Lab 3 - Report

TNM098 - Advanced Visual Data Analysis

Isabelle Rosenquist (isaro242)
Sandra Pettersson (sanpe282)

The task

The task is split into two minor subtasks. The first will focus on comparison between images and the second will focus on comparison between texts. Both tasks was done in MATLAB.

Image comparison

The data for the first task consists of twelve images of various themes and sizes, see figure 1. The purpose of the task is to find similarities in the images by calculating different features in each image and comparing them with each other.

Figure 1: The twelve different images in the database.

Method

First of all, each image was loaded into the program. There were then five different features taken in to consideration when comparing the images; the colour content, the colour distribution around both the center point and around five other points and the luminance around the center point and around five other points. The colour content calculates the mean value for the red, green and blue channel. The colour distribution is calculated by taking the mean value for a square around either the center point or squares around five different points in in the image. The luminance is calculated in the same way but instead of the three colour channels it is converted to the Lab colour space [1]. When comparing it is then possible to either compare by using all of these methods or just one of them.

Image 7 was compared with the other images in the data set and the result ranked from best to worst match. No weighting was applied to the features when comparing the images.

Result

The ranking of the eleven remaining images in the data set can be seen in Table 1. The images ranked the same way can be seen in Figure 2.

Table 1: Image 7 compared with all images in the data set, ranked from best to worst match.

Best match	7	9	12	10	3	5	6	11	8	1	4	2	Worst match

Figure 2: Image 7 compared with all images in the data set, ranked from best to worst match.

Conclusion

The best match for image 7 is image 9. When analyzing these images further you can see that both images share a darker color palette that contain some blue elements. The worst matches are either too bright or have a color palette that are more red and yellow. From this it is possible to say that both the color and brightness affect the comparison.

No weighting of features was implemented to see what features would be more prominent. When comparing all the images with each other, the most prominent features are the the colour distribution and the luminance, both around five points.

Text comparison

The second set of data is instead ten different text files. All of these contain a certain amount of sentences where some texts contain plagiarism. The purpose of the task is to find in which texts there is a plagiarised sentence and a plagiarised paragraph is hidden.

Method

All ten text files are read and formatted. The text formatting includes striping the text from different characters, blank spaces, line breaks and capital letters. The texts are then split into sentences on dots, question marks and exclamation points. The texts are the compared with each other. When two matching sentences are found the first step is to check if this is the start of the paragraph. This is done by checking if the following sentences also match. If they do this is repeated until they no longer match and the paragraph is found and saved. If the sentence is not a part of the paragraph the longest one is saved as the sentence.

Result

The plagiarised paragraph *“Dr. Rutherford was pacing, with surgical precision, up and down my den. He looked slightly more self-possessed than the day before and seemed to be in excellent physical condition. I guessed at the contour beneath my wadded black silk dressing gown and re-considered my original plan to throw him bodily out of the house for having come without my invitation.”* is found in text 2 and 8.

The plagiarised sentence *“I tried out the scales and found that my involuntary host weighed over 195 pounds--a good deal of it around the middle.”* is found in text 2 and 6.

The computation time was measured 10 runs, for which an average of 7.26 seconds was calculated. The different computation times can be seen in Table 2.

Table 2: The computation time for 10 runs.

Computation time (sec):	6.75	6.94	6.84	6.93	7.35	7.23	7.18	7.61	8.17	7.63

Conclusion

The plagiarized sentence is found by just comparing them one by one. The problem that was encountered was that each word or phrase will also count as plagiarism. The actual plagiarism had to be longer than just a few words so therefore when a longer sequence is found it is more likely to be correct.

The plagiarized paragraph was checked at the same time, if two sentences match and the next two does as well, the paragraph is found. The reason for checking for both at the same time is to save just that, time.

The computation time is not that long but with a better computer it could of course be shorter. The comparison is what takes most time since the sentences are compared with each other over and over again to find the correct one.

References

[1] Wikipedia. *CIELAB color space*. [ONLINE] Avaliable at: https://en.wikipedia.org/wiki/CIELAB_color_space [Accessed 10 November 2019].