TECHNO
Vol.20, No.1, April 2019, Hal. 53~64
P-ISSN: 1410-8607, E-ISSN: 2579-9096

SEARCHING SIMILARITY DIGITAL IMAGE USING COLOR HISTOGRAM

Wahyu Wijaya Widiyanto[1], Kusrini[2], Hanif Al Fatta[3]
[123]Department of Informatics, University AMIKOM Yogyakarta, Jl. Ring Road Utara, Condong Catur, Sleman, Yogyakarta, Central Java, 55283, Indonesia

Informasi Makalah

Dikirim, 07 Januari 2019
Direvisi, 11 Januari 2019
Diterima, 30 April 2019

Kata Kunci:

Computer Vision
Similiarity
Euclidean Distance
Grayscale
Histogram

INTISARI

Dalam era globalisasi dan modern seperti saat sekarang ini teknologi informasi banyak dimanfaatkan dalam bidang pendidikan, perdagangan, peternakan, pertanian bahkan hingga ke sektor hukum. Salah satu cabang ilmu dalam bidang teknologi informasi yang berkembang pesat adalah *computer vision.* Salah satu peran penting *computer vision* dalam kehidupan sehari-hari adalah digunakannya *computer vision.* Hal tersebut bisa diterapkan dalam hal *face recognition, object detection*, serta bisa diterapkan untuk melakukan pengelompokan citra berdasarkan urutan kemiripan citra tersebut, kemampuan dari *computer vision* diterapkan untuk memudahkan pekerjaan manusia dalam melakukan seleksi dari beberapa gambar untuk mencari gambar yang paling mirip. Dalam penelitian ini diuraikan mengenai proses pencarian kemiripan sebuah citra dengan citra lainnya melalui beberapa tahap alur penelitian, metode yang digunakan adalah menggunakan nilai RGB yang telah di konversi ke *grayscale*, kemudian dilakukan penghitungan jarak *euclidean distance* untuk menentukan berapa nilai kedekatan sebuah citra sedangkan perhitungan akurasi kinerja algoritma menggunakan confusion matrix. Proses uji coba pencarian menghasilkan tingkat akurasi 0,42, presisi 0,42 dan recall 1 dari 1000 dataset dan diambil 30 data acak yang diuji. Ditemukan gambar yang berbeda dalam warna dan bentuk tetapi ketika dikonversi menjadi histogram data tersebut memiliki kesamaan yang cukup tinggi dengan kueri. Kelemahan dari penelitian ini adalah gambar yang memiliki histogram yang mirip dengan kueri ditampilkan sebagai gambar yang serupa meskipun kenyataannya adalah gambar yang sangat berbeda dari warna dan bentuk.

Keyword:

Computer Vision
Similiarity
Euclidean Distance
Grayscale
Histogram

ABSTRACT

In the era of globalization and modernization, as now, information technology is widely used in the fields of education, trade, animal husbandry, agriculture and even to the legal sector. One branch of science in the field of information technology that is growing rapidly is computer vision. One of the important roles of computer vision in everyday life is the use of computer vision. This can be applied in terms of face recognition, object detection, and can be applied to group images based on the order of similarity of the image, the ability of computer vision is applied to facilitate human work in selecting from several images to find the most similar images. In this study described the process of finding the similarity of an image with other images through several stages of research flow, the method used is to use RGB values that have been converted to grayscale, then the euclidean distance is calculated to determine the value of proximity of an image while calculating performance accuracy algorithm using confusion matrix. The search trial process resulted in an accuracy rate of 0.42, precision of 0.42 and recall 1 of 1000 datasets and 30 random data were taken. Found images that differ in color and shape but when converted into histograms the data has a fairly high similarity to the query. The disadvantage of this research is that images that have histograms similar to queries are displayed as similar images even though the reality is that images are very different from colors and shapes.

