# INTERPRETING PAINTINGS USING IMAGE SEMANTIC SEGMENTATION AND DECISION TREES

JASPER HAVEN S. BRIONES, Ateneo de Davao University
KENNETHE ANN Q. MINA, Ateneo de Davao University

Although understanding art is subjective, there are factors which can be used as basis of interpretation. These are subject matters or objects and the colors present in the painting. There are several methods done in order to extract the objects and colors present in the paintings, such as using a convolutional neural network model for semantic segmentation, a color quantization-based algorithm for color extraction, and a random forest classifier to determine the potential emotions in the painting. A tenfold cross-validation is performed to increase the accuracy of the model. Realism painting dataset is used.

General Terms: Semantic Segmentation, Color Extraction, Convolutional Neural Networks, Random Forest Classifiers

Additional Key Words and Phrases: Plutchik's Wheel of Emotions

## 1. INTRODUCTION

### 1.1 Background of the Study

Interpreting paintings is not a simple task. It requires to go beyond what our eyes have perceived. They do not look to what the artist is trying to express, but solely judging the surface of the painting. Moreover, it would be much more difficult if the painting expresses a deeper meaning which is harder to interpret for other or most people. Furthermore, each one of us has a different opinion in defining what expressions did the artist convey resulting to a possible off-track from what they want to express. Because of this, the purpose of the artist in creating their arts would be in vain and the message they are trying to pass would not reach to the people looking at their works.

It would be feasible if there is a tool that would assist people looking to different works of art in interpreting the artist's flow of emotions. Such a tool would make the viewers of the painting have insights that they could never think about, thus having another perspective that would let them appreciate it more. Moreover, it would be easier for them to interpret these forms of art if the tool provides emotions present from the subject matter.

Image Semantic Segmentation is the process of understanding and recognizing an image by pixel level. It extracts features like shape, or color by dividing it into regions with boundaries in defining the objects present in an image. There are various approaches in this process, one popular approach is the use of Neural Networks.

Decision Tree is a diagram that branches out the possible outcomes of a certain input, which gives out a tree-like figure. This is commonly used when the factors affecting the outcomes are conditional statements. Each branch represents a statistical probability as to how the input should be interpreted. Random Forest is an ensemble of decision trees. The trees created in random forest will provide outputs in which majority vote is done to determine the final decision. Random forest, based on various studies, shows great performance in multi-class problems.

In order to accomplish such a task, this study uses the Image Semantic Segmentation approach in getting the possible subject matters seen and depicted by the painting, and Decision Trees for weighing and choosing the set of emotions present in the artwork. Furthermore, in light to address the people such as viewers and critics interpreting artworks by artists whom have expressed their creativity, the proponents pursued this study for handing out assistance to them. Through this study, it would greatly ease their tasks in terms of time efficiency and work efficiency.

## 1.2 Problem Statement

This study aims to classify paintings based on the emotions depicted in the image. This process is done through (i) a semantic segmentation approach, specifically convolutional neural network, (ii) a color quantization-based algorithm for color extraction, and (iii) a machine learning ensemble method, the Random Forest.

This study sought to address the following questions:

1. How to build a dataset of realism-themed paintings suitable for the study?
2. Is semantic segmentation the appropriate approach for feature extraction?
3. How to apply Random Forest in building a classification model?
4. How to evaluate the performance of the classification model?

## 1.3 Objectives

The general objective of this study is to be able to identify the possible emotions present by combining the features extracted from semantic segmentation and color extraction, and then feeding it to the random forest classifier as input.

The following are the specific objectives of the study:

1. Explain how to build a suitable realism-themed painting for the study.
2. Determine whether semantic segmentation is appropriate approach for feature extraction.
3. Explain how to apply Random Forest in building a classification model.
4. Explain how to evaluate the performance of the classification model.

## 1.4 Significance of the Study

The main beneficiaries of this study are those who are in need of assistance when it comes to interpreting and in getting more insights of a painting. Critics or those who are aspiring to become critics in the field of art may use this study to have a broader idea on what the painting is about. Future researchers may also use this study to build a system that would categorize the paintings based on emotions. It would be useful for art curators that has a huge collection of paintings.

## 1.5 Scope and Limitations

The list of emotions will be based on Plutchik's wheel of emotions. Only the 8 basic bipolar emotions are taken. The dataset for training and testing will only contain realism-themed paintings. These will be taken from WikiArt, Google Images, or on other websites that provides public realism-themed datasets. The experts or annotators for constructing the ground truth will be selected individuals from the Ateneo de Davao University Humanities division.

# 2. REVIEW OF RELATED LITERATURE

## 2.1 Interpretation

### 2.1.1 Plutchik's Wheel of Emotions

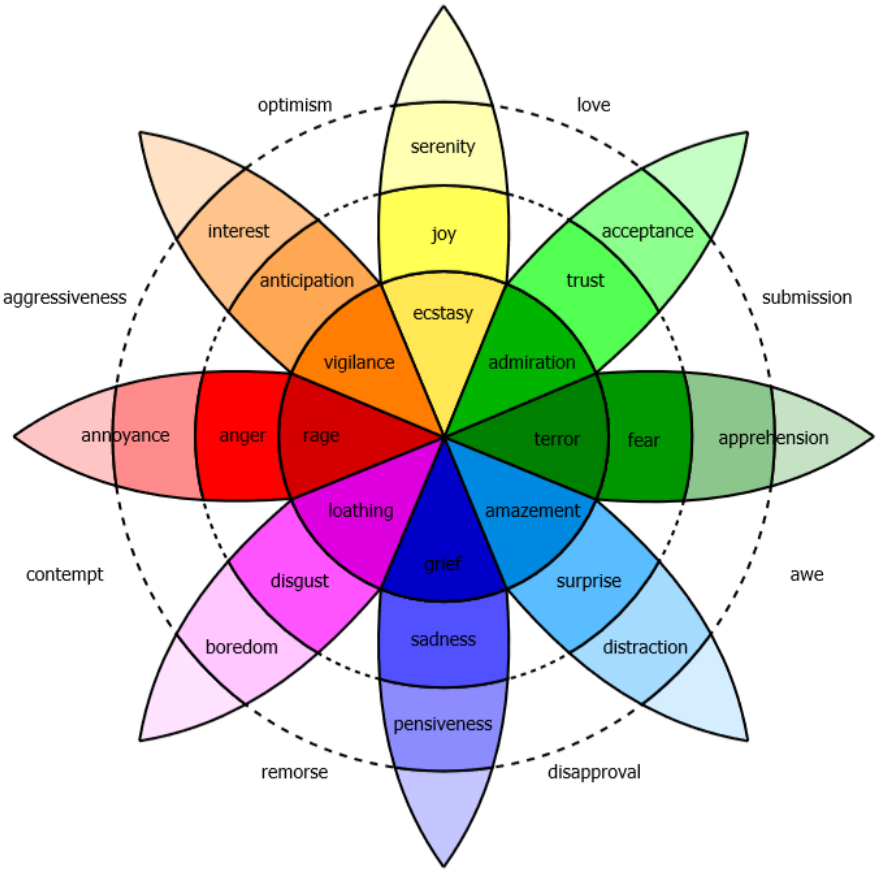



Fig. 1. Circumplex structure of emotion concepts based on similarity scaling and on the semantic differential *[1].*

The wheel above shows the correlation or relatedness of each emotions from one another. The 8 basic bipolar emotions, as stated and was proposed in 1958 by Robert Plutchrik, are as follows: joy, sorrow, anger, fear, acceptance, disgust, surprise and expectancy. Dr. Plutchik came with this circumplex after reviewing various studies of the similarity of emotions.

