TECHNO
Vol.20, No.1, April 2019, Hal. 65~75
P-ISSN: 1410-8607, E-ISSN: 2579-9096

IMAGE SIMILARITY SEARCHING USE MULTI PART CUTTING AND GRAYSCALE COLOR HISTOGRAM

Sofyan Pariyasto[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*:*

Pencarian Kemiripan Citra
Histogram Citra Grayscale
Jarak Ecludiance

INTISARI

Pemanfaatan tekonologi informasi dalam kehidupan sehari-hari terus meningkat dengan begitu pesatnya. Hal ini tidak terlepas dari peran para peneliti khususnya dalam bidang teknologi informasi. Teknologi informasi sudah menjadi kebutuhan sehingga banyak dimanfaatkan dalam bidang pendidikan, perdagangan, peternakan bahkan hingga ke sektor pertanian. Salah satu kendala yang ada yaitu dibutuhkan manusia untuk melakukan pengecekan terhadap seluruh aktivitas yang melibatkan sistem informasi terutama pada saat terdapat data dalam bentuk citra. Permasalahan yang muncul biasanya dibutuhkan manusia untuk melakukan pengecekan dan pemilahan item yang selama ini biasanya dilakukan oleh manusia. Hal tersebut melatar belakangi penelitian ini untuk membantu mengurangi aktivitas yang melibatkan manusia. Proses pencarian kemiripan gambar dalam computer vision bisa dimanfaatkan dalam beberapa bidang seperti, pendidikan, retail, dan bidang lainnya. Dalam bidang pendidikan computer vision dapat dimaanfaat untuk proses absen otomatis memlaui face recognition, dalam hal retail bisa dimanfaatkan untuk proses sorting melalui object detection. Proses pencarian kemiripan antara citra yang menjadi queri dan citra dataset akan menjadi pokok penelitian, serta proses penghitungan kemiripian antara queri dan dataset akan dibahas langkah demi langkahnya.Metode yang digunakan dalam proses pencarian yaitu dengan menghitung jarak terpendek antara citra queri dengan dataset. Langkah-langkah yang dilakukan yaitu dilakukan fitur ektrasi kemudian dilakukan konversi warna RGB ke warna abu-abu. Tahapan berikutnya dilakukan pemotongan citra menjadi empat bagian yang selanjutnya akan dihitung jarak ecludiannya. Dibagian akhir akan hitung kinerja algoritma menggunakan metode confusion matrik, sehingga didapatkan hasil uji yang berupa, error rate, presisi, dan akurasi. Proses ujicoba menggunakan 30 data dengan menggunakan 1000 dataset. Dalam hasil uji didapatkan informasi yang berupa recal sebesar 1, akurasi 0.66 dan presisi 0.66.

Keyword*:*

Image Similarity
Histogram Image Grayscale
Ecludiance Distance

ABSTRACT

The use of information technology in everyday life continues to increase so rapidly. This is inseparable from the role of researchers, especially in the field of information technology. Information technology has become a necessity so that it is widely used in the fields of education, trade, livestock and even to the agricultural sector. One of the obstacles that is needed is to check all activities involving information systems, especially when there is data in the form of images. Problems that arise are usually needed by humans to check and sort items that have been carried out by humans. This is the background of this research to help reduce activities involving humans. The process of finding the similarity of images in computer vision can be used in several fields such as education, retail, and other fields. In the field of computer vision education can be utilized for the automatic absence process through face recognition, in terms of retailing, it can be used for sorting through object detection. The process of finding similarities between images that are queries and dataset images will be the subject of research, and the process of calculating similarities between queries and datasets will be

discussed step by step. The method used in the search process is by calculating the shortest distance between query images and dataset. The steps taken are the extraction feature and then RGB to gray color conversion. The next stage is to cut the image into four parts which will then be calculated the distance of the ecludian. The final part will calculate the performance of the algorithm using the matrix confusion method, so that the test results are in the form of error rates, precision, and accuracy. The trial process uses 30 data using 1000 datasets. In the test results obtained information in the form of recall of 1, 0.66 accuracy and 0.66 precision.

