saliency-enhanced input image is given to the network while training. This allows the model to learn the saliency related aesthetic features to a significant extent than the generic global/local aesthetic features. The saliency features thus obtained are used either independently or in combination with the global/local features to analyse the role of saliency in image aesthetics. This sub section discusses the approach for formation of saliency enhanced images that will be later utilized for training. From here on, this saliency enhanced image is called as *saliency-image* in this document.

The saliency-image is a version of the input image where the regions of the image that are more salient have higher intensity as compared to the ones that are not. This version

of the image ensures that more importance is given to the salient regions of the image than the insalient ones. i.e the neurons connected to the salient regions in the image receive higher input activation than the ones that are not (applicable only for the Convolutional Layers). This also ensures that the network learns the saliency-related aesthetic features to a significant extent than the global/local features without modifying the network architecture.

In order to generate the saliency-images, the saliency map of each image is generated and is then convolved with the image. In this work, the standard Graph Based Visual Saliency Model (GBVS) [31] is employed to extract the saliency map of each input image. Putting this into a mathematical form, the saliency-image I_s is generated as follows,

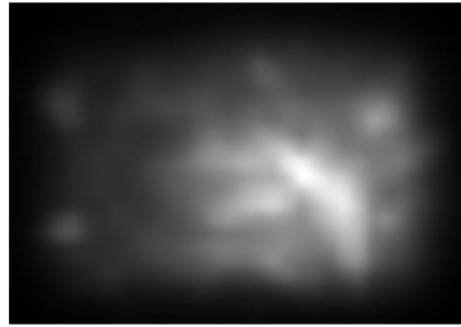
$$I_s = I_o * S \dots \dots \dots (4)$$

where, I_o is the original image, S is the generated saliency map and (*) represents the convolution operation.

The approach is explained with an example shown in figure 9. Figure 9a. shows the original image (I_o) which is picked from the AVA Data-set and figure 9b shows the corresponding saliency map. It can be noticed from fig. 9b that more than 70% of the image pixels are less than 30% salient. Masking out of all the image regions whose saliency value is less than 0.3 results in the saliency-image 9c. However, this variant of the saliency-image contains a large number of “0” valued pixels, which results in a bias when passed to the neural network. i.e patches that are blacked out to a larger extent result in less number of neuron activations therefore resulting in a small value of the output probabilities.



(a)



(b)



(c)



(d)



(e)



(f)

Figure 14: Variations of saliency-images for different saliency operations. (a) Input Image on which the operations are performed. (b) Saliency Map generated by GBVS (c) Saliency-image generated by masking out of pixels with saliency value less than 0.3 (d-e-f) Saliency-image generated by convolution of the input image and the saliency for $\alpha = 1, 0.2$ and 0.05 .

For this reason, instead of completely masking out the pixels, a convolution of the saliency map with the image is considered. This results in the figure 9d in the above example. However, most of the pixels are either completely black or on the low end of the image pixel value range in this version of the saliency-image as well. Therefore, a function of saliency map is used to convolve the image instead of direct usage of saliency map, i.e

$$I_s = I_o * f(S) \dots \dots \dots (5)$$

where f is chosen to be an exponential function in this work. i.e

$$f(S) = S^\alpha \dots \dots \dots (6)$$

Here, α is a parameter which can take any value between [0,1]. A value of '0' for α is equivalent to direct usage of the image without saliency information and a value of '1' is equivalent to direct convolution of the saliency map with the image. The values between (0,1) is similar to convolving the image with a saliency map that has the pixel values replaced by comparatively high pixel values as defined by the function. Please note that the saliency values, i.e the values of S are in the range [0,1] and not [0,255]. Thus raising an exponent of it which is in (0,1) range results in a higher function value than the input. Figure 10 is plotted for the readers' understanding of the function. As shown in the figure, a lower value of α leads to a saliency map that has more pixels on the higher end of the image pixel value range than a higher value of α . Figure 10 plots the curves for the modified saliency map values for three values of α , 1, 0.2 and 0.05. The lesser the value of α , the closer the new image(I_s) becomes equal to the original image(I_o). This means that a lower value of α results in a rather darker image than a higher value of α . An example of the resulting image for the values of $\alpha = 0.2$ and 0.05 are given in figures 9e and 9f respectively.

