School of Computer Science and Engineering

Hash match for similar images

Team Project #2

Course Title	Image Processing
Professor	Choi Kwang Nam
Major	Software
Date	2021. 11. 10
Team Name	YSWL
20150413	Yoo Seungshin
20151224	Seo Jeonghyeon
20191335	Wang ShengPeng
20192455	Lee JooYoung

1. 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.

2. 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.

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 is a representative method of measuring distance. This is often the case for measuring the distance between points and points learned in the mathematical class.

Hamming distance is quite different from the concept generally known. Hamming distance is to check whether the two values of comparison match exactly. For this reason, the condition that the two digits of the two values must be the same in order to use the Hamming distance follows.

Without above two methods, there is also Manhattan distance as a typical method of measuring distance. This is a measurement method that calculates the sum of the difference of each dimension's absolute number.

3. 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.

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 Euclid 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.

4. 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