Experimental Evaluation of Basic Similarity Measures and their Application in Visual Information Retrieval

Miroslav Marinov, Yordan Kalmukov, Irena Valova
Computer Systems and Technologies
University of Ruse
Ruse, Bulgaria

***Abstract*—Searching for similar images is an important feature for image databases and decision support systems in various subject domains. However, it is essential that search results are sorted by degree of similarity in reverse order. This paper presents a comparative analysis of four existing similarity measures and experimentally tests whether they could be used to calculate similarity between images. Metrics could be evaluated by comparing their results to the cumulative human perception of similarity between the same images, obtained by real people. However, this introduces a lot of subjectivism due to non-uniform judgement and evaluation scales. The paper presents a more objective approach - checks which measure performs best in retrieving more images, containing objects of the same type. Results show all four measures could be used to calculate similarity between images, but Jaccard's index performs best in most cases, because it compares features vectors positionally and thus indirectly consider shape, position, orientation and other features.**

Keywords*—Content Based Image Retrieval (CBIR); image search and ranking; similarity measures; image databases

I. Introduction and Related Work

With the development of the Internet and information technology, it has become possible to store and process larger volumes of data, with more and more data in the form of images. This is the basis of the great interest in the approaches and algorithms for image organization, search and retrieval. Naturally, storing large volumes of images requires a new and efficient image retrieval approach. There are two main approaches to image organization and storage - text descriptions, keywords or labels (known as text-based image retrieval) [1] and content based image retrieval [2], [3]. The use of text descriptions is a slow and time consuming process (because of the need a person to describe images with text, not from a computational point of view), so algorithms for content based image retrieval are of greater scientific interest. The main characteristics used in these algorithms are color [4], [5], shape [6], texture [7], spatial features and their combinations [8]. Color is one of the most basic and at the same time distinctive features that hardly changes when you rotate, reduce or increase the size or when changing the orientation of the images. Therefore, the use of color or the color distribution in images at CBIR is the most popular approach among researchers, and yet it is not exhausted and is still subject of interest.

The typical architecture of CBIR systems consists of two main elements. The first is related to the feature extraction of the images and their storage, organization and indexing. The second concerns the assessment of the similarity between the query image and the images in the database. What similarity measures to use and how to assess their suitability for the specific application?

One of the major problems with assessing the visual similarity of images is that there is no classification to use as a criterion. Therefore, it is not possible to make an accurate assessment of the results of the application of the various methods for assessing the similarity of images. It is not possible to use user evaluation (through surveys or any other methods) as the subjective factor in the evaluation is too important and there are undoubtedly huge differences in similarity ratings made by different people, even on a small sample of images. All this requires the search for automatic and without human intervention criteria for assessing similarity.

II. Goal and Motivation

The aim of this paper is to test whether four popular similarity measures (not specially designed for image comparison) could be used to calculate similarity between images. We have tried to do it in our previous paper [9] by comparing the results of similarity measures to the cumulative human perception of similarity between the same images, obtained from an online survey. However, we encountered an enormous problem then – non-uniform judgment and evaluation scales used by the individual respondents.

The survey was designed so that a query image was shown next to a set of sample images, and users were required to specify the exact value of similarity (in their own opinion) in percentage between the query and each image within the set. Since we used nominal, rather than ordinal scale, we have got quite high non-uniformity between individual answers. For example, a respondent specified the similarity between the query and the image X is 80%. Another respondent specified 95% for the same pair of images, while a third respondent specified 40%. Averaging answers having high discrepancies as the above mentioned, could not guarantee reliability and accuracy of obtained "human perception of similarity". So the latter could not be reliably used as a reference.

