Signature Similarity Search Using Cluster Image Retrieval

Pandapotan Siagian[1], Herry Mulyono[2], Erick Fernando[3]
Department of Computer System in STIKOM DB of Jambi
Jl. Sudirman No. 1 Thehok Jambi, Indonesia
[1] siagian.p@gmail.com
[2] herry_mulyono@yahoo.com
3 erick_88@yahoo.com

***Abstract*— Significant image search in the database of signatures on previous research is still 65%, because the data stored that result capture camera [13, 14,]. In this paper, we do the retrieval pattern signatures automatically, the algorithm that is used to find data points based on contributions. Similarity search algorithms will optimize the intra cluster and claster signatures in 3 classes i.e., Gx, Gy, gt. image capture-based content, which can calculate the similarity between the shape and texture of the images and do a grouping of pictures with minimum Euclidean distance consideration. CBIR is a set of techniques for taking pictures of semantically relevant than just image database based on recommended sources of images automatically. Performance evaluation method is now done with precision and recall for a different database. Response time to find most of the signatures from 3 grade database, giving the effect of a 78% accuracy rate higher.**

***Keywords*— signature digital, k-mean clustering; calculate similarity; CBIR image; gradient magnitude Prewitt**

I. INTRODUCTION

Results of the capture image with a digital camera and saves the data into the database signature, the decline of the quality of the data. So in search of less accurate in finding the Indigo difference of the two pieces of the same image. Search in the database signature significantly on previous research is still 65% [14]. This research, we do two stages of uptake patterns automatically and the Signature Algorithm used to cluster the search based on the contribution of the data point. Content-based image retrieval of k-mean clustering algorithm. Similarity search algorithms will optimize the intra-cluster and claster of signatures in 3 classes i.e., Gx, Gy, gt. search system, content-based image retrieval, which is able to calculate the similarity between the shape and texture of the image.

Significant image search in the database of signatures on previous research is still 65%, because the data stored that result capture camera. In this paper, we do the retrieval signature patterns automatically and cluster Algorithm used to search based on contribution data points. Content-based image retrieval of k-mean clustering algorithm. Similarity search algorithms will optimize the intra-cluster and claster of signatures in 3 classes i.e., Gx, Gy, Gt. Search system, content-based image retrieval, which is able to calculate the similarity between the shape and texture of the image. The K Means Clustering Algorithm is then used to cluster the group of images based on feature vector of images by considering the minimum Euclidean distance.CBIR is a set of techniques for taking pictures of semantically relevant than just image database based on the features of the source image automatically.

One of the main tasks for the system of CBIR is counting in common, compare, extracting the signature feature of each image based on pixel values and defines rules for comparing images. This feature becomes the representation of images to measure the similarity with the other images in the database. The performance evaluation of the present method is done by Precision and Recall for different databases. Compute the difference compared with the image features components for other image descriptors.

II. LITERATURE REVIEW

A Signature Files

Faloutsos & Christodoulakis (1987), A signature file is a file that stores a signature record for each Image in the database. Each signature has a fixed size of b bits representing terms. A simple encoding scheme goes as follows. Each bit of a document signature is initialized to 0. A bit is set to 1 if the term it represents appears in the document[1,11]. A signature s1 matches another signature s2 if each bit that is set in signature s2 is also set in s1. Since there are usually more terms than available bits, there may be multiple terms mapped into the same bit. Such multiple-to-one mappings make the search expensive since a image that matches the signature of a query does not necessarily contain the set of keywords of the query. Improvements can be made by firs by first performed frequency analysis, stemmed, and by filtering stop words, and then use a hashing technique and superimposed coding technique to encode the list of terms into bit representation [15]. Nevertheless, the problem of multiple-to-one mappings still exists, which is the major disadvantage of this approach.

B Digital Signature

In this paper, paper-based authentication implementation document presented. The integrity of the message text and the author of the document can be verified by using a digital signature and QR codes. The proposed Methodology can be automated or semi-automated. It's a semi-automatic when OCR is inaccurate and requires the user to visually compare text messages on paper and obtained from the QR code; however, this method provides convenience for users dealing with large amounts of data [1, 10].

C Content Based Image Retrieval (CBIR)

The images are very rich in the content such as in colour, texture, and shape information which are presented in them. Retrieving images based on colour similarity is achieved by

computing a colour histogram for each image that identifies the proportion of pixels within an image holding specific values (that humans express as colors) [4,13].

