Automation of control the state of graphic collections and of content selection for design tasks in the printing

N E Proskuriakov¹, B S Yakovlev¹, N N Arkhangeskaia¹, K V Dementyeva²
¹Tula State University, 92, Lenin Ave., Tula, 300012, Tula, Russia
²Mordovian state university of N.P. Ogarev, Saransk, Russia

E-mail: vippne@mail.ru

Abstract. Results of pilot studies of use of average color for problems of support of graphic collections in current state and selection of graphics for the color parameters set by the user – search of duplicates of images, selection of graphics due to data entry of color in the RGB format and also search of illustrations in the set color range or the specific image are provided. Based on the conducted researches it was established that under certain conditions the indicator of average color of the image rather precisely finds duplicates of illustrations and can be applied to selection of the same images. The research showed that the JPG format, despite the low opportunities in transfer of color, can be considered the optimum format for search of duplicates of digital images. Also, it was defined that BMP, TIFF, PNG behave equally in tasks of finding of duplicates, optimum deviations from average color for the solution of this task were revealed.

1. Introduction
As a result of development of means of digitization, increasing volumes of hard drives, their bigger availability and high-speed performance and also information technology development there are non-standard and new methods of solution of tasks connected with maintenance of digital collections in the optimum state, selection of graphics software. Most articles are devoted to problems of pattern recognition, artificial intelligence, search, and the systems of images storage. However, now insufficiently studied question is search of images in the customer's sample. Also, there are problems with search of duplicates of illustrations and with storage and protection of collections. Search of duplicates of images is insufficiently qualitatively carried out in cases when photos have the different sizes and can be reflected across, verticals or at the same time in both directions. Similar questions are still relevant and partially considered in works [1], [2]. Issues of storage and data protection, especially large collections, were considered also in the work [3]. Besides, designers, color correctors and technologists working in the Web-to-Print format often should solve problems of selection of optimum color of the scene and design depending on preferences of the customer. Still enough correct methods and recommendations for the solution of the similar tasks, especially providing automation, are not developed. The closest on subject is the technology described in work on optimization of storage of images in databases and search queries considering desirable color of the illustration [4, 5]. In these works, the problem of reduction of volume of the database of images due to reduction of the binary illustrations inserted into it, and their replacement by parameters
of color of the image was solved. However, this method is oriented to the solution of the problem of optimization of entering of data in the database. The essence of the method was that it is necessary to enter data not for each image separately, and to integrate them in expanded groups depending on colors. In this work the color model HSL, HSB and the H and S parameters which were entered in the final database was used.

In our opinion, the current problem in the analysis of images and maintenance of collections in current state is still the insufficient adequacy of different techniques, in particular, use of average color for the solution of problems of selection of graphics and search of duplicates.

The solution of the problems described above will have the impact on selection of graphics for multimedia, periodicals, Web-to-Print technology and also on process of searching of optimal solutions for the end user on the basis of the preferable template, "drawing sample", but not the indication of ranges of colors. Similar approach will allow visually correctly, and it is clear for the operator to set parameters of selection of templates.

Also, this approach can influence process of "brainstorming" at graphic design, the concept of registration of the product due to selection of similar references.

This approach is especially demanded in the large regional editions and publishing houses which are engaged in production of the offset product with the long-term created image banks, and thanks to the considered program their systematization and development is noncritical.

For example, the total amount of the photo archive of the regional newspaper «Izvestiya Mordovia» exceeds 5 Terabyte, and search of the image having adequate printing quality takes the long time, despite short deadlines of delivery of the newspaper required.

The program will become especially relevant for the holding regional editions which are published in the different cities with different content. For example, in the line of the newspapers «Pro Gorod» [6], which are published by RNTI-holding, the big general image bank is open for all regional editions which number contains several tens. It is replenished with photographers from different points of the country and is the characteristic example of big corporate storage of the unique visual content which is in fact the commercial asset. For such storages the search of technological decisions on fast selection and processing of images is conducted today.