Korespondensi Penulis:

Sofyan Pariyasto
Department of Informatics, University AMIKOM Yogyakarta
Jl. Ring Road Utara, Condong Catur, Sleman, Yogyakarta
Central Java, 55283, Indonesia
Email: sofyanpariyasto@gmail.com

1. INTRODUCTION

The use of information technology in everyday life has never been released from the application of algorithms, methods and theories that have been generated from the contributions of the researchers developed. One of the contributions published in computer vision is a number of theories that have been found in the process of finding image similarity.

The search process for image similarity can be done by comparing Images to me can be done by comparing Feature to Feature (F2F), while in the search process will be calculated the distance (distance) of each existing image[1]. The image processing process is done by processing the color feature data, because the color feature has very winning information[2].

Images performed in the image search process Similarities to digital images are by utilizing information stored in images or commonly known as Content Based Image Retrieval (CBIR). CBIR is a method that is widely used in the process of finding image similarity that utilizes information from content contained in the image. These content are form information, texture information and other information that can be extracted in vector form[3].

Data in an RGB image will be numeric, the data is arranged in a matrix [4]. The matrix formed from images with size m*n is as follows.

(1)

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

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

2. METHODS

The process of finding image similarities in everyday life is very often found, whether we are aware or unconsciously. This image search is commonly used in good and rotten fruit selection processes, pattern recognition processes, face matching processes, and book search processes in libraries. In the whole process of the activity unwittingly carried out matching the similarity of the results of the visualization with the data stored in the memory, so that the results obtained were the same or resembled between the memory and the results of visualization.

2.1. Research Flow

The steps taken in the research process begin with the image selection process as a query, followed by the conversion of the image to gray. After the conversion, a histogram of the image becomes the query

which is compared to the image that is the dataset. The lowest distance image is the image that is most similar to the query entered, the search flow process looks like the following figure 1.

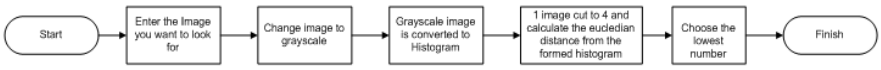

Figure 1. Research Flow

2.2. Grayscale Color

The Grayscale color is formed from color points between 0 and 1 where zero is black and 1 is white. Gray is produced by mixing black and white[5]. The process of converting RGB colors to gray can be done using the following equation[6].

$$G = 0.2126 * R + 0.7152 * G + 0.0722 * B \quad (2)$$

The image conversion process in equation 1 will produce a new image with gray as shown in the following figure 2.

Figure 2. Results of RGB image conversion to Gray

All images that will be counted pixel value will be converted into grayscale color.

2.3. Image cutting

Cutting the image in the process of looking for image similarity is done to compare each part in an image between the query and the dataset. Cutting the image is done by cutting the image into four parts consisting of the upper left, upper right, lower left and lower right, as shown in the following figure 3.

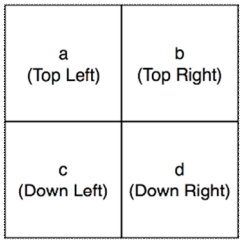

Figure 3. Display of image pieces

To get the size of the center line, the equation will be used as follows.

$$h = \frac{1\,x\,H}{2} \quad (3)$$

$$w = \frac{1\,x\,W}{2} \quad (4)$$

Information:
H : Overall image width
H : The width of the new image from the cut
W : Overall image length
W : The length of the new image from the cut

The process of cutting the image that has been done will produce four new images as shown in the following figure 4.

Figure 4. The image cutting results into 4 parts

2.4. Grayscale Color Histogram

The grayscale color histogram that is formed is the result of the calculation of the number of color components between 0 and 255. In the search process carried out it will produce four histograms which are the results of four images that have been cut in the previous stage. The four histograms of the image that has been cut off look like the following figure 5.