Colour searches will usually involve in comparing the colour of histograms, though this is not the only technique in practice. Texture measures look for visual patterns in images and how they are spatially defined. The identification of specific textures in an image is achieved primarily by modelling texture as a two-dimensional grey level variation [4,14]. The relative level brightness of pairs of pixels are computed such in the degree of contrast, regularity, coarseness and directionality that may be estimated. The shape does not refer to the shape of an image, but to the shape of the particular region that is being sought out.

We will make shapes often be determined firstly by applying segmentation with method Prewitt gradient magnitude edge detection to an image. In our cases, the accurate shape detection with method Prewitt gradient magnitude edge detection will require human intervention because methods like segmentation are very difficult to automate complete. Here are some discussions about shape extractions using gradient magnitude edge detection masks, like in Prewitt gradient operators.

1) Shape Feature

Shape is the most important and most powerful feature used for image classification, indexing and retrievals. Shape information extracted using histogram for edge detection. The edge information in the image is obtained by using the Prewitt edge detection [13,14].

In shape, we will segmentation of two Prewitt gradient magnitude edge detection are two images which at each point contain the horizontal and vertical derivative approximation techniques. Shapes representations can be generally divided into two categories, they are Boundary based and Region-based, see Fig. 1.

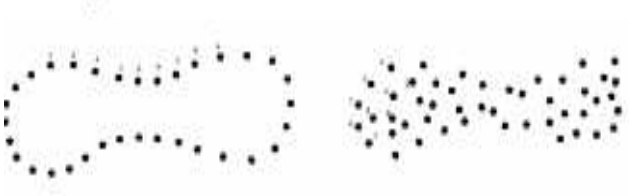


Fig. 1 Region based Images

2) Prewitt Edge Detection Technique

The Prewitt operator performs spatial gradient magnitude measurement in an image. The applying convolution K to pixel group p can be represented as [13,14]:

$$N(x,y) = \sum_{k=-1}^{1} \sum_{j=-1}^{1} K(j,k)\, p(x-j, y-k) \tag{1}$$

The Prewitt Edge Detector uses two convolution kernels, one is to detect changes in vertical contrast (hx) and the other is to detect horizontal contrast (hy). Fig. 2. shows the Prewitt Edge Detector uses two convolution kernels, one is to detect changes in vertical contrast (hx) and the other is to detect horizontal contrast (hy).

$$h_x = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \qquad h_y = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

Fig. 2. Detect vertical contrast (hx) and Detect horizontal contrast (hy)

Pandapotan Siagian (2013), these kernels are designed to respond maximal to the edges, running vertically and horizontally, relative to the pixel grid, one kernel for each of the two perpendicular orientations. The kernels can be applied separately to the input image, to produce separate measurements of the gradient component in each orientation (call these *Gx* and *Gy*) [13,14].

We are can combine altogether to find the magnitude of the gradient at each point and the orientation of that gradient [14]. Typically, the steps used to find the similarity of gradient magnitude at each point in an input it fingerprints and signatures can be seen as follow:

a) The image is fingerprint and signatures in format png with image size are 400 x 400 results.
b) A data fingerprint and signatures on the shapes often be determined firstly by applying segmentation with method prewitt gradient magnitude edge detection. This method uses two convolution kernels, one is to detect changes in vertical contrast (hx) and the other is to detect horizontal contrast (hy). is stored in the directory
c) Prewitt edge detector uses a pair of 3×3 convolution masks, one is to estimate the gradient in the x-direction (columns) and the other is to estimate the gradient in the y-direction (rows).
d) A convolution mask is usually much smaller than the actual image. As the result, the mask is slid over the image, manipulating a square of pixels at a time.
e) If we define A as the source image, and Gx and Gy are two images which at each point contain the horizontal and vertical derivative approximations, then the masks will be marked as follows :

-1	0	1	-1	-1	-1
-1	0	1	0	0	0
-1	0	1	1	1	1
	Gx			Gy	

The magnitude of the gradient is then calculated using the formula:

$$|G| = \sqrt{G^2 x + G^2 y} \tag{2}$$

Approximate magnitude can be calculated using:

$$|G| = |Gx| + |Gy| \tag{3}$$

The angle of the orientation of the edge (relative to the pixel grid) which is giving rise to the spatial gradient is given by: $\theta = \arctan(Gy / Gx)$, when,

$$G_x = \delta f/\delta x,\ Gy = \delta f/\delta y \tag{4}$$

C Performance Evaluation of CBIR Systems

R. Xu and D. Wunsch (2005), Content-based image retrieval system is first evaluated in terms of retrieval effectiveness. In order to evaluate effectiveness of retrieval systems, two well known metrics, precision and recall are used :

Precision = (the number of retrieved images that are relevant) / (The number of retrieved images)

Recall = (the number of retrieved images that are relevant)/ (The total number of relevant images).

