UDC 004.925.2
DOI: 10.15587/1729-4061.2024.298744

IMPROVING THE TECHNOLOGY FOR CONSTRUCTING A SOFTWARE TOOL TO DETERMINE THE SIMILARITY OF RASTER GRAPHIC IMAGES

Yevhen Hrabovskyi
Corresponding author
PhD, Associate Professor*
E-mail: maxmin903@gmail.com
Dmytro Bondarenko
PhD, Associate Professor
Department of Information Systems**
Igor Kobzev
PhD, Associate Professor*
*Department of Computer Systems and Technologies**
**Simon Kuznets Kharkiv National University of Economics
Nauky ave., 9-a, Kharkiv, Ukraine, 61166

The object of research is the process of searching and analyzing images of raster graphics. In the context of this work, the problem of the lack of an effective and fast procedure for determining the similarity of images was solved.

The technology for improving the construction of a software tool for determining the similarity of raster graphics images by devising a procedure for determining the similarity of images using a hash code that corresponds to all variants of the image, regardless of size and aspect ratio, is reported. The image features of raster graphics were systematized.

A procedure for determining the similarity of images using a hash code is proposed. This procedure makes it possible to process all possible variants of the image, regardless of the size and aspect ratio. The resulting indicator of the proposed procedure is the value of the hash codes.

It is proposed to use the mathematical apparatus of fuzzy logic by introducing linguistic variables to estimate the similarity index. A comparison of the numerical values of similarity, obtained on the basis of the use of information systems, and the linguistic values revealed in the survey process was carried out. Threshold values were obtained that make it possible to assess the degree of similarity of the images.

Based on the proposed algorithm, a prototype of the information system for determining the similarity of images of raster graphics has been designed. As a result of the calculation of the numerical characteristics of the improvement of the technology of constructing a software tool for determining the similarity of images of raster graphics, the value of the precision indicators was approximately 0.89 and the completeness was 0.8. The advantage of the proposed technology for determining the similarity of images over known analog technologies is illustrated by the amount of RAM of the developed software, which is 210 MB

Keywords: raster graphics, image similarity, software, linguistic variables, hash code, algorithm

Received date 01.12.2023
Accepted date 16.02.2024
Published date 28.02.2024

***How to Cite**: Hrabovskyi, Y., Bondarenko, D., Kobzev, I. (2024). Improving the technology for constructing a software tool to determine the similarity of raster graphic images. Eastern-European Journal of Enterprise Technologies, 1 (2 (127)), 16–25. doi: https://doi.org/10.15587/1729-4061.2024.298744*

1. Introduction

Image retrieval has received much attention in recent decades. First of all, it is connected with the rapid growth of the volume of multimedia information, and as a result, with the emergence of problems of searching and navigating through it. One of the approaches to solving the task of finding visually similar images is search by content. It is based on the analysis of numerous characteristics of image pixels, that is, it takes into account color and texture characteristics, the shape of objects, the composition of the scene, and more.

Despite the variety of proposed algorithms for solving this problem, the quality of the search leaves much to be desired. This is primarily due to the "semantic gap", which is the main problem of content-based image retrieval. The "semantic gap" is the fact that people and machines compare images completely differently: a person compares semantic content (semantics), and machines are based on the comparison of features that describe the visual characteristics of an image (color, texture, shape of objects).

The task of image search by content is relevant and its solution is potentially applicable in many areas. For example, this is a search in large collections or on the Internet. Search engines usually do not take into account the content of the image and search for keywords based on the context of the image. The use of content-based search methods could provide an opportunity to search a larger volume of data, as well as improve its quality.

Corporations such as Google, Microsoft, and Amazon develop and support similar image search products to enable users to find related images on the Internet. Google Images (a Google product) and Bing Image Search (a Microsoft product) are two popular Internet search services that allow users to find similar images on the Internet. They use a number of technologies to analyze and compare visual features of images. In particular, Google Images uses computer vision and machine learning algorithms to visually identify objects and features in images. Bing Image Search uses proprietary image processing and machine learning technologies. This is achieved through integration with Microsoft Azure technologies, including the Computer Vision API, which provides additional image analysis and recognition capabilities. Amazon Rekognition, offered by Amazon, makes it possible to construct image search engines based on similarity and other parameters. Amazon Rekognition easily integrates with Amazon S3 for convenient storage and processing of large volumes of images.

Also, determining the similarity of images to establish the similarity of images and diagnosis is an important task in modern medicine. In the medical field, technologies such as Deep Learning and neural networks, Transfer Learning, Image Registration, medical APIs, and cloud services are used to determine the similarity of images.

Usually, the determination of the similarity of bitmap images is based on the use of various algorithms and methods of image processing. These methods are feature extraction, wavelet transformation, machine learning methods, structural methods, segmentation, calculating a color histogram for each image and comparing their histograms to determine similarity. But at the same time, it is very difficult to choose the exact algorithm or method that would make it possible to match all image options, regardless of the size and aspect ratio. As a result, there is an urgent need to devise such a method that corresponds to all image options, regardless of size and aspect ratio. When trying to go beyond these limitations for operational image processing, objective difficulties arise. These difficulties are associated with the lack of automated tools that allow optimizing the process of processing graphic illustrations, in particular, increasing the efficiency of determining the similarity of images. In addition, existing methods do not always correspond to human perception of image similarity.

Overcoming these difficulties should be carried out by devising an improvement technology for constructing a software tool for determining the similarity of images of raster graphics. The essence of information support in this sense is to determine the similarity of images using the appropriate information system. The central link of this advanced technology is the method of determining the similarity of images using a hash code that corresponds to all variants of the image, regardless of size and aspect ratio.

