Kyffin Williams: Digital Analysis of Paintings

Report Name	Progress Report
Author (User Id)	Alexander D. Brown (adb9)
Supervisor (User Id)	Hannah M. Dee (hmd1)
Module	CS39440
Date	November 7, 2012
Revision	0.1
Status	Draft
Word Count	?

Contents

List of Figures

List of Tables

1 Project Summary

Sir John "Kyffin" Williams was a landscape painter from Wales whose work was predominantly based in Wales and Patagonia. His work, with associated metadata collected by Gareth Lloyd Roderick and the National Library of Wales, allows for some interesting analysis; particularly that of temporal or geological data for a given painting.

Temporal analysis will be the focus of this project as it allows for a diverse range of techniques; from statistical analysis of RGB values of the paintings to looking at the length and style of paintbrush strokes.

Whilst it would be nice to be able to predict the age of a painting with no known year, it is far more interesting to try to guess the year of paintings where the date is known. This allows for cross-validation of the techniques employed to guess the year. This project will use leave-one-out cross-validation for simplicity as the data set is small enough to allow this computationally expensive technique.

2 Current Progress

2.1 Design

An initial UML class diagram is depicted by figure 1. This design will allow additional techniques, both analysis and machine learning based, to be added easily. Command line arguments, or later a GUI front-end, will be used to specify which techniques should be used.

The GUI elements depicted are used for visualising the analysed data or results of the machine learning algorithms.

2.2 Colour Space Statistical Analysis

With the design solidified, I begun to work on the statistical analysis of digital images. The easiest of these techniques is to read the image pixel by pixel, taking the various intensities at that pixel, then averaging them out over the whole image.

The most common colour space used is RGB (Red, Green, Blue), where each pixel holds three different intensities, one for each colour with a value of 0 to 255.

Using this analysis I plotted a graph of all paintings with a known year against the different intensities (see figure 2).

Another popular colour space used in image processing is HSV (Hue, Saturation, Value) as it shows variations in colour, saturation and brightness separately (unlike RGB where all three values are affected by changes in brightness). It was a simple matter of changing the existing code for RGB so that it would analyse HSV values instead.

I also decided it would be an improvement to average all values in the same year together to help show changes as time goes on (see figure 3).

2.2.1 Distance Measures

Part of the analysis needed to include a way of measuring the distance of two outputs from the associated technique; for points in space distance measures such as Manhattan distance (figure 4) or euclidean distance (figure 5). This distance measure is then used by the machine learning module

2.2.2 Brushstroke Analysis

After these techniques are implemented there is a lot of space to develop new techniques specific to both painting analysis and to Kyffin Williams. One very interesting technique is to