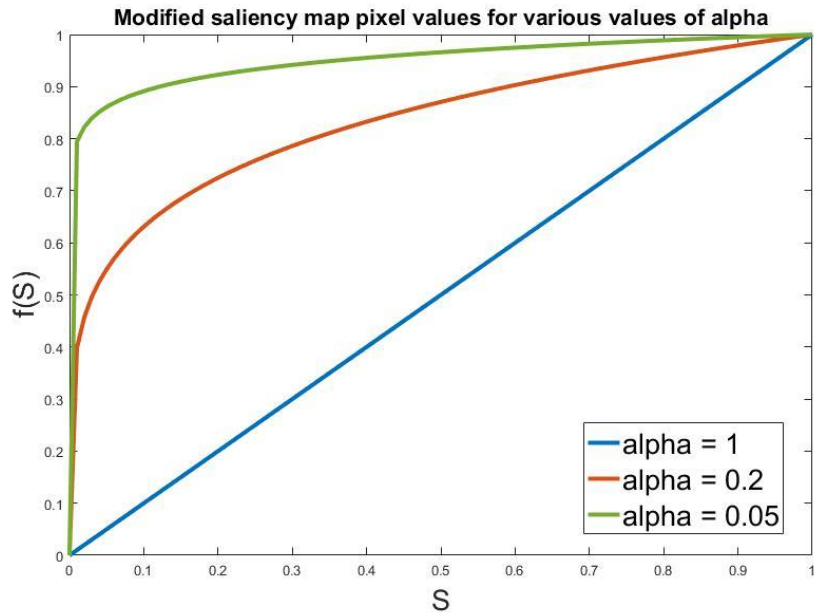


Figure 15: Modified saliency map pixel values for various values of α .

An optimum value of α results in saliency-images that capture the saliency information at best simultaneously resulting in lesser dark area in the image.

Saliency-images thus generated are used to train a deep convolutional neural network using the relative labels. More details about the experiments and the results are provided in the next section.

4.3 EXPERIMENTS AND RESULTS:

Two sets of experiments are performed to analyze if the performance of the system can be enhanced by incorporating the saliency-related information.

4.3.1 Training entirely on Saliency-images:

The intention of this set of experiments is to allow the network to see only the saliency-images thus making the model learn aestheticism entirely from the saliency-

related features. Therefore, the saliency-images are inputted into the network stand-alone, without inputting the original images.

A	Ranking Accuracy
1	58.04 %
0.2	60.27 %
0.05	60.78 %

Table 3: Performance of the system for various values of α

An identical model as presented in chapter III is used for this experiment, with the exception that the global and the local patches are extracted from the saliency-images instead of the original ones. The saliency images are extracted for all the images in the training and the test sets and used for this experiment. The same relative labels as described in chapter III are used for training. A total of 20,000 image pairs are used for training, 3,000 for validation and the remaining for testing. The ranking performance is evaluated on the exactly same test data-set that is used in chapter III. The experiment is repeated for three different values of α , 1, 0.2, 0.05.

Training of the model is performed using stochastic gradient descent with learning rate as 0.01 and regularizer as 0.05. Learning rate is reduced by 25% percent after each 10 epochs and trained until no significant improvement in validation accuracy is observed.

Table 3. summarizes the results obtained by training the models for various values of alpha. The training accuracy increased with decrease in α as expected. A higher value of α results in a significant darkening of the image thus resulting in poor results. As the value of α decreases, the image becomes more and more bright leading to an increased performance. Additionally, the low accuracy of ~60% can be attributed to the fact that the saliency related features are dominant in the network with little or no global/local image features. Further, the model is predicting the rank of the images with ~60% accuracy which depicts that the model learnt some kind of features than pure random guessing. However, it cannot be assured that the features learnt are saliency related aesthetic features. It might be the case that the model learnt generic image features (or some aesthetic related features) which results in such kinds of performance. In any case, training a network by the saliency-images stand-alone degrades the performance of the system to a significant extent.

