Hash Match

Hash algorithm is used to compare the similarity of images

Course Title	Image Processing
Professor	Kwang Nam Choi
Team Name	YSWL
Date	2021. 11. 10

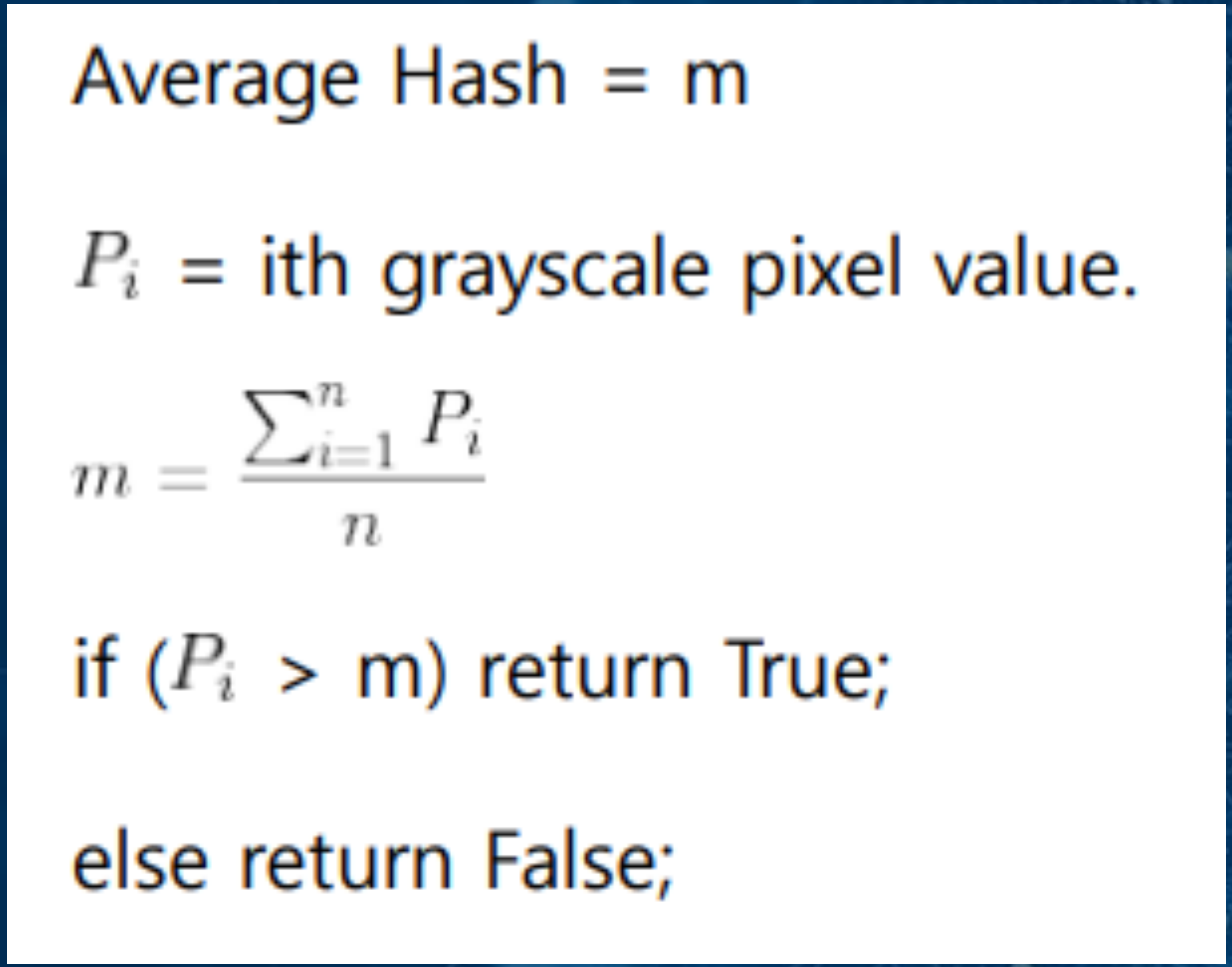
Reasons of choosing this topic

Everyone knows that this is the era of big data, even if they are not a computer major. The amount of data we encounter is increasing exponentially, which has the advantage of being able to obtain various paths for desired information. However, as the amount of data increases, the disadvantages naturally come, and some of them is just junk result.This can be easily felt through search engines such as Google and Naver.

Survey

The most important function used in this project is image matching. Image matching helps determine whether the two images have similar characteristics or are the same by comparing two different images. Image matching is divided into Template Matching and Average Hash Matching. Find the same image using templates, which is prepared images is Template Matching. The function adopted by the project is Average Hash Matching.

*Hash Matching



Average Hash = m

P_i = ith grayscale pixel value.

$$m = \frac{\sum_{i=1}^{n} P_i}{n}$$

if (P_i > m) return True;

else return False;

The principle of average hash matching is simple. First, calculates the average value of all pixels of the image and if the value of each pixel is greater than the average, it is changed to 1, and if less than the average, changed to 0. Similarity measurements use a method of measuring the distance between two points, and measurements include Euclidean distance and Hamming distance.