Korespondensi Penulis:

Wahyu Wijaya Widiyanto
Department of Informatics, University AMIKOM Yogyakarta
Jl. Ring Road Utara, Condong Catur, Sleman, Yogyakarta
Central Java, 55283, Indonesia
Email: wahyu.wijaya@students.amikom.ac.id

1. INTRODUCTION

In the era of globalization and modernization, as now, information technology is widely used in the fields of education, trade, farming, agriculture and even to the legal sector. One branch of science in the field of information technology that is rapidly developing is computer vision, where its application is widely used for needs relating to the ability to see. One of the important roles of computer vision in everyday life is the use of computer vision in making selections from several images to find the most similar images. This can be applied in terms of face recognition, object detection, and can be applied to group images based on the order of the image's similarity. Face recognition is one of the biometric studies. Until now face recognition is still an interesting and challenging field of research. Face recognition has been widely used in applications such as system security, credit card verification, criminal identification etc [1].

Histograms represent a popular representation of features in computer vision. Examples of applications include: object detection, human detection, texture analysis and tracking. Histograms encode the distribution of irregular spatial measurements in an area. More formally, histograms are defined as numeric arrays, where each element (termed bin) matches the frequency calculation of the range of values (eg image intensity, color, gradient orientation, etc.) in the given image or subset. From a probabilistic point of view, a normalized histogram can be seen as a function probability distribution. In terms of intensity and color-based histograms, this histogram shows invariance of the translation, rotation in the plane, and changes slowly without aircraft rotation, changes in object distance and occlusion. In addition, the lighting invariant can be realized by changing the input image using appropriate transformations before histogram construction (eg, normalized RGB, HSV, YUV color spaces) [2]. In the image processing process there is basic information that can be processed from an image that is in the form of color features, features in the form and features in the form of textures. The color feature in an image is a feature that is quite dominating because of the feature sensitivity information is obtained about the viewpoint of an image, the translation of an image and the rotation of an image [3]. Data that has been extracted from an image can usually be in the form of numerical data that is ready for calculation, because data is generated in a digital image shaped matrix with length m and width n, where n itself is the size in pixels.

2. METHODS

2.1. Research Flow

From the research made, the plot is first the image data that will be searched for similarities are entered into the prototype, the image in the form of RGB is converted to grayscale then the image formed from grayscale is converted again to the histogram, from the histogram calculation is done to find the closest distance from some images look for similarities, the results of calculations are entered into the database then in the closest distance filter from the image that has similarities and the final results are displayed filtered image. The flow of research is more clearly seen in Figure 1:

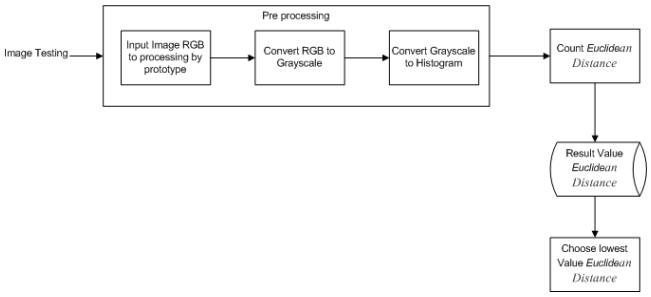


Figure 1. Research Flow

2.2. Degree of Gray (Grayscale)

Digital images can also be expressed in two-dimensional matrices with. x and y are pixel coordinates in the matrix and f is the degree of intensity of the pixel [4]. The matrix formed from images with size m X n is as follows:

$$f(x,y) = \begin{bmatrix} f(0,0) & f(0,1) & \dots & f(0,n-1) \\ f(1,0) & f(1,1) & \dots & f(1,n-1) \\ f(2,0) & f(2,1) & \dots & f(2,n-1) \\ \dots & \dots & \dots & \dots \\ f(m-1,0) & f(m-1,1) & \dots & f(m-1,n-1) \end{bmatrix} \quad \text{.....(1)}$$

Information:
m = Number of lines in an image
n = Number of columns in an image

Grayscale are color pixels that are in a gradation range between black and white. Graysclae is a blend of minimum black and minimum white color [5].

The process of converting RGB images to Graysclae can be done using the following equation [6].

G = 0.2126*R + 0.7152*G + 0.0722*B .....(2)

The image conversion process like the above equation will produce a new image with Graysclae color as shown in Figure 2 below**:**

Figure 2. Converting RGB images to Grayscale

2.3. Grayscale Color Histogram

The image histogram refers to the probability mass function of the image intensity. This is extended to color images to capture the combined probability of the intensity of three color channels [7]. The gray histogram formed from an image consists of 256 points on the x axis consisting of numbers 0 to 255. The y-axis contains the number of repetitions of each color on the x-axis [8].

