Tool Image Searching Based Canopy K Means Clustering Using Web Programming Study Case MNC Publishing

Ariadi Retno[1*], Wilda Imama[2], Puspa Kirana[3], Vit Zuraida[4]
Department of Information Technology, Politeknik Negeri Malang, Indonesia
*faniri4education@gmail.com

Abstract.

This research builds a tool for e-commerce search based on image input searching where the study case in this research is in MNC Publishing. The purpose of this research is to help users find the book that they want by inputting the image in the tool and then finding the book by the image that they find, and the result is linked to the detail of the book that they find. This research uses Canopy K Means Clustering to cluster the image and count the value of centers in system machine learning in this application to use clustering to minimize sum error compared by the count of centers value result by counting average images each cluster with reduced data images with value error loose distance each cluster until N iteration. Images normalization with reduced matrix from size 242193 reduce to 100 dimensions of matrix each data in web programming with 500 images result matrix input [500x100] dimensions system counts the image searching from size matrix [500x242193] become [500x100] dimensions of matrix for high computation with Canopy K Means Clustering. Testing with grayscale, the application got the result clustering 100% true for clustering from 50 data images and 77,4% from 500 data images.

Keywords: Canopy K Means Clustering; Image Book, Matrix, PHP, My SQL, DIA Modelling

1. Introduction

Application for image searching has already been developed in many applications actually for E-Commerce. Image searching in E-Commerce makes user easier to find the product that they want by inputting the image from a product that they want to find. Besides searching by text, E-Commerce development to search by image can help reduce redundant search results to display for customers online. Application E-Commerce can be applied in the sale of the product for Company products, Home Industry products, Book stores, etc. To reduce computation in search with data image, we can use the clustering method [1][2] and the data used for clustering learning is very influenced by the computation error data can be changed to grayscale by the system and can be added to few filters for add learning data. E-commerce build makes user customers easily find products they want by application web, so they can reach information from anywhere and get the product by system so they don't need to go to store place, which can reduce customer cost. That's why E-Commerce become important in business. The transaction can be much more than the offline transaction. So, E-Commerce tries to make user customer easier to get information on the application web.

The facility in E-Commerce like search tools usually used based on text input and the user finds a product by text input, the system will show the product that customers find. Tool by input image usually known as Image Searching [3][4][5][6][7][8] is a tool for how customers get information product by input image and then system will show product. The system in Image Searching will count the value of pixel and similarity of pixel and picture in the product database, then the system will show all products that have a range close value that's clustered product with the input image. There are view E-Commerce use tool image processing, the difficulty in the application is computation image, and computation in image searching [9][10][11] make the application search slowly. To reduce this problem the picture needs to be minimized with a reduced size dimension, but this reduced dimension influences in recognition of the picture in the system.

This research uses a clustering method with Canopy K Means Clustering to cluster the value image center of the product, in this case application with MNC Publishing Book Store. This research uses filtered because image data is used in the database only one from the data bookstore, so more characteristic data by filtered image. The tool in this research by choose picture in computer data, and then the system will count the clustered value that the input image has a minimized relation value and then show the information of the book in system MNC Publishing.

2. Method

The data in this research is an image from MNC Publishing from the cover of the Book Store, and then data filtered because only one data each cover. Machine learning in this system counts the center value from image data using Canopy K Means Clustering, and before counting the center value, data must normalize which reduction dimension. Data in this research become 500 data images filtered from 50 covers where the system already filtered each cover with 10 filter images using PHP programming. Reduce dimension from actual size image [434x607] if change to dimension become 242193, if use this dimension for high computation in Canopy K Means Clustering needs management memory in PHP programming, so to reduce the computation change image by resize function become [10x10] pixel and its very significant value and must be influenced in result clustering.

Management memory in matrix data is influenced by dimension reduction and management file in application. After normalization data, the data image clustered, and the result center value for testing data. In testing data, each data will be tested in the tool system one by one different from learning data where all data count in machine learning use Canopy K Means Clustering.

Statistical value in Canopy K Means Clustering each iteration must count center value and average new center value each iteration. Each iteration must average all data that is included in each cluster and use the Canopy K Means Clustering formula to get a new value center for each iteration. The system will display the result from the input image by the user, and then display details of the book from the MNC Publishing link web. The most difficult in this application is the computation in clustering, and how to normalize the system for learning data and each testing data one by one. The system model in this application with application DIA and flowchart shows the design of the system.

3. Result and Discussion

Desain Modelling System

The center value can get from Canopy K Means Clustering starting from the initial value and then counting each iteration by Canopy K Means Clustering before learning computation with Canopy K Means Clustering data must normalized. Figure 1 is the flow of the system in this research for admin and user with their priority in system use tool DIA.