Devising the specified technology will provide an effective toolkit for improving the information support of publishing and printing. In addition, improving the technology of constructing a software tool for determining the similarity of raster graphics images will create conditions for improving the competitiveness of multimedia web studios and obtaining certain effects from the introduction into production. In particular, an increase in the number of customers of illustrated editions can be ensured due to the improvement of the quality of graphic image processing.

These aspects indicate the relevance of research aimed at devising a technology for constructing a software tool for determining the similarity of images of raster graphics.

2. Literature review and problem statement

The task to determine the similarity of images of raster graphics was addressed by a number of researchers. Thus, in [1], a systematic review and discussion of various approaches to constructing a final illustration of a prepress design is offered. But this work does not take into account the specificity of determining the similarity of images.

Study [2] proposes a dynamic curve color model for image processing. However, this study does not provide an opportunity to consider the specificity of image processing automation.

Work [3] reports analysis of color transfer in the process of image processing. However, this work does not include a description of the parameterization of features of raster images.

Work [4] presents a systematic review and discussion of the problems of quality control of printing illustrations. But in this work, there is no focus on informational support of the process of determining the similarity of images of raster graphics.

The method of automated balancing of a vector illustration and its software implementation are proposed in [5]. At the same time, the specific features of the features of the raster graphics image, which should be taken into account in the process of determining the similarity of the images, are not considered.

The mechanisms of using different color models in the process of determining the similarity of images of raster graphics are considered in [6]. However, in this paper there is no algorithm for determining the similarity of images.

Study [7] provides recommendations for automated image processing for certain prepress tasks. Examples of automated image processing are given. However, this research lacks a prototype software tool for determining the similarity of raster graphics images.

Work [8] considers the development of a global image similarity metric called "central similarity". However, this work does not include a description and detailed consideration of the image texture at different scales.

Paper [9] considers the specificity of processing graphic raster images in the context of constructing mobile online assistants. However, this study does not include recommendations for the assessment of the similarity index.

Study [10] offers a comparative analysis of metrics of perceptual similarity of images. However, the question of defining the conceptual provisions of the method for determining the similarity of raster graphics images in this study remains open.

Aspects of improving the quality of graphic images in the process of publishing publications are considered in work [11]. However, this scientific work does not take into account the need to formalize certain variables in order to determine the similarity of images.

The general drawback of the above-described approaches to determining the similarity of raster graphics images is the lack of a method of determining similarity that corresponds to all variants of the image, regardless of size and aspect ratio. In addition, the lack of an efficient and fast software tool for determining the similarity of images is also a drawback of the analyzed approaches. Therefore, there is a need to devise a method for determining the similarity of images of raster graphics and construct a suitable software tool based on it.

3. The aim and objectives of the study

The purpose of our work is to improve the design of a software tool for determining the similarity of images of raster graphics. This will make it possible to improve the information support of illustration processing by devising a method for determining the similarity of images using a hash code that corresponds to all variants of the image, regardless of the size and aspect ratio.

To achieve the goal, the following tasks were solved:

– to analyze and systematize features of the image of raster graphics;

– to devise a procedure for determining the similarity of images using a hash code;

– to define the concept of similarity and introduce linguistic variables to estimate the similarity indicator;

– to implement the improved technology of constructing a software tool for determining the similarity of images of raster graphics.

4. The study materials and methods

The object of research is the process of searching and analyzing images of raster graphics.

During the research, the following hypothesis was formulated: it is possible to determine the similarity of images using a hash code that corresponds to all variants of the image, regardless of the size and aspect ratio.

The following research methods were chosen:

– generalization – for the formation of conceptual provisions of the procedure for determining the similarity of images of raster graphics. The choice of generalization method is justified by the variety of bitmap graphics formats (for example, JPEG, PNG, BMP). The generalization method works effectively with different types of raster images;

– classification – with the aim of highlighting and substantiating the main features of a bitmap image, which should be taken into account in the process of determining the similarity of images. The choice of this method is justified by the need to highlight classification features of the image;

– deduction – for parameterization of features of bitmap images. The deduction method is chosen based on its ability to identify key image features;

– analysis and synthesis – in order to consider the texture of the image at different scales. The choice of these methods is explained by the fact that analysis can provide a local view of texture, paying attention to small details, while synthesis can provide a global view, determining the overall structure;

– fuzzy logic – in the process of entering linguistic variables to evaluate the image similarity index. Images and notions of similarity between them may be subjective and may contain vagueness. Fuzzy logic allows modeling this fuzziness in language and evaluations, allowing for more flexible consideration of linguistic expressions.

– expert assessment – to determine linguistic variables in order to determine the similarity of images. Linguistic variables often have vague and subjective characteristics. Expert judgment provides an opportunity to model this uncertainty and account for differences in expert judgments. In addition, experts can take into account contextual features that can affect the perception of image similarity, such as lighting, perspective, time, and other factors.

Precision and Recall were chosen as the numerical characteristics of the improvement of the technology of constructing a software tool for determining the similarity of images of raster graphics. This choice is justified by the fact that the use of both metrics makes it possible to balance the errors of the system and obtain an objective assessment of its effectiveness in the task of determining the similarity of images. In particular, the use of these metrics is very important for the medical field where false identification of dissimilar images (for example, X-ray images) as similar could have great negative consequences.

Precision determines what proportion of all objects marked by the model are actually positive:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}, \tag{1}$$

where True Positives (TP) is the number of images that the program correctly identified as similar; False Positives (FP) – the number of images that the program mistakenly identified as similar.