Additionally, the huge degradation of performance can be attributed to the minute amount of darkness induced into the image or the changes in pixel intensity values. This demonstrating that an intact set of global/local features might be significant in determining the aestheticism of an image. Therefore, I try to incorporate all the features into one model and calculate the accuracy which is presented in the next sub-section.

4.3.2 Incorporating saliency into the existing model:

Another experiment is performed by training a model similar to the one described in chapter III, with the exception that the current model has three columns in each channel instead of two. The three columns take the Global, Local and the Saliency-Image as the inputs respectively. The expectation from this experiment is that the model now learns

saliency related features explicitly along with the global and the local ones. This model is trained with saliency-images generate only for a single value of $\alpha = 0.2$.

Further, this model is trained on relative labelled data with 20,000 image pairs for training. Three versions of each input image global view, local view and the saliency-image are inputted into the network. 3,000 image pairs are used for validation and the rest for testing. Training is performed in a similar way like mentioned in the previous section with stochastic gradient descent algorithm until convergence.

This model gave a ranking accuracy of 68.91% when tested on the same 19,414 image pairs used for testing the ranking architecture. The decrease in performance can be attributed to the possibility of the model directing into a different local-minima while training.

Additionally, this experiments shows that the addition of a third column to the network, i.e the saliency column did not improve the performance of the system. It can be said that the automatically learnt global features would have much efficiently captured the saliency related information, thus making the addition of a separate saliency module redundant. The improvement in accuracy in [30] can be attributed to the fact that the hand-crafted global features may not have sufficient captured the saliency information which was later captured by the approach presented in [30]. Therefore, it can be believed that incorporating saliency information into the system has no significant effect on the ranking performance, thus proving the addition of this module redundant.

CHAPTER 5

CONCLUSION

Many applications such as image retrieval, image enhancement require that images are ranked automatically based on their aesthetic features, thus attributing to little/no human involvement in execution of certain time-consuming tasks. Inspired from these applications, this work intends to attempt the task of ranking of images based on aesthetics. While most of the previous work in this area attempted to categorize the images into beautiful/non-beautiful classes, this work introduced the novel problem called *Relative Aesthetics* which deals with ranking images based on aesthetics instead of classifying them.

The main contributions of this work are as follows. A novel data-set is created which consists of image pairs and their relative labels, in order to attempt the problem of Relative Aesthetics. A new double channel deep convolutional neural network model is built and trained on this relatively labelled data. This approach facilitated for 1. Automatic feature extraction attributed to the high representation power of deep learning and 2. Better learning achieved from relative training than absolute training as mentioned in [16]. This model is trained and tested for the tasks of ranking and binary classification with improved results than a similar model trained on the classification data. This shows that a model trained with relative labels learnt better aesthetic feature representation.

Additionally, some analysis has been done on enhancing the performance of the model using saliency related information. Saliency-images are generated by convolving the original images with the Graph Based Visual Saliency maps and are used for training both stand-alone and in combination with the global and local features of the original image. In both the experiments, no performance enhancement is achieved showing that

either the saliency image features are already learnt by the model thus making adding of a separate saliency column redundant, or the saliency information as given to the network in the form of saliency-images is irrelevant in determining the aestheticism of an image.

Overall, this work is an attempt at trying to learn better aesthetic features that can be used for various tasks like classification and ranking. Results show that the proposed approach learnt better features than the state-of-art approaches. As a part of future work, I try to analyze the AVA data-set by performing some subjective evaluations. Additionally, I would like to analyze the drawbacks of the proposed approach and try to build efficient aesthetic feature learning models.

