SEGMENTATION OF LUNGS ON *BRONCHUS* FROM IMAGE CT SCAN USING *K-MEANS CLUSTERING* AND *WATERSHED*

Bagus Perdana[1]

[1] Teknik Informatika - Universitas Komputer Indonesia
Jl. Dipatiukur 112 - 114 Bandung
E-Mail : Bagusperdanaactive@gmail.com

ABSTRACT

Some of the technology used in the healthcare field are Chest X / X-ray, MRI *(Magnetic Resonance Imaging)* and CT Scan. Many doctors who use CT scans as a tool to scan a disease or symptoms that exist in certain organs.

One image processing techniques used to overcome the high level of noise is a *watershed* method. *Watershed* method used in CT scan image of *the bronchus* to detect patterns in the images in order to identify the *bronchus* in kedaan normal or not normal. *Watershed* method used to detect the characteristics of whatever form and Define whether the patterns are formed from the *watershed* method is normal or not normal *bronchus. K-means clustering* method will be used to detect these patterns and will be categorized into two categories: normal and not normal.

From the research, the analysis of image processing, training and testing methods *watershed* and *k-means clustering* to identify patterns CT scan image of *the bronchus* to obtain information *bronchus* state of normal and abnormal. As for the accuracy of the application of the method was 76.6% and can recognize the image of CT scan properly.

1. PRELIMINARY

Bronchus is an airway that is formed from two parts of the trachea at a height of approx fifth thoracic vertebrate, has a similar structure to the trachea and coated by the same cell type. *Bronchus* goes toward the bottom and side to the lungs and branched into two, namely the right *bronchus* and *bronchus* left [1].

Some of the technology used in the healthcare field are Chest X / X-ray, MRI *(Magnetic Resonance Imaging)* and CT Scan. Many doctors who use CT scans as a tool to scan a disease or symptoms that exist in certain organs. A CT scan has several weaknesses, among others, have noise is high and this is because exposure to X-ray radiation that is used around 4% of X-ray radiation used in radiology / x-rays, thereby reducing the accuracy of the picture and there is a picture that should no be recorded due patients when recording moving images.

1.1. image processing

To continue with the introduction of an object owned by the texture of the object then dibutuhkanlah a technique for doing so, such technique is the image processing techniques. An image processing and image analysis processing that involve visual perception. This process has the characteristics of data input and output information in the form of images. The term digital image processing is generally defined as a two-dimensional image processing by a computer. In a broader definition, digital image processing also includes all two-dimensional data.

1.2. Elements Image

In the image processing object used is an image of a specific object in which the image contains a number of basic elements. The basic elements in the manipulation of the image processing, these elements are color, brightness, contrast, contour, form, texture, and movement time and the detection and recognition.

1.3 Patterns

Pattern is the entity defined and identifiable through its features *(features).* These characteristics are used to distinguish a pattern with other patterns. A great characteristic is having distinguishing characteristic is high, so grouping patterns based on the features of which can be done with high accuracy.

1.3. Watershed

Create M1, M2, M3, ..., MR became a collection point coordinates in an area with a minimum value of an image g (x, y). Pictures here generally had an image that has been processed by pre-processing first.Then create a C (Mi) into sets of coordinates on the charging area which has a

relationship with the minimum area Mi (considered charging area and the minimum area forming components that are connected).Notation min and max are used to mark the minimum and maximum values of g (x, y). Then create a new T [n] into a collection of coordinates (s, t) where g (s, t) <n, so it can be defined:

In geometry, T [n] is a set of coordinates of g (x, y) which is under the area g (x, y) = n. Topography is then replenished by the addition of an integer ranging from n = min up to n = max. At each addition of n, the algorithm needs to know the number of points under the "depth" n. In general, the area under g (x, y) = n be black or a value of 0 and above it are colored white or a value of 1.

Then create Cn (Mi) into sets of coordinates of points on the charging area relating to the minimum area Mi filled the stage n. Cn (Mi) can be seen as a binary image when using the equation:

In other words Cn (Mi) = 1 at the location (x, y) if (x, y) ∈ C (Mi) and (x, y) ∈ T [n], other than that Cn (Mi) = 0. From here it can be said that and

C [n-1] is a subset of C [n] and C [n] is a subset of T [n] then C [n-1] is a subset of T [n]. From here it was found that each component is connected from C [n-1] has one connected component of T [n]. If Q is a collection of connected components in T [n], then for each component connected q ∈ Q [n], then there are three possibilities:

1. q ∩ C [n-1] is empty

2. q ∩ C [n-1] contains one connected component of C [n-1]

3. q ∩ C [n-1] contains more than one connected component of C [n-1]

If the condition c occurs then charging will cause the two areas into one, then the point on C [n-1] belong dam or watershed line.

1.4. Parameter Image Value

1.4.1. Area

To facilitate the detection and understanding of the characteristics on an object. Suppose only on an image of the object being examined there and directly applicable in a more simple way. O notation in the following persamaanpersamaan is a point or pixel of an object in a binary image, while the index of *x* and *y* is the location of that point in the matrix of size *m x n.*

$$A = \sum_{x=1}^{m} \sum_{y=1}^{n} O(x, y)$$

1.4.2. Perimeter

Perimeter is the outermost part of an object adjacent to Pixels or puksel-puksel of backgrounds. Therefore perimeter has several different definitions but has the same purpose.