Recall determines what proportion of all truly positive objects the model was able to identify:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}, \tag{2}$$

where False Negatives (FN) is the number of images that the program did not recognize as similar, although they should have been recognized.

By combining both metrics, a comprehensive assessment of the effectiveness of the image similarity detection system is obtained. For example, if a system has high precision but low recall, this may indicate that it is too conservative and identifies only obviously similar objects, leaving out less obvious ones. On the other hand, if there is high recall but low precision, this may indicate that the system identifies many objects as similar, even though they are not really similar.

Thus, the use of both metrics makes it possible to balance the errors of the system and obtain an objective assessment of its effectiveness in the task of determining the similarity of images.

5. Research results regarding the improvement of the technology for determining the similarity of images of raster graphics

5. 1. Analysis and systematization of image features

Feature analysis of a bitmap image can include a variety of methods and techniques to identify key image characteristics. In the process of analyzing the features of raster images, the following parameters should be taken into account:

1. Color characteristics:

1) color histogram, i.e., analysis of color distribution in the image;

2) average color – determination of the average color of the entire image;

3) color profiles – study of color space and distribution.

2. Textural characteristics:

1) texture matrices – use of GLCM (Gray Level Co-occurrence Matrix) to determine texture characteristics;

2) wavelet transformation – texture analysis based on wavelet transformation;

3) local binary samples – definition of local texture samples0.

3. Geometric characteristics:

1) dimensions and proportions – determination of dimensions and proportions of objects in the image;

2) moments of the image – use of moments to describe the shape and position of objects.

4. Frequency characteristics:

1) spectral analysis – determination of the frequency spectrum in the image;

2) frequency filtering – using filters to highlight or hide certain frequencies.

5. Contextual characteristics:

1) pixel dependencies – analysis of dependencies between pixels to identify areas and structures;

2) unification of objects – definition of the united areas and their characteristics.

6. Features of the form:

1) contour characteristics – determination of the shape of objects based on contours;

2) elliptical features – analysis of elliptical features of objects.

7. Local features: reference points – use of key points to determine the characteristics of image areas.

These features can be used individually or in combination to construct a detailed image description. Feature analysis helps in the recognition and comparison of images, which is an important step in solving many tasks of bitmap processing and analysis.

In image processing, computer vision, and related areas, the moment of an image is equal to a specific weighted average value (moment) of the intensities of the image pixels.

Image moments are useful for describing objects after segmentation. Simple image properties that can be found from image moments include area (or total intensity), centroid, and information about its orientation.

Like other tasks in computer vision, such as recognition and detection, the latest neural network-based search algorithms are vulnerable to both candidate and query adversarial attacks. It is shown that the obtained ranking can be drastically changed only by small perturbations, imperceptible to humans.

Due to the fact that the search for similar images is based on content, and not on description, we shall consider only those features of images that can be automatically calculated using image processing methods:

– attributes of color;

– attributes of texture;

– attributes of form.

The most common representation of color is a color histogram (color distribution histogram).

The approach is that the color space is divided into intervals and for each interval the proportion of pixels from the given interval is calculated. The process of dividing the color space into a limited number of color ranges is called quantization. Calculation of color histograms, which effectively characterize the global and local distribution of colors in the image, does not require significant hardware resources. In addition, they are resistant to turns, and also change slightly with changes in scale and viewing angle.

To consider the texture of the image at different scales, one can use wavelet analysis, which consists in decomposing the signal by basis functions. Basis functions (wavelets) are built on the basis of the generating wavelet by shifting and scaling. The original image is taken, and the first projection of the signal is constructed (convolution with the first basis function), then the difference of the original signal with the received one is calculated and the second projection of this difference is constructed (convolution with the second basis function), etc. Moreover, each basis function is a shift of the previous, stretched $2n$ times (n characterizes the scale). Thus, as a result, we get a rough version of the image.

Another important feature for comparing images is the shape of objects. The simplest features are the center of gravity of the figures, the area, the direction of the main axis, etc. There are also more complex methods that represent the figures in more detail, they can be divided into two classes:

1) a regenerative description of the shape of the image, according to which it is possible to completely restore the image of the shape (for example, Fourier attributes);

2) non-restorable description of the shape (descriptive features) by which it is possible to distinguish a given shape from other shapes but it is not possible to completely restore the picture image.

5. 2. The procedure for determining the similarity of images using a hash code

Since perceptual primitives, i.e., distinctive features, must be used to compare images, it has been proposed to use a hash function representing the average value of low frequencies.

This choice is justified by the fact that in an image, high frequencies provide detail, while low frequencies show structure. A large, detailed photo contains many high frequencies. There is no detail in the very small image, so it consists entirely of low frequencies.

Since the goal is to save time and computing power, the fastest way to get rid of high frequencies is to downscale the image. And you don't need to worry about the proportions, just reduce the image to $n*n$ sizes. The dimensions should be as small as possible in order to speed up image processing, but the image should not be too small.

Experimental optimization of the graphic raster image to remove high frequencies, by reducing the graphic raster image, was performed using the image size reduction algorithm and low-pass filtering to remove high-frequency noise. This algorithm included the following stages:

1. Loading the initial graphic bitmap to be optimized.

2. Setting the filtering parameters. Parameters for low-pass filtering were chosen, such as the size of the filter kernel and the degree of sweep (sigma) for the Gaussian filter. These parameters determine the level of smoothness and preservation of details in the image. The size of the Gaussian filter kernel is an odd number that defines the width and height of the kernel. In our case, the size of the filter kernel was 5. The degree of expansion (sigma) determines the degree of expansion of the Gaussian filter. Larger sigma values result in more filter sweep and more smoothing. This parameter should be adjusted manually depending on the noise level and desired effect. In our case, the value of sigma was 0.5.

