Journal of Engineering Technology
Volume 6, Issue 1, Jan. 2018, PP. 548-558

RESEARCH ARTICLE -ENGINEERNG TECHNOLOGY

Image clustering of complex balinese character with DBSCAN algorithm

Oka Sudana[1], Darma Putra[2], Made Sudarma[3], Rukmi Sari Hartati[4], Ayu Wirdiani[5]

[1]Udayana University, Bali, Indonesia

Abstract: Introduction, The Balinese writing is unique in its almost identical form, and some writings are distinguished by a single line stroke. The Balinese writing is complex, in the form of a combination of several characters in a syllable group with the position of surrounding the main character in Balinese script called the sound usage. The main characters generally have a combination of follower characters in front, behind, above and / or under the main characters; even in each position can contain combinations of more than one character, thus forming a model that is much more complex than the Latin script. **Methodology**: DBSCAN (Density-Based Spatial Clustering of Application with Noise) algorithm is suitable for clustering process. DBSCAN has an algorithm that builds high-density areas into clusters and finds clusters of any kind in a spatial database containing noise inside. The clustering process is as an early stage in Balinese Optical Character recognition (OCR) System on Kakawin Books into poetry in Latin Letters. Trials using a sample image of Balinese script are taken from Kakawin Ramayana Book. The process begins with binary, followed by cropping automatically to get rows per line of writing. After that they are processed with Clustering Process to get the character objects. Variations in minimum point value (*minpts*) and epsilon (*eps*) values. **The results** obtained by DBSCAN Algorithm with the minimum value of points 2 and 3, epsilon = sqrt (2) and sqrt (3) succeeded in clustering with error percentage below 3%. **Conclusion:** DBSCAN algorithm is very good for conducting Complex Balinese Handwriting Process.

Keywords: clustering, Balinese Character, DBSCAN, OCR, Kakawin Ramayana.

1. Introduction

The Balinese writing is unique in its almost identical form, and some writings are only distinguished by a single line streak [1]. The Balinese writing is complex, in the form of a combination of several characters in a syllable group with the position of surrounding the main characters in Balinese script called the sound usage. The main character generally has a combination of follower characters in front, behind, above and / or under the main character, even in each position can contain combinations of more than one character, thus forming a model that is much more complex than Latin writing.

Image segmentation is a technique to divide an image into several parts (regions) where each region has similarity attributes. One of the techniques in image segmentation is by clustering. Cluster-based image segmentation uses multidimensional data to group image pixels into multiple clusters. Generally, those pixels are clustered based on pixel distance proximity [2]. DBSCAN (Density-Based Spatial Clustering of Application with Noise) algorithm is suitable for clustering process. DBSCAN is an algorithm that builds high-density areas into clusters and finds clusters of any kind in a spatial database containing noise inside. The clustering process as an early stage in Balinese Optical Character Recognition (OCR) System on kakawin

Books into poetry in Latin Letters. Trials using a sample image of Balinese script are taken from Kakawin Ramayana Book.

Much research has been done in the field of character recognition along with the stages of the process. A variety of methods are used in each of these stages, especially in the reprocessing and Segmentation stages. The Zone-based approach is used to classify and recognize the Telugu Handwriting Character (South India) [3]. Research on the Cielab Method and Projection Profile is used to perform the image segmentation of Balinese script on Lontar. The thresholding process with Local Adaptive Thresholding and Thinning Processes using the Zhang-Suen Method is used in the reprocessing stage [4]. The research uses Fuzzy C-Means clustering method with object in the form of Korean and British Board Image. The research was conducted to detect the text area in brand image using edge profile technique and continued with segmentation process using Fuzzy C-Means clustering technique which resulted in clustering technique Fuzzy C-Means able to cluster each region that has been distinguished into text area (black) and background area (white) [5]. Research based on LDA Method can classify Balinese Script on papyrus with good accuracy. LDA is able to distinguish each character of Balinese Script even if just a little different [6]. Research using Fuzzy C-Means clustering with an object such as the image of Korean and English brand boards. The study was conducted to detect the text area on the image of the brand board using the profile edge techniques and is continued to the process of segmentation using Fuzzy C-Means clustering technique that produces Fuzzy C-Means clustering techniques capable clustering each region which has been divided into a text area and the background area [7]. Research based on word segmentation method for handwritten Korean text lines. It uses gap information to separate a text line into word units, where the gap is defined as a white-run obtained after a vertical projection of the line image. Each gap is classified into a between-word gap or a within-word gap using a clustering technique. Its take up three gap metrics - the bounding box (BB), run-length/Euclidean (RLE) and convex hull (CH) distances - which are known to have superior performance in Roman-style word segmentation, and three clustering techniques - the average linkage method, the modified MAX method and sequential clustering. An experiment with 498 text-line images extracted from live mail pieces has shown that the best performance is obtained by the sequential clustering technique using all three gap metrics [8].