REFERENCES

- [1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Studying aesthetics in photographic images using a computational approach,” in *ECCV*. Springer, 2006, pp. 288–301.
- [2] Y. Ke, X. Tang, and F. Jing, “The design of high-level features for photo quality assessment,” in *IEEE CVPR*, vol. 1, 2006, pp. 419–426.
- [3] S. Bhattacharya, R. Sukthankar, and M. Shah, “A framework for photoquality assessment and enhancement based on visual aesthetics,” in *The 18th ACM international conference on Multimedia*, 2010, pp. 271–280.
- [4] Y. Luo and X. Tang, “Photo and video quality evaluation: Focusing on the subject,” in *ECCV*. Springer, 2008, pp. 386–399.
- [5] S. Dhar, V. Ordonez, and T. L. Berg, “High level describable attributes for predicting aesthetics and interestingness,” in *IEEE CVPR*. IEEE, 2011, pp. 1657–1664.
- [6] W. Luo, X. Wang, and X. Tang, “Content-based photo quality assessment,” in *IEEE ICCV*, 2011, pp. 2206–2213.
- [7] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, “Aesthetic quality classification of photographs based on color harmony,” in *IEEE CVPR*, 2011, pp. 33–40.
- [8] P. O’Donovan, A. Agarwala, and A. Hertzmann, “Color compatibility from large datasets,” in *ACM Transactions on Graphics (TOG)*, vol. 30, no. 4, 2011, p. 63.
- [9] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien, “Scenic photo quality assessment with bag of aesthetics-preserving features,” in *The 19th ACM international conference on Multimedia*, 2011, pp. 1213–1216.
- [10] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, “Assessing the aesthetic quality of photographs using generic image descriptors,” in *IEEE ICCV*, 2011, pp. 1784–1791.
- [11] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, “Rapid: Rating pictorial aesthetics using deep learning,” in *The ACM International Conference on Multimedia*, 2014, pp. 457–466.
- [12] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, “Deep multi-patch aggregation network for image style, aesthetics, and quality estimation,” in *IEEE ICCV*, 2015, pp. 990–998.

- [13] Z. Wang, F. Dolcos, D. Beck, S. Chang, and T. S. Huang, “Braininspired deep networks for image aesthetics assessment,” arXiv preprint arXiv:1601.04155, 2016.
- [14] J. Yan, S. Lin, S. B. Kang, and X. Tang, “A learning-to-rank approach for image color enhancement,” in IEEE CVPR, 2014, pp. 2987–2994.
- [15] N. Murray, L. Marchesotti, and F. Perronnin, “AVA: A large-scale database for aesthetic visual analysis,” in IEEE CVPR, 2012, pp. 2408–2415.
- [16] D. Parikh and K. Grauman, “Relative attributes,” in IEEE ICCV, 2011, pp. 503–510.
- [17] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [19] D. Ciresan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” in IEEE CVPR. IEEE, 2012, pp. 3642–3649.
- [20] F. Chen, H. Yu, R. Hu, and X. Zeng, “Deep learning shape priors for object segmentation,” in IEEE CVPR, June 2013.
- [21] Y. Sun, X. Wang, and X. Tang, “Deep convolutional network cascade for facial point detection,” in IEEE CVPR, June 2013.
- [22] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition,” arXiv preprint arXiv:1310.1531, 2013.
- [23] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, “Recognizing image style,” in *Proceedings of the British Machine Vision Conference.*, 2014.
- [24] A. Kovashka, D. Parikh, and K. Grauman, “Whittlesearch: Image search with relative attribute feedback,” in IEEE CVPR, 2012, pp. 2973–2980.
- [25] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Sackinger, and R. Shah, “Signature verification using a siamese time delay neural network,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 04, pp. 669–688, 1993.
- [26] O. Litzel, *On Photographic Composition*. Amphoto, 1975.

- [27] W. Niekamp, "An exploratory investigation into factors affecting visual balance," *ECTJ*, vol. 29, no. 1, pp. 37–48, 1981.
- [28] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, 2011.
- [29] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, 2011.
- [30] Wong, Lai-Kuan, and Kok-Lim Low. "Saliency-enhanced image aesthetics class prediction." 2009 16th IEEE International Conference on Image Processing (ICIP). IEEE, 2009.
- [31] Harel, Jonathan, Christof Koch, and Pietro Perona. "Graph-based visual saliency." *Advances in neural information processing systems*. 2006.