3. Reducing the image size. A downscaling algorithm such as resampling was used to reduce the size of the image. In the context of the given task, it is appropriate to reduce RAM requirements and speed up loading.

4. Low-pass filtering. A low-pass Gaussian filter was applied to remove high-frequency noise and smooth the image. Filtering helps reduce artifacts and improve overall quality.

5. Evaluation of optimization results, taking into account image quality and size. A comparison with the original image was performed and it was determined whether they meet the requirements of the given task.

As a result of the application of this algorithm, it was experimentally determined that it is optimal to reduce the size of the image to 8×8 pixels (Fig. 1). Thus, the hash code will match all variants of the image, regardless of size and aspect ratio.

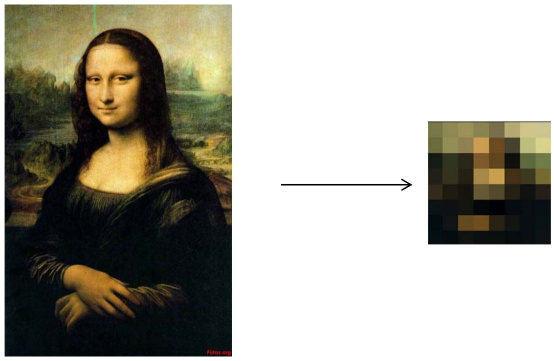

Fig. 1. Getting rid of high frequencies by reducing the high frequencies of an image

The next step is the conversion of the received image in gray scale (Fig. 2). This is necessary in order to reduce the size of the hash code three times: from 64 pixels (64 red, 64 green, and 64 blue) to a total of 64 color values.

Next, it is necessary to determine the average value of brightness. After that, the image is binarized, that is, those pixels whose brightness value is greater than the average brightness value are taken as 1, and the others as 0.

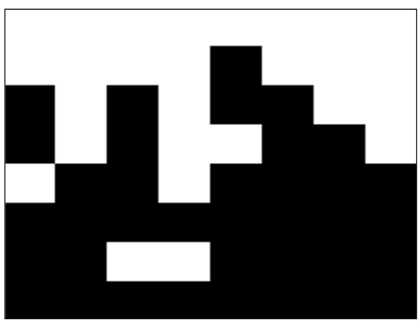

Fig. 2. Image binarization

After that, the hash code is built. The resulting 64 values of 1 and 0 are translated into one 64-bit hash code value, which is 8a0303769b3ec8cd.

The resulting hash code values must be compared. We proposed the hypothesis that the similarity of two images should be determined using the Hamming distance.

When comparing two image hash codes, a zero distance means that they are most likely the same images (or variations of the same image). Next, the received image similarity indicators are converted into percentages.

5. 3. Definition of the concept of similarity and introduction of linguistic variables to estimate the indicator of similarity

Since assigning the value of a variable in words, without using numbers, is more natural for a person, it will be appropriate to introduce linguistic variables for the similarity indicator.

Before entering the linguistic variables for the similarity indicator, it is necessary to define the concept of similarity.

Similarity is defined as a fuzzy equivalence relation, or, as it is also called, a similarity relation S on a set X. Similarity is mathematically defined as a function μ_S: $X\times X\rightarrow[0,1]$, called the membership function of the fuzzy relation S, which takes values on the segment [0, 1] and satisfies the following properties:

1) symmetry:

$$\mu S(x, y)=\mu S(x, y),\ \forall x,y\in X; \quad (3)$$

2) reflexivity:

$$\mu S(x, y)=1,\ \forall x\in X; \quad (4)$$

3) transitivity:

$$\mu S(x, z)\geq\vee(\mu S(x, y)\wedge\mu S(x, z)),\ \forall x,y,z\in X, \quad (5)$$

where the symbols $\vee$ and $\wedge$ denote, respectively, the operations of taking the maximum and minimum values, and the expression $X\times X$ is the Cartesian product of the object set by itself.

Thus, the expression μ_S (x, y)=0,7, x, $y\in X$ or, using the algebraic form of notation $S(x, y)$=0,7, x, $y\in X$, can be interpreted as follows: "the object x belonging to the set X is similar to the object y belonging to the set X at power 0.7".

The value 0 in the right-hand side of the equality means the complete difference of objects x and y, and the value 1, respectively, means the complete identity of the considered objects. It should be noted that the set X is an ordinary distinct set of objects in its classical sense. The symmetry condition (3) of relation S is interpreted as "object x is similar to object y to the same extent that object y is similar to object x", and the reflexivity condition (4) of relation S is interpreted as "object x is identical to itself". The transitivity condition (5) of the relation S was called by Professor Zadeh the property of fuzzy transitivity and interpreted as follows: "The similarity of x and z has at least the same value as the similarity of x and y or the similarity of y and z".

The similarity relation, which is determined by conditions (3) and (4), which has a natural interpretation, and which was designated $S2$ in the special literature in order to distinguish it from the similarity relation S, divides the subject area X into intersecting classes. For the similarity ratio, the expression μ_S (x, y)=0,7, x, $y\in X$ can be interpreted as follows: "the degree of similarity of objects x and y belonging to the set X is equal to 0.7." The degree of similarity of elements can be indicated not only by a number from the interval [0, 1] but also by language expressions. These linguistic expressions are linearly ordered among themselves, such as "very weak similarity", "weak similarity", "medium similarity", "strong similarity", "very strong similarity". In the example, when using a linguistic variable, the original expression can be interpreted as follows: "objects x and y belonging to the set X are very similar."