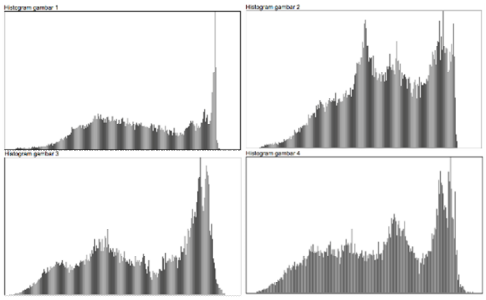

Figure 5. The formed histogram

3. RESULTS AND ANALYSIS

In the search trial process, the data taken from the Wang 1k dataset is used. In the dataset there are a thousand images divided into ten groups of categories.

3.1. Query Feature Extraction

In looking for images that are most similar to queries, several stages are interrelated with each other. The first step is extracting pixel values in an image. The extraction process will produce a numeric RGB value. The next RGB value will be changed to the gray image as shown in the following figure 6.

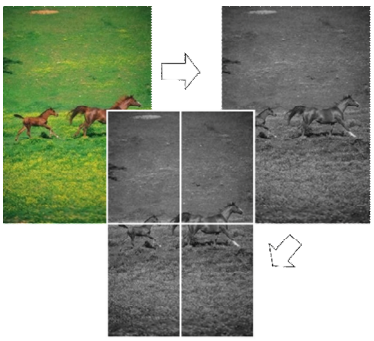

Figure 6. Image conversion and cutting process

From the results of cutting the image as shown in figure 6, the respective gray color values will be calculated so as to produce numbers such as table 1 below.

TABLE 1
IMAGE COLOR VALUE OF GRAY QUERY

Image	Color	Part 1	Part 2	Part 3	Part 4

757.jpg	75	13,295	28,21	51,063	58,637
757.jpg	76	14,35	25,763	57,593	63,63
757.jpg	77	13,352	29,32	68,835	71,103
757.jpg	78	17,734	33,323	65,573	70,106
757.jpg	79	18,57	35,951	67,817	63,223
757.jpg	80	25,896	45,322	63,939	74,858
757.jpg	81	22,786	41,375	59,943	81,279
757.jpg	82	18,643	41,124	79,131	71,682
757.jpg	83	20,47	43,938	70,268	78,688
757.jpg	84	19,422	56,95	70,617	93,598
757.jpg	85	21,618	66,312	79,007	100,468
757.jpg	86	29,495	76,676	75,863	103,768
757.jpg	89	46,643	88,03	92,209	111,771
757.jpg	90	47,514	86,512	101,237	118,013

From table 1, four histograms will be formed that will be compared to the images in the dataset. The histogram image formed looks like Figure 7 below.

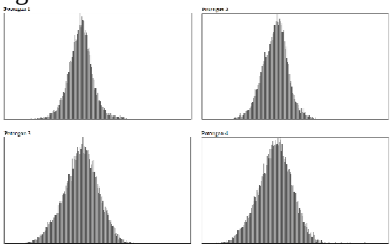

Figure 7. Histogram query

3.2. Query Distance and dataset

The distance calculation process is done by cutting the image into Query and Image which becomes a Dataset into four parts, then counting between datasets and queries entered, note the following figure 8.

Figure 8. Image Query with Search Results

Each image piece will then be compared one by one to look for similarities based on the lowest eccentric distance. The equation used to find the shortest distance between the image and the dataset is as follows[7].

(5)

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

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

The histogram that is formed from the queri image part 1 (a. top left) and the dataset looks like Figure 9 below.

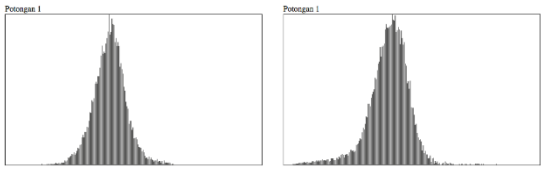

Figure 9. Histogram part 1, Query and dataset

The search for eccentric distance looks like table 2 below.