To test the four similarity measures (Jaccard's index, Euclidean distance, City block distance and Chi-squared dissimilarity) and evaluate how good they are, we decided to use an alternative more objective approach. Inspired by the Top-N accuracy, we applied a similar evaluation approach. We defined a set of 200 images – 50 red roses, 50 tomatoes, 50 red apples and 50 red peppers. The colors of all images are similar – red (roses, fruits, vegetables) and green (leaves). The idea is to check which measure performs best in correct classification of retrieved items (precision) for a specific level of recall. Let's say we are looking for a rose and the system is set up to return 10 results. Then the best similarity measure will be the one that returns most roses out of these 10 results and just a few (or preferably none at all) tomatoes, apples and peppers. However, since there are 50 images of roses, we run 50 queries (every image is used as a query) and average their respective precision for the specified level of recall (top-N results).

This study is important in order to determine which of these four basic similarity measures performs best in searching for images. Results will allow to design and develop an improved universal image retrieval system that could correctly find similar images in various subject domains, or even a system that could automatically select the best similarity measure for a given subject domain by itself.

III. Experimental Environment and Evaluation

The experimental CBIR system used is described in details in [9] and [10]. Briefly, the formation of a feature vector for each image is a sequence of the following actions:

- Each image is divided into 32 by 32 blocks in both width and height dimensions (Fig. 1 to 4).
- All pixels in all the blocks are converted from RGB to one of our 64 primary colors. How these 64 colors were selected and the process of color transformation is described in our previous research [9], [10].
- The dominant color (out of 64 selected colors in our proposed and used color scheme) in each block is determined based on the number of pixels of each color. This dominant color is associated with this block, and the other colors in it are ignored.

In other words, we use just a single color code to substitute multiple pixels per block. In this way, the enormous image color content is reduced to a feature vector with 1024 (32 by 32) color codes. Results of such quantization and color substitution are showed on Fig. 1 to 4. That allows fast image processing and similarity searching. Also, it improves recall as well. The system does the same color analysis for both the image query and the image set and computes such feature vectors for each graphic file. Based on set-theoretic or algebraic methods and similarity measures such as Jaccard Index, Euclidean Distance, City Block Distance and Chi-Square Dissimilarity described in [9] we calculate the similarity factor between the query and each result. At the end, the system returns a sorted list of similar images (Fig. 9).

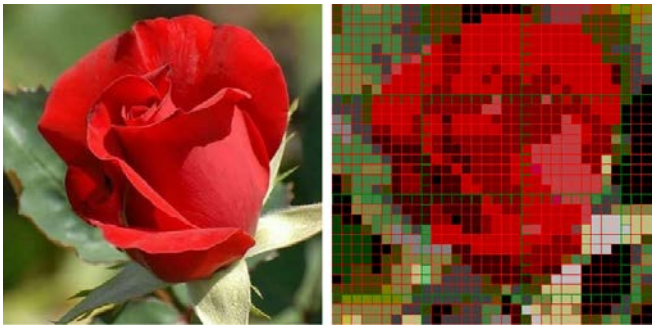

Fig. 1. Original Rose Image Processed by the Application (Left) and its Quantized Image by 32x32 Blocks with Only One Dominant Color Per each Quantization Block (Right).

Fig. 2. Original Tomato Image Processed by the Application (Left) and its Quantized Image by 32x32 Blocks with Only One Dominant Color Per each Quantization Block (Right).

Fig. 3. Original Pepper Image Processed by the Application (Left) and Its Quantized Image by 32x32 Blocks with Only One Dominant Color Per each Quantization Block (Right).

Fig. 4. Original Apple Image Processed by the Application (Left) and its Quantized Image by 32x32 Blocks with Only One Dominant Color Per each Quantization Block (Right).

A set of 200 images (as described earlier) is used in our study. They all have common visual or color characteristics, but are divided in four separate groups - roses, tomatoes, apples and peppers. The feature vectors are stored in the database and each of these 200 images is used as a query in the experiment. It is known which image is from which of the four groups and keeps track of how many images there are in the returned result in the first n similar images from the same group. The first 3, 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50 results are examined and the average of the number of images returned from the same group is calculated for every query. This is repeated for each of

the similarity measures with the idea of checking the degree of accuracy/adequacy for each of the measures.