For a query q, the data set of images in the database that are relevant to the query q is denoted as R(q), and the retrieval result of the query q is denoted as Q(q). The precision of the retrieval is defined as the fraction of the retrieved images that are indeed relevant for the query using :

$$pre = \frac{IQ(g)IR(q)I}{IR(q)I} \quad (5)$$

The recall is the fraction of relevant images that is returned by the query using :

$$\text{Re}-Call = \frac{IQ(g)IR(q)I}{IR(q)I} \quad (6)$$

Usually, a tradeoff must be made between these two measures since improving one will sacrifice the other. In typical retrieval systems, recall tends to increase as the number of retrieved items increases; while at the same time the precision is likely to decrease. In addition, selecting a relevant data set R(q) is much less stable due to various interpretations of the images. Further, when the number of relevant images is greater than the number of the retrieved images, recall is meaningless. As a result, precision and recall are only rough descriptions of the performance of the retrieval system.

D Contribution Based Clustering

J. Mac Queen (1967) and L. Kaufman (1990), Partitional clustering aims at partitioning a group of data points into disjoint clusters optimizing a specific criterion [5,7,8]. When the number of data points is large, a brute force. enumeration of all possible combinations would be computationally expensive. Instead, heuristic methods are applied to find the optimal partitioning. The most popular criterion function used for partisional clustering is the sum of squared error function given by

$$E = \sum_{j=1}^{k} \sum_{x \in Ci}^{x} (x - mi)^2 \quad (7)$$

where k is the number of clusters, Ci is the i^{th} cluster, x is a data point and mi is the centroid of the i^{th} cluster.

A cluster is a collection of data points that are similar to one another within the same cluster and dissimilar to data points in other clusters [2,3,6]. Clustering is a method of unsupervised classification, where data points are grouped into clusters based on their similarity. The goal of a clustering algorithm is to maximize the intra-cluster similarity and minimize the inter-cluster similarity. Clustering algorithms can be broadly classified into five types: 1. Partitional clustering, 2. Hierarchical clustering, 3. Density-based clustering, 4. Grid-based clustering and 5. Model-based clustering [5].

Partitional and hierarchical clustering are the most widely used forms of clustering. In partition clustering, the set of n data points are partitioned into k non-empty clusters, where $k \leq n$. In the case of hierarchical clustering, the data points are organized into a hierarchical structure, resulting in a binary tree or dendogram [16].

K- means Clustering Algorithm

Mac Queen (1967), K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem [6]. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grouping is done. At this point we need to re-calculate k new centroids as binary centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centred. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{x} \left\| x_i^{(j)} - Cj \right\|^2, \quad (8)$$

where $\left\| x_i^{(j)} - Cj \right\|^2$, is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre ,, is an indicator of the distance of the data points from their respective cluster centred. The algorithm is composed of the following steps :1) Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids, 2) Assign each object to the group that has the closest centred, 3) When all objects have been assigned, recalculate the positions of the K centroids, and 4) Repeat Steps 2 and 3 until the centroids no longer move[9].

It results in separation of the objects into groups from which the metric to be minimized can be calculated. Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster center. The k-means algorithm can be run multiple times to reduce this effect. K-means is a simple algorithm that has been adapted to many problem domains and a good candidate for extension to work with fuzzy feature vectors [17].

III IMPLEMENTATION

In our system, the digital Signature is obtained by the process of encode, decode web-based and use a paint pen. Any personal Data will be stored in three classes, namely class Gx, Gy, gt. application system perform a two-stage process i.e., the process of learning and classification. The input process of learning is the learning process with the input images stored in a database with three classes of data, results of pre-process by the method of gradient prewitt which aims to increase (i) feature extraction accuracy of the k-means algorithm (ii) an increase in the uptake of better accuracy, distance and similarity retrieval quick. This system is a web-based application. The system will calculate and display similarity 3 pictures at once, see in Fig. 3.