### 2.1.2 Image Interpretation

Image interpretation is the analysis of a certain image that one must understand its context before drawing conclusions afterwards. There should be understanding of the main subject matter depicted in the image having to derive the semantics of the interconnected relationships from its main context.

Based on a recent research from Aditya *et al.* [2], the proponents of that study faced a problem in the area of image understanding under computer vision that commonly most approaches would only describe the salient aspects of an image. It would not describe all the aspects with reasonings to back up the connections of its contents that are present. From this, they were motivated to model an architecture that is based on how humans would interpret an image. Moreover, this human

perception of interpreting consists of an interaction of both visual input and language that would come afterwards, thus giving out the semantics and understanding.

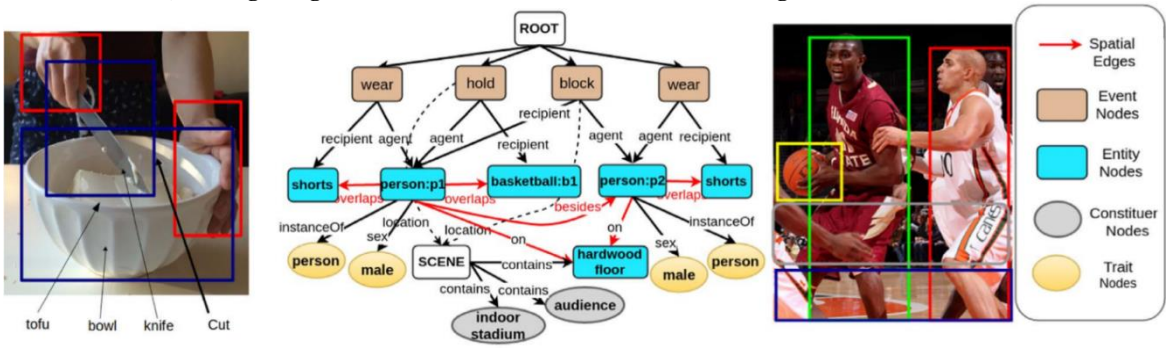



Fig. 2a. Corresponding ideal SDG encoding semantic, ontological, and spatial relations.

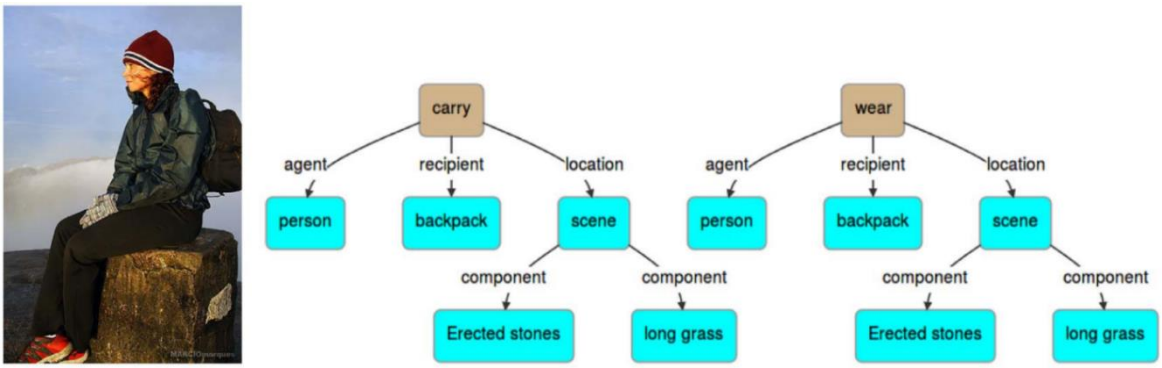



Fig. 2b. An example of an SDG.

From this study, they have used the approach called Scene Description Graph (SDG). This approach is defined as a graph labeled directly representing objects, actions, regions, as well as their attributes together with the concepts inferred and semantic, ontological, and spatial relations. Furthermore, SDG depicts the semantics of a given scene, and having an integration of direct visual knowledge and background common sense knowledge, Moreover, SDG has a similar structure in comparison to semantic structure of sentences, thus having an interaction between vision and natural language.

They had concluded their study to an extent where their evaluation from their generated output (sentences) is quite thorough and relevant. Their output was considered not as informative as the studies that have used existing neural approaches. Furthermore, they had ended it having said their proposed architecture of work can be used to properly elaborate the results shown and evaluate its error sources, be it from their visual detection, knowledge base or reasoning module.

Our study would differentiate from our method in terms of their uses a graph to formulate the semantics and relationships detected from their visual detection module while we would not use this kind of approach.

#### 2.1.2.1 Image Caption Generation

Karpathy *et al.* [3] introduces a problem where they recognized the remarkable ability of how humans can describe an image at first glance while existing and previous visual recognition models
are having difficulty in accomplishing the same level as how humans would do. They also mentioned that even though there are a lot of convenient models in labeling images with a

fixed set of visual categories, it still has a great restriction compared to the vast and numerous descriptions that a human can think of.

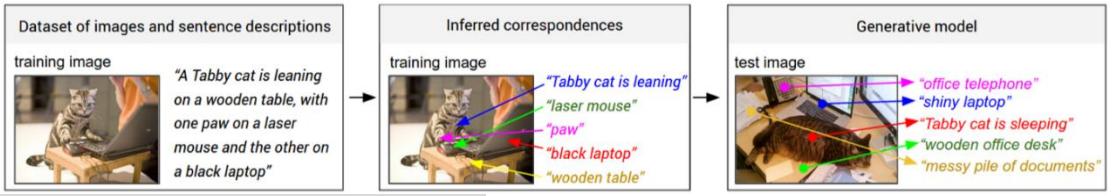



Fig. 3. Overview of the approach by Karpathy *et al.*

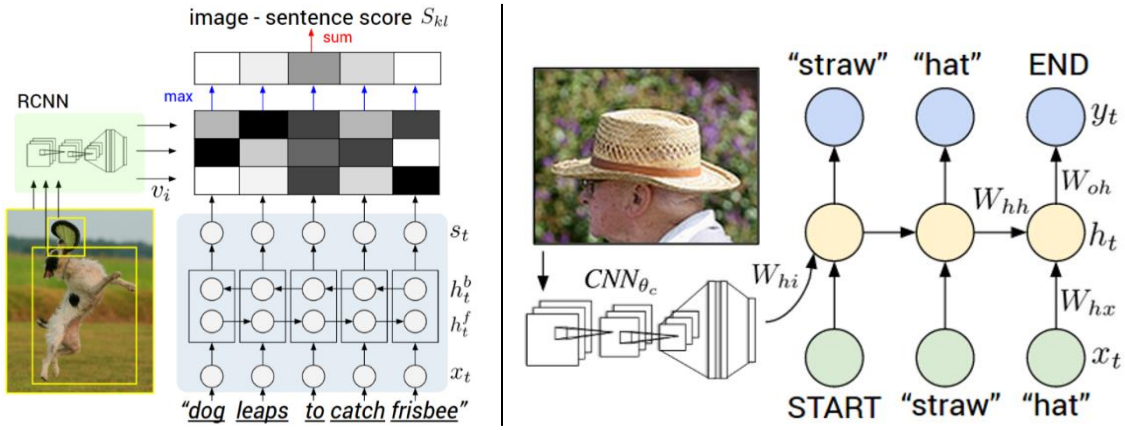



Fig 4. CNN-BRNN architecture model (left). Multimodal Recurrent Neural Network architecture model (right).