Let's take the group of images of roses for example. It is checked for each rose image, run as a query, how many of the top-N returned results are roses as well. The test is repeated for all 50 rose images and then the average precision value is determined as a proportion of the returned rose images to all returned images (e.g. top-3 results). So, if for "rose image 1" as a query, we have 3/3 returned rose images (i.e. all returned images are roses), for "rose image 2" as a query, we have 2/3 rose images (i.e. 2 images are roses indeed and the third image is another object), and for "rose 3", we have 1/3 rose images returned (i.e. 1 image of a rose and two other objects), then the average precision for top-3 results is (3 + 2 + 1) / (3 + 3 + 3) = 6 / 9 = 2 / 3.

In the experiment, this is done with 50 image queries from each of the four groups and the top-3, -5 ... results are examined. In this way, we can track how the mean precision changes for each similarity measure, based not only on individual queries, but on all 50 queries from each image group. The results are shown in Tables I, (for the set of roses, run as queries), II (for the set of apples), III (for the set of peppers) and IV (for the set of tomatoes).

TABLE I. AVERAGE RESULTS FOR EACH SIMILARITY MEASURE BASED ON ALL 50 ROSE QUERY IMAGES

TOP X RESULTS	Jaccard's index	City block distance	Euclidean distance	Chi-squared dissimilarity
TOP 3 RESULTS	2.6 / 3	2.06 / 3	2.06 / 3	1.82 / 3
TOP 5 RESULTS	4.00 / 5	3.08 / 5	3.16 / 5	2.78 / 5
TOP 10 RESULTS	7.26 / 10	5.50 / 10	5.22 / 10	4.56 / 10
TOP 15 RESULTS	10.08 / 15	7.62 / 15	7.20 / 15	6.24 / 15
TOP 20 RESULTS	12.74 / 20	9.34 / 20	8.82 / 20	7.80 / 20
TOP 25 RESULTS	15.30 / 25	10.98 / 25	10.36 / 25	9.00 / 25
TOP 30 RESULTS	17.96 / 30	12.58 / 30	11.98 / 30	10.36 / 30
TOP 35 RESULTS	20.40 / 35	14.20 / 35	13.76 / 35	11.92 / 35
TOP 40 RESULTS	22.92 / 40	16.06 / 40	15.54 / 40	13.22 / 40
TOP 45 RESULTS	25.06 / 45	17.96 / 45	17.24 / 45	14.64 / 45
TOP 50 RESULTS	27.58 / 50	19.46 / 50	19.14 / 50	16.04 / 50

TABLE II. AVERAGE RESULTS FOR EACH SIMILARITY MEASURE BASED ON ALL 50 APPLE QUERY IMAGES

TOP X RESULTS	Jaccard's index	City block distance	Euclidean distance	Chi-squared dissimilarity
TOP 3 RESULTS	1.76 / 3	1.76 / 3	1.62 / 3	1.60 / 3
TOP 5 RESULTS	2.38 / 5	2.20 / 5	2.04 / 5	1.98 / 5
TOP 10 RESULTS	3.82 / 10	3.42 / 10	3.22 / 10	3.34 / 10
TOP 15 RESULTS	5.24 / 15	4.60 / 15	4.46 / 15	4.58 / 15
TOP 20 RESULTS	6.66 / 20	5.92 / 20	5.46 / 20	5.70 / 20
TOP 25 RESULTS	8.04 / 25	7.08 / 25	6.50 / 25	6.96 / 25
TOP 30 RESULTS	9.20 / 30	8.24 / 30	7.74 / 30	8.18 / 30
TOP 35 RESULTS	10.70 / 35	9.44 / 35	8.72 / 35	9.46 / 35
TOP 40 RESULTS	11.84 / 40	10.68 / 40	9.92 / 40	10.74 / 40
TOP 45 RESULTS	13.06 / 45	11.84 / 45	11.36 / 45	12.08 / 45
TOP 50 RESULTS	14.30 / 50	13.08 / 50	12.36 / 50	13.40 / 50