Similarly, the use of DBSCAN Algorithm for clustering has been widely used in various fields including for clustering on the image. BSCAN is specifically reviewed and compared to its use with other algorithms for various problem areas, providing excellent results [9]. Research on the DBSCAN algorithm works well enough to perform color image segmentation that contains a lot of noise. [10] The DBSCAN algorithm combined with Morphological Operator is used to segment the image of prostate cancer, which works quite well by detecting and isolating the affected image area prostate cancer [11]. The DBSCAN algorithm is also used to detect blood vessels in the retina combined with the Back Propagation Neural Network. Applications were made to detect the features of Diabetic Retinopathy disease in the patient's retina. The result of this research is 2 that is accuracy based on wound and accuracy based on the picture. The accuracy of disease detection in 20 images yields 100% sensitivity and accuracy results from disease-based disease detection with 92% sensitivity [12]. The combination of DBSCAN with Fuzzy Classifier for Diabetic Retinopathy image clustering result with 90% accuracy [13]. Development of DBSCAN method from a traditional model which is called VDBSCAN that varies eps value simultaneously in accordance with *k-dist* plot shows very effective results used for high or uneven data density variations [14]. The DBSCAN algorithm also provides good results then used to segment the lines of visual sensors in various automated parking assistant systems [15]. The DBSCAN method has also become one of the recommended alternative methods of reconstruction and CR of literary works using Devnagari Letters [16]. Devnagari Letters have similarities with the model of Balinese Character, so it is expected that the use of DBSCAN on clustering image Balinese script also gives good results.

Journal of Engineering Technology
Volume 6, Issue 1, Jan. 2018, PP. 548-558
RESEARCH ARTICLE -ENGINEERNG TECHNOLOGY

2. Research Method

Clustering process conducted in this system aims to get the clusters of Balinese script formers per line from the image input image scanning Kakawin Book. Clustering process is a preparatory step towards the feature extraction stage in Optical Character Recognition System (OCR), the process begins with scanning the page Book Kakawin Ramayana. Results scanned with the file format. Bmp, jpg or other image formats. This file is inputted to the system, for Pre-Processing through Binary Process which generates binary image (black and white). Then it followed by Cropping Process automatically to get lines of script. After that go into Clustering Process by using DBSCAN Algorithm, so get object characters that become input on Extraction Process Feature for OCR System Script Balinese to Latin script. The general stages performed in this system are as seen in System Overview in Figure 1.

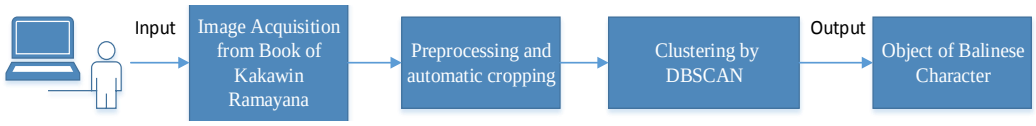

Figure. 1. System Overview

2.1 Image Acquisition

The data samples used as test materials are Kakawin Ramayana and Mahabharata Books, which use Balinese script in Old Javanese. Kakawin Book is scanned by page with format .bmp or .jpg. A partial example of the image input of Kakawin Ramayana Books page 4 is shown in the following Figure 2:

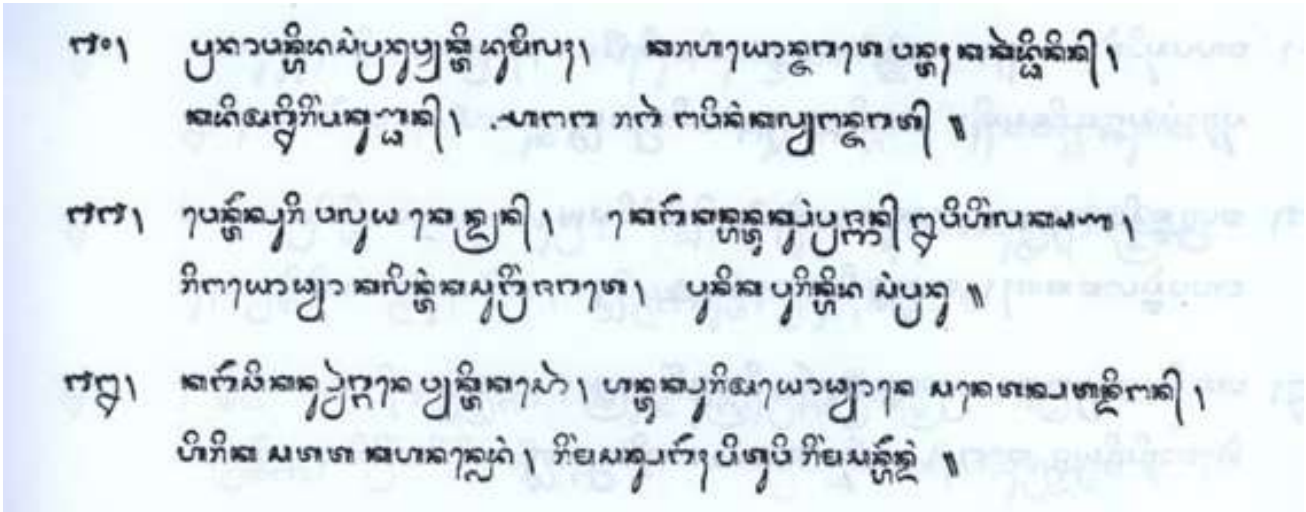

Figure 2. Scanned page from Ramayana Kakawin Books.

2.2. Preprocessing and Automatic Cropping

Preprocessing begins with the process of converting the input image into a binary image, either from the form of a color image or gray scale image. After that cropping process is done to get the group of rhyme lines from the input image. The basic principle of this cropping process is to read each line of image from scratch, then

calculate the total black pixels, if the total black pixel is more than the noise tolerance constant, then the line is grouped as the beginning of the line containing the letters. This step is repeated until the line gets a total black pixel less than noise tolerance. The beginning of the line and the end of the line are recorded for cropping the line group of characters. Avoiding noise so as not to be regarded as a line groups, used at least a Balinese script as a control. If the character group size is less than the minimum value of the average character of Balinese script, then it is decided not as a line of poetry. The process is repeated up to the end of the input image line.

Some of the functions or tools available in MATLAB are used to process images, including functions for conversion to binary images, cropping images and others. The detail algorithm used in the Cropping Process line from kakawin poem is as follows:

```
function [pot_row] ← crop_citra(image_source)
    set min_height_char;
    set toleransi_noise;
    image1 ← read_image(image_asal);
    read height_image (image1);
    read width_image (image1);
    sizeblok ← [height_image width_image];
    fungsi ← @(block_struct)im2bw(block_struct.data,
         graythresh(block_struct.data));
    BW1 ← blockproc(image1,sizeblok,fungsi);
    nrow ← 0; jum_row ← 0;
    while nrow <= height_image
        image_ ← imcrop(BW1,[0,nrow,width_image,0]);
        jum_black ← width_image - sum(image_);
        if jum_black >= toleransi_noise
            first_row ← nrow;
            height_crop ← 0;
            while ((jum_black >= toleransi_noise) |
                       (height_crop <= min_height_char))
                nrow++; height_crop++;
                image_ ← imcrop(BW1,[0,nrow,width_image,0]);
                jum_black ← width_image - sum(image_);
            end
            end_row ← nrow;
            jum_row++;
            if height_crop > min_height_char
                file_name_result ←
                   strcat('baris_aksara',int2str(jum_row),'.jpg');
                image_crop ←
                     imcrop(BW1,[0,first_row,width_image,height_crop]);
                imwrite (image_crop,file_name_result);
```

```
            end
        end
        nrow++;
        pot_row ← jum_row;
end
```

2.3. Clustering

Clustering process begins by finding the position of black pixels (the points of composing Balinese character objects) from the previous cropping image. Furthermore, clustering process using DBSCAN Algorithm with variation of minimum point value (minpts) and epsilon (eps) is applied; the result is in the form of Balinese characters that form Kakawin poem from the input image. The DBSCAN method places high density areas that are separated from each other by low density areas. There are two input parameters to form clusters ie Eps and MinPts. Eps gives a one cluster radius and MinPts gives a minimum number of points, this is for spatial datasets and objects in two-dimensional space [2] [10].
The core and border in DBSCAN Concept (Eps = 1 and Minpts = 6) is as shown in Figure 3.