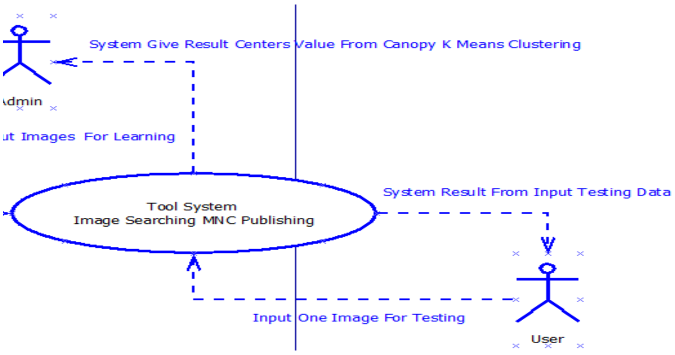

System Give Result Centers Value From Canopy K Means Clustering
\dmin
ıt Images For Learning
Tool System
Image Searching MNC Publishing
System Result From Input Testing Data
Input One Image For Testing
User

Figure 1. User Priority in System Application Image Searching in MNC Publishing

Figure 1 is the priority user admin can input data learning in the system database, in this research admin inputs 50 images and saves the path folder in the database. The tool system in this research will count the 50 images and then filter them be 500 images for variation center value characters, the result is that 50 center values from each cluster will be shown for admin from Canopy K Means Clustering. For the user customer, the customer inputs only one image in the system, and then the system counts by the center value that is already got from the system admin so the comparison only to 50 centers not to 500 images becomes the purpose of Canopy K Means Clustering, and then the system will link to MNC Publishing detail about input cover book that user already chosen.

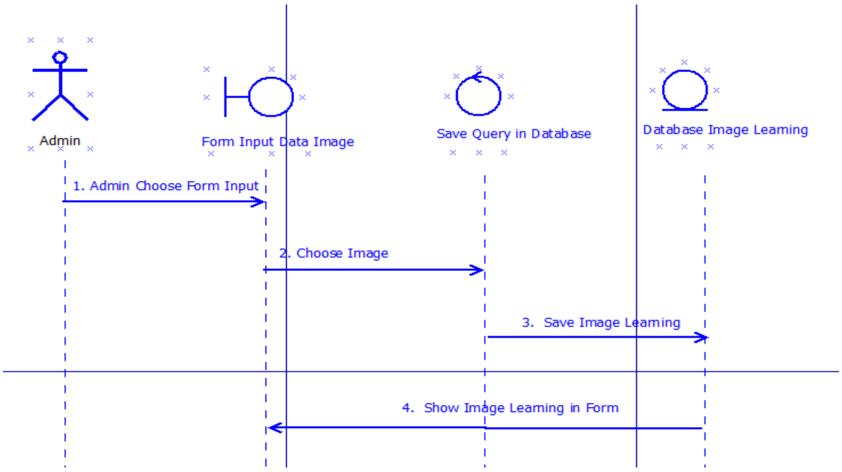


Figure 2. Sequence Diagram Admin input images learning

Figure 2 is the sequence admin in input images in the system by input form and then image learning will be saved in the database. Input images in the application one by one chosen by admin and then save by query in database My SQL.

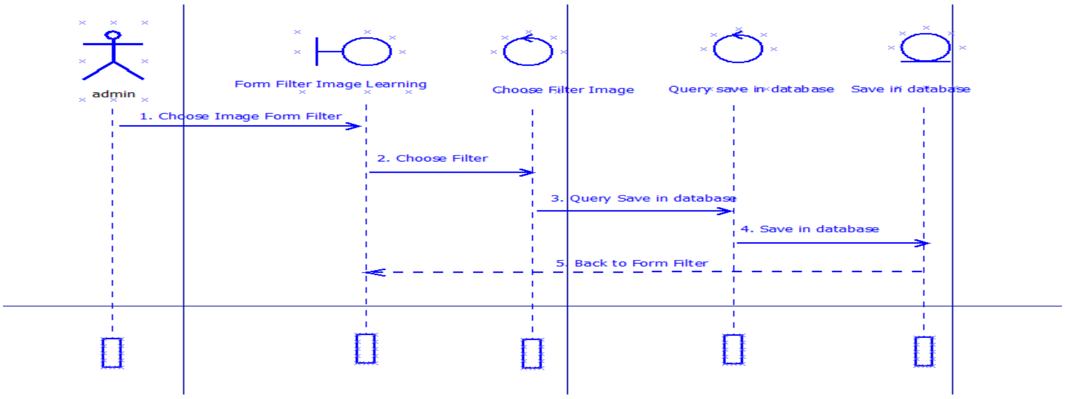


Figure 3. Sequence Diagram Admin filter images

Figure 3 shows a sequence diagram for the admin to filter image each image, and this form displays 10 filters for each data beginning with the admin choosing the form filter, then choosing an image, and choosing the filter in the system, and the result filter saves in the database. This process is the same for all filter images, from grayscale, filter means removal, filter brightness, filter contrast, filter colorize, filter Gaussian blur, filter smooth, rotation -5 degrees, rotation +4 degrees.

