AUTOMATED SYSTEM FOR SEARCHING IDENTICAL DIGITAL IMAGES

D. Sc., prof., A.E. Filatova, stud. A.O. Hrichaniuk, NTU "KhPI", Kharkiv

Many computer and laptop owners, who are fond of photography or store a large number of images, photos or screenshots, sometime or other faced the fact that there are dozens of similar files on the hard disk, which occupy a large amount of memory. The total size of secondary files can reach several gigabytes. Finding duplicate digital images is the first step to freeing up disk space. At the same time, searching for and deleting the same photos, pictures and screenshots manually is a rather tedious and time-consuming task. Therefore, the goal of the research is to create an automated system for searching for duplicate digital images.

To achieve the goal, a method of searching for duplicates based on digital image processing methods was proposed.

The basis of the duplicate search method is the discrete cosine transform (DCT) of the corresponding color planes of digital images, as well as the halftone component. To reduce the time of calculations and bring the images to the same size, it is suggested to reduce the size of the images to 100×100 pixels. The conversion of color images into halftone is performed as follows: $Y = 0.299R + 0.587G + 0.114B$, where *Y* is intensity per halftone image; *R*, *G*, *B* – red, green and blue components of the color image, respectively. The search for identical images is offered by calculating a vector of values of halftone DCT coefficients and each color component, while using only the first 21 DCT coefficients for the halftone component and the first 6 DCT coefficients for each color component. The value of each coefficient is divided by 10 and rounded to a whole value to reduce the amount of information. Thus, each image is described by a dimension vector of 39 values: $\vec{x} = (x_1, x_2, \ldots, x_j, \ldots, x_{39})^T$. The classification of a new image is performed using the Fix-Hodges method ("nearest neighbors" method) for $k = 1$, that is, the new image belongs to the class Ω_l ($l \in \overline{1, N}$), for which the "nearest neighbor" of the new image ω: $R(\omega, \omega_l) = \min\limits_{i \in \overline{1, N}} R_i(\omega, \omega_i)$, where $R_i = \sqrt{\sum\limits_{k=1}^{39} (x_j - x_{ij})^2}$ is the distance between the new image ω and image ω_i in the database; x_j, x_{ij} – values of the *j*-th feature measured in images ω and ω_i respectively; *N* is the number of images in the database.

Thus, the thesis proposed a method for finding duplicate digital images based on digital image processing and pattern recognition methods.