In order to determine the linguistic variables, expert evaluation was carried out. Graphic designers from multimedia web studios TessLab (Kharkiv, Ukraine), Studio-Site (Kyiv, Ukraine), Art Lemon (Kharkiv, Ukraine) acted as experts.

They were asked to compare several collections with images and determine their similarity scores ("very weak similarity", "weak similarity", "medium similarity", "strong similarity", "very strong similarity").

In Fig. 3, images from the "Portraits" collection are shown, which were offered for comparison.

The results of the assessment are given in Table 1 (the following designations are proposed: a – "very weak similarity", b – "weak similarity", c – "medium similarity", d – "strong similarity", e – "very strong similarity").

Table 1 demonstrates that images 1 and 2 are very similar, 1 and 3 are moderately similar, 1 and 4 are weakly similar, 1 and 5 are strongly similar, 1 and 6 are very weakly similar.

Fig. 3. Images offered for comparison: 1 — large-scale image; 2 — image of a slightly reduced scale; 3 — illuminated image with superimposed background; 4 — image of another person; 5 — image with an overlaid layer mask; 6 — image of an animal on a light background

Table 1

Results of expert evaluation

Image Expert No.	1 and 2	1 and 3	1 and 4	1 and 5	1 and 6
1	e	c	b	d	a
2	e	c	b	d	a
3	e	d	b	c	a
4	e	c	b	d	a
5	e	c	b	d	a
6	e	c	b	d	a
7	e	c	b	d	a
8	e	c	b	d	a
9	e	c	b	d	a
10	e	c	b	d	a
11	e	c	b	d	a
12	e	c	b	d	a
13	e	d	b	c	a
14	e	d	b	c	a
15	e	c	b	d	a
16	e	c	b	d	a
17	e	c	b	d	a
18	e	c	b	d	a
19	e	c	b	d	a
20	e	d	b	c	a

If we compare the numerical similarity values obtained using software tools and the linguistic values found in the survey process, we can conclude that:

– "very weak similarity" – [0; 49];
– "weak similarity" – [50; 59];
– "average similarity" – [60; 74];
– "strong similarity" – [75; 89];
– "very strong similarity" – [90; 100].

5. 4. Implementation of the improved technology for constructing a software tool to determine the similarity of images of raster graphics

5. 4. 1. Development of a prototype algorithm for determining the similarity of images

The developed prototype of the system for determining the similarity of images should have the following functions:

– image hashing;
– calculation of image similarity;
– visualization of the results.

Before constructing a prototype, it is necessary to develop an algorithm that will allow a clear understanding of the procedure for determining the similarity of images.

The block diagram of the algorithm is shown in Fig. 4.

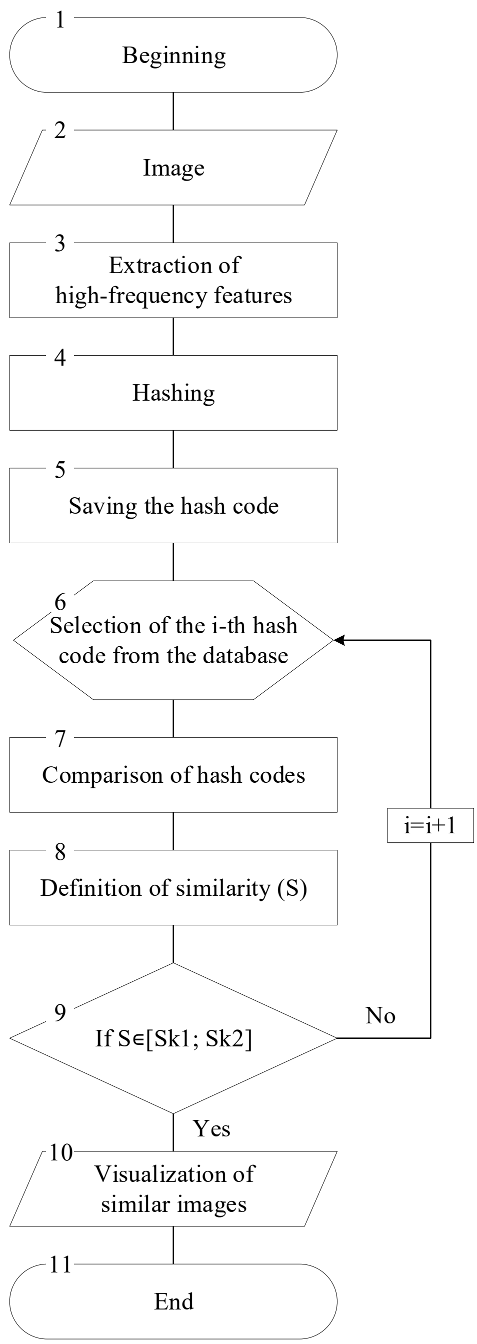

Fig. 4. Block diagram of the algorithm of the prototype of the system for determining the similarity of images

The input data is an image, that is, the user uploads an image to the program that s/he wants to save in her/his collection.

After that, the program analyzes the image and removes high-frequency features. The next step is to get the hash code of the image and store it in the database. This is necessary in order to minimize information and facilitate its processing.

After that, the prototype selects a hash code from the database and compares it with the hash code of the image that was uploaded by the user, thus determining the similarity.

The similarity index that falls within the interval specified by the user is stored together with the file name.

After all iterations have been performed, that is, all hash codes in the database have been checked, the results will be displayed on the screen.