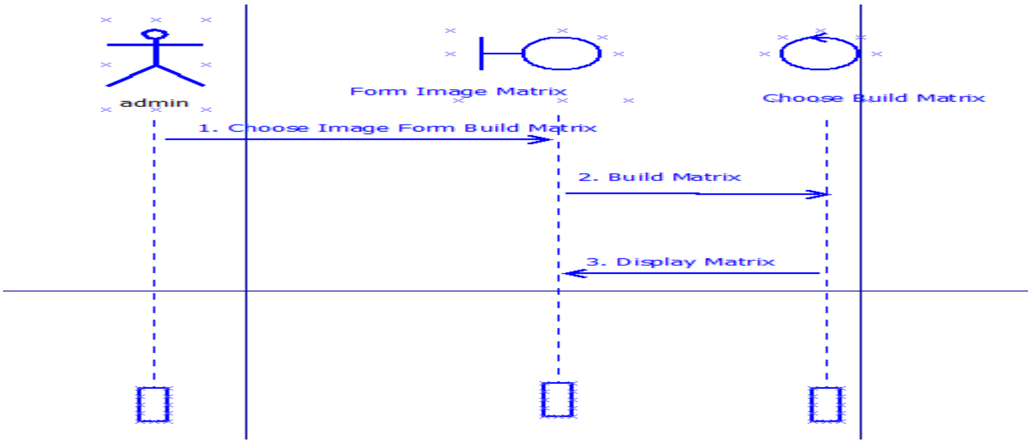


Figure 4. Sequence Diagram Admin Build Matrix

Figure 4 shows the sequence diagram to build a matrix from all image data in the database, the result is a matrix size of 500 rows and 100 columns, so the size of the matrix is [500x100]. This matrix will be computed with high computation because in Canopy K Means Clustering uses iteration and calculation of every data in each cluster.

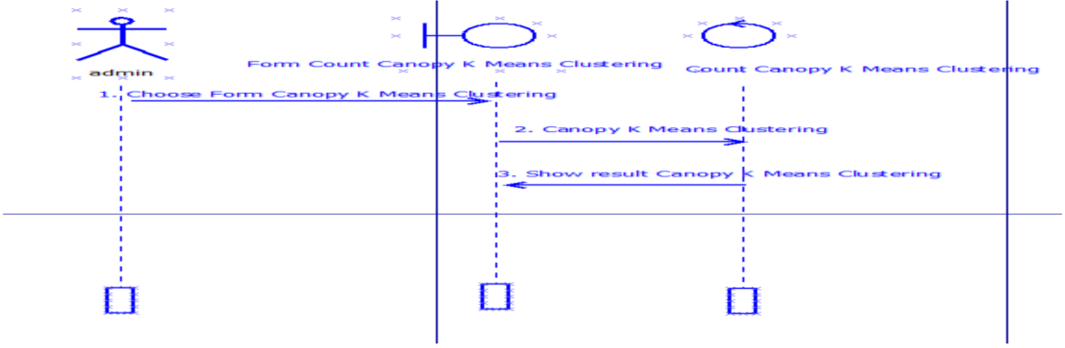

admin
Form Count Canopy K Means Clustering
Count Canopy K Means Clustering
1. Choose Form Canopy K Means Clustering
2. Canopy K Means Clustering
3. Show result Canopy K Means Clustering

Figure 5. Sequence Diagram Count Canopy K Means Clustering

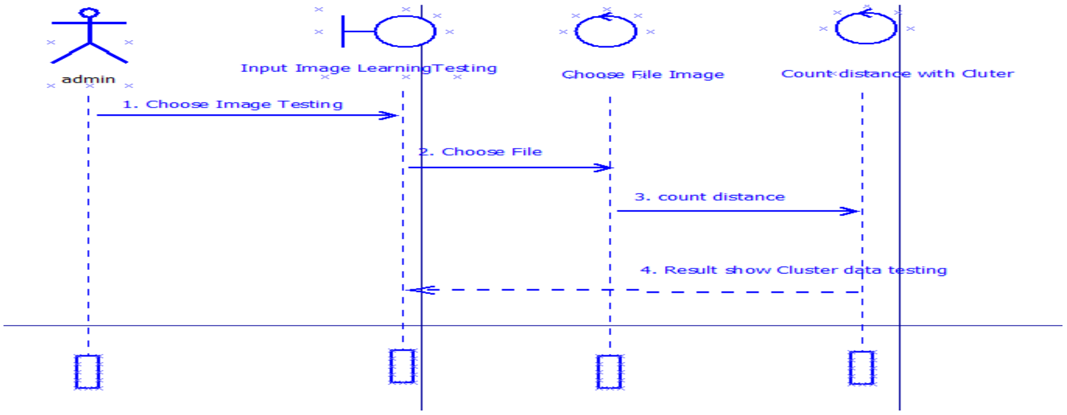

admin
Input Image LearningTesting
Choose File Image
Count distance with Cluter
1. Choose Image Testing
2. Choose File
3. count distance
4. Result show Cluster data testing

Figure 6. Sequence Diagram Testing Image Data

Figure 5 is a sequence diagram in count Canopy K Means Clustering in the system, with input data images 500 data and dimension 100, computation in each iteration, and at last get the result value of cluster. Figure 6 shows user input data testing, one data image for each test data, and then the system will normalize data and count the value with the center clustered in this research have 50 clustered and then the result will show in form after that link detailed of input image will be shown.

