TNM098 - Advanced Visual Data Analysis

Lab3 - Content based analysis of data

Authors:
David Tran - davtr766
Oscar Nord - oscno829

I. Introduction

The final lab in the course TNM098, Advanced Visual Data Analysis, consists of two parts and therefore the report is divided into two sections which will cover the approach taken in order to solve the problem and the result. The main focus of this lab is content based data analysis.

II. Part 1

A. Approach

The first part consists of analyzing the content of twelve different images and ranking them based on the following features:

- Color content
- Color distribution around the central point
- Luminance distributions around one or several points
- Edge positions and orientations

The implementation is written in MATLAB because of convenient image libraries.

B. Result

The program reads all images and calculate the mean values of the red, green and blue channels in each image. The second step is to calculate the mean value of the color distribution inside a 100x100 pixels grid at the center of the images. The relative luminance in each image is calculated by first converting the gamma-compressed *RGB* values to linear *RGB*, and using the formula

$$Y = 0.299 * R + 0.587 * G + 0.114 * B$$

to calculate the luminosity of the image. Where green light have the most contribution and blue light have the least to the overall intensity of the image.

The edge position and orientation is calculated by using the Sobel Operator [1] and the normalized *X* and *Y* gradient to detect edges and orientation in the images.

The feature vector containing the color content, color distribution, luminance and the position and orientation of edges was then compared to the first image in the set. The results where as following:

Comparison between feature vectors the conclusion was that the first and eight image was the closest match when comparing color content, color distribution and luminance but not when analyzing the edges orientations and position where the closest match was the fourth image.

When applying a weighting to the feature vectors of the most important features (color content, color distribution and luminance) and compensating for exposure variation in the images the closest match for the first image in the series was the eight image as seen in figure 1.

III. Part 2

The second part of the lab introduces content based data analysis of text files. Given ten text files, two of the ten text files contains plagiarized content from one of the other files. The plagiarized content is a sentence and a whole paragraph. The aim of this part is to write a program that will identify the plagiarized sections and the source and destination text files.

A. Approach

C++ is the programming language used in this part for performance reasons and convenient standard library. Pre-processing of the text files are necessary to solve this problem in an efficient way. The first step is to go through each text file and separate each sentence in the paragraph, convert to lower case and remove all punctuations. The next

Fig. 1: Closest match when comparing feature vectors. First image above and the eight one at the bottom

step is to create a hashed dictionary (hash table) and go through all the text files and generate converted numerical output files for each text file depending on the position of the word in the hashed dictionary.

The last step is to compare one converted file with every converted files except itself to see if the sentences match. If they do, the hashed dictionary is used to print out the words from the matching sentence. To get the percentage of plagiarized content with respect to number of sentences, the amount of plagiarized sentences are divided with the total amount of sentences in the text file.

B. Result

The program works by letting the user specify the text file that the user wish to be checked for plagiarization. The hardware used is a MacBook Pro Retina, Late 2013 with an Intel Core i5 2.4GHz processor, Intel Iris 1536 MB graphic card and 8 GB 1600 MHz DDR3 memory.

The results where the following:

- The file *02.txt* has a whole paragraph that is identical with a paragraph in the file *08.txt*. The text file contains about 3.6% plagiarized content and checking for plagiarization algorithm took about 0.96 seconds.
- The file *06.txt* has a whole sentence that is identical with a sentence in the file *02.txt*. The text file contains about 1.16% plagiarized content and checking for plagiarization algorithm took about 0.78 seconds.

Therefore, the two of the ten files that contains plagiarized content is *02.txt* and *06.txt*.

REFERENCES

[1] Sobel operator, Accessed: 2018-04-30 http://hlevkin.com/articles/SobelScharrGradients5x5.pdf