From this, they were motivated to make a model in generating deep descriptions of images. They aimed to formulate a model design where it is rich enough in reasoning the contents of an image. Furthermore, to accomplish their goals, they have used Convolutional Neural Networks (RNN) over sentences, and a model that would align the previous two models through the use of multimodal embedding. Lastly, using those alignments from describing a Multimodal Recurrent Neural Network architecture, it learns to generate novel descriptions of image regions.

They concluded their study that it outperforms its retrieval baselines from its evaluation of performance on both full-frame and region-level experimentations. In addition, a study by Vinyals *et al.* [4] that tackles about the fundamental problem in artificial intelligence that relates with computer vision. Having said that researchers from computer visions aim to describe an image with a deep sense of semantic analysis. Their study uses deep CNN as an encoder for their image classification tasks and from having reached the last hidden layer will be then used as input for their RNN as a decoder in generating sentences or captions from an image. They called their approach or model as Neural Image Caption (NIC).

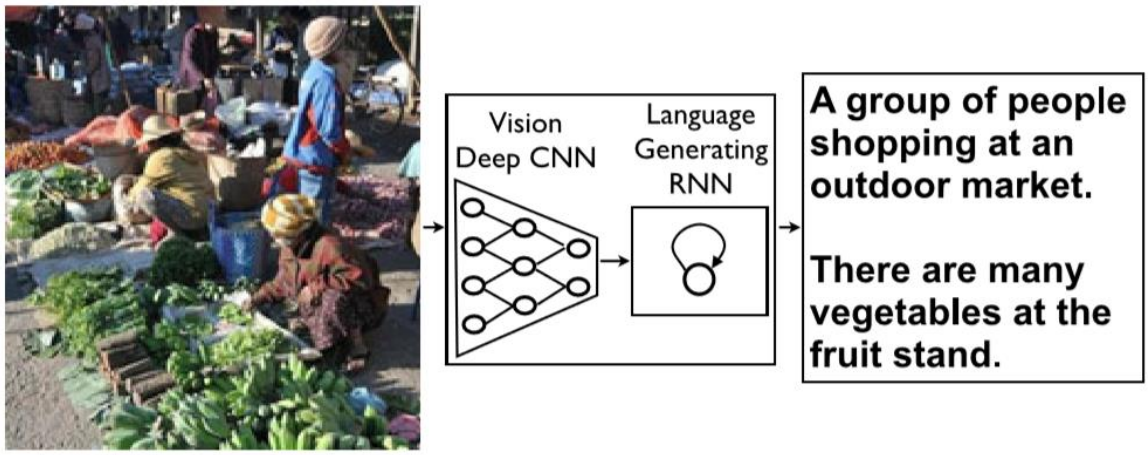



Fig 5a. Neural Image Caption model; vision CNN followed by a language generating RNN.

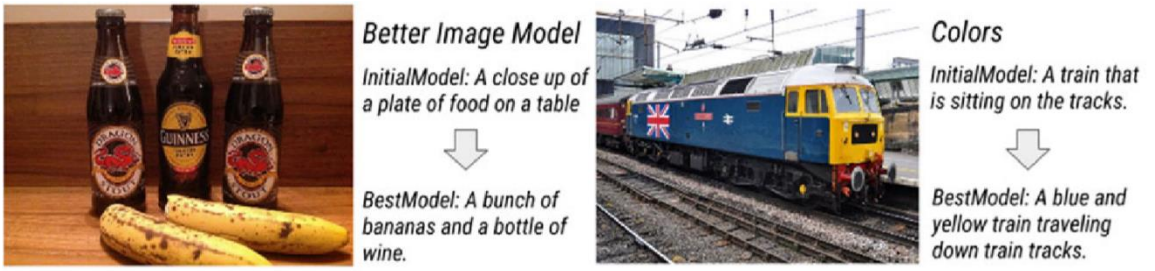



Fig 5b. Comparison of Vinyal *et al.*'s initial model and best model.

Furthermore, their study had quite good and accurate results from their initial model, and after having made their best model from their participation of a contest called 2015 MS COCO Challenge, they have reached first in rank from automatic and human evaluation. The researches for this study said descriptions of an image through natural language while our method would have this of integration.

Of the two studies about image caption generation being reviewed, they used an integration of models between convolutional neural networks (CNN) and recurrent neural networks (RNN) in generating descriptions of an image through natural language while this work will not include this kind of integration.

#### 2.1.2.2 Painting Scene Recognition using Homogenous Shapes

A study by Condorovici *et al.* [5] was conducted to address the problem of analyzing the semantics of a painting by automatic detection of the represented scene type. It had been discussed in this study that there are three possibilities in the analyzing level of interpretation. First, it would consist of the pictorial information such as the technique, thickness of brush strokes, type of painting material, and color composition from a low-level analysis. Second, the mid-level would focus on the specific objects, types of painting, or subject. Third and last, the background data would be looked upon which consists historical events or artist and period in general. From this and its related literatures, they had situated their analysis as mid-level for their system.

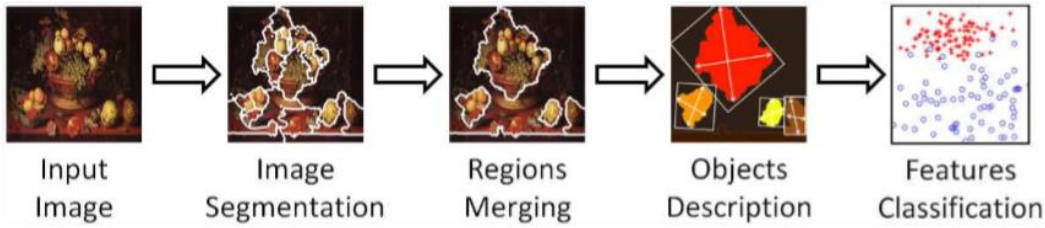



Fig 6. Overview of the algorithm by the study of Condorovici *et al.*

The perspective of their approach is perceptual and based on the Gestalt principles. They proposed their solution to model this kind of perspective as to how humans would perceive objects. To achieve this kind of model, they have used image segmentation to extract the basic components of the input that is used. After having segmented the image, they will merge the regions detected based on the Gestalt principles. Moreover, they will determine the object description which consists of color space, feature selection, object selection, and the number of objects. Lastly, they had made a classifier (a bagged ensemble of 25 decision trees) for the detection of scene type which had resulted a 67.4% accuracy as its best classifier result.