Canopy K Means Clustering Method

The clustering method is implemented in many applications such as clustering in data with numeric characteristics such as clustering for academic transcripts whose students have a probability get a scholarship, clustering data in stores like the most products with high transactions or products with the lowest transaction, etc. Clustering method for image data [12][13][14][15] must be normalized data first.

Normalized data

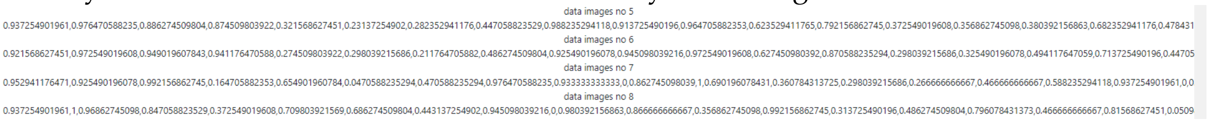
Normalized data in this research have a few phases as below:

1. Input the actual size data image from the folder computer.
2. Filter to gray image.
3. Resize the image as the same dimension for all images if the size matrix is [10x10] all images should be resized in dimension [10x10].

That's normalized data in this research with the purpose that all images have the same dimension for getting accurate values in computation.

Filter data

Filter data [16] in this research is influenced for accuracy and adds characteristic data in data learning, this research has 10 filter-use functions in PHP programming. All figures, have a characteristic value for clustering data and each pixel starts from 0 until 1 and between them already normalized 0 to 255 become 0 to 1 in the system as Figure 7.



```
                                   data images no 5
0.937254901961,0.976470588235,0.886274509804,0.874509803922,0.321568627451,0.23137254902,0.282352941176,0.447058823529,0.988235294118,0.913725490196,0.964705882353,0.623529411765,0.792156862745,0.372549019608,0.356862745098,0.380392156863,0.682352941176,0.478431
                                   data images no 6
0.921568627451,0.972549019608,0.949019607843,0.941176470588,0.274509803922,0.298039215686,0.211764705882,0.486274509804,0.925490196078,0.945098039216,0.972549019608,0.627450980392,0.870588235294,0.298039215686,0.325490196078,0.494117647059,0.713725490196,0.44705
                                   data images no 7
0.952941176471,0.925490196078,0.992156862745,0.164705882353,0.654901960784,0.0470588235294,0.470588235294,0.976470588235,0.933333333333,0,0.862745098039,1,0.690196078431,0.360784313725,0.298039215686,0.266666666667,0.466666666667,0.588235294118,0.937254901961,0,0
                                   data images no 8
0.937254901961,1,0.96862745098,0.847058823529,0.372549019608,0.709803921569,0.686274509804,0.443137254902,0.945098039216,0,0.980392156863,0.866666666667,0.356862745098,0.992156862745,0.313725490196,0.486274509804,0.796078431373,0.466666666667,0.81568627451,0.0509
```

Figure 7. Example one of image with normalized data value in user administrator.

A characteristic filter need in this system because only one covers data so to add characteristic data in this research use a filter. Before learning in Canopy K Means Clustering, must build a data matrix in one variable, and all 500 data must change to a data array and save in one variable, the size matrix is [500x100].

Canopy K Means Clustering

Algorithm Canopy K Means Clustering [17][18] method as bellow:

1 Choose Loose distance and Tight distance.
2 Count each cluster except Loose distance.
3 Count center.
4 Clustered data with data more than Tight distance.
5 Remove Loose distance data from each cluster.
6 Looping from no 2 until N iteration.