TABLE 2
QUERY DISTANCE AND PART 1 DATASET

Color	Query	Dataset	(k-k')	$(k-k')^2$
111	237,77	298,48	-60,71	3.685,29
112	245,34	265,5	-20,16	406,39
113	237,96	269,05	-31,09	966,72
114	196,37	275	-78,63	6.182,36
115	203,66	270,21	-66,55	4.428,26
116	191,86	241,67	-49,8	2.480,52
117	182,19	268,8	-86,61	7.501,10
118	156,61	253,84	-97,23	9.453,88
119	148,34	240,46	-92,12	8.486,57
120	126,35	245,61	-119,27	14.224,37
121	117,64	202,17	-84,53	7.145,94
122	112,7	198,11	-85,41	7.294,04
123	82,37	163,13	-80,76	6.521,90
124	88,54	167,69	-79,15	6.264,66
125	88,75	152,07	-63,32	4.009,76
126	70,14	168,42	-98,28	9.658,19
127	75,31	115,38	-40,07	1.605,57
128	53,19	114,95	-61,77	3.815,34
129	54,12	114,51	-60,38	3.646,17
130	55,07	100,46	-45,39	2.060,61

$d = \sqrt{128066,68}$

$d = 457,86$

The comparison of the histogram for the 2nd cut (b. top right) looks like the following figure 10.

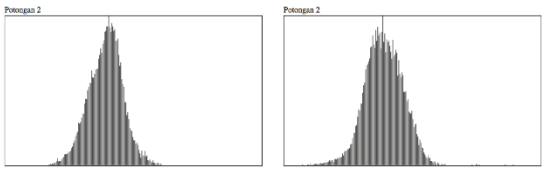

Figure 10. Cut 2 Histogram, Query and dataset

Calculation of the distance of the ecludiance formed looks like table 3 below.

TABLE 3
QUERY DISTANCE AND DATASET PART 2 DATASET

Color	Query	Dataset	(k-k')	$(k-k')^2$
111	251,22	232,95	18,27	333,83
112	255,62	226,9	28,72	825,02
113	222,3	180,23	42,06	1.769,33

114	200,31	211,17	-10,86	117,99
115	207,56	184,07	23,49	552,01
116	173,36	177,23	-3,86	14,93
117	163,68	187,92	-24,24	587,41
118	158,32	181,6	-23,28	541,97
119	127,84	148,51	-20,67	427,1
120	127,77	147,74	-19,97	398,85
121	101,11	149,65	-48,55	2.356,80
122	107,18	129,72	-22,54	508,04
123	93,97	133,54	-39,57	1.566,04
124	75,19	108,26	-33,06	1.093,17
125	71,62	106,33	-34,71	1.204,75
126	69,19	104,36	-35,17	1.237,26
127	57,61	110,16	-52,56	2.762,25
128	51,95	93,84	-41,89	1.754,82
129	50,51	77,25	-26,74	714,99
130	44,07	82,21	-38,14	1.454,63

$d = \sqrt{103362,22}$

$d = 321,49$

For the third image piece (c. bottom left), a histogram will be formed which is a comparison between the image that becomes the query and the image that is the dataset. The histogram that is formed from the image fragment part 3 looks like the following figure 11.

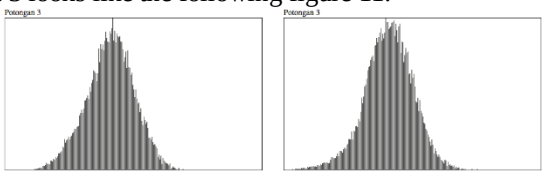

Figure 11. Histogram section 3, Query and dataset

Calculation of the ecludian distance formed from figure 11 looks like table 4 below.