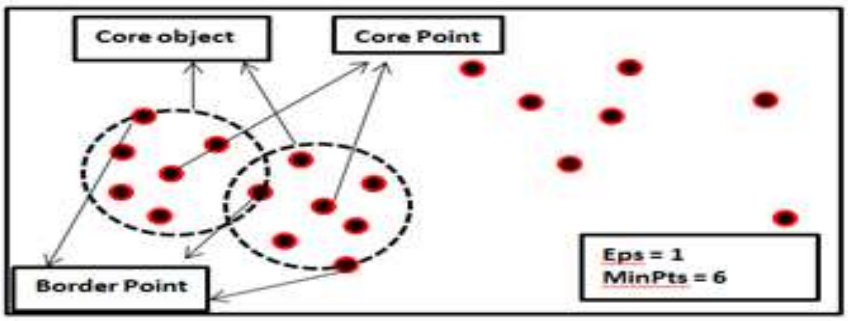

Figure 3. Core and Border in DBSCAN Concept.

The algorithm sequence of DBSCAN generally has 6 steps:

1. Select the initial *p* point randomly.
2. Take all points that are density reachable to point *p*.
3. If *p* is the core point, then the cluster is formed.
4. If *p* is a border point, nothing is a density-reachable relationship of *p* and DBSCAN will visit the next point of the database.
5. Continue the process until all the points have been processed. The results obtained do not depend on the sequence of process points taken.

The detailed algorithm of this Clustering Process is as follows:

```
Set nama filecitra;
Set warna ← 0;
Set Epsilon (eps); Set Minimal Point (minpts);
count_sloka ← crop_citra(filecitra);
for t ← 1:count_sloka
    set nama_file_siap = strcat('baris_aksara',int2str(t),'.jpg');
```

Journal of Engineering Technology
Volume 6, Issue 1, Jan. 2018, PP. 548-558
RESEARCH ARTICLE -ENGINEERNG TECHNOLOGY

```
    BW1 ← imread(nama_file_siap);
    [pos_xy] ← cari_posisi_warna(BW1,warna);
    [idx, isnoise] ← Oka_DBSCAN(pos_xy, eps, minpts);
    a ← max(idx); s ← size(BW1);
    cluster_result ← zeros(s(1),s(2));
    set matrix (cluster_result) with color_white;
    temp ← size(idx);
    for row ← 1:temp(1)
        row_ ← pos_xy(row,1);
        col_ ← pos_xy(row,2);
        cluster_result (row_,col_) ← idx(row,1);
    end
   imshow(cluster_result,colorcube);
end
```

3. Result and Analysis

3.1. Automatic Cropping

An example of the result of automatic cropping for images in Figure 2 is as you see in the following figure 4. It is shown that the system is very well able to do cropping to get poetry line.

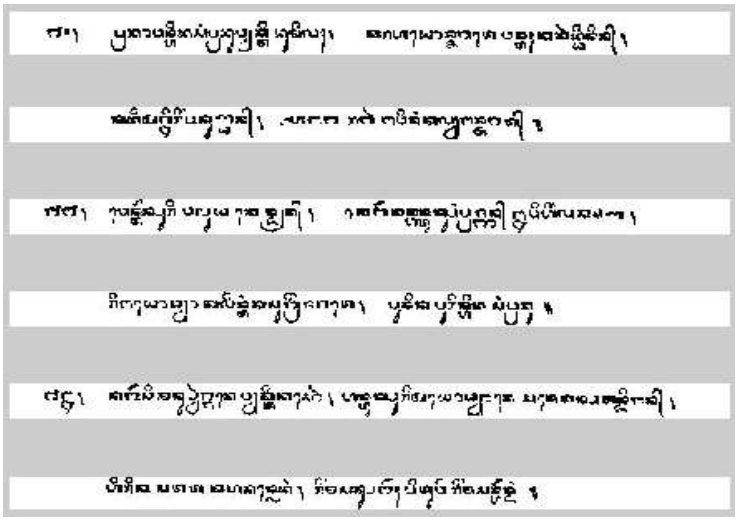

Figure 4. Automatic cropping results example.