There was a disadvantage of the study that the classification of a scene types had a result of confusion such as the scene types between nude and still life, a landscape and cityscape, thus having led to a decrease of the overall accuracy. But nonetheless, the results were still acceptable for all scene types.

Finally, they had concluded their study that it outperforms the GIST solution from their own dataset, thus successfully classifying scene types (portrait, nude, landscape, cityscape, and still life) from a 500 images database.

This study in regards to our method would contrast in terms of the Gestalt principles and theory for the researchers of this study would use this as basis for extracting the features present in the image such as objects and the image properties.

#### 2.1.2.3 Domain Adaptation for Enhancing Deep Networks Capacity to De-Abstract Art

A study by Badeo *et al.* [6] tackles on how neural system approaches can have a grasp of understanding art and able to recognize the painting's genre (subject) from both abstract and artistic scenes. Moreover, differentiating the abstraction level achieved by deep convolutional neural networks and human performance. Also, having the concern of whether neural networks can pass through the abstraction limitation of painting and able to recognize the subject depicted by the painting or its scene type.

To resolve this issue, they have developed a state-of-the-art CNN from scratch using Residual Network (ResNet) as their architecture having 34 layers, and WikiArt as their database consisting approximately 80,000 digitized paintings for the training process. Having said their method, they reached an accuracy of 61.64% as their top accuracy from 26 classes which ran through the process 5 times. But when a process of augmentation was integrated, their accuracy had an increase of approximately 2% which resulted to a 63.58% accuracy. In conclusion, they ended their study that CNN are similar to humans when analyzing art with an abstraction that the deeper the abstract, the difficult it is to recognize the painting's genre. Also, CNN learns better if more works of art are presented into the training set.

#### 2.1.2.4 Genre and Style based Painting Classification

A study from Agarwal *et al.* [7] as reviewed by Badea *et al.* [6] of having a similar approach as the study of Condorovici *et al.* [5] (classical feature plus classifier) tested theirs with a 1500 images database. Having to address the problem of feature extraction on paintings, they had aim to classify them based on their genres and styles. The results they came up with had an

accuracy of 84.56% for genre classification that consisted 6 genres as their scope, and 62.37% accuracy for style classification having made 10 styles within the bounds of their study.

From these related literatures, we proponents have decided to pursue an approach where we will use Image Semantic Segmentation as to extract the objects and features present in the painting image at the field of visual analytics. On the other hand, we will use Decision Trees to integrate the relationships of the features from a pre-trained dataset of known specific category of subject matters in giving out a possible list of ideas and semantics depicted from the chosen images, specifically digitized paintings.

## 2.2 Computer Vision

Computer vision enables machines to generate descriptions or descriptions or interpretations of an image input. Doing so is not a simple task as it requires the right method or way of training where a single factor may significantly affect the results. Applications in this field includes facial recognition, object detection, and in different variations of segmenting tasks.

### 2.2.1 Semantic Segmentation

The aim of segmentation in images is to locate groups of pixels that go together, in other words, understanding the image in pixel-level [8] [9] [10] [11] [12]. Adding semantics in the process of segmentation means that segmented objects are assigned to a particular class. Most of the results of previous studies under this category showed accuracy where each research used different approach.

#### 2.2.1.1 Convolutional Neural Networks

According to Lecun [13], convolutional neural networks are neural networks that recognize patterns visually at a pixel image and that there is only a little amount of preprocessing needed. The patterns that a CNN can recognize has a lot of variations such as handwritten characters or simple geometric transformations, and those patterns can either be distorted or clear.

#### 2.2.1.2 Stochastic Gradient Descent

A review from the study of Andersson *et al.* [14] defines stochastic gradient descent (SGD) as an algorithm that is the simplified version of back propagation where it minimizes the result of the loss-function (a function that computes the difference between the expected value the predicted value), and calculates the result computed with respect to all weights and updates of the weights accordingly to have an optimized result. SGD, in comparison, only takes a subset of the whole dataset, thus having an approximation of the true gradient computed from the whole dataset and is faster.

$$\min_{w} \frac{1}{N} \sum_{n=1}^{N} L(\boldsymbol{p_n}, \boldsymbol{y_n} | W)$$

Where $N$ is the total number of samples for training, $\boldsymbol{p_n}$ is the predicted output, $\boldsymbol{y_n}$ is the desired output, $L$ is the loss-function, and $W$ is the set of weights and biases.

#### 2.2.1.3 Batch Normalization

The same study by Andersson *et al.* [14] defines batch normalization as the process of normalizing the amplification is called internal covariance shift. Moreover, batch normalization addresses the constant needs of the network in adapting from the new distributions being distributed inside. Although SGD is efficient enough in normalizing the

data, but constant updates must be processed as each layer goes by for the better performance in reducing the number of trainings.

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

The equation above is for computing the mean of the mini-batch.

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2$$

The variance of the mini-batch is computed using the formula above.

$$\hat{x}_i \leftarrow \frac{(x_i - \mu_{\mathcal{B}})}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$

Normalization of data distributed is then computed.

$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \mathbf{BN}_{\gamma,\beta}(x_i)$$

Finally, the scaling and shifting of data is computed.

#### 2.2.1.4 Region Convolutional Neural Networks

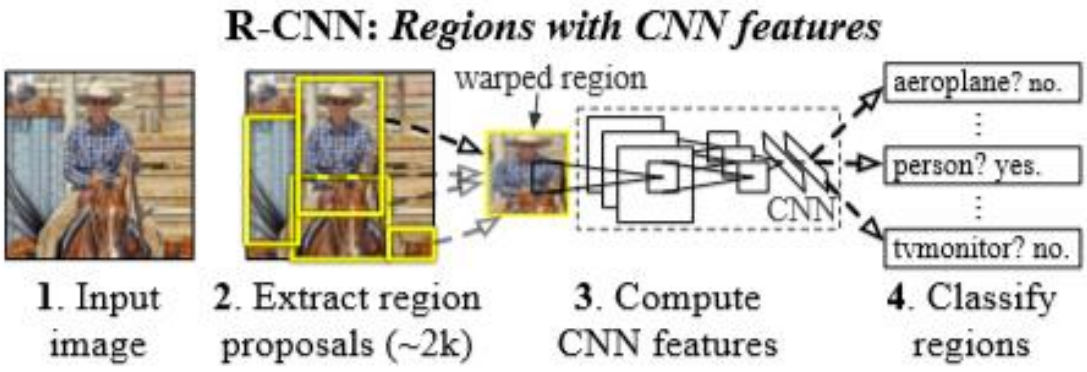



Fig 7. Proposed process flow of Donahue *et al.*'s system

A study by Donahue *et al.* [15] has presented an algorithm, a simple and scalable object detection type, that gives 30% of relative improvement from the previous state-of-the-art results on PASCAL VOC 2012. Their approach was to divide the image into regions which is considered to be a proposal for the localization and segmentation of the objects that are to be detected. From this region proposal, around 2000 would be made and each of those proposals would have a computation in getting the features used from a large CNN. Furthermore, it classifies those regions using a class-specific linear SVM's. It resulted to a mean average precision (mAP) score of 53.3%.