Canopy K Means clustering is based on K Means Clustering [19][18][20][21]with mathematics formula as below

$$X = \{x_1, x_2, x_3, \dots, x_N\} \quad (1)$$

Where X is variable for all data images, x is variable for each image, N is sum of images data

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left| x_{j-c_i} \right|_2 \quad (2)$$

Where J is variable of sum error for all clusters, sum of clusters is variable k, i is variable index from cluster 1 untill cluster k, $C_i$is a variable for all elements in cluster i, x_i is a variable for data in includes cluster i, x_{j-c_i} is a variable count error distance from each data to each cluster i.

x_i=($x_{i1}, x_{i2}, x_{i3}, \dots, x_{im}$), m is the dimension of cluster normalization, i is an index of the cluster.

x_j=($x_{j1}, x_{j2}, x_{j3}, \dots, x_{jm}$), j is index of images in this research.

$$d(x_i, x_j)_= \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{im} - x_{jm})^2} \quad (3)$$

Where $d(x_i, x_j)$ is distance from each data x_j with each cluster x_i, for all dimension data images and clusters.

$$\text{Procentage} = \frac{\text{Number true searching images}}{\text{Number off all images}}. \quad (4)$$

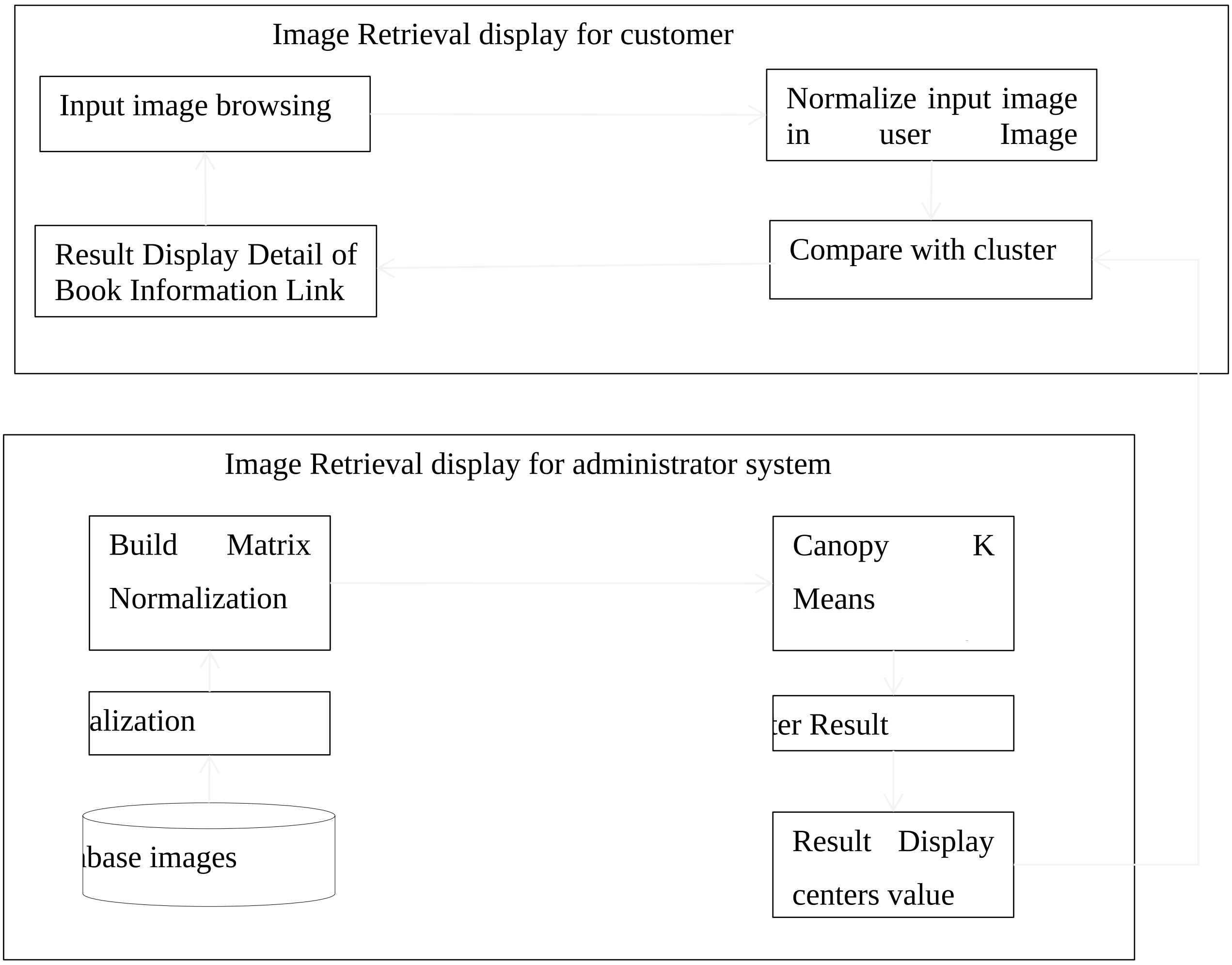


Image Retrieval display for customer
Input image browsing
Normalize input image in user Image
Compare with cluster
Result Display Detail of Book Information Link
Image Retrieval display for administrator system
Build Matrix Normalization
Canopy K Means
alization
er Result
ıbase images
Result Display centers value

Figure 8. Image retrieval image searching for customer and admin in the system.