5. 4. 2. Construction of a prototype of a software tool for determining the similarity of images

The development of a prototype system for determining the similarity of images may involve the use of different libraries and approaches. The program code in the Python language is given in [12].

This code uses OpenCV to work with images and compare by structural similarity.

The OpenCV library was used to determine the similarity of images using key points and descriptors.

A practical implementation of the technology of constructing a software tool for determining the similarity of images of raster graphics has been carried out.

Fig. 5 shows a typical architecture based on a content-based image retrieval system – the CBIR system (CBIR – Content-based image retrieval). It performs two main functions: data entry and request processing.

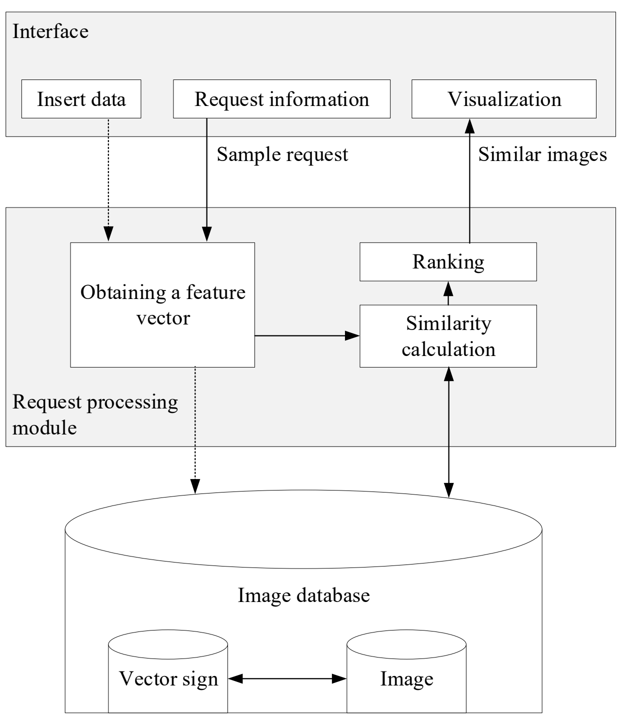

Fig. 5. Typical CBIR system architecture

The data entry subsystem is responsible for extracting the relevant features of the image and saving them in the image database (dotted arrows). This process is usually done automatically.

The processing of requests, in turn, has the following structure: the interface allows the user to ask the necessary request according to the sample and visualize the search result – similar images. The query processing module extracts the feature vector from the sample and evaluates the similarity between the image request and the database of existing images. Next, it ranks the degree of similarity of the images with the query and performs feedback, that is, it transmits the most similar images to the front-end module.

Thus, the constructed prototype of the system for determining the similarity of images should include the following subsystems:

– image database;
– image feature extraction subsystems;
– base of visual attributes;
– user interface;
– a subsystem for determining the similarity of images, which corresponds to the human perception of the similarity of images;
– a subsystem for evaluating and ranking the similarity of images.

To develop a software prototype interface, you can use the Tkinter library, which is a standard for constructing a graphical user interface in Python.

The constructed interface of the developed prototype software tool for determining the similarity of images is shown in Fig. 6.

In the menu bar, the user can perform some settings and read help and information about the author.

Next, you need to adjust the similarity level. The level of similarity can be selected from the drop-down list where it is specified in the form of linguistic variables. Five similarity levels can be selected: very strong similarity, strong similarity, medium similarity, weak similarity, and very weak similarity. If it is more convenient for the user to set the level of similarity in the form of percentages, s/he can enter similarity limits.

The user must then upload the image they wish to add to the collection. Next, the similarity of the uploaded image with the images in the database is determined. The results are displayed on the right side of the form, namely the similarity level and the file name.

If the user is satisfied with the result, s/he can save the image in the database.

A prototype software tool was implemented to determine image similarity on the Google Landmarks dataset. A total of 110 images were processed using the developed software tool. A fragment of the result of the prototype of a software tool for determining the similarity of images is shown in Fig. 7.

According to formulas (1) and (2), the calculation of the numerical characteristics of the improvement of the technology of constructing a software tool for determining the similarity of images of raster graphics was performed. The values of the True Positives (TP), False Positives (FP), and False Negatives (FN) indicators were obtained based on the involvement of experts who were graphic designers from Global Digital Technology LLC (Balta, Odesa oblast, Ukraine). As a result of expert assessment, the following values were obtained:

– True Positives (TP)=80;
– False Positives (FP)=10;
– False Negatives (FN)=20.