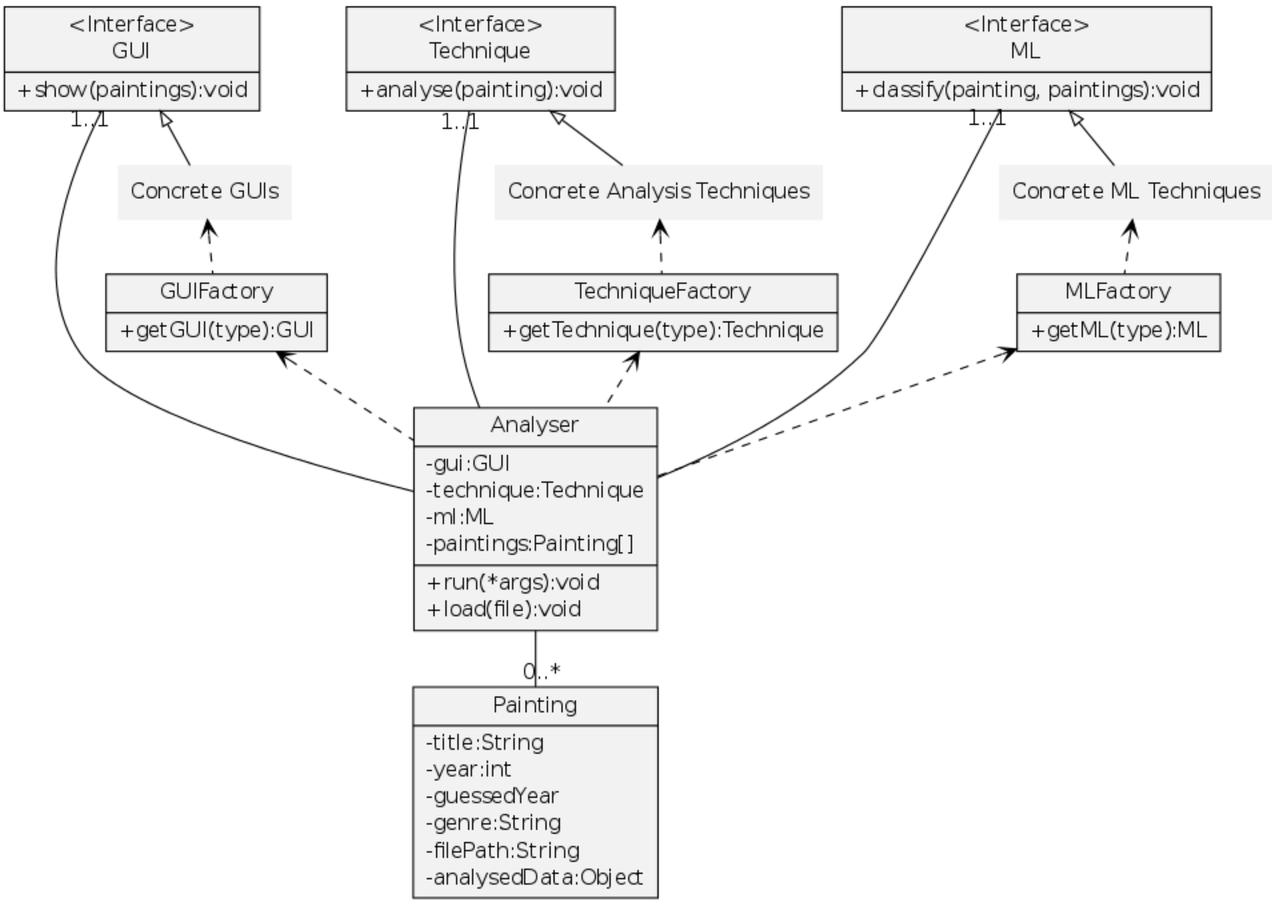

Figure 1: Initial Design

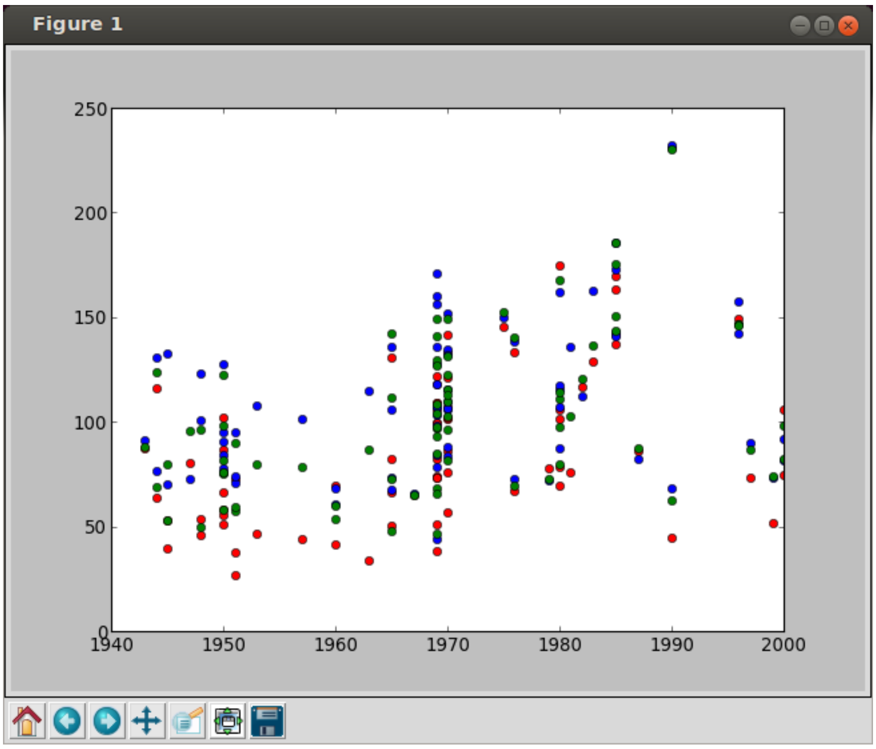

Figure 2: Mean RGB intensities by year.