2. Problem definition
The task of determination of average color has the known problem - it is process speed. It was analyzed and partially solved in the work [7]. As this work shows, the most important condition of reduction of time of definition is the size of the image as it increases the number of pixels that directly influences process speed. In this work it was shown that Python together with dynamic library PIL and the server version of Linux OS, analyze the image much quicker than the applications written for Windows.

In the work [7] methods were described, and software solutions are created which substantially will facilitate implementation of the objects set in this article.

During this work it is necessary to solve the following problems: to analyze the possibility of use of average color of the image for search of duplicates, to show the possibility of selection of similar illustrations for «drawing sample» and also in the parameters of color entered by the user in the color model RGB, to develop the program complex for the solution of these tasks, capable to work without access to the Internet.

3. Theory
Today search of images on similar or on color scheme is carried out by means of search services, for example, by Yandex, Google.

However, in this case we have the number of problems:
1. We cannot influence the selection of results, adjusting certain parameters as it is necessary for us. This option of setup is unavailable to us.
2. We cannot carry out the search in the collections which are already available. It means that it is necessary to do graphics selection in a new way every time, to spend traffic and time for this task.

3. Not all search services are able to look for images on the picture specified by the user.

4. Use of the found images on the Internet most often can relate to copyright infringement by the editorial offices of newspapers.

That is why it was decided to create a software package so that it performs tasks in the process of work without using Internet resources and is independent.

The software package was created on the basis of two decisions: earlier prepared program written on Python 3 and PIL-libraries, which was in detail described in work [7] and also the developed software in the C#, intended directly for search of duplicates and selection of the images, suitable for the criteria specified by the user.

This software works in two modes:
1. The explicit indication of initial color according to the color model RGB.
2. The choice of «drawing sample» based on which similar drawings will be picked up.

Also, expansion of range of selection of images due to input of the minimum and maximum limits of the deviation of color in channels R, G, B of rather original color is provided.

For testing of operability of our software for search of duplicates and selection of graphics the collection of images consisting of 2000 photos of the different sizes and proportions was prepared.

PNG was chosen as the main format of graphic files. Illustrations were taken from results of requests on popular search services. In the equal proportion images of the nature, declines, deserts, views of the cities, space, winter were the part of the collection.

For this collection average color was analyzed with the help of our software in the Python 3 language, working at Ubuntu Server LTS. As a result, the text file with the indication of the name of the file, data on average color was received. This file was applied further as the databank to further search of graphics and duplicates.

For the solution of the problem of selection of similar images based on the specified color the simple method of restrictions – minimum and maximum tolerances from basic color was used. This task is simple, software configuration is subjective. It is exclusively creative process for the person – to think up restrictions on colors for the selection of images.

The solution of the problem of selection of graphics on the set image differs from described earlier. Color is set here automatically based on the analysis of average color of the selected image. The person can set deviations, but in the limited ranges. They need to be revealed and specified on the experiment result.

In case of search of duplicates of illustrations, it is necessary to directly select the image which it is necessary to check and run for search. During the experiment it is necessary to establish whether the file format, the image dimensional change and turn of the illustration down, across, or in both directions influence this process at once. This position is very important as programs for search of duplicates badly cope with strongly reduced or turned and reflected images. The reflected images often cannot be found just because programs of search analyze not all photo, but its parts. Respectively at reflection across the given part on the original will not match the part of the reflected image. This method is often applied to bypass the automatic verification of images, video on «plagiarism», for example, on the video hosting of YouTube. In the theory, the problems described above should not be at application of the way of determination of average color according to all image. However, deviation limits from basic color are not clear. The experiment must reveal it.