TABLE 4
QUERY DISTANCE AND PART 3 DATASET

Color	Query	Dataset	(k-k')	$(k-k')^2$
111	270,25	289,5	-19,25	370,51
112	264,7	276,69	-11,98	143,59
113	262,68	252,96	9,72	94,41
114	264,26	284,11	-19,84	393,69
115	254,67	289,93	-35,26	1.243,44
116	253,88	231,94	21,93	481,08
117	238,64	254,29	-15,64	244,76
118	214,7	252,19	-37,48	1.404,94
119	223,46	213,8	9,65	93,2
120	234,66	227,77	6,89	47,45
121	202,14	241,06	-38,92	1.515,07
122	236,2	232,45	3,75	14,06
123	203,09	234,36	-31,26	977,4
124	202,34	189,55	12,79	163,51
125	188,59	191,08	-2,49	6,19
126	192,54	171,61	20,93	438,24
127	168,58	198,73	-30,15	909,27
128	159,13	182,68	-23,55	554,48
129	144,51	190,65	-46,14	2.129,25
130	159,09	165,76	-6,66	44,39

$d = \sqrt{58940,32}$

$d = 242,77$

Calculations are also performed on the fourth image piece (d. bottom right) so as to produce a new histogram, which is then compared with the dataset as shown in the following figure 12.

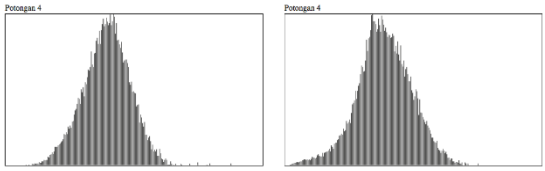

Figure 12. Histogram part 4, Query and dataset

Calculation of the ecludian distance formed from figure 12 looks like the following table 5.

TABLE 5
QUERY DISTANCE AND PART 4 DATASET

Color	Query	Dataset	(k-k')	$(k-k')^2$
111	248,82	222,96	25,86	668,69
112	249,48	246,21	3,28	10,74
113	240,59	220,16	20,44	417,77
114	233,91	208,34	25,57	653,67
115	245,66	238,92	6,74	45,41
116	224,16	201	23,16	536,32
117	204,72	196,68	8,03	64,55
118	208,12	208,53	-0,41	0,17
119	210,72	176,45	34,27	1.174,76
120	210,81	191,38	19,43	377,45
121	203,21	195,06	8,15	66,44
122	195,46	181,95	13,51	182,58
123	165,95	167,53	-1,59	2,52
124	160,33	167,83	-7,5	56,27
125	166,89	176,72	-9,83	96,73
126	146,97	149,9	-2,92	8,54
127	147,25	144,52	2,73	7,46
128	144,81	130,21	14,6	213,27
129	143,22	137,9	5,33	28,37
130	117,84	151,29	-33,45	1.119,02

$d = \sqrt{105346,73}$

$d = 324,57$

3.3. Search result

The search process with the image query 757.jpg produces several images that are considered similar, as shown in the following figure 13.

Figure 13. The most similar search results.

From figure 13, there are 12 image that show the shortest distance between the query and the dataset, as shown in the following table.

TABLE 5
NEARBY QUERIES AND DATASET

Query	Database	Distance
757.jpg	533.jpg	1001,58
757.jpg	705.jpg	1004,08
757.jpg	349.jpg	1005,30
757.jpg	793.jpg	1006,00
757.jpg	728.jpg	1006,93
757.jpg	702.jpg	1014,74
757.jpg	370.jpg	1019,47
757.jpg	783.jpg	1034,64
757.jpg	813.jpg	1034,65
757.jpg	227.jpg	1035,20
757.jpg	727.jpg	1044,95
757.jpg	172.jpg	1047,95

3.4. Algorithm Performance Measurement

To measure the performance of the algorithm, sampling is done by taking each of the two images in each category to produce 30 random images.