A. Learning Process

Learning process is the input of a collection of learning software image that has been known as the class label. Process code, decode signature is shown in Fig. 4 and the learning process system from administrator is shown in Fig. 5 and Fig. 6.

Learning Process input learned there step the process of digital data (fingerprint, signature), consist of:

a) The digital signature is obtained by the process of encode, decode web-based and use a paint pen.

b) Digitalize data in a data store in Portable Network Graphic (PNG) format with image size is 400 x 400 results.

c) Prewitt gradient magnitude uses two convolution kernels, one is to detect changes in vertical contrast (hx) and the other is to detect horizontal contrast (hy). The data is stored in the directory, the signatures and image results prewitt gradient magnitude.

d) Save the signatures data with a file name according to the name you have.

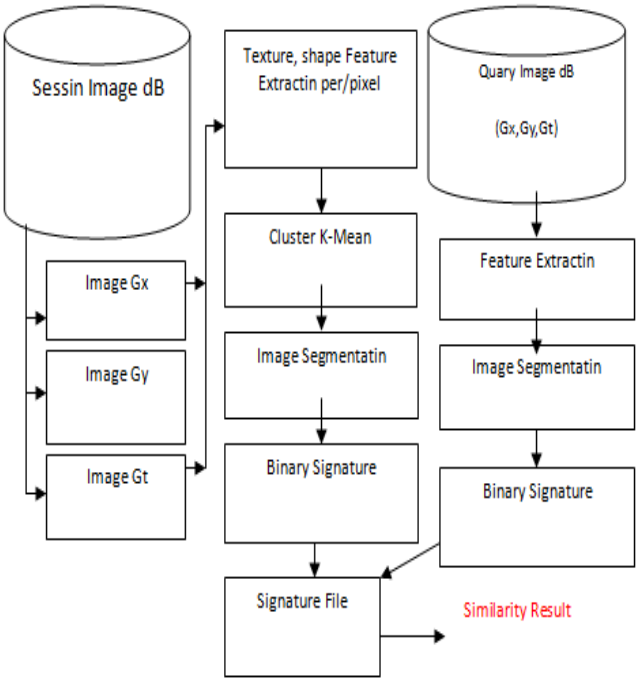


Fig. 3 System Overview

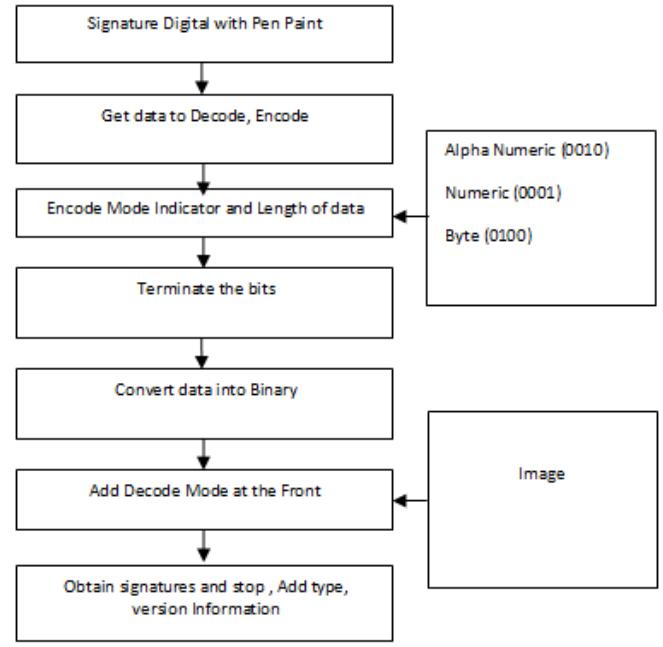


Fig. 4 Signature Decode Process

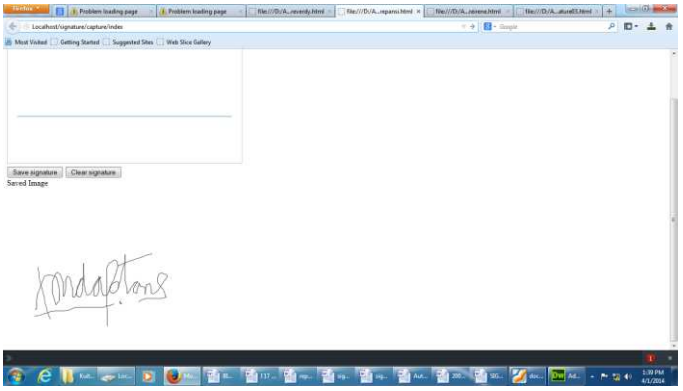


Fig. 5 Input Data Signature Digital

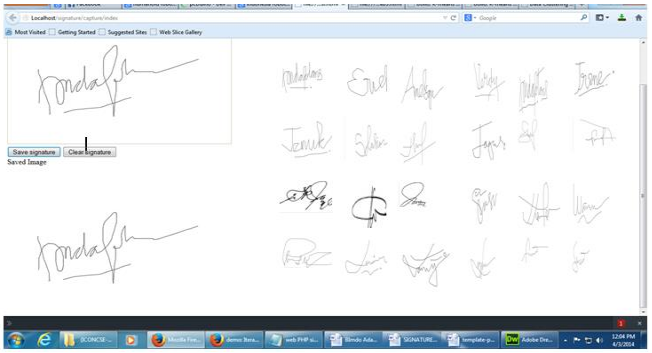


Fig. 6 Data Signature

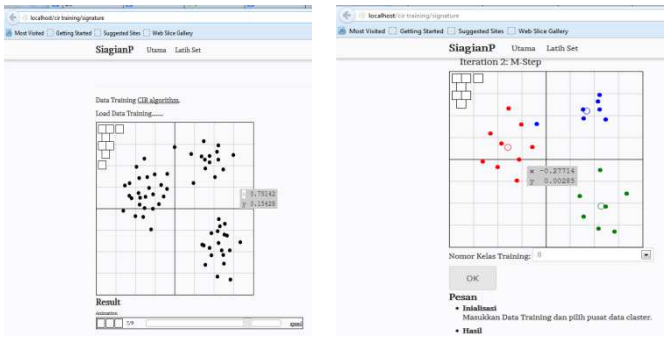
B Classification

The algorithm consists of the following steps: (1) where K points in space represented by any class i.e. Gx, Gy, Gt specific value points represent initial group centroids per

class, and (2) defines each object to the group that has the closest centroid, (3) when all is set, recalculate the position of the K centroids, (4) determine the distance of each object to the centroid. and (5) repeat step 2 and 3 until the centroids no longer move. Because the same data is displayed for ID and a signature that has the same pattern. The pattern shown is the minimum, maximum and average the digital signature of each individual.