Based on the results of the test for the cropping process of 22 pages Kakawin Ramayana Book's scanning result, each page consisting of 20 lines of poetry, 100% success to get the lines that should be obtained, which are as many as 440 lines of poetry. This method was also successfully used for cropping in other studies [17,18,19].

Journal of Engineering Technology
Volume 6, Issue 1, Jan. 2018, PP. 548-558
RESEARCH ARTICLE -ENGINEERNG TECHNOLOGY

3.2. Clustering with DBSCAN

The result of the clustering process succeeded in obtaining the cluster of Balinese Character object objects, which are represented by different colors as seen in Figure 5 and Figure 6. The combination of colors representing the image object represents the cluster index stored in the *idx* variable.

Figure 5. Clustering example result in one verse poetry line.

Figure 6. Clustering example result in some verse poetry lines.

The test is done by using sample data from scanning the pages on Kakawin Ramayana Book. Variations in the number of characters in the line characters are from 20 to 60 characters. Minimal Point Variation range (*minpts*) is from 1 to 11 and variations in Epsilon Value (*eps*) from 1 to 5, with detailed focus on sqrt (2) and sqrt (3). The example of a test result for aksara1.jpg file with 20-character forming objects is shown in Table 1. The cells in a table marked with green indicate the number of clusters generated in the combination of *minpts* 2 and 3, and eps sqrt (2) and sqrt (3) ie 20 objects in accordance with many objects should, so the percentage of success in this trial is 100%.

Journal of Engineering Technology
Volume 6, Issue 1, Jan. 2018, PP. 548-558
RESEARCH ARTICLE -ENGINEERNG TECHNOLOGY

Table 1.a. Clustering test result from file aksara1.jpg with 20 characters' objects.

Total Object	**20**	**Minpts**										
Eps		1	2	3	4	5	6	7	8	9	10	11
	1	23	22	22	96	176	0	0	0	0	0	0
	sqrt(2)	20	**20**	**20**	31	38	42	72	68	71	0	0
	sqrt(3)	20	**20**	**20**	31	38	42	72	68	71	0	0
	2	19	19	19	19	19	22	28	37	59	67	60
	3	14	14	14	14	14	15	15	15	17	18	22
	4	11	11	11	11	11	11	11	11	11	12	13
	5	11	11	11	11	11	11	11	11	11	11	11

Table 1.b. Procentage clustering test result from Table 1.a.

Total Object	**20**	**Minpts**										
Eps		1	2	3	4	5	6	7	8	9	10	11
	1	15%	10%	10%	380%	780%	100%	100%	100%	100%	100%	100%
	sqrt(2)	0%	**0%**	**0%**	55%	90%	110%	260%	240%	255%	100%	100%
	sqrt(3)	0%	**0%**	**0%**	55%	90%	110%	260%	240%	255%	100%	100%
	2	5%	5%	5%	5%	5%	10%	40%	85%	195%	235%	200%
	3	30%	30%	30%	30%	30%	25%	25%	25%	15%	10%	10%
	4	45%	45%	45%	45%	45%	45%	45%	45%	45%	40%	35%
	5	45%	45%	45%	45%	45%	45%	45%	45%	45%	45%	45%

An example of a test result for a 36 characters aksara4.jpg file is like Table 2. The cells in a table marked with green indicate the number of clusters generated in the combination of *minpts* 2 and 3, and *eps* sqrt (2) and sqrt (3) are 36 objects in accordance with many objects should, so the percentage of success in this trial is 100%.

Table 2.a. Clustering test result from aksara1.jpg file with 36 characters' objects.