Figure 8 shows normalization images for the user customer and user admin for resizing images in the same dimension size for all images. Images in image retrieval for user admin clustered 500 data images, and for user customer find an image by one image input customer. Centers value from user customer get from result Canopy K Means Clustering from user admin. In user administrator, besides normalization image with resize image for reduction dimensions, there is management memory in PHP Programming because in Canopy K Means Clustering needs high computation with some iteration. The system administrator displays the result of cluster values, and for user customer shows the detailed information of the cover book to MNC Publishing web by clustered testing image results. Figure 9 shows filtered images used in this research, there are 10 filters used to add characteristics data in this research.

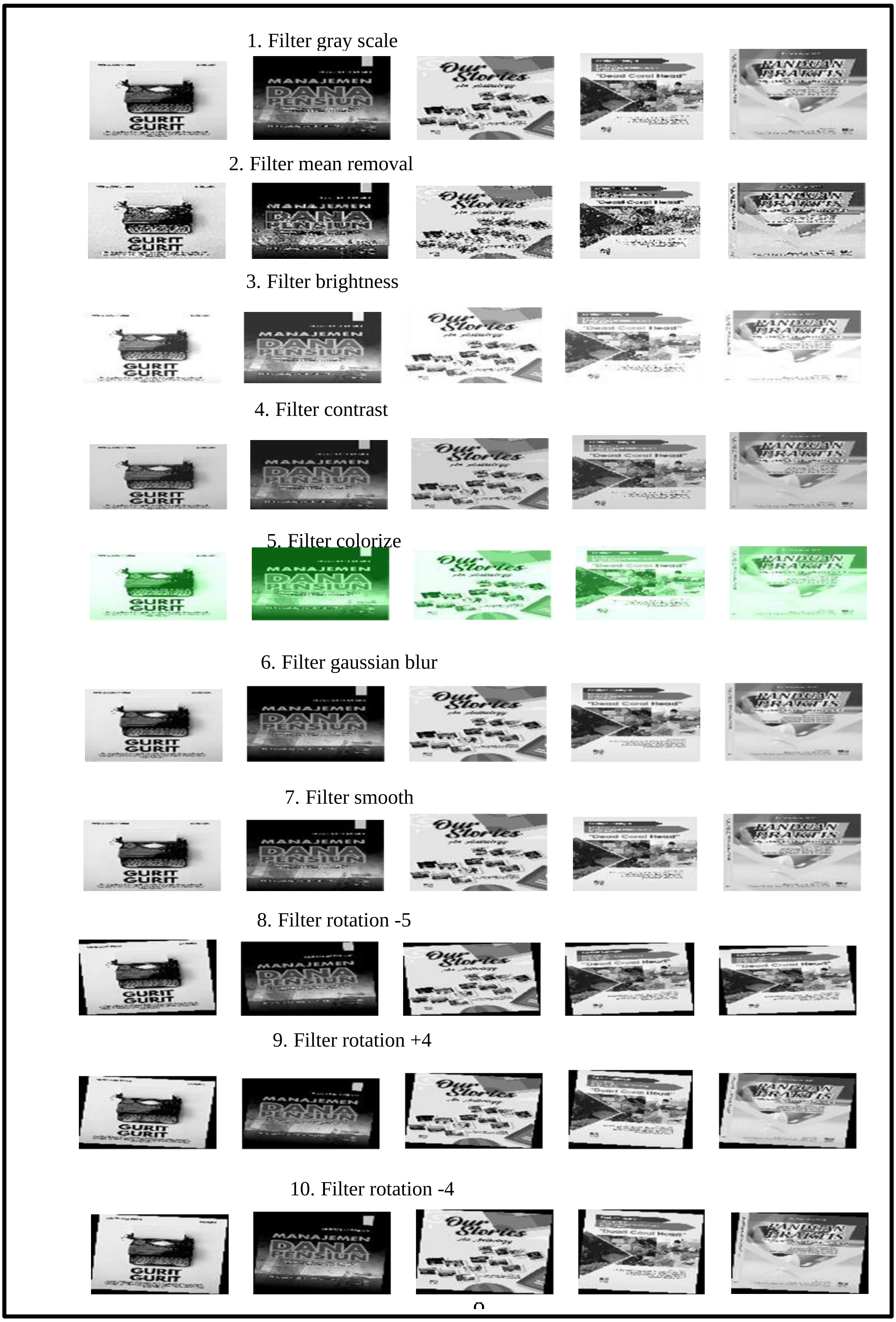


Figure 9. 10 Filters in Tool System

Application Result

Result from Canopy K Means Clustering for 500 data clustered as Table 1 below:

Table1. Result Clustered data

cluster	true	false	cluster	true	false	cluster	true	false
1	7	3	21	8	2	41	10	0
2	7	3	22	8	2	42	8	2
3	7	3	23	10	0	43	10	0
4	10	0	24	10	0	44	9	1
5	8	2	25	9	1	45	9	1
6	10	0	26	10	0	46	8	2
7	10	0	27	10	0	47	10	0
8	9	1	28	8	2	48	10	0
9	10	0	29	5	5	49	10	0
10	10	0	30	7	3	50	10	0
11	9	1	31	10	0			
12	10	0	32	2	8			
13	9	1	33	5	5			
14	10	0	34	10	0			
15	10	0	35	6	4			
16	8	2	36	6	4			
17	7	3	37	10	0			
18	10	0	38	8	2			
19	9	1	39	10	0			
20	10	0	40	10	0			

Table 1 is testing from 500 data images and each cluster has 10 image data from testing from 500 data image and each cluster has 10 image data from filtering and cover of books using Canopy K Means Clustering with 50 center comparison and pixel [10x10] dimension each image and very significant from actual size matrix. Results from a few clusters are 100%, and many clusters have 80% to 90%. The percentage from cluster result is influenced by detail characters of pixel image after resizing the image become [10x10] and filtering characteristics. Although the significantly reduced dimension with the resize matrix, the system can recognize pixels, which means characteristic data in each cluster is enough if count clustering uses size [10x10] because clustering is high computation so resize is one of the ways to reduce dimension. Figure 10 shows the sum of each cluster data from 500 data based on Table 1, the picture shows the sum of images clustered to true cluster and the sum of images clustered to false clustered.

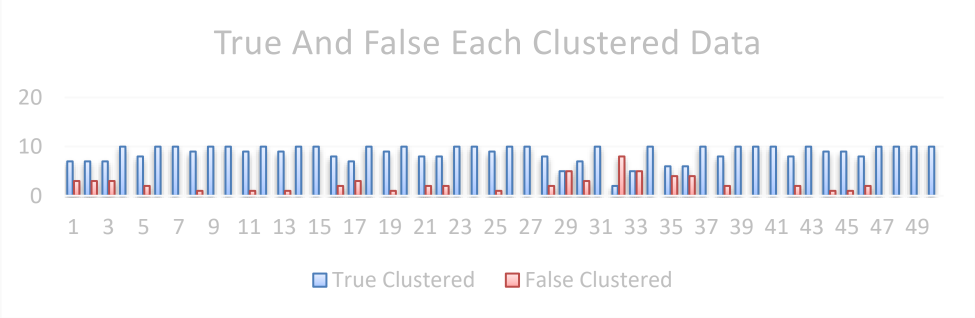


Figure 10. Graphic true and false for each cluster

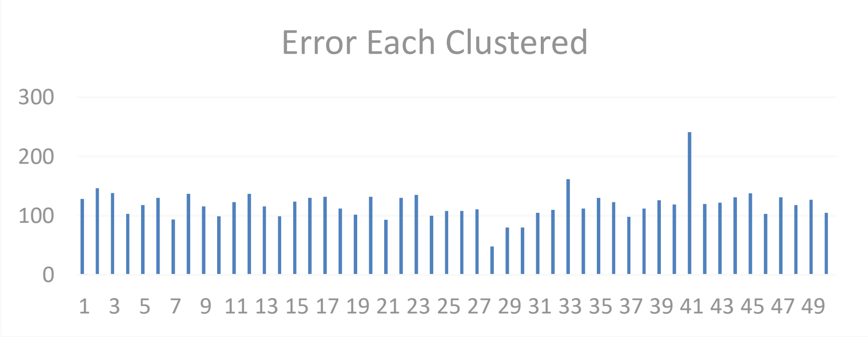
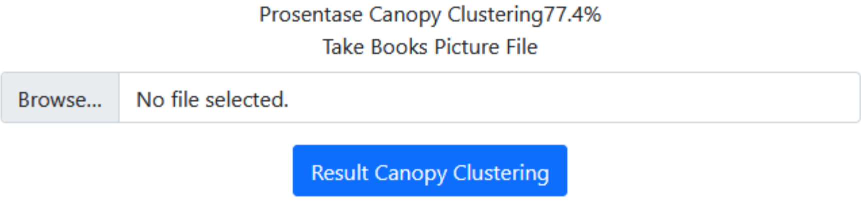


Figure 11. Graphic sum error for each cluster


Prosentase Canopy Clustering77.4%
Take Books Picture File

No file selected.

Figure 12. Form Tool Image Searching user test

Figure 11 is the sum error for each cluster in the system, size of clusters resulting from Canopy K Means Clustering is [50x100]. Figure 12, shows a tool for user customer input images and then will link to MNC Publishing details of the cover book.

4. Conclusion

The result from this clustered in a web application has image search result clustering 77,4 % for 500 data and for data gray scale has image search result clustering 100% with reduced size from actual size dimension 242193 become 100 dimensions it means the computation system for clustering data has true value with computation some iterations, size of normalization dimension and filter influenced centers data value and influenced the accuracy.