IV. RESULTS AND DISCUSSION

For the proposed method, experiments for testing in an environment Windows XP and PHP web Platform. Experiments conducted by various databases for different sizes. Signature image uploads from sub-direectory., Gx, Gy, gt. consisting of 3 classes.



(a)Set Data Training (b) Result

Fig. 6 Training of k-mean Cluster

A *Clusters Of Min-Max*

The grouping feature of extrasi for 3 classes of data i.e. Gx, Gy, gt. Grouping is done for the separation of objects into groups that can calculate the metrics to be minimized. Though it can be proved that the procedure will always end, k-means algorithm does not always find the most optimal configuration functions, the purpose of the global minimum. This algorithm also significantly sensitive to initial cluster Center is chosen randomly. The test results contained in Table I.

TABLE I
CLUSTERING OF MIN-MAX

Class Image	K-Means cluster Signature	
	Min	Max
Gx	12.5%	Gx
Gy	13%	Gy
Gt	16%	Gt

B *Precision And Recall*

The success rate of the classification system for the same data in testing data in the database can be evaluated by using performance measures, re-call and precision. To measure the ability of a system to capture all relevant models, while the precision of measuring the ability of the system to take only the relevant model.

Below are the results of precision and recall for different databases. The K-means algorithm can be run multiple times to reduce this effect. K-means is a simple algorithm that has been adapted to many domain problem and a good candidate for the extension to work with vector gradient feature.

$$pre = \frac{Numberofre\ levantCIRs\ ignature}{Tota\ ln\ umberofCIR\ signature}$$

$$Re-call = \frac{Numberofre\ levantCIRs\ ignature}{Tota\ ln\ umberofCIR\ signature\ int\ heDatabase}$$

a. Database of 50 and number of Clusters varying from 2 to 24 :