Total Object	**36**	**Minpts**										
Eps		1	2	3	4	5	6	7	8	9	10	11
	1	39	37	37	102	132	0	0	0	0	0	0
	sqrt(2)	36	**36**	**36**	40	49	62	61	60	37	0	0
	sqrt(3)	36	**36**	**36**	41	50	63	62	61	38	0	0
	2	30	30	30	30	30	39	51	58	51	47	37
	3	10	10	10	10	10	10	12	12	14	20	24
	4	8	8	8	8	8	8	8	8	8	8	9
	5	5	5	5	5	5	5	5	5	5	5	5

Journal of Engineering Technology
Volume 6, Issue 1, Jan. 2018, PP. 548-558
RESEARCH ARTICLE -ENGINEERNG TECHNOLOGY

Table 2.b. Procentage clustering test result from Table 2.a.

Total Object	20	Minpts										
Eps		1	2	3	4	5	6	7	8	9	10	11
	1	8%	3%	3%	183%	267%	100%	100%	100%	100%	100%	100%
	sqrt(2)	0%	**0%**	**0%**	11%	36%	72%	69%	67%	3%	100%	100%
	sqrt(3)	0%	**0%**	**0%**	14%	39%	75%	72%	69%	6%	100%	100%
	2	17%	17%	17%	17%	11%	8%	42%	61%	42%	31%	3%
	3	72%	72%	72%	72%	72%	72%	67%	67%	61%	44%	33%
	4	78%	78%	78%	78%	78%	78%	78%	78%	78%	78%	75%
	5	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%

The average error gained from the overall test is as shown in Table 3 below:

Table 3. The Performance of Clustering with DBSCAN.

Total Object	1865	Minpts										
Eps		1	2	3	4	5	6	7	8	9	10	11
	1	8.5%	5.9%	6.1%	188.1%	325.5%	97.5%	98.7%	96.4%	97.8%	98.0%	97.6%
	sqrt(2)	3.6%	**2.4%**	**2.3%**	17.9%	35.5%	56.1%	91.6%	81.5%	46.0%	84.5%	85.2%
	sqrt(3)	3.7%	**2.5%**	**2.3%**	18.4%	35.9%	57.2%	92.5%	82.1%	46.4%	84.1%	85.0%
	2	26.7%	27.1%	29.2%	29.0%	21.4%	20.4%	23.5%	44.4%	58.4%	62.5%	34.9%
	3	60.9%	63.0%	61.5%	61.0%	61.6%	60.8%	58.3%	56.9%	52.4%	43.3%	35.3%
	4	63.9%	63.9%	63.9%	63.9%	63.9%	63.9%	63.9%	63.9%	62.7%	61.7%	60.6%
	5	65.9%	65.9%	65.9%	65.9%	65.9%	65.9%	65.9%	65.9%	65.9%	65.9%	65.9%

Table 3 shows the average performance of the system in clustering the image taken from the scanning of Kakawin Ramayana Book, which are 1865 objects (in 44 lines) as many object properly. The system gives the best performance at the minimum value of points 2 to 3 and epsilon sqrt (2) and sqrt (3) ie the average error of 2.3% to 2.6%, or the performance of 97.6% success, and the sharply increasing tendency of error occur starting from *minpts* 4, 5 and so on. Similarly, for the epsilon value from 2, 3 and so on, the errors that occur increase sharply. The result of this experiment is almost in the same success rate as compared with the use of Text Region Segmentation Method, with the success of segmentation of 97.48% [18]. This clustering result is better than using Fuzzy C-Means Algorithm (FCM) which the failure percentage is 13% [7]. Latin Writing segmentation has a success rate close to 100% by looking for a separator between characters in the writings [19].

4. Conclusion

Based on the experiments that have been done, it can be concluded that DBSCAN algorithm is very suitable for clustering process in Bali Complex Image. The results obtained by DBSCAN Algorithm with the minimum value of points 2 and 3, epsilon = sqrt (2) and sqrt (3) succeeded in clustering with error percentage below 3%. The process begins with binary, followed by cropping automatically to get rows per line of writing. After that it is continued by Clustering Process to get the character objects. The clustering process is as an early stage in Balinese Optical Character Recognition (OCR) System on Kakawin Books into poetry in Latin Letters. The DBSCAN algorithm is very well used for clustering for the Balinese Letter image model such as handwriting, signature, medical image and others

References

[1] Agung B.W., I Gede Rudy Hermanto, Retno Novi. "Pengenalan Huruf Bali dengan menggunakan Metode Modified Direction Feature (MDF) dan Learning Vector Quantization (LVQ)". "in Konferensi Nasional Sistem dan Informatika 2009", Institut Teknologi Bandung.