TABLE III. AVERAGE RESULTS FOR EACH SIMILARITY MEASURE BASED ON ALL 50 PEPPER QUERY IMAGES

TOP X RESULTS	Jaccard's index	City block distance	Euclidean distance	Chi-squared dissimilarity
TOP 3 RESULTS	1.80 / 3	1.38 / 3	1.38 / 3	1.44 / 3
TOP 5 RESULTS	2.32 / 5	1.82 / 5	1.82 / 5	1.70 / 5
TOP 10 RESULTS	4.00 / 10	2.62 / 10	2.54 / 10	2.48 / 10
TOP 15 RESULTS	5.62 / 15	3.70 / 15	3.60 / 15	3.58 / 15
TOP 20 RESULTS	6.94 / 20	4.94 / 20	4.96 / 20	4.80 / 20
TOP 25 RESULTS	8.46 / 25	5.90 / 25	6.14 / 25	5.84 / 25
TOP 30 RESULTS	10.14 / 30	7.00 / 30	7.10 / 30	7.08 / 30
TOP 35 RESULTS	11.42 / 35	8.46 / 35	8.14 / 35	8.44 / 35
TOP 40 RESULTS	12.78 / 40	9.36 / 40	9.08 / 40	9.50 / 40
TOP 45 RESULTS	14.16 / 45	10.54 / 45	10.34 / 45	10.80 / 45
TOP 50 RESULTS	15.46 / 50	11.58 / 50	11.46 / 50	11.96 / 50

TABLE IV. AVERAGE RESULTS FOR EACH SIMILARITY MEASURE BASED ON ALL 50 TOMATO QUERY IMAGES

TOP X RESULTS	Jaccard's index	City block distance	Euclidean distance	Chi-squared dissimilarity
TOP 3 RESULTS	1.92 / 3	2.04 / 3	1.92 / 3	1.86 / 3
TOP 5 RESULTS	2.78 / 5	3.02 / 5	2.78 / 5	2.64 / 5
TOP 10 RESULTS	4.90 / 10	5.12 / 10	4.38 / 10	4.72 / 10
TOP 15 RESULTS	6.62 / 15	7.30 / 15	6.18 / 15	6.44 / 15
TOP 20 RESULTS	8.78 / 20	8.74 / 20	7.74 / 20	8.16 / 20
TOP 25 RESULTS	10.34 / 25	10.32 / 25	9.24 / 25	9.88 / 25
TOP 30 RESULTS	11.82 / 30	11.92 / 30	10.76 / 30	11.68 / 30
TOP 35 RESULTS	13.58 / 35	13.60 / 35	12.36 / 35	13.06 / 35
TOP 40 RESULTS	15.34 / 40	14.90 / 40	13.92 / 40	14.44 / 40
TOP 45 RESULTS	16.66 / 45	16.46 / 45	15.50 / 45	15.78 / 45
TOP 50 RESULTS	18.88 / 50	17.74 / 50	16.78 / 50	16.92 / 50

Results are also presented graphically on Fig. 5 (for the set of rose query images), Fig. 6 (the set of apple query images), Fig. 7 (the set of pepper query images) and Fig. 8 (the set of tomatoes query images).