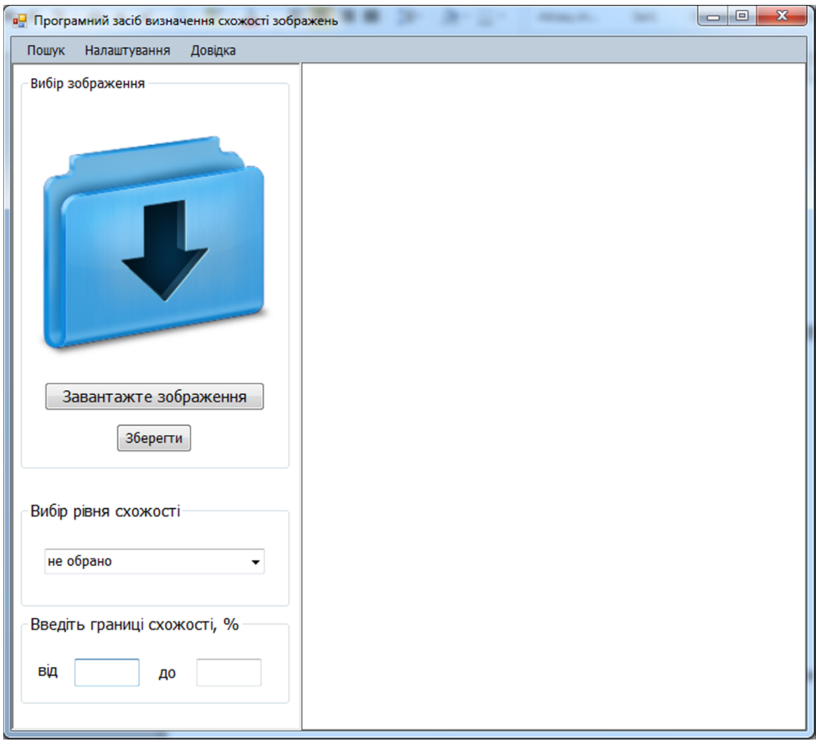

Fig. 6. Prototype interface of the image similarity software

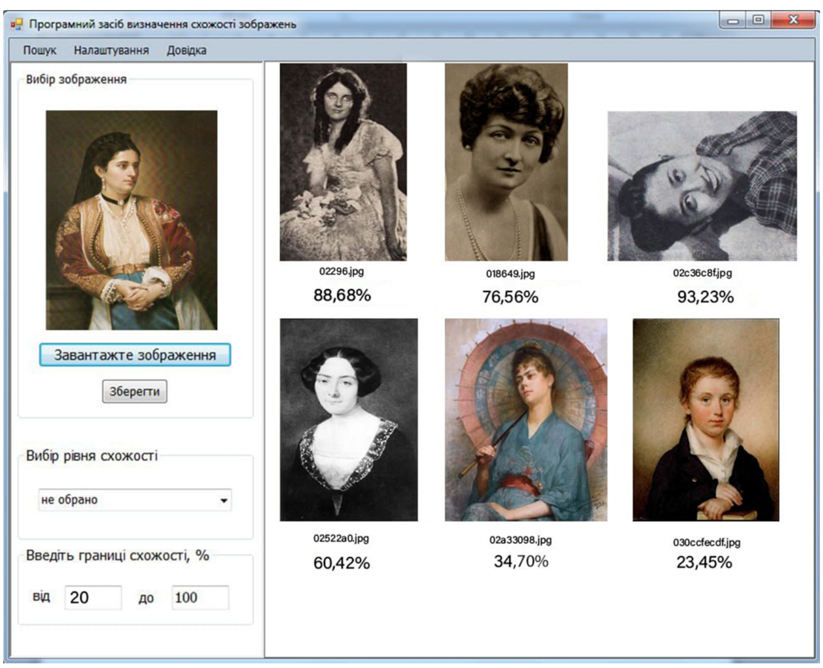

Fig. 7. The result of the prototype of a software tool for determining the similarity of images

The results of the calculations are as follows:

$$\text{Precision} = \frac{80}{80+10} \approx 0.89,$$

$$\text{Recall} = \frac{80}{80+20} \approx 0.8.$$

So, as a result of the calculations, the Precision is approximately 0.89, and the Recall is 0.8. These metrics can be interpreted as follows: the program correctly recognizes similarities in 89 % of cases (Precision), and in samples that are really similar, the program recognizes them in 80 % of cases (Recall).

6. Discussion of results of improving the technology of constructing a software tool for determining the similarity of images of raster graphics

As part of this study, an improved technology for constructing a software tool for determining the similarity of raster images has been devised. In the process of devising the technology, the key features of the image were considered, which should be taken into account in the process of determining the degree of similarity of raster drawings. The proposed technology is a continuation of our research on the development of a procedure of information support for publishing and printing.

As a characteristic that confirms the superiority of the proposed technology for determining the similarity of images over known analog technologies, the amount of RAM is used. After all, the amount of RAM is an important characteristic of computer systems that can affect various aspects of productivity and efficiency. The developed prototype of the software tool for determining the similarity of images requires 210 MB of RAM. The amount of RAM used by the VGG16 model of Convolutional Neural Networks technology is approximately 410 megabytes. And the services Microsoft Azure Computer Vision, Amazon Rekognition, AWS DeepLens, Google Cloud Vision API, Google Similar Images require significant amounts of RAM, from 1 GB and above.

Possible areas of practical application of the proposed methodology include:

– computer graphics programming;

– assessment of the quality of graphic images;

– informational support for publishing and printing.

The proposed technology solves the problem of informational support for the assessment of the quality of graphic images in the aspects of determining the similarity of images using a hash code. This is achieved thanks to the image hashing procedure that we devised and explained using the corresponding results of the designed prototype of the software tool for determining the similarity of images (Fig. 6). In turn, the software prototype was constructed on the basis of the proposed algorithm of the system prototype for determining the similarity of images (Fig. 4) and using the typical architecture of the CBIR system (Fig. 5).

The results of the current study are adequate within the scope of operations for determining the similarity of raster graphic images.

A feature of the proposed technology in comparison with works [2, 5, 7] is the use of the method of determining the similarity of images using a hash code.

The advantages of the proposed technology for constructing a software tool for determining the similarity of bitmap images are:

– taking into account the opinions by leading specialists of multimedia web studios regarding the selection of linguistic variables for determining the similarity of images (in the process of expert evaluation, Table 1);

– the use of a hash code that corresponds to all variants of the image, regardless of the size and aspect ratio (by getting rid of the high frequencies of the image by reducing it, Fig. 1, and further binarizing the image, Fig. 2);

– the presence in the developed software of a subsystem for determining the similarity of images, which corresponds to the human perception of the similarity of images (Fig. 7).