4. Results of experiments

Researches

Experiments were carried out on two independent computers. Average color of 2000 images was received by the program written in the Python 3 language, PIL libraries on the server version of Ubuntu 20.04.2 LTS (GNU/Linux 4.15.0-50-generic x86_64); processor: Intel (R) of Xeon (R) of CPU X 5460 @ 3.16 GHz; random access memory: 8 GB; system type: 64-digit version. The second
computer was on the basis of Windows 10 Pro OS of version 20H2; processor: Intel (R) of Core (TM) i9-10900F CPU @ 2.80 GHz 2.81 GHz; random access memory: 16 GB; system type: 64-digit system. Identification of technical features of the computer, which influenced process, was not the purpose of experiments. The definitions of average color given on process acceleration, the choice of the optimum size of the image for these purposes were in detail described in the work [7]. In our opinion, the configurations of both computers have no effect on the results of the experiment.

As it was mentioned earlier, the experiment was made on selection of 2000 photos. However, some photos were the same illustration played the duplicate role. For check of impact of reflection (turn) of the image concerning the original and influences of the graphic format and extent of reduction of the size of the image we prepared variants of the initial illustration (figure 1).

This image has the geometrical sizes 1024×768 pix, resolution 75 DPI, the format – BMP. It was kept from the initial BMP format in the graphic JPG, TIFF, PNG formats.

For the experiment the original file decreased from initial, 100% of the size gradually, with the step of reduction of 10%, up to 90% of reduction of the image of rather original size. As a result, we got 10 illustrations for each of the formats. Then each of the received images was reflected consistently: across, down, and across and verticals at the same time. As a result, we got the set for check of extent of finding of duplicates containing 160 images duplicates.

All changes were carried out in the Photoshop program, when saving the received files, the highest quality was chosen. Images were saved without compression, with standard settings of Photoshop.

Change of the sizes of illustrations was carried out with the preservation of proportions of drawings, by change of height of the drawing.

![Image](image1.png)

**Figure 1.** The initial image used to assess the possibility of search of duplicates

During the experiment results of selection of duplicates from the collection of images which are given in table 1 were received. We considered that for this task deviations from the base color should be small - within 15 units. Deviations of values of color channels (R, G, B) from basic color in the lower or greater side were decided to make identical.

Based on data from table 1, in the MATLAB program dependences for the BMP and JPG formats (figure 2, 3) were constructed. The BMP format behaves like the TIFF and PNG formats. The JPG format has the deviation in statistics of search results of duplicates.

Also researches on finding similar images were conducted. It is worth explaining that this approach can be carried out in several ways: search by a color preset by a person, or search in the specific image. Results of work of the program for both approaches are given in figure 4-9.

Based on data from table 1, in the MATLAB program dependences for the BMP and JPG formats (figure 2, 3) were constructed. The BMP format behaves like the TIFF and PNG formats. The JPG format has the deviation in statistics of search results of duplicates.
Also researches on finding similar images were conducted. It is worth explaining that this approach can be carried out in several ways: search by a color preset by a person, or search in the specific image. Results of work of the program for both approaches are given in figure 4-9.

Table 1. Search results for duplicates by the average color method, considering deviations from original average color

| Percent reduction, % | Deviations (0 – 15) |
|----------------------|---------------------|
| 100 (reference)      | 156 156 156 156 156 156 156 156 156 156 156 156 156 156 156 |
| 90                   | 156 156 156 156 156 156 156 156 156 156 156 156 156 156 156 |
| 80                   | 156 156 156 156 156 156 156 156 156 156 156 156 156 156 156 |
| 70                   | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 60                   | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 50                   | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 40                   | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 30                   | 156 156 156 156 156 156 156 156 156 156 156 156 156 156 156 |
| 20                   | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 10                   | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |

BMP format

| JPEG format |
|-------------|
| 100 (reference) | 156 156 156 156 156 156 156 156 156 156 156 156 156 156 156 |
| 90           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 80           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 70           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 60           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 50           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 40           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 30           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 20           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 10           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |

PNG format

| TIFF format |
|-------------|
| 100 (reference) | 156 156 156 156 156 156 156 156 156 156 156 156 156 156 156 |
| 90           | 156 156 156 156 156 156 156 156 156 156 156 156 156 156 156 |
| 80           | 156 156 156 156 156 156 156 156 156 156 156 156 156 156 156 |
| 70           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 60           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 50           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 40           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 30           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 20           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
| 10           | 160 160 160 160 160 160 160 160 160 160 160 160 160 160 160 |
Figure 2. Schedule of dependence of reduction of the size of the image on deviations of initial color for the BMP format

Figure 3. Schedule of dependence of reduction of the size of the image on deviations of initial color for the JPG format

Figure 4. Result of the selection of similar images on the entered color in the RGB format (R: 100 G: 200 B: 160) with deviations +/-50 units on each channel
Figure 5. Result of the selection of similar images on the entered color in the RGB format (R: 245 G: 245 B: 245) with deviations upward by 5 units and downward by 50 units on each channel

Figure 6. Result of the selection of similar images on the entered color in the RGB format (R: 245 G: 245 B: 245) with deviations upward by 5 units and downward by 60 units on each channel

Figure 7. Result of the selection of similar images for the specified image (R: 77 G: 68 B: 66) with deviations upward and downward by 5 units on each channel
Discussion of results
Based on the analysis of graphic and tabular data for the problem of search of duplicates of files was revealed:
1. The BMP, PNG, TIFF formats behave equally.
2. Optimal conditions for search of duplicates is the situation at which the image is reduced by 50% of the original. In all studied formats this reduction of the size holds average position between incomplete finding of duplicates and the lower bound.
3. At the BMP, PNG, TIFF formats sharp deterioration in search of duplicates at 70% observed reduction of the image (figure 2). In our opinion, it relates to sharp change of set of range of colors in
the image. However, it is possible to confirm it only with repeated repetition of the experiment with different test objects that is beyond the scope of this article. It certainly requires additional researches as with further reduction of the original recovery of former indicators of detection of duplicates is sometimes observed. We consider this area incorrect for prediction and we recommend choosing the range of reduction of the image ranging from 30…50%.

4. The JPG format has proven to be the most stable by search of duplicates of images. During the experiment jumps were not observed. Everything occurred smoothly and with traceable dynamics (figure 4).

5. If we exclude "strange" behavior of the BMP, PNG, TIFF formats at 70% reduction of the image, then dynamics comes to light – up to 20% not all duplicates come to light, then at 30…60% – the stable zone with detection of all duplicates then the unstable zone in which recession of percent of detection of duplicates (figure 4) is possible is observed.

6. For the BMP, PNG and TIFF formats deviation in color channels do not influence search of duplicates in cases when the size remains without changes (original) and at 10% and 20% reduction.

7. For all studied formats if the deviation is equal to zero, it is impossible to find the full list of duplicates. Within deviations 1…15 units on each channel all duplicates were equally and in full at reduction of the drawing by 30…60% concerning the original.

8. The images found in figure 6 demonstrate that in this set of images there are duplicates that were accidentally included in the sample. Partly it speaks about operability of the method.

9. In cases when not all duplicates of images were found, the list of not found duplicates was almost identical, except the cases when the image was reduced by 90%, namely:

- 52 duplicates – were not found cases when images were reduced by 30, 40, 50, 60, 80, 90%.
- 92 duplicates – were not found JPG formats in cases when images were without dimensional change and reduced by 10, 20 and 70%.
- 156 duplicates – were not found JPG formats in cases when images were reflected at reduction of the sizes by 90%.
- 12 duplicates – were not found JPG formats in cases when images were reduced by 90%.
- 4 duplicates – were not found BMP, PNG, TIFF formats in cases when images were reduced by 90%. There were only JPG format illustrations.

The last situations (when images were reduced by 90%) with search of duplicates were extreme and cannot be considered further as 90% reduction of the image give the big error in comparison with the original picture.

Based on search results of similar images, it was revealed:

10. The most exact result is yielded by the selection according to the photo. Deviations on channels at the same time can be the minimum and equal 5 units.