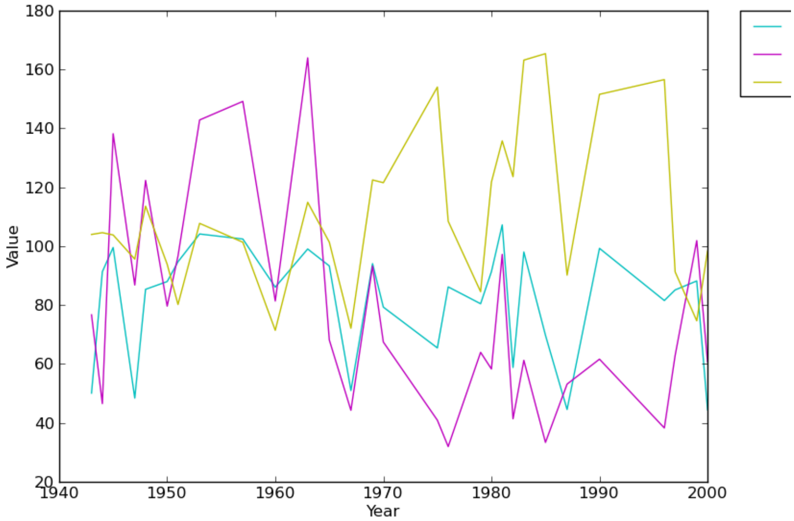

Figure 3: Mean HSV intensities by year.

$$d = \sum_{x=0}^{X} |a_x - b_x|$$

X: All dimensions present in both a and b.
a: The first point.
b: The second point.

Figure 4: Manhattan Distance

$$d = \sqrt{\sum_{x=0}^{X} (a_x - b_x)^2}$$

X: All dimensions present in both a and b.
a: The first point.
b: The second point.

Figure 5: Euclidean Distance

$bestDistance \leftarrow \infty$
for $x = 1 \rightarrow X$ **do**
  if $distance(a, x) < bestDistance$ **then**
    $bestDistance \leftarrow distance(a, x)$
  end if
end for

Where:
a is the example to classify.
X is the list of all training examples.

Figure 6: 1-Nearest Neighbour

analyse the brush strokes the artist made on the painting. It is clear from looking at his work that he dramatically reduces the number of brush strokes he makes as time goes on.

According to the AUTHENTIC project "The digital extraction of brushstrokes proceeds in two steps: (I) contour enhancement, and (II) quantification of brushstroke shape" [**?**]. They perform contour enhancement by applying a circular filter to the image, this filter picks out the characteristic parallel contours of brushstrokes. For the works of Kyffin Williams this may not work as well as he typically used a pallet knife to paint with, instead of a regular paintbrush.

Once these contours are enhanced you can then fill a closed area for each brushstroke, reduce this to a thin line and finally fit some form of polynomial function to this line. Again, for the Kyffin Williams project the third part of this may not be applicable as the AUTHENTIC project specialises in ensuring the authenticity of a painting and does so by verifying that the brushstrokes in one painting match the style of Van Gogh. Even looking at the number of visible brushstrokes in Kyffin Williams' work may be enough to classify a painting.

2.3 Machine Learning Goals

The Machine Learning (ML) module is designed to classify analysed paintings to the training examples in order to get the approximate year the were created.

Machine Learning on it's own is a large and complex field, so the techniques used here may be simple to ensure the project completes on deadline.

2.3.1 K-Nearest Neighbour Classification

The simplest solution to the classification problem is to take the nearest neighbour of a given, analysed, painting and set the year of the new example to the one of the existing example. This is highly dependant on the type of analysis, noise in the training examples, etc.

It is, however, very simple to implement (see Figure 6), with simple modification one can easily change 1-Nearest Neighbour to K-Nearest Neighbour (see Figure 7) and approach with different values for K to give a potentially more accurate answer. K-Nearest Neighbour also has options for interpreting the result, by taking the modal result or a mean of all the results.

3 Planning

Your Bibliography - REMOVE for final version

You need to include an annotated bibliography. This should list all relevant web pages, books, journals etc. that you have consulted in researching your project. Each reference should include an annotation.

$kBest \leftarrow []$
for $x = 1 \rightarrow X$ **do**
  $cur \leftarrow distance(a, x)$
  for $k = 1 \rightarrow K$ **do**
    if $cur < kBest[k]$ **then**
      $temp \leftarrow kBest[k]$
      $kBest[k] \leftarrow cur$
      $cur \leftarrow temp$
    end if
  end for
end for

Where:
a is the example to classify.
X is the list of all training examples. K is the number of nearest neighbours to check.

Figure 7: K-Nearest Neighbour

The purpose of the section is to understand what sources you are looking at. A correctly formatted list of items is sufficient. You might go further and make use of bibliographic tools, e.g. BibTeX in a LaTeX document, could be used to provide citations, for example [**?**] [**?**] [**?**, 99-101] [**?**]. The bibliographic tools are not a requirement, but you are welcome to use them.

You can remove the above *Your Bibliography* section heading because it will be added in by the renewcommand which is part of the bibliography. The correct annotated bibliography information is provided below.