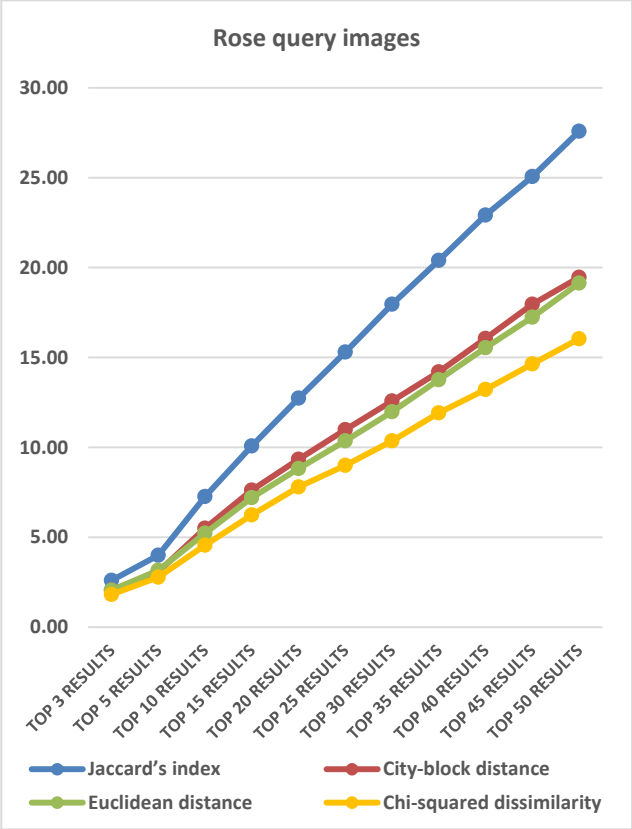

Fig. 5. Average Number of Returned Images, Containing *Roses* (y-axis), Per Similarity Measure and Number of Returned Results (x-axis).

It clearly seems that the Jaccard's index significantly outperforms (retrieves more images containing an object of the same type) all other similarity measures for the set of roses (see Fig. 5). However, this is not the case for tomatoes (Fig. 8), for example, although they have the very same colors. The in-depth analysis of the similarity measures themselves reveals the reason - the Jaccard's index calculates similarity between two images by positionally comparing the dominant colors block by block. All other described measures utilize colors globally. Taking local color distribution into account allows considering not just colors, but shapes and local details as well.

Jaccard's index performs better with roses rather than tomatoes, because the rose's flower consists of multiple individual leafs that reflect light differently and creates dark shadows between leafs (Fig. 1), while tomatoes are singular convex rounded objects (Fig. 2). So, accounting position of the shadows, the Jaccard's index can more easily and reliably guess if the red object in the center of the image is a rose or something else. However, distinguishing a convex red tomato from a convex red apple is much more difficult. That is why Jaccard's index outperforms all other similarity measures for the set of roses (due to additional surface features – shadows between leafs) and the set of peppers (due to the oblong shape), but achieves less better (but still better) results for apples, and no improvement for tomatoes image set.

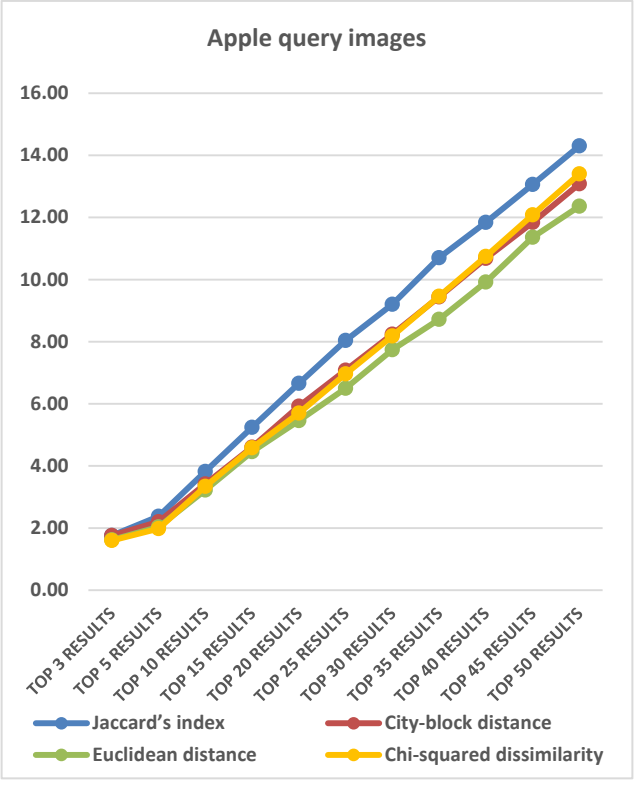

Fig. 6. Average Number of Returned Images, Containing *Apples* (y-axis), Per Similarity Measure and Number of Returned Results (x-axis).