The disadvantage of our research is that the results of the prototype of the software tool for determining the similarity of images could sometimes be sensitive to changes in lighting, angles, image sizes, etc. This shortcoming can be eliminated in the future by using more complex architectures of neural networks that can automatically learn useful features.

In the process of using the proposed results, the following limitations of a subjective nature may be imposed:

– the use of expert evaluations to determine the indicators of similarity of images can lead to the subjectivity of the obtained results;

– the list of image characteristics may change depending on the specific conditions of a certain graphic design task.

Further areas of research may involve:

– development and optimization of deep models of raster graphics analysis;

– devising a procedure for constructing neural networks for automatic study of high-level features and traits that determine similarities between images;

– improvement of pixel comparison algorithms and structural characteristics of bitmap images.

7. Conclusions

1. This paper systematizes the features of the raster graphic image based on the principle of dividing them into different categories depending on the measurement technique, analysis, and interpretation of the information they provide. As a result, such characteristics of a raster image as color characteristics, texture characteristics, geometric characteristics, frequency characteristics, contextual characteristics, shape features, local characteristics were obtained. Systematized features allow constructing a detailed description of the image. In addition, the analysis of these features allows the recognition and comparison of graphic raster images.

2. We have proposed a procedure for determining the similarity of images using a hash code, which allows processing all possible variants of the image, regardless of the size and aspect ratio. The resulting indicator of the proposed procedure is the value of the hash codes. As a result of the implementation of the specified procedure, a 64-bit value of the hash code was obtained, which is equal to 8a0303769b3ec8cd.

3. It is proposed to introduce linguistic variables to estimate the similarity index. Numerical values of similarity, obtained on the basis of the use of information systems, and linguistic values revealed in the survey process were compared. As a result, threshold values were obtained that make it possible to assess the degree of similarity of images. Namely, very weak similarity" – [0; 49]; "weak similarity" – [50; 59]; "average similarity" – [60; 74]; "strong similarity" – [75; 89]; "very strong similarity" – [90; 100].

4. The working algorithm of the prototype for determining the similarity of images has been developed, which makes it possible to regulate the procedure of image processing. Based on the algorithm, a prototype of the information system for determining the similarity of images of raster graphics has been developed. The developed prototype allows automation of the process of searching for similar images, as a result of which the time spent by the user on searching for similar images is saved. The value of the numerical characteristics of the improvement of the technology of constructing a software tool for determining the similarity of images of raster graphics was obtained. Namely, the precision index was approximately 0.89, the recall index was 0.8. A characteristic that confirms the superiority of the proposed technology for determining the similarity of images over known analog technologies is the amount of RAM of the developed prototype, which is 210 MB.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study and the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

All data are available in the main text of the manuscript.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

References

1. Sun, W.-T., Lin, Y.-J. (2020). Systematic Review and Discussion on Final Artwork of Prepress Design. Education and Awareness of Sustainability. https://doi.org/10.1142/9789811228001_0175
2. Cho, S., Byun, H. (2012). Dynamic curve color model for image matting. Pattern Recognition Letters, 33 (7), 920–933. https://doi.org/10.1016/j.patrec.2011.04.012

3. Gu, C., Lu, X., Zhang, C. (2022). Example-based color transfer with Gaussian mixture modeling. Pattern Recognition, 129, 108716. https://doi.org/10.1016/j.patcog.2022.108716
4. Cahyadi, T., Susanto, A., Riyono, D. (2021). Control of packaging print quality with an integrated production flow system in prepress. KREATOR, 2 (1). https://doi.org/10.46961/kreator.v2i1.283
5. Al'boschiy, O., Dorokhov, O., Hrabovskyi, Y., Naumenko, M. (2022). Automated balancing method of vector Illustration and its software implementation. Bulletin of the Transilvania University of Brasov. Series III: Mathematics and Computer Science, 177–192. https://doi.org/10.31926/but.mif.2022.2.64.1.12
6. Buehring, J., Bishop, P. C. (2020). Foresight and Design: New Support for Strategic Decision Making. She Ji: The Journal of Design, Economics, and Innovation, 6 (3), 408–432. https://doi.org/10.1016/j.sheji.2020.07.002
7. Roudn , P., Dr kov , M. (2020). Use of prepress automation in the Czech Republic and examples of automated processing for selected prepress tasks. Proceedings - The Tenth International Symposium GRID 2020. https://doi.org/10.24867/grid-2020-p72
8. Yuan, L., Wang, T., Zhang, X., Tay, F. E., Jie, Z., Liu, W., Feng, J. (2020). Central Similarity Quantization for Efficient Image and Video Retrieval. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr42600.2020.00315
9. Hrabovskyi, Y., Kots, H., Veretelnyk, K. (2023). Development of a methodology for creating an on-line layout assistant for electronic publications for mobile devices. EUREKA: Physics and Engineering, 5, 164–176. https://doi.org/10.21303/2461-4262.2023.003105
10. Vallez, C., Kucharavy, A., Dolamic, L. (2022). Needle in a Haystack, Fast: Benchmarking Image Perceptual Similarity Metrics At Scale. arXiv. https://doi.org/10.48550/arXiv.2206.00282
11. Hrabovskyi, Y. (2023). How to Enhance the Management and Quality of Electronic Publications? Materials Research Proceedings. https://doi.org/10.21741/9781644902691-50
12. A fragment of the Python program code of a prototype information system for determining the similarity of images. Available at: https://drive.google.com/file/d/17C1gE5CwJipYZ3rzr3thx1dGYKfbgVOB/view