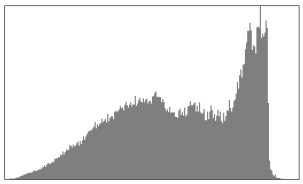


Figure 3. The formed histogram

3. RESULTS AND DISCUSSION

3.1. Prototype of Image Similarity Test

Based on previous research in the similarity of the image to the search process based on shape and color, a sequencing process is based on the threshold value of the sample image through the use of the threshold algorithm [9], and get the comparison between the threshold value and aggregation value almost the same. Conversely, if it approaches 0, the comparison becomes very different. Other studies mention color composition can be displayed in the form of a histogram that represents the distribution of the number of pixels for each color intensity in the image. In determining the level of maturity of apples, it can be determined based on the composition of the color, with the results of experiments on programs that have been made show that the image that has similar color image distribution exactly has the difference in distance

equal to zero [10]. For current research using a simple prototype in the search process, prototype like image 4. which consists of several menus using the php programming language, namely:

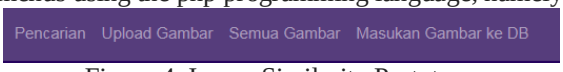


Figure 4. Image Similarity Prototype

From Figure 4 above, the search menu is used to search for images that are similar to the image being tested, the image upload menu is used to add data sets, the menus of all images are used to display all images that have been inputted to the prototype, menus to insert images to DB are used to store image that has been uploaded to the database.

3.2. Extraction Query Feature

The initial stage in the image search process is to extract features that are in an image that become queries. The extraction process is carried out to obtain RGB value information in an image in the form of numbers, which will then be converted from RGB to grayscale, as shown in figure 5 below using the formula in equation (1):

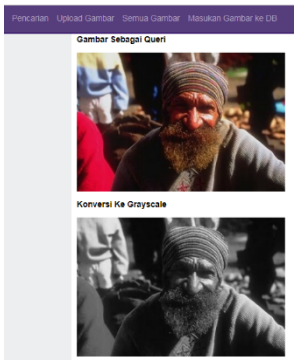


Figure 5. The Process of Converting RGB Images to Grayscale

While the results of calculations stored in the database are shown in table 1 below (from the 1000 datasets tested are shown in table 1 only 19 sample data):

TABLE I
GRAYSCALE QUERY IMAGE COLOR VALUE

Figure	Color	Value
18.jpg	9	29,257
18.jpg	10	80,5847
18.jpg	11	115,233
18.jpg	12	168,54
18.jpg	13	109,056
18.jpg	14	93,0221
18.jpg	15	91,8076
18.jpg	16	89,9738
18.jpg	17	97,166
18.jpg	18	101,488
18.jpg	19	84,8416
18.jpg	20	85,1393
18.jpg	21	87,8334
18.jpg	22	82,5851
18.jpg	23	85,4489
18.jpg	24	85,1631
18.jpg	25	85,2882
18.jpg	26	79,7214
18.jpg	27	78,1257

3.3. Query Distance and Dataset

To get optimal results, the distance between Queries and Datasets is calculated. The equations used to calculate the distance are as follows [11]:

$$d = \sqrt{(k - k')^2 + (k - k')^2 + ... + (n - n')^2} \quad \text{.....(3)}$$

Information:
d = Distance to be searched
k = Number of pixels in one Queri color value
k '= Number of pixels in one dataset color value
n = Number of pixels in one color value Query n
k '= Number of pixels in one dataset color value n

From Figure 5 a histogram is generated and an eccentric distance calculation is performed to determine the distance between the image that is the query and the closest search results as shown in the following figure 6:

Figure 6. Histogram Query and Dataset

3.4. Search Results

The search process with query image 18.jpg produces several images that are considered similar, as shown in Figure 7 below after accuracy is calculated from the sum of all the pixels of the sum of results in the root using the equation (3) above:

Figure 7. The Search Results That Are Considered The Most Similar Based Nearest Distance

From Figure 10, there are 12 images that are the shortest distance between the query and dataset (1000 image data in the database).