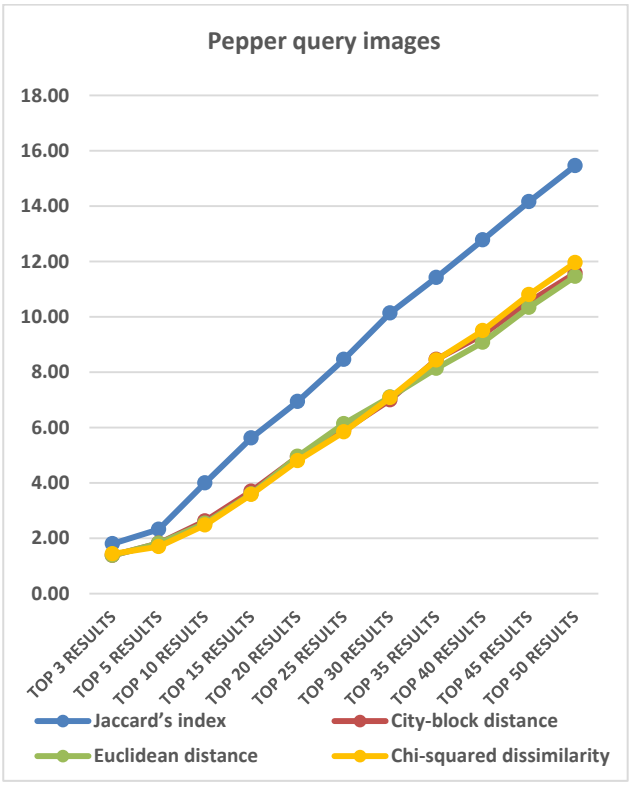

Fig. 7. Average Number of Returned Images, Containing *Peppers* (y-axis), Per Similarity Measure and Number of Returned Results (x-axis).

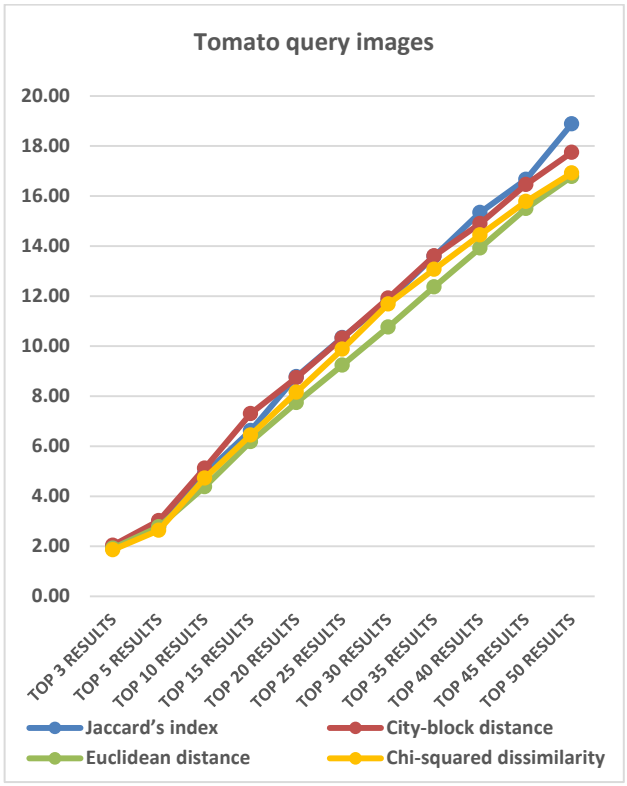

Fig. 8. Average Number of Returned Images, Containing *Tomatoes* (y-axis), Per Similarity Measure and Number of Returned Results (x-axis).