#### 2.2.1.5 Deep Convolutional Neural Networks

Islam *et al.* [16] used deep convolutional neural networks to perform dense image labeling. The approach of the said study involves two dense image labeling tasks, the semantic segmentation, which aims to label objects according to its category at a pixel level, and geometric labeling, which aims to label each pixel according to its geometrical class. They had 5 different configurations in the experimentation. Finally, they concluded that their model outperformed all the baseline methods.

#### 2.2.1.6 Fully Convolutional Networks

Shelhamer *et al.* [17] proposed to use fully convolutional networks (FCN) for semantic segmentation as it improves the performance of the machine compared to the previous state-of-the-art studies that used CNN. Results of their study, as shown in Table I., had the best outcome compared to the two. It had 20% relative improvement.

Table I. Results of the model Shelhamer *et al.* built for semantic segmentation.

| | mean IU VOC2011 test | mean IU VOC2012 test | inference time |
|---|---|---|---|
| *R-CNN* | 47.9 | - | - |
| *SDS* | 52.6 | 51.6 | ~50 s |
| ***FCN-8s*** | **62.7** | **62.2** | **~175 ms** |

Moreover, the advantage of this study from the previous studies [18,19,20,21,22,23,24,25] that only uses CNN is that instead of having fully connected layers in the last layers before classification, it uses convolutional layers having a kernel size of 1x1. Following that, to recover from the spatial loss gained from those convolutional layers, unpooling layers are placed. As a result, output maps are of the same resolution or dimensions as the input from the model. Another advantage is this study also defines a new architecture called *skip architecture*. This architecture skips the connections between the pooling and unpooling layers in honing the semantics further and the spatial precision of the output [12].

##### 2.2.1.6.1 SegNet

Another study that uses FCN as basis for semantic segmentation, addressing to design an efficient architecture for better road and indoor scene understanding, is proposed by Vijay *et al.* [24] having architecture of encoder-decoder, consisting of 13 convolutional layers at each network. The layers of these networks are not fully connected; hence they are only convolutional. With this kind of structure, they were able to retain higher resolution maps at the deepest encoder output to significantly lessen the number of parameters (from 134M to 14.7M), thus SegNet becoming easier to train, this lead to results achieving high scores in road scene understanding on the well-known and large datasets.

#### 2.2.1.7 Texton Forest

To understand the idea behind texton forest, we must first define what are textons. Textons, according to Julesz [26], refers to the fundamental micro-structures are present in natural images and are considered as the atoms of pre-attentive human visual perception. Which means that these are the things that humans subconsciously process before the conscious mind takes place.

A study conducted by [27] built a model that would categorize and segment objects according to their classes. They used semantic texton forests (STFs) to give an image-level output or description. Each pixel has a segmentation forest acting on them to improve speed.

#### 2.2.1.8 Conditional Random Field

Zeng *et al.* [28] used dense semantic segmentation to give a more accurate result in describing the concept of an image based on the objects and attributes present in the said image than the proposed approach by other studies. The said study used conditional random field (CRF) to model the relationships of the attributes and objects to label at the pixel level. It had an average label accuracy improvement by 42% on object class segmentation. It was also stated that when both factors are not jointly used, such as removing pixel-level, region-level, or both of them, the accuracy reduces by 5%, 4.4%, and 10.1% respectively. This means that the attributes serve a significant role in object segmentation. The overall study had an average label-accuracy of 61.4% (aNYU dataset).

#### 2.2.1.9 Dataset Ratio

A study by Horváth *et al.* [29] prepared their training and validation, and test image dataset in mixing their own annotated images and from public repositories in a ratio of 60% and 40% respectively. Their study was to aim in recognizing roads accurately through the use of fully convolutional neural networks for a self-driving car having to participate in a Shell Eco Marathon competition by 2018.

Another study by Badea *et al.* [6] uses WikiArt as their dataset for training their model. Out of the 79,434 images used having 26 classes, 75,302 were for training, and 4,132 for evaluation and testing. This would show a 80%:20% ratio for their training and testing set respectively.

### 2.2.2 Color Extraction

A proposed study by Thyagharajan *et al.* [30] extracts the dominant colors present in an image by the process of quantization—a pre-processing phase for images before any mathematical operations are performed—using Hidden-value learned K-mean Clustering. In this study, they had used the color space HSV (Hue, Saturation, Value) in getting the pre-dominant color from its original color model which in this case had used RGB. Furthermore, they had chosen HSV from the reason that this color model gives give a better extraction of hue since it depends on the saturation and brightness values of the image. Their form of extraction display is through a histogram chart where it showcases what color is dominant from the input image hence having indexed the images according to its highest color histogram value.

**ALGORITHM 1:** Prevalent Color Indexing

1. Begin
2. Normalize the EMK quantized(I x J) image, RGB values between [0 – 1] by dividing the values with 255
3. Let $P_x$ be the maximum of normalized RGB value and $P_n$ be the minimum value of normalized RGB
4. The Brightness value V be $P_x$ ; Define D = $M_x$ - $M_n$
5. If $P_x$ = 0 The Saturation value S = 0; At this condition Hue is undefined and the pixel is Grey in color it won't reveal the true color of the pixel
6. Else S = D / $P_x$
7. H is also defined if D = 0
8. H is computed in degree by
    a. If $P_x$ = R and G > B then H = (60*(G-B))/D

b. If $P_x$ = R and G < B then H = (360+60*(G-B))/D (adding the 36- degree to the value)
c. If $P_x$ = G then H = (60*(2+(B-r))/D
d. Else H = (60*(4+(R-G))/D

9. H can be also be computed in decimal value by
   a. (√3 (G – B)) / ((2 * R) – G – B)
10. Compute the total number of color band pixel for 6 color class {Red, Yellow, Green, Cyan, Blue, Magenta} using the H matrix value as shown:
    a. Let P = Total number of pixels in the image
    b. for i : = 0 to I
    c. for j : = 0 to J
    d. if ( H(i,j) == ( -1 to .083) || .9167 to -1) $R_x$++
    e. else if ( H(i,j) == ( .083 to 0.25) $Y_x$++
    f. else if ( H(i,j) == ( .25 to 0.416) $G_x$++
    g. else if ( H(i,j) == ( .416 to 0.5833) $C_x$++
    h. else if ( H(i,j) == ( .5833 to 0.75) $B_x$++
    i. else if ( H(i,j) == ( .75 to 0.91673) $Y_x$++
    j. end
    k. end
    l. End
    m. Then compute the total number of each color pixel as
    n. Red color pixel = $R_x$ / P
    o. Yellow color pixel $Y_x$ / P
    p. Green color pixel $G_x$ / P
    q. Cyan color pixel $C_x$ / P
    r. Blue color pixel $B_x$ / P
    s. Magenta color pixel $M_x$ / P
11. Prevalent color = Maximum {Red, Yellow, Green, Cyan, Blue, Magenta} of the HSV converted image
12. Save the images as per there maximum color band value in their respective color classes
13. End

This algorithm counts the color classes present per pixel in analyzing each of the pixel's hue value, and then outputs the values of the color classes present in the image, thus giving out also the prevalent color class.