5. References

[1]. Basar, S., Ali, M., Ruiz, G. O., Zareei, M., Waheed, A., and Adnan, A. 2020. Unsupervised Color Image Segmentation: A Case of RGB Histogram Based K-Means Clustering Initialization. *Journal Pone* 0240015

[2]. Bakhthemmat, A. and Izadi, M. 2020. Decreasing the Execution Time Of Reducers By Revising Clustering Based on The Futuristic Greedy Approach, *Journal of Big Data*.

[3]. Bhoir, S. V. and Patil, S. 2021. A Review on Recent Advances in Content-Based Image Retrieval A Review on Recent Advances in Content-Based Image Retrieval used in Image Sear used in Image Search Engine. *Library Philosophy and Practice* (e-journal) 5617.

[4]. Dewan, J. H. and Thepade, S. D. 2021. Feature Fusion Approach for Image Retrieval with Ordered Color Means Based Description of Key Points Extracted Using Local Detectors. *Journal of Engineering Science and Technology* 16(1) p482 – 509.

[5]. Al-Jubouri, H. A. 2019. Content-Based Image Retrieval: Survey. *Journal of Engineering and Sustainable Development* 23(3).

[6]. S. Dhinakaran. 2020. Study of Content-Based Image Retrieval Using Data Mining Techniques, *IJSRCSEIT* 6(3).

[7]. Jaiswal, A. K., Liu, H. and Frommholz, I. 2019. Effects of Foraging in Personalized Content-based Image Recommendation. *Proceedings of EARS: The 2nd International Workshop on Explain a b*.

[8]. Hu, Y. P, Yin, H., Han, D., and Yu, F. 2019. The Application of Similar Image Retrieval in Electronic Commerce. *Scientific World Journal* Volume Article ID 579401.

[9]. Khan, M. F., Monir, M., and Naseem, I. 2021. Robust Image Hashing Based on Structural and Perceptual Features for Authentication of Color Images. *Turkish Journal of Electrical Engineering & Computer Sciences*.

[10]. Larijani, M. R., Ardeh, E. A. A., Kozegar, E. and Loni, R. 2019. Evaluation of Image Processing Technique in Identifying Rice Blast Disease in Field Conditions Based on KNN Algorithm Improvement By K-Means. Food Science and Nutrition. 7(12), p 3922-3930.

[11]. Lin, X., Gokturk, B., Sumengen, B. and Vu, D. 2008. Visual search engine for product images, *Proceedings of SPIE - The International Society for Optical Engineering*.

[12]. Sotomayor, C. G., Mendoza, M., Castañeda, V., Farías, H., Molina, G., Pereira, G., Härtel, S., Solar, M., and Araya, M. 2021. Content-Based Medical Image Retrieval and Intelligent Interactive Visual Browser for Medical Education, Research and Care, Diagnostics, Diagnostics. 11(8) p. 1470.

[13]. Yu, M. and Liu, X. 2011. Computer Image Content Retrieval considering K-Means Clustering Algorithm. *Mathematical Problems in Engineering*, Article ID 7914842.

[14]. Xu, Z., Li, L., Yan, M., Liu, J., Luo X., Grundy J., Zhang, Y. and Zhang, X., A. 2021. Comprehensive Comparative Study of Clustering-Based Unsupervised Defect Prediction Models. *The Journal of Systems & Software*.

[15]. Xu, Y., Hu, X., Wei, Y., Yang, Y. and Wang, D. 2019. A Machine Learning Dataset For Large-Scope High-Resolution Remote Sensing Image Interpretation Considering Landscape Spatial Heterogeneity. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W13, p731–736.

[16]. Campos, G. F. C., Mastelini, S. M., Aguiar, G. J., Mantovani, R. G., Melo, L. F., and Barbon, S. 2019 Machine Learning Hyperparameter Selection for Contrast Limited

Adaptive Histogram Equalization. *EURASIP Journal on Image and Video Processing* 59.

[17]. Sagheer, N. S. and Yousif, S. A. Canopy with K-Means Clustering Algorithm for Big Data Analytics. *AIP Conference Proceedings* 2334, 070006,

[18]. Ghazal, T. M., Hussain, M. Z., Said, R. A., Nadeem, A., Hasan, M. K., Ahmad, M., Khan, M. A., and Naseem, M. T. 2021. Performances of K-Means Clustering Algorithm with Different Distance Metrics. *IASC*, 30(2) 2021.

[19]. Faizan, M., Zuhairi, M. F., Ismail, S. and Sultan, S. 2020. Applications of Clustering Techniques in Data Mining: A Comparative Study. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 11(12).

[20]. Kasliwal, N. N., Lade, S., and Prabhune, S. S. 2012. Introduction of Clustering by using K-means Methodology. *International Journal of Engineering Research & Technology (IJERT)*, 1(10).

[21]. Liu, R. 2022. Data Analysis of Educational Evaluation Using K-Means Clustering Method. *Computational Intelligence and Neuroscience*.