*Euclidean Distance

*Two-dimensional space

$$\rho = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \ , |X| = \sqrt{x_2^2 + y_2^2}.$$

*Three-dimensional space

$$\rho = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2},$$

$$|X| = \sqrt{x_2^2 + y_2^2 + z_2^2}.$$

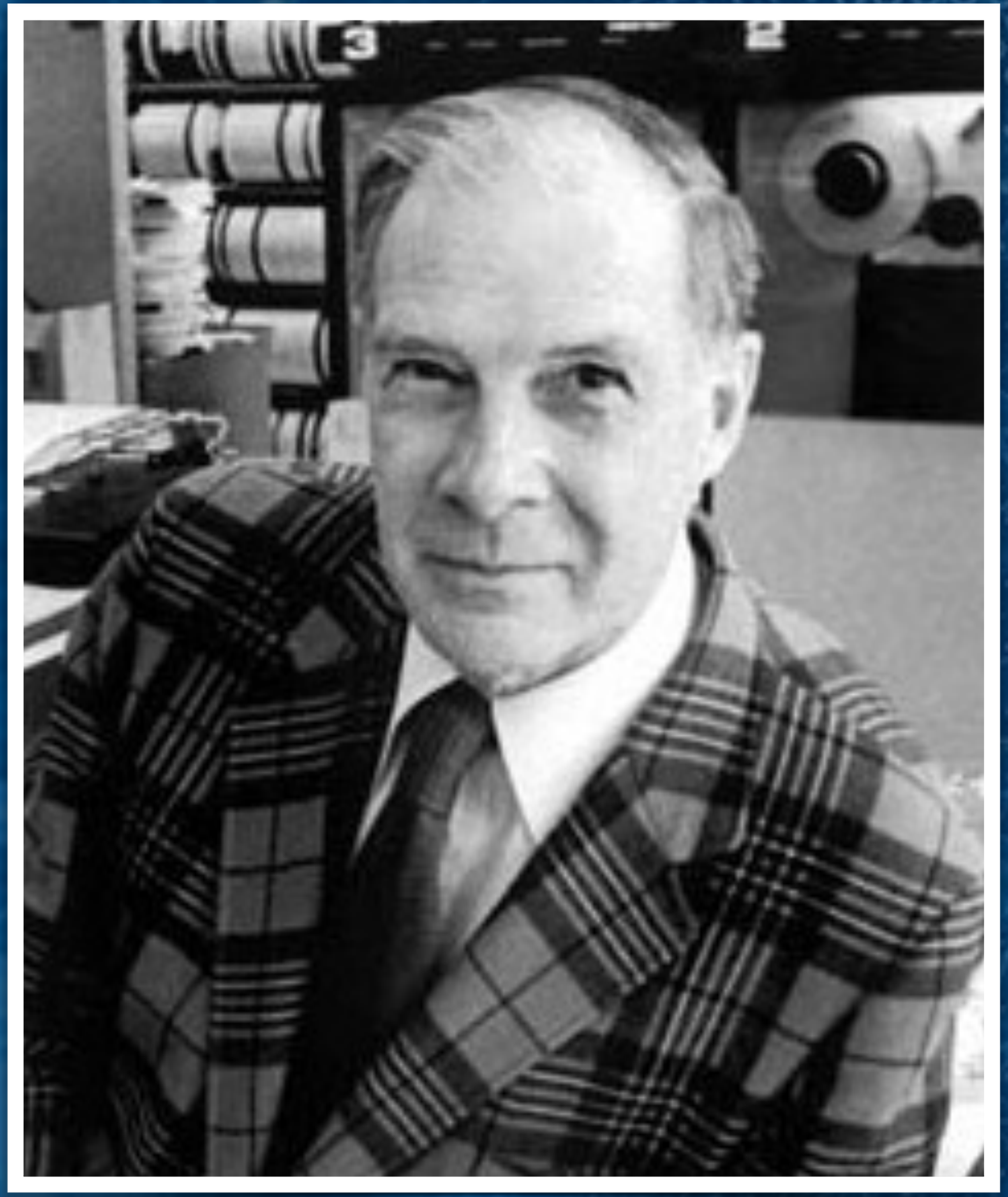
*N-dimensional space

$$d(x, y) := \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

*Hamming Distance

Hamming distance refers to the number of different bits corresponding to two characters (of the same length). We use D (x,y) to represent the Hamming distance between two characters x and y. Perform XOR on two strings and count the number that results in 1. This number is the Hamming distance.

The higher the vector similarity is, the smaller the corresponding Hamming distance is. For example, 10001001 and 10010001 have two different bits.



Plan

The average value of all pixels is obtained as a preliminary task for image matching. By comparing the average value and each pixel, we need to change it to 0 and 1. By using distance measurement, similarity was checked. It should be checked through the values of each row, and the size of the image must be the same regardless of Hamming or Euclidean method for comparison. For this reason, both the input image and the data images should be resized to a predetermined size and proceeded at the same size. Hamming distance will be used as the measurement technique.

Result

In the case of Euclid, it was confirmed that there were serious points that could not be used. For example, in the case of 0111111 and 1000000, all pixels are exactly the opposite, but the difference in Euclide is only 1, so in Euclid it should be similar image, but it does not actually.

The shorter the Hamming distance, the less difference it has from the existing image, so if the data image satisfy with a predetermined distance, it will be selected.

Schedule

11/3(Wd) ~ 11/10(Wed): Survey & Plan

11/11(Thu) ~ 11/13(Sat): Brainstorm and collect input & data images

11/14(Sun) ~ 11/24(Wed): Program each own part.

11/25(Sun) ~ 11/27(Sat): Review and solve extra problems which have occurred.

11/28(Sun) ~ 12/5(Sun): Final report, ppt and project demo

12/6(Mon) ~ 12/9(Thu): Prepare for presentation

The End

Teammate:

유승신	20150413
이주영	20192455
왕생봉	20191335
서정현	20151224