### 2.2.3 Classifiers

Image categorization, also known as image classification, is used to group images of the same class.

#### 2.2.3.1 Naïve Bayes Classifier

This classifier is based on the theorem of Bayes, also known as the Bayesian Classifier. In Bayesian terms, X is considered as "proof". On a set of n attributes, they are described by the measurements made. Let F be some hypothesis, such that X is a data tuple that belongs to class L. In the environment of classification problems, the aim is to determine P(F|X), holding the probability of hypothesis F, given the "proof" or the observed tuple data X. P(F|X) is the posterior probability of F by the condition of X [31].

$$\mathcal{P}(\mathcal{F}|\mathcal{X}) = \frac{\mathcal{P}(\mathcal{X}|\mathcal{F})_{\mathcal{P}(F)}}{\mathcal{P}(\mathcal{X})}$$

The posterior probability is approximated as shown above.

A study by Kadar *et al.* [32] uses the naïve bayes classifier for the criteria assessment of quality classifications of a certain candidate for a particular position of company X. From their study, their classifier classified the HR quality classifications per candidate according to divisions of criteria based on the aspect of the assessment. The study turned out having a candidate matched a certain criterion after being classified by the classifier based on the quality classifications of the HR of company X.

#### 2.2.3.2 Random Forest

Random forest is an ensemble of trees. Various multi-class studies used random forest classifiers which gave them high accuracies. A module from `scikit-learn.org` provides a method for different ensemble learning classifiers including a random forest classifier.

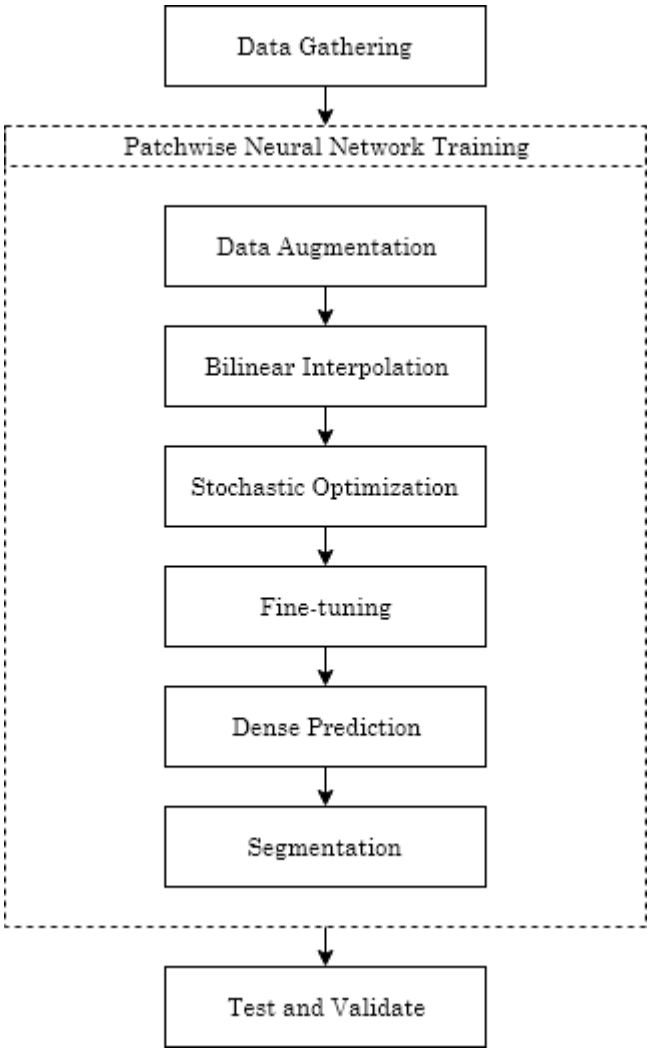
**ALGORITHM 2:** Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import make_classification

X, y = make_classification(n_samples=1000, n_features=4,
                           n_informative=2, n_redundant=0,
                    random_state=0, shuffle=False)
clf = RandomForestClassifier(max_depth=2, random_state=0)
clf.fit(X, y)