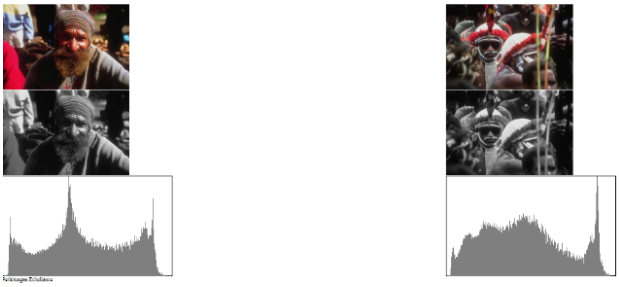


Figure 8. The Search Results That Are Considered Most Similar to the Nearest Distance (image 83.jpg)

58

Figure 9. The Search Results That Are Considered Most Similar to the Nearest Distance (image 2.jpg)

Figure 10. The Search Results That Are Considered Most Similar to the Nearest Distance (image 949.jpg)

Figure 11. The Search Results That Are Considered Most Similar to the Nearest Distance (image 114.jpg)

Figure 12. The Search Results That Are Considered Most Similar to the Nearest Distance (image 25.jpg)

Figure 13. The Search Results That Are Considered Most Similar to the Nearest Distance (image 93.jpg)

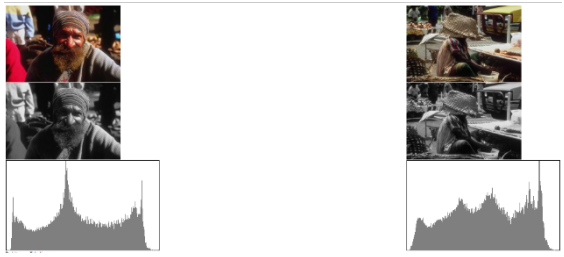


Figure 14. The Search Results That Are Considered Most Similar to the Nearest Distance (image 84.jpg)

Figure 15. The Search Results That Are Considered Most Similar to the Nearest Distance (image 825.jpg)

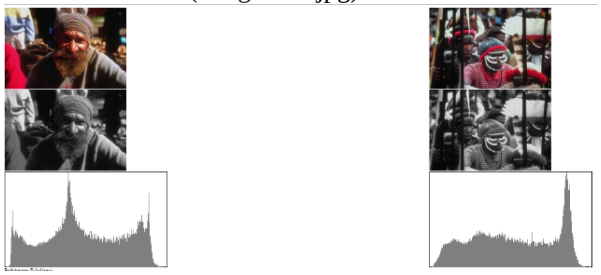


Figure 16. The Search Results That Are Considered Most Similar to the Nearest Distance (image 55.jpg)

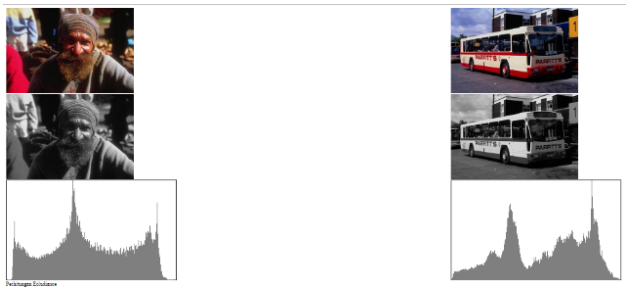
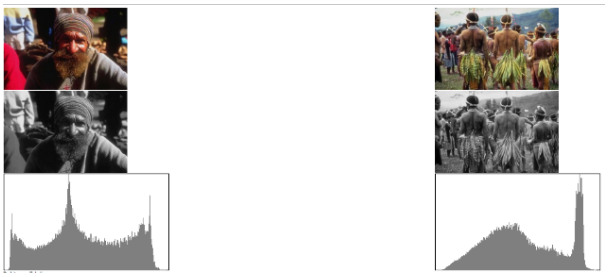


Figure 17. The Search Results That Are Considered Most Similar to the Nearest Distance (image 341.jpg)

Figure 18. The Search Results That Are Considered Most Similar to the Nearest Dis 59 (image 94.jpg)