TABLE 6
DATA SAMPLE FOR PERFORMANCE MEASUREMENT

			Result	
No	Query	dataset	Actual	Prediction
1	10.jpg	39.jpg	T	T
2	15.jpg	76.jpg	T	T
3	84.jpg	57.jpg	F	T
4	105.jpg	16.jpg	F	T
5	175.jpg	98.jpg	F	T
6	180.jpg	173.jpg	T	T
7	211.jpg	216.jpg	T	T
8	245.jpg	686.jpg	F	T
9	264.jpg	242.jpg	T	T
10	316.jpg	364.jpg	T	T
11	338.jpg	641.jpg	F	T
12	359.jpg	338.jpg	T	T
13	437.jpg	474.jpg	T	T
14	455.jpg	462.jpg	T	T
15	494.jpg	442.jpg	T	T
16	533.jpg	547.jpg	T	T
17	536.jpg	513.jpg	T	T
18	572.jpg	534.jpg	T	T
19	619.jpg	695.jpg	T	T
20	626.jpg	603.jpg	T	T
21	685.jpg	627.jpg	T	T
22	747.jpg	750.jpg	T	T
23	757.jpg	774.jpg	T	T
24	788.jpg	743.jpg	T	T

25	838.jpg	871.jpg	F	T
26	855.jpg	144.jpg	F	T
27	869.jpg	866.jpg	F	T
28	911.jpg	325.jpg	F	T
29	965.jpg	976.jpg	F	T
30	984.jpg	939.jpg	T	T

The algorithm used to measure algorithm performance is as follows[8][9][10][11].

(6)

$$Precision = \frac{TP}{TP + FP}$$

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

$$Recall = \frac{TP}{TP + FN}$$

TABLE 7
CONFUSION MATRIX

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

From table 6, the value of TP=20, FP=10, FN=0, and TN=0. The results of the calculation produced are as follows

$$Precision = \frac{20}{20 + 10} = 0.66$$

$$Accur = \frac{20 + 0}{20 + 0 + 10 + 0} = 0.66$$

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

The process of measuring algorithm performance uses the confusion matrix method so that it is produced[12]. Value Precision rate of 0.66, 0.66 Accuracy, and 1 Recall.

4. CONCLUSION

From the research conducted it can be concluded that, in the trial process using a 1k dataset that was converted into RGB images. The image is then cut into 4 parts which are then compared to the dataset. The search process requires between 20 and 30 minutes, with hardware specifications Proc 2.4 Ghz Intel Core I5, 8GB ram, OS MacOS High Siera 10.13.6. The search trial process produces an accuracy rate of 0.66, precision 0.66 and recall 1.There are images that differ in color and shape but when converted into histograms they have a fairly high similarity to the query. The weakness of this research is that images that have histograms similar to queries are displayed as similar images even though the reality is a very different image.

5. REFERENCES

[1] L. Liu, M. Yu, and L. Shao, “Image Similarity Search,” vol. 46, no. 11, pp. 1–11, 2016.
[2] A. Baita, B. S. W, and A. Sunyoto, “Logo Retrieval Berdasarkan Ekstraksi Multifitur,” *Magistra*, no. 98, pp. 53–59, 2016.
[3] C. Digital, “Citra Digital,Warna, CBIR, RGB, HSV, Sistem, Histogram. 1.,” vol. 19, no. 1, 2016.

[4] F. Flaurensia, T. Rismawan, R. Hidayati, and J. S. Komputer, "PENGENALAN MOTIF BATIK INDONESIA MENGGUNAKAN DETEKSI TEPI CANNY DAN TEMPLATE MATCHING," vol. 4, no. 2.
[5] M. Pada and C. A.- Qur, "SISTEM PENDETEKSI POLA LAFADZ ALLAH DAN."
[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. Li and B. Lu, "An adaptive image Euclidean distance," vol. 42, pp. 349–357, 2009.
[8] S. P. M. Choi, S. S. Lam, K. C. Li, and B. T. M. Wong, "Learning Analytics at Low Cost : At-risk Student Prediction with Clicker Data and Systematic Proactive Interventions," vol. 21, pp. 273–290, 2018.
[9] M. A. Nugroho and H. A. Santoso, "KLASIFIKASI DOKUMEN KOMENTAR PADA SITUS YOUTUBE MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR (K-NN)," no. 5.
[10] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," *Inf. Sci. (Ny).*, 2016.
[11] A. M. Abdel-zaher and A. M. Eldeib, "PT US CR," *Expert Syst. Appl.*, 2015.
[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.