print(clf.feature_importances_)
print(clf.predict([[0,0,0,0]]))
```

In the module *sklearn.ensembles*, Random Forest Classifier is already implemented. It can be used by doing the method call *RandomForestClassifier()* [33].

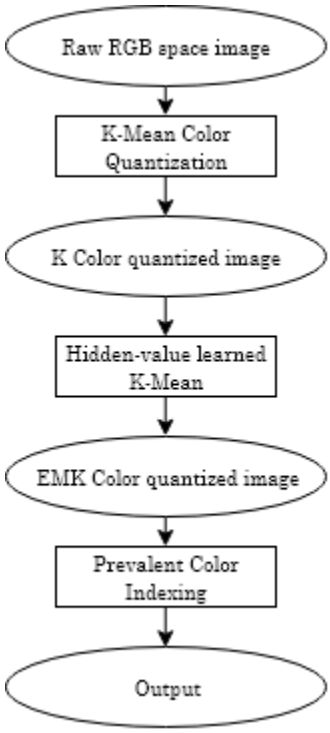
## 2.3 Theoretical Framework

### 2.3.1 Semantic Image Segmentation




Here, the researchers of this study present a theoretical framework for the process of semantic image segmentation based on the study of Shelhamer *et al.* [17] . They have used patchwise as their training style for correcting class imbalances, faster training, and effectivity. In the training process, data is augmented first, then proceeds to bilinear interpolation—implementing both forward and backward passes of convolution to connect coarse outputs to dense pixels. Afterwards, it goes through the process of optimization where error value must be at the minimum. Furthermore, fine-tuning layers by back-propagation, and pixel-wise labelling prediction was implemented before giving a segmented output image.

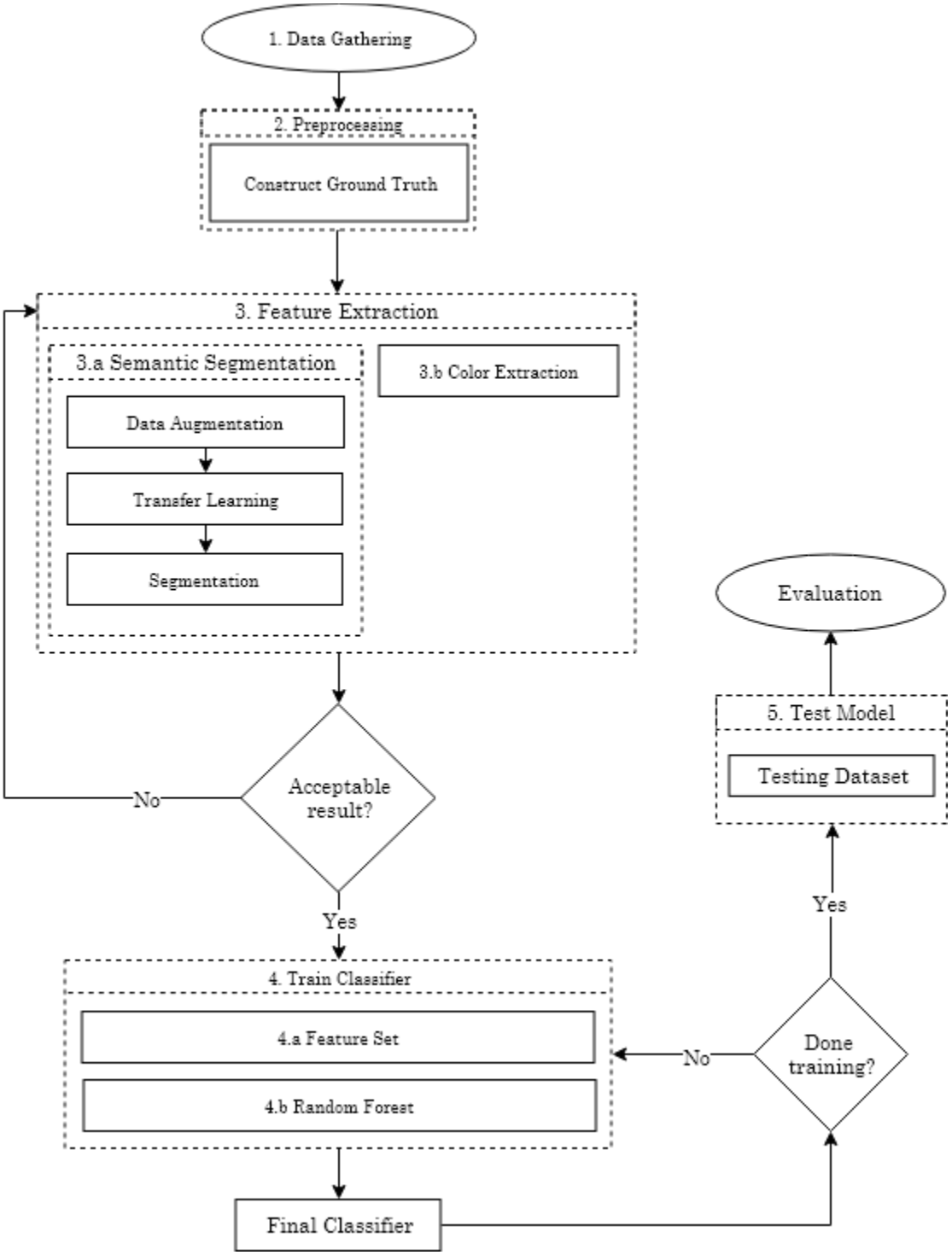
### 2.3.2 Prevalent Color Extraction and Indexing




This is a theoretical framework based on the study of Thyagharajan *et al.* [30]. The study presents 3 algorithms for color quantization and extraction. The algorithm K-Mean Color Quantization accepts a raw RGB space image as input, and after the process has taken place, it would then result to color quantized image having K number of colors. This output will then be fed to an algorithm called Hidden-value learned K-Mean Clustering (EMK) which covers the degradation of the previous algorithm, and outputs then a final color quantized image. After this process, the output of the previous two algorithms being chained together will then be used as an input for indexing the dominant colors present in the quantized image. After having determined the prevalent color class of an image, it would then index it according to its output.

# 3. RESEARCH DESIGN AND METHODOLOGY

## 3.1 Conceptual Framework



## 3.2 Methodology

In this research, it will undergo these phases:

1. Data Gathering
2. Preprocessing
   a. Construct Ground Truth
3. Feature Extraction
   a. Semantic Segmentation
      i. Data Augmentation
      ii. Transfer Learning
      iii. Segmentation
   b. Color Extraction
4. Train Classifier
   a. Feature Set
   b. Random Forest
5. Test Model
6. Evaluation

### 3.2.1 Data Gathering

The data gathered or collected will be in the form of digitized paintings that are available on the Internet. This data will be collected at a website called WikiArt [6] that offers a dataset of their own scope—by artist, style, period, etc.—and the style chosen will be *realism*. In addition, WikiArt has 13,982 paintings of realism from their collection, thus the researchers have decided to select 3,000 digitized paintings since resources are limited for this research. Furthermore, the dimensions of each data are not fixed, thus having the flexibility that these digitized paintings can be of any size, and that the architecture used supports this kind of condition.

### 3.2.2 Preprocessing

Since the dataset that will be used in this research is built by getting the paintings online from various websites, the data collected is not annotated as to what kinds of emotions do the paintings depict. Therefore, constructing the ground truth is required.

#### 3.2.2.1 Construct Ground Truth

In constructing ground truth, the data gathered will be annotated by at least 2 persons from the Ateneo de Davao University Humanities division. Cohen's Kappa Statistic is then applied to measure the interrater reliability. The paintings will be annotated based on what emotions are present in the painting. The list of emotions will be taken from Robert Plutchik's wheel of emotions [1].

### 3.2.3 Feature Extraction

This phase is split in to two parts. One is for the semantic segmentation, the other for color extraction.

#### 3.2.3.1 Semantic Segmentation

Through the use of FCN model by Shelhamer *et al.* [17], semantic segmentation is done on the image input. Multiple iterations done in [17] will also be used to determine how each one would perform on paintings.

##### 3.2.3.1.1 Data Augmentation

Due to the dataset being limited considering the resources also being bounded by a certain capacity, data will then be augmented. This process will consist of having flips (horizontal, or vertical) and rotations by a set of degrees in achieving good performance and accuracy in compensation for the limited resources. Moreover, as said by [6], data augmentation can also improve recognition performance, thus achieving an increase of accuracy.

#### 3.2.3.1.2 Transfer Learning

This method will be used in this research due to the limited resources that the researchers have. The researchers will use the same deep architecture as Shelhamer *et al.* [17]. Basically, what transfer learning does is to use a model that was trained from a large dataset previously and will be then used in having another unique dataset, freezing the lower layers for they are the feature selectors, and creating new layers for the model to adapt and learn from your own dataset, thus creating a bottleneck view for the model in the classification phase. Since the chosen NN was only trained from natural images [17], and the researchers are to use images with artistic styles, so they had thought that this process would be necessary for having a good accuracy for the classification process. As to why the researchers have chosen this kind of architecture is because first that it supports arbitrary sizes for input images among the layers. Second, it is said by [12] that the said architecture is the basic method used amongst the later deep models proposed after their study, and it is better to use it since it is best to start as said at the basics before accomplishing something. Third, aside that this architecture give a great improvement of performance in the segmentation tasks, it also proves that CNN can learn dense class predictions (dense pixel-wise classifying) for semantic segmentation [12].

#### 3.2.3.1.3 Segmentation

This process consists of the neural network flow and its definitions that this study will use and adapt. Moreover, it will also consist in defining the hyperparameters that will be used in the training phase as well as updating the weights initialized. Also, ensuring there is proper pixel-wise classification for dense prediction and output feature maps through the process of unpooling or simple bilinear interpolation considering the domain used is different from what the model was trained to predict and classify [17,6].

After the segmentation process, if the segmented output having went through pixel-wise labelling or classification is inaccurate or not acceptable, the process of segmenting will then be repeated having the updated weights or variables to be modified at the upper layers of the network—upper layers are only modified and not the lower layers since the lower layers are already the feature selectors, thus freezing their weights and variables as is during training—and will start again from the process of data augmentation.