Figure 19. The Search Results That Are Considered Most Similar to the Nearest Distance (image 5.jpg)



The results of calculation of the Query Distance and Nearby Dataset can be seen in table 2. below:

TABLE II
NEARBY QUERIES AND DATASET IMAGE 18.jpg

Query	Database	Distance
18.jpg	83.jpg	322051,5
18.jpg	2.jpg	362957,2
18.jpg	949.jpg	390157,8
18.jpg	114.jpg	465294,3
18.jpg	25.jpg	467122,7
18.jpg	93.jpg	483212,5
18.jpg	84.jpg	483405,8
18.jpg	825.jpg	483951,6
18.jpg	55.jpg	504945,2
18.jpg	341.jpg	522342,3
18.jpg	94.jpg	539630,2
18.jpg	5.jpg	544187,1

While other random examples for searching datasets based on similarity and the closest distance can be seen in the picture and table below:

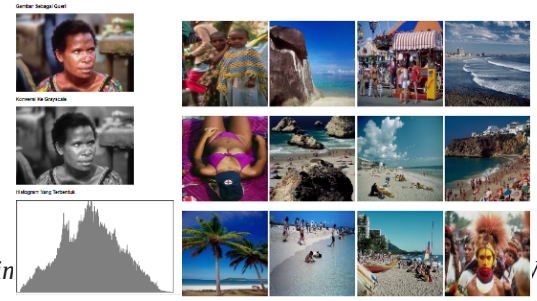


Figure 20. Random Dataset Image 1.jpg Along With Similar Image Search Results

The results of calculation of the Query Distance and Nearby Dataset can be seen in table 3. below:

60

TABLE III
NEARBY QUERIES AND DATASET IMAGE 1.jpg

Query	Database	Distance
1.jpg	238.jpg	434926,7
1.jpg	267.jpg	453750,8
1.jpg	134.jpg	485970,8
1.jpg	33.jpg	487198,4
1.jpg	204.jpg	491947,8
1.jpg	222.jpg	506973,1
1.jpg	355.jpg	523500,3
1.jpg	259.jpg	526240,9
1.jpg	255.jpg	576141,9
1.jpg	215.jpg	586233,8
1.jpg	224.jpg	597792
1.jpg	238.jpg	434926,7

Figure 21. Random Dataset Image 67.jpg Along With Similar Image Search Results

The results of calculation of the Query Distance and Nearby Dataset can be seen in table 4. below:

TABLE IV
NEARBY QUERIES AND DATASET IMAGE 67.jpg

Query	Database	Distance
67.jpg	1.jpg	313000,8
67.jpg	971.jpg	344804
67.jpg	754.jpg	362126,2
67.jpg	708.jpg	401922,3

67.jpg	238.jpg	412815,3
67.jpg	236.jpg	434253,6
67.jpg	66.jpg	441742,7
67.jpg	98.jpg	442505,9
67.jpg	224.jpg	455203,1
67.jpg	264.jpg	456046
67.jpg	891.jpg	464533,5
67.jpg	20.jpg	472918,1

61

Figure 22. Random Dataset Image 104.jpg Along With Similar Image Search Results

The results of calculation of the Query Distance and Nearby Dataset can be seen in table 5. below:

TABLE V
NEARBY QUERIES AND DATASET IMAGE 104.jpg

Queri	Database	Distance
104.jpg	531.jpg	120693,6
104.jpg	568.jpg	162749,7
104.jpg	583.jpg	234236,1
104.jpg	583.jpg	234236,1
104.jpg	574.jpg	257729,6
104.jpg	266.jpg	260615,2
104.jpg	109.jpg	282342,6
104.jpg	218.jpg	289609,8
104.jpg	510.jpg	350325
104.jpg	253.jpg	373948,4
104.jpg	198.jpg	373975,2
104.jpg	219.jpg	389983

Figure 23. Random Dataset Image 179.jpg Along With Similar Image Search Results

The results of calculation of the Query Distance and Nearby Dataset can be seen in table 6. below:

62

TABLE VI
NEARBY QUERIES AND DATASET IMAGE 179.jpg

Queri	Database	Distance
179.jpg	177.jpg	319408,7
179.jpg	272.jpg	323952,2
179.jpg	273.jpg	336507,2
179.jpg	274.jpg	368173
179.jpg	161.jpg	371678,5
179.jpg	295.jpg	407217,9
179.jpg	346.jpg	419712
179.jpg	11.jpg	426531,6
179.jpg	28.jpg	450426,1
179.jpg	382.jpg	451314,2
179.jpg	173.jpg	453624,9
179.jpg	191.jpg	460973,8

3.5. Measurement of Algorithm Performance

Performance measurement of a study is very important, this is done in order to obtain information on how high the accuracy of an algorithm when compared with other algorithms, so also in this study to be able to see how high the accuracy of the algorithm used, Confusion Matrix method is used [12]. In the confusion matrix, the results of the trial results will be divided into two classes, positive class and negative class. Where the positive class contains the correct test results that are considered true (true positive) and the correct test results are considered wrong (true negative). Whereas in the Negative class there is a wrong trial result (false positive) and the wrong test result is false (false negative).

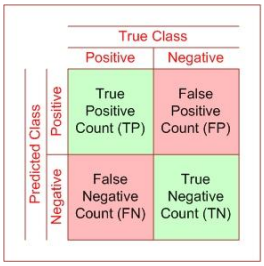


Figure 20. Classes in Confusion Matrix

In the Confusion Matrix, the equation used to calculate the accuracy of a method looks like this:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$$

$$Precision = \frac{TP}{FP+TP} * 100\%$$

$$Recall = \frac{TP}{FN+TP} * 100\% \quad \text{.....(4)}$$

Where:
1. TP is True Positive, which is the amount of positive data that is correctly classified by the system.
2. TN is True Negative, which is the amount of negative data that is correctly classified by the system.
3. FN is False Negative, which is the amount of negative data but is incorrectly classified by the system.
4. FP is False Positive, which is the number of positive data but incorrectly classified by the system.

63

From the research conducted to measure the performance of the algorithm, sampling was carried out by taking each of the two images in each category so that 30 random images were generated using the formula (4) and the results as below:

TABLE VII
DATA SAMPLE FOR PERFORMANCE MEASUREMENT

No	Query	dataset	Result	
			Actual	Prediction
1	1.jpg	67.jpg	T	T
2	81.jpg	257.jpg	F	T
3	133.jpg	142.jpg	T	T
4	167.jpg	170.jpg	T	T
5	206.jpg	282.jpg	T	T
6	277.jpg	222.jpg	T	T
7	309.jpg	336.jpg	T	T
8	341.jpg	339.jpg	T	T
9	416.jpg	28.jpg	F	T
10	422.jpg	254.jpg	F	T
11	533.jpg	152.jpg	F	T
12	591.jpg	15.jpg	F	T
13	617.jpg	126.jpg	F	T
14	656.jpg	106.jpg	F	T
15	711.jpg	102.jpg	F	T
16	767.jpg	114.jpg	F	T
17	831.jpg	100.jpg	F	T
18	893.jpg	106.jpg	F	T
19	900.jpg	10.jpg	F	T
20	953.jpg	100.jpg	F	T
21	11.jpg	53.jpg	T	T
22	27.jpg	1.jpg	T	T
23	55.jpg	325.jpg	F	T
24	119.jpg	114.jpg	T	T
25	661.jpg	109jpg	F	T
26	703.jpg	108.jpg	F	T
27	881.jpg	102.jpg	F	T
28	965.jpg	1.jpg	F	T
29	974.jpg	10.jpg	F	T
30	10.jpg	15.jpg	T	T

Based on table 7 above the scenario performed for testing is shown in table 8 below:

TABLE 8
CONFUSION MATRIX

		Prediction	
		Similar	Not Similar
Actual	Similar	TP	FN
	Not Similar	FP	TN

From table 8, the value of TP=20, FP=11, FN=0, and TN=0. The results of the calculation produced are as follows

$$Accuracy = \frac{20+0}{20+0+11+0} * 100\% = 0.42$$

$$Precision = \frac{20}{20+11} * 100\% = 0.42$$

$$Recall = \frac{20}{0+20} = 1$$

Based on the calculation of the confusion matrix algorithm above, it can be concluded that the accuracy is 0.42, the precission is 0.42, and the recall is 1 of the random data of 30 datasets.