TABLE II
PRECISION AND RECALL FOR DATABASE OF 2 TO 24

N0. Of Cluster	Precision	Re-call
2	100	100
4	87.89	92,44
8	83.76	75.29
10	80.28	70.78
12	78.67	74.39
14	71.38	71.25
16	73.78	73.69
18	74.21	74.10
20	75.67	75.60
22	74. 66	74.69
24	70.12	69.20

b. Database is varied from 30 to 100 and number of Clusters equal to 7:

TABLE III
PRECISION AND RECALL FOR DATABASE OF 30 TO 100

N0. Of Cluster	Precision	Re-call
30	74.73	100
40	73.73	72.89
60	73.17	74.67
70	72.54	72,45
80	71.67	71.39
80	71.38	71.25
100	73.78	73.69

c. Database of 200 and number of Clusters varied from 4 to 10

TABLE IV
PRECISION AND RECALL FOR DATABASE OF 4 TO 10

N0. Of Cluster	Precision	Re-call	N0. Of Cluster
4	74.73	100	4
5	73.73	72.89	5
6	73.17	74.67	6
7	74.54	74,45	7
8	74.67	75.39	8
9	76.38	75.25	9
10	63.71	63.69	10

V. CONCLUSIONS

In this paper, we have the algorithm how K is combined with a combination of K means algorithm with Prewitt Filter. integrated clustering algorithm for image classification is tested with a different image that image signatures with a different gradient.

We found that then performs well compared to before. The algorithm is robust and highly effective in producing the desired classification in particular in the field of pattern recognition as a fascinating area as shown by the results of the experiment.

Different neural network algorithm in the future can be used to classify images of signatures on gradient prewitt.

REFERENCES

[1] Digital Signature Standard (DSS), FIPS PUB 186-3,2013.

[2] E. Forgy, Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications , Biometrics, vol. 21, pp. 768–780, 1965.

[3] Höppner, F. Klawonn, and R. Kruse, Fuzzy cluster analysis:

[4] Methods for classification, data an alysis, and image recognition, New York, Wiley, 1999.

[5] John Eakins and Margaret Graham, Content Based Image Retrieval. Chatpter 5.6, pg 36-40, University of Northrumbia at New Castle, October 1999.

[6] J. Han and K. Micheline, "Data mining concepts and techniques," Morgan Kauffman, 2006.

[7] J. MacQueen, Some methods for classification and analysis of multivariate observations, in Proc . 5th Berkeley Symp., vol. 1, pp.281–297, 1967.

[8] L. Hong, A. Jain & S. Pankanti. 1999. *Can Multibiometrics Improve performance*, Proceedings of AutoID 99, pp. 59-64.

[9] L. Kaufman and P. Rousseeuw, Finding groups in data: An introduction to cluster analysis, Wiley, 1990.

[10] Minakshi Banerjeea, MalayK. Kundua,b, Pradipta Majia,b, Content-based image retrieval Using visually significant point features,Elsevier, Fuzzy Sets and Systems 160, 3323–3341,2009.

[11] Maykin Warasart, and Pramote Kuacharoen, Paper-based Document Authentication using Digital Signature and QR Code. *4^{TH} International Conference on Computer Engineering and Technology (ICCET 2012),2012.*

[12] Public Key Cryptography for the Financial Services Industry. 2005. The Elliptic Curve Digital Signature Algorithm (ECDSA), ANSI X9.62.

[13] Prewitt, J. A Computational Approach to Edge Detection. IEEE Trans. Pattern Analysis and Machine Intelligence,1986.

[14] Pandapotan Siagian, Perancangan Autentikasi Login E-Library Dengan Template Matching (Studi Kasus: STIKOM DB Jambi). Prosiding Konferensi Nasional Sistem Informasi (KNSI), ISSN 978-602-17488-0-0, Lombok, 2013.

[15] Pandapotan Siagian, Development System with the Attendance of Content Based Image Retrieval (CBIR), Information Systems International Conference (ISICO)Bali, 2013.

[16] R.L. van Renesse, Paper-based document security–A Review, in European Conf. on Security and Detection,1997.

[17] R. Xu and D. Wunsch, Survey of clustering algorithms, IEEE Transactions on Neural Networks, Vol.16, Issue 3, pp. 645– 678, May 2005.

[18] Swapna Borde, Dr. Udhav Bhosle, Image Retrieval Using Steerable Pyramid, International Journal of Computer Applications (0975-8887),Volume 38-No. 7, January 2012.