11. The selection of images on the entered color the person yields good result only at considerable deviations of channels from original color.

12. With increase in the deviation selection of the found images increases. However, so far as concerns the problem of selection of images on specified by the user, the range of deviations must be much more, than in the case with finding of photos on the specified template.

13. The analysis of figure 7-9 shows that it is possible to select similar images for the template set by the user, including similar for creation. Figure 9 is especially indicative. On it, we see that the sample includes images with a sunset and a similar general color.

14. Selections in figure 4-6 show how variable the problem of selection of images of the specified color can be, without template. In figure 6 deviations upwards is equal to 5 units and downwards - 60 units. However, the general color is rather sustained. Therefore, in such option of use of this method the problem of search becomes exclusively creative process.

6. Conclusions

1. It is shown that average color of the image can be applied to search of duplicates and selection of similar images.
2. The most suitable graphic format for search of duplicates is JPG as it was the most stable in this situation.
3. It is established, that the optimum percent of reduction of the image at which there is the full detection of duplicates – the range from 60 to 50% of reduction from the original size.
4. Deviations from initial color in problems of search of duplicates is important. In our opinion, the deviation in 5 units is optimum.
5. The search of duplicates offered by us does not depend on reflection (turn) of the image and the illustration dimensional change.
6. Search of similar images works much better when the image template is specified. At the same time deviations from initial color are required to be small.
7. The situation when the initial color for search of similar images is entered, it usually becomes the creative task, is without special recommendations. It is possible only to advise to enter considerable deviations from basic color.
8. Increase in the deviation from initial color expands selection. Despite it, the general color of similar images is selected quite well, it is subjective and cannot be accurately estimated.
9. This method can be applied as additional tools during the work with collections. It cannot become the basis for the organization and classification of collections. They need to be sorted by subjects, the description, keywords. And neural networks which are capable to recognize that is represented in the drawing will be suitable for these tasks, of course, more. However, the method of sorting according to average color can be used as additional tools for sorting within subject-specific collections. It can be used also as the separate tool for search of graphic data necessary for the user.
10. Selection of sketches, graphics, versions of projects for the drawing set by the user can be demanded in situations when it is difficult to explain to the client the initial idea. In fact, the client in this case indicates the reference, which, in his opinion, is the most suitable.
11. This way of selection of graphics and search of duplicates can work well in the combination to standard opportunities of Web-to-Print technology, namely in situations of selection of templates for the required image of the user. It will also be in demand in large regional editions and publishing houses with constantly replenished and long-term collections of graphics.

7. References
[1] Shicong Cao 2018 Review of Image Expression of Content-based Image Retrieval Technology Journal of Physics: Conf. Series 1087 062045.
[2] Nor Asma Mohd Zin, Rozianiwati Yusof, Saima Anwar Lasahri, Aida Mustapha, Norhalina Senan and Rosziati Ibrahim 2018 Content-Based Image Retrieval in Medical Domain: A Review Journal of Physics: Conf. Series 1019 012044.
[3] Proskuriakov N E and Yakovlev B S 2019 Determination of parameters of the Automated backup process of digital data for printing houses and publishing houses without use of external network technologies transformations Journal of Physics: Conference Series, 1210(1) 012116.
[4] Juli Rejito, Atje Setiawan Abdullahi, Akmal, Deni Setiana and Budi Nurani Ruchjana 2017 Image indexing using color histogram and k-means clustering for optimization CBIR in image database Journal of Physics: Conf. Series 893 012055
[5] Mirvakhabova L, Pukalchik M, Matveev S, Tregubova P and Oseledets I 2018 Field heterogeneity detection based on the modified FastICA RGB-image processing Journal of Physics: Conf. Series 893 012009.
[6] Pro City of Saransk [Electronic resource]. https://twitter.com/PG13Saransk
[7] Proskuriakov N E, Yakovlev B S and Arkhangelskaia N N 2019 Determination of parameters affecting the calculation time of the image average color Journal of Physics: Conf. Series 1260 072017.