Fig. 9. Example of Top-50 Similarity Results based on Rose Image Query and Jaccard's Index Sorted by Degree of Similarity in Descending Order.

IV. Conclusion

Results from the series of experiments show that:

- All described similarity measures (Jaccard's index, Euclidean distance, City block distance and Chi-squared dissimilarity) could be used to calculate similarity between images. They all provide similarity within the range of [0, 1] and allow search results to be sorted by similarity in reverse order.
- When searching by color content, and consider colors globally, then Euclidean distance, City block distance and Chi-squared dissimilarity produce commensurate results. That is clearly noticeable on Fig. 5 to 9. The main difference between these metrics is in the magnitude of the calculated value. However, the relationships between the calculated similarity factors remain the same, regardless of which one of these three similarity measure is used. It should be noted here that exactly the relationships between similarity factors, rather than the absolute values themselves, create the order of the search results.
- In contrast to all other similarity measures, Jaccard's index compare feature vectors positionally, so it takes into account not just colors, but also their spatial distribution. As a result, it indirectly considers shape, position, orientation and other features.
- When objects have specific features on their surfaces or irregular (e.g. oblong) shape, then Jaccard's index significantly outperforms other similarity measures. That is easily noticeable on Fig. 5 for the set of roses and Fig. 7 for the set of peppers.
- In general, when there is no a-priori information about the image database, the Jaccard's index seems the best single similarity measure between images. This statement is supported by the data in all tables and figures.

Acknowledgment

This paper is supported by project 2022–EEA–01 "Analysis of big data processing algorithms and their application in multiple subject domains", funded by the Research Fund of the "Angel Kanchev" University of Ruse.

References

[1] Liu, Y., Zhang, D., Lu, G., & Ma, W. Y. (2007). A survey of content-based image retrieval with high-level semantics. Pattern recognition, 40(1), 262-282.

[2] Long, F., Zhang, H., & Feng, D. D. (2003). Fundamentals of content-based image retrieval. In Multimedia information retrieval and management (pp. 1-26). Springer, Berlin, Heidelberg.

[3] Shivamurthy R C, Procedures Design and Development of Framework for Content Based Image Retrieval, International Journal of Advanced Research in Engineering and Technology, 12(1), 2021, pp. 1167-1180.

[4] Chugh, H., Gupta, S., Garg, M., Gupta, D., Juneja, S., Turabieh, H., ... & Kiros Bitsue, Z. (2022). Image retrieval using different distance methods and color difference histogram descriptor for human healthcare. Journal of Healthcare Engineering, 2022.

[5] Ashraf, R., Ahmed, M., Jabbar, S., Khalid, S., Ahmad, A., Din, S., & Jeon, G. (2018). Content based image retrieval by using color descriptor and discrete wavelet transform. Journal of medical systems, 42(3), 1-12.

[6] Xu, G., Xiao, K., & Li, C. (2019). Shape description and retrieval using included-angular ternary pattern. Journal of Information Processing Systems, 15(4), 737-747.

[7] Bu, H. H., Kim, N. C., Park, K. W., & Kim, S. H. (2019). Content-based image retrieval using combined texture and color features based on multi-resolution multi-direction filtering and color autocorrelogram. Journal of Ambient Intelligence and Humanized Computing, 1-9.

[8] Mistry, Y., Ingole, D. T., & Ingole, M. D. (2018). Content based image retrieval using hybrid features and various distance metric. Journal of Electrical Systems and Information Technology, 5(3), 874-888.

[9] Marinov M., I. Valova, Y. Kalmukov, "Comparative Analysis of Existing Similarity Measures used for Content-based Image Retrieval", 2019 X National Conference with International Participation (ELECTRONICA), Sofia, Bulgaria, 16 - 17 May 2019.

[10] Marinov, M., Valova, I., & Kalmukov, Y. (2020). Design and implementation of the CBIR system for academic/educational purposes. In 2020 International Conference Automatics and Informatics (ICAI) (pp. 1-4). IEEE.