.

4. CONCLUSION

The process of finding image equations is quite fast from the tested dataset using 1000 dataset images with calculations using RGB values that have been converted to grayscale, from grayscale converted to histogram to obtain eucludian distance calculation values, the eucludian distance value is calculated to determine the proximity value of an image so obtained the equation of the image sought.

The search process takes between 1 minute and 1.5 minutes, with hardware specifications of the 2.40 Ghz Intel Core I5, 2GB RAM processor. The search trial process produced an accuracy rate of 0.42, precision of 0.42 and recall of 1 of 1000 datasets and 30 random data were taken. Found images that differ in color and shape but when converted into histograms the data has a fairly high similarity to the query. The disadvantage of this study is that images that have histograms similar to queries are displayed as similar 64 es even though the reality is that images are very different from color and shape.

5. REFERENCE

[1] E. Budianita, J. Jasril, and L. Handayani, "Implementasi Pengolahan Citra dan Klasifikasi K-Nearest Neighbour Untuk Membangun Aplikasi Pembeda Daging Sapi dan Babi," *J. Sains dan Teknol. Ind.*, vol. 12, no. Vol 12, No 2 (2015): Juni 2015, pp. 242–247, 2015.

[2] "HISTOGRAM-BASED SEARCH : A COMPARATIVE STUDY Mikhail Sizintsev , Konstantinos G . Derpanis Department of Computer Science and Engineering Toronto , ON , Canada Faculty of Business and Information Technology University of Ontario Institute of Technology Os," *Proc. 21st IEEE Conf. Comput. Vis. Pattern Recognit. - CVPR '08*, 2008.

[3] A. Baita, B. S. W, and A. Sunyoto, "Logo Retrieval Berdasarkan Ekstraksi Multifitur," *Magistra*, no. 98, pp. 53–59, 2016.

[4] C. Kavitha, D. Rao, and D. Govardhan, "Image retrieval based on color and texture features of the image sub-blocks," *Int. J. …*, vol. 15, no. 7, pp. 33–37, 2011.

[5] N. C. Santi, "Mengubah Citra Berwarna Menjadi Gray¬Scale dan Citra biner. Jurnal Teknologi Informasi DINAMIK," *J. Teknol. Inf. Din.*, vol. 16, no. 1, pp. 14–19, 2011.

[6] J. Mukherjee, I. K. Maitra, K. N. Dey, S. K. Bandyopadhyay, D. Bhattacharyya, and T. H. Kim, "Grayscale conversion of histopathological slide images as a preprocessing step for image segmentation," *Int. J. Softw. Eng. its Appl.*, vol. 10, no. 1, pp. 15–26, 2016.

[7] J. Sangoh, "Histogram-Based Color Image Retrieval," *Psych221/EE362 Proj. Report, Stanford Univ.*, pp. 1–21, 2008.

[8] S. Kusumaningtyas and R. A. Asmara, "Identifikasi Kematangan Buah Tomat Berdasarkan Warna Menggunakan Metode Jaringan Syaraf Tiruan (Jst)," *J. Inform. Polinema*, vol. 2, no. 2, pp. 72–75, 2016.

[9] A. H. Rangkuti, N. Hakiem, R. B. Bahaweres, A. Harjoko, and A. E. Putro, "Analysis of image similarity with CBIR concept using wavelet transform and threshold algorithm," *IEEE Symp. Comput. Informatics, Isc. 2013*, no. June 2017, pp. 122–127, 2013.

[10] C. Iswahyudi, "Prototype Aplikasi Untuk Mengukur Kematangan Buah Apel," *J. Teknol.*, vol. 3, pp. 107–112, 2010.

[11] J. Li and B. L. Lu, "An adaptive image Euclidean distance," *Pattern Recognit.*, vol. 42, no. 3, pp. 349–357, 2009.

[12] E. R. Ariyanto, "Implementasi Deteksi Citra Pornografi Berbasis Model Warna YCbCr dengan Metode Perbaikan C4.5 dan Shape Descriptor Untuk Filter Upload Foto di Media Sosial," pp. 1–6.