[2] Darma Putra. "Pengolahan Citra Digital". Andi Offset. 2010.

[3] N. Shobha Rani, Sanjay Kumar Verma, Anitta Joseph. "A Zone Based Approach for Classification and Recognition of Telugu Handwritten Characters". International Journal of Electrical and Computer Engineering (IJECE). 2016; 6(4): 647-653

[4] Sutramiani Ni Putu, Darma Putra I Ketut Gede, Sudarma Made. "Local Adaptive Thresholding pada Preprocessing Citra Lontar Aksara Bali". Jurnal Teknik Elektro. 2015;14(1):27-30.

[5] Made Sudarma. "Identifiying of the Cielab Space Color for the Balinese Papyrus Characters". TELKOMNIKA Indonesian Journal of Electrical Engineering. 2015; 13(2): 321-328.

[6] Made Sudarma, Sri Ariyani, Manuh Artana. "Balinese Script's Character, Reconstruction using Linear Discriminant Analysis". Indonesian Journal of Electrical Engineering and Computer Science. 2016; 4(2):479-485.

[7] Jonghyun Park, Toan Nguyen Dinh, and Gueesang Lee. "Binarization of Text Region based on Fuzzy Clustering and Histogram Distribution in Signboards". "in World Academy of Science, Engineering and Technology". 2008; 43:85-90.

[8] S.H. Kim, S.Jeong, Guee-Sang Lee, C.Y. Suen. "Word Segmentation in Handwritten Korean Text Lines based on Gap Clustering Techniques", "Six International Conference on Document Analysis and Recognition 2001", IEEE Xplore August 2002.

[9] Xianjin Shi, Wanwan Wang, Chongsheng Zhang. "An Empirical Comparison of Latest Data Clustering Algorithms with State-of-the-Art", "Indonesian Journal of Electrical Engineering and Computer Science". 2017; 5(2): 410-415.

[10] Atrayee Dhua, Debjani Nath Sarma, Sneha Singh, Bijoyeta Roy. "Segmentation of Images using Density-Based Algorithm", "International Journal of Advanced Research in Computer and Communication Engineering". 2015; 4(5).

[11] R. Manavalan, K. Thangavel. "TRUS Image Segmentation using Morphological Operators and DBSCAN Clustering", "in World Congress on Information and Communication Technology (WICT)". IEEE Xplore January 2014.

[12] Shantala Giraddi, Jagadeesh Pujari, Shraddha Giraddi. "Exudates Detection with DBSCAN Clustering and Back Propagation Neural Network", "International Journal of Computer Applications". 2014; 86(19): 16-20.

[13] Shantala Giraddi, Jagadeesh Pujari. "Automated Detection of Exudates using DBSCAN Clustering and Fuzzy Classifier", "International Journal of Advanced Research in Computer Science". 2012; 3(7).

[14] Peng Liu, Dong Zhou, Naijun Wu. "VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise". "in International Conference on Service Systems and Service Management 2007". IEEE Xplore, July 2007.

[15] Soomok Lie, et al. "Directional-DBSCAN: Parking-slot detection using a clustering method in around-view monitoring system". "in Entelligent Vehicles Sysposium (IV), IEEE 2016". IEEE Xplore August 2016

[16] Kunal Ravindra Shah and Dipak Dattatray Badgujar. "Devnagari handwritten character recognition (DHCR) for ancient documents: A review", "in IEEE Conference on Information & Communication Technology (ICT) 2013". IEEE Xplore, July 2013.

[17] Azadeh Nazemi, Iain Murray, and David A Mc Meekin. "Practical Segmentation Methods for Logical and Geometric Layout Analysis to Improve Scanned PDF Accessibility to Vision Impaired", "International Journal of Signal Processing, Image Processing and Pattern Recognition". 2014; 7(4): 23-35.

[18] Ayatullah Faruk Mollah, Nabamita Majumder, Subhadip Basu and Mita Nasipuri. "Design of Optical Character Recognition System for Camera-based Handheld Devices", "IJCSI International Journal of Computer Science Issues". 2011; 8(4,1): 283-289.

[19] Namrata Dave. "Segmentation Methods for Handwritten Character Recognition", "International Journal of Signal Processing, Image Processing and Pattern Recognition". 2015; 8(4): 155-164.