### 3.2.3.2 Color Extraction

Colors present in the image is then extracted in order to determine the possible emotions depicted in the image. The algorithm that will be used is adapted from the study by [30] wherein they first did two quantization processes to get the RGB values of the image and performed indexing to convert RGB values to HSV. HSV is the preferred values as it is closer to human perception.

### 3.2.4 Train Classifier

During this phase, the classifier model is trained. The extracted features from section 3.2.3 is fed onto the random forest classifier. Tenfold cross-validation is performed to increase the accuracy of the model.

Considering what does tenfold cross-validation do, cross-validation must first be explained. A cross-validation process is that out of the whole dataset, a set of sample data (usually 90% of size) will be used as a training set, and the remaining will be used as the testing set. But this is only for one iteration process. The number of iterations done will be defined on how many n cross-validations must be done before getting any final classifier accuracy output. If it is said that a 10 cross-validation must be used due to a limited dataset, there would be 10 iterations where each iteration, one of the sample dataset will be used as a testing set. In other words, every sample set is a testing set for the whole n iteration cross-validation processes.

#### 3.2.4.1 Feature Set

Features extracted from section 3.2.3 will be joined together to serve as input for the next section, 3.2.4.2.

#### 3.2.4.2 Random Forest

A random forest classifier is constructed. Random forest is known to have a good performance when it comes to multi-class problems. An existing module `sklearn.ensemble` has a method for a random forest classifier. This will be used for this research.

### 3.2.5 Test Model

After training, the data will then be tested. During this phase, the classifier will determine the potential emotions depicted by the painting. A tenfold cross-validation is performed to increase the accuracy.

### 3.2.6 Evaluation

The results of 3.2.5 is then evaluated based on the ground truth constructed. This will determine whether the classifier did in fact determine the right emotions present in the painting. The results will be recorded and be used to determine the accuracy of this study.

REFERENCES


1 Plutchik, R. *The emotions*. University Press of America, 1991.

2 Aditya, S., Yang, Y., Baral, C., Aloimonos, Y., and Feruller, C. *Image understanding using vision and reasoning through scene description graph*. 2017.

3 Karpathy, A. and Fei-Fei, L. *Deep visual-semantic alignments for generating image descriptions*. 2014.

4 Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Transactions on Patter Analysis and Machine Intelligence*, 39, 4 (April 2017), 652-663.

5 Condorovici, R., Florea, C., and Vertan, C. *Painting scene recognition using homogenous shapes*. Bucharest, Romania, 2013.

6 Badea, M., Florea, C., Florea, L., and Vertain, C. *Can we teach computers to understand art? domain adaptation for enhancing deep networks capacity to de-abstract art*. 2017.

7 Agarwal, S., Karnick, H., Pant, N., and Patel, U. Genre and style based painting classification. *In IEEE Winter Conference on Applications of Computer Vision* (2015), 588-594.

8 Yang, W., Zhou, Q., Fan, Y. et al. *Deep Context Convolutional Neural Networks for Semantic Segmentation*. 2017.

9 Peng, C., Zhuang, X., Yu, G., Luo, G., and Sun, J. *Large Kernel Matters - Improve Semantic Segmentation by Global Convolutional Network*. 2017.

10 Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. Pyramid Scene Parsing Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CPVR)* ( 2017).

11 Lin, G., Milan, A., Shen, C., and Reid, I. *RefineNet: Multi-Path Refinement Networks with Identity Mappings for High-Resolution Semantic Segmentation*. The University of Adelaide, 2016.

12 Tao, Y., Wu, Y., Zhao, J., and Guan, L. Semantic segmentation via highly fused convolutional network with multiple soft cost functions. *arXiv:1801.01317* (January 2018).

13 LeCun, Y. *LeNet-5, convolutional neural networks*. 2018.

14 Andersson, V. *Semantic segmentation - using convolutional neural networks and sparse dictionaries*. Linköping University, SE-581 83 Linköping, Sweden, 2017.

15 Donahue, J., Darell, T., Malik, J., and Girshick, R. *Rich Feature Hierachies for Accurate Object Detection and Semantic Segmentation*. University of California, Berkeley, 2013.

16 Islam, M.A., Bruce, N., and Wang, Y. *Dense Image Labeling Using Deep Convolutional Neural Networks*. University of Manitoba, Victoria, 2016.

17 Shelhamer, E., Long, J., and Darell, T. *Fully Convolutional Networks for Semantic Segmentation*. University of California, Berkeley, 2016.

18 Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* (2012), 1097-1105.

19 Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., and Rabinovich, A. Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014), 580-587.

20 Girshick, R., Donahue, J., Darell, T., and Malik, J. *Rich features heirarchies for accurate object detection and semantic segmentation*. 2014.

21 Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 779-788.

22 Szegedy, C., Reed, S., Erhan, D., Anguelov, D., and Ioffe, S. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441v3* (2015).

23 Li, Y., He, K., and Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Advances in Neural Information Processing Systems* (2016), 379-387.

24 Vijay, B., Kendall, A., and Cipolla, R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *arXiv:1511.00561v3 [cs.CV]* (Oct 2016), 1-14.

25 Noh, H., Hong, S., and Han, B. Learning deconvolution network for semantic segmentation. *Proceedings of the IEEE International Conference on Computer Vision* (2015), 1520-1528.

26 Julesz, B. Textons, the Elements of Texture Perception, and their Interactions. *Nature*, 290, 5802 (March 1981), 91-97.

27 Shotton, J., Johnson, M., and Cipolla, R. *Semantic Texton Forests for Image Categorization and Segmentation*. 2008.

28 Zeng, S., Cheng, M., Warrel, J., Sturgess, P., Vineet, V., Rother, C., and Torr, P. *Dense Semantic Image Segmentation with Objects and Attributes*. Columbus, 2014.

29 Horváth, E., Pozna, C., and Ballagi, Á. Road recognition using fully convolutional neural networks. *Bulletin of the Transilvania University of Braşov*, 10, 59 (June 2017).

30 Thyagharajan, K. K. and Minu, R. I. Prevalent color extraction and indexing. *International Journal of Engineering and Technology (IJET)*, 5, 0975-4024 (Dec 2013), 4841-4849.

31 Sobhani, F. M. and Madadi, T. Studying the suitability of different data mining methods for delay analysis in construction projects. *Applied Research in Industrial Engineering*, 2, 1 (June 2015), 15-33.

32 Kadar, J. A., Agustono, D., and Napitupulu, D. Optimization of candidate selection using naive bayes: case study in company x. *Journal of Physics: Conference Series*, 954.

33 Pedregosa, F., Gael, V., Gramfort, A. et al.

34 O'Shea, K. and Nash, R. An Introduction to Convolutional Neural Networks. *CoRR*, abs/1511.08458 (2015).

35 Hariharan, B., Arbel, P., Girshick, R., and Malik, J. Simultaneous Detection and Segmentation. In *European Conference on Computer Vision (ECCV)*. Berkeley; Colombia, 2014.

36 Thoma, M. A Survey of Semantic Segmentation. *CoRR*, abs/1602.06541 (2016).

37 Yang, Z., Zhang, Y., Rehman, S., and Huang, Y. *Image Captioning with Object Detection and Localization*. Tsinghua University, Beijing, 2017.