1.4.3. Compactness

The combination of several geometric properties of objects can be used to analyze the shape of the object through the calculation of a dimensionless form factor. Compactness *(compactness)* of an object can be measured through analysis of dimensionless form factor *isoperimetric* inequality represented by the following equation.

$$C = \frac{4 .\pi .A}{P^2}$$

Where C, A and P, respectively *compactness,* Area and Perimeter.

5. K-Means Clustering Method

Data Clustering is one method of Data Mining that is without direction (unsupervised). There are two types of data clustering is often used in the process of grouping data that is hierarchical (hierarchical) and nonhierarchical clustering of data (non-hierarchical) clustering of data. K-Means is one method of non-hierarchical clustering of data that seek to partition the data into the form of one or more clusters / groups. This method of partitioning the data into clusters / groups so that the data that have the same characteristics are grouped into the same cluster and data that have different characteristics are grouped into another group. The purpose of the data clustering is to minimize the objective function that is set in the process of clustering, which generally seeks to minimize variations in a cluster and maximize inter-cluster variation.

Data clustering using the K-Means method is generally done with the basic algorithm as follows [8]:

1. Determine the Number of Cluster

2. Allocate data into random cluster

3. Calculate Centroid / average of the data of each cluster

4. Allocate each data to the centroid / Average nearby

5. Go back to step 3, if there is still data switch cluster centroid or if the value changes, there is above a specified threshold value or if the value changes to the Objective Function used above the specified Threshold value

2. RESEARCH

2.1. Analysis System

Analysis system is a means used to identify and evaluate problems, obstacles faced and needs to be expected in a system. This analysis is needed as a basis for the design stage of the system. The results of the analysis of this system will produce a wide range of conclusions and recommendations that could be used to base the design of the application to be built.

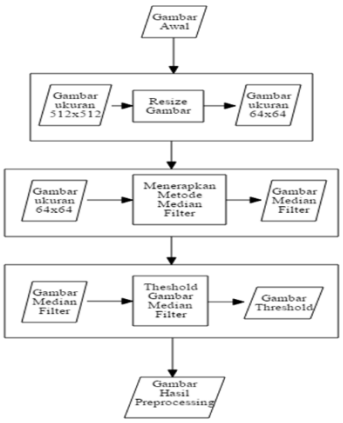
2.2. Analysis Dataset

This study uses a CT scan image dataset bronchus human lungs that are the basis of image data to JPEG format with the average image size is 64x64.

2.3. Analysis Methods

The analysis methods used to determine the process flow of a method used can be incorporated into applications built. This application development using *watershed* method for image segmentation process and *k-means clustering* for image classification that has been segmented.

2.3.1. Citra Latih



Gambar 1 Flowchart Pre-Processing Citra

The image is an image that included training in order to fill the training data that has no ability to add adan the accuracy of the methods used as for the image of the train used is the image of a CT scan of the bronchi.

Imagery training which will be entered in order to get some treatment process train *noise* in an image can be eliminated so as to make it easier for further processing. The process is as follows:

1. Resize
2. Median Filter
3. thresholding

The results of the processing and in use for the watershed segmentation method.

2.3.1.1. Resize

After the incoming image to resize and will soon be the size 64x64. Then I'll go into the next stage of *median filtering*.

2.3.1.2. Median Filter

To reduce the noise contained in the initial image *preprocessing* is the first stage in applying the method of median filter.

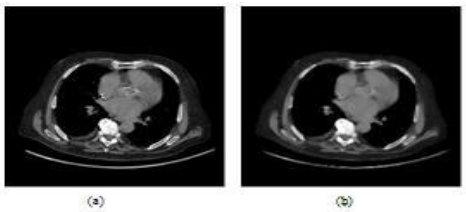


Figure 2 Median Filtering (a) Initial image (b) Where the Image Results Median Filter Noise Figure The Smooth And Lost

median filter penghalusannya seen that the process occurs in edge areas of the image. So that despite the image smoothing, *median filter* more towards "soften" the image there. It may reduce the computation by finding extreme values that need to be changed, or by limiting the median value is used.

How to sort the values on 3x3 matrices which exist in the example image of Figure 3.4 (a) and convert the picture into a matrix and showed the following results :

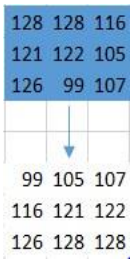


Figure 3 Process Median Filter

2.3.1.3. Threshold

The next step is to apply *thresholding* method to change the image of the beginning into a binary. In equation (2.2)

$$f(x,y) = \begin{cases} 0, & x < 103 \\ 1, & x \geq 103 \end{cases}$$

T = level threshold

namely by filtering the numbers in the matrix into two values: 0 and 1. Level *threshold* rated 103. Value 103 is taken as the value is very appropriate to obtain information if the *threshold* value using a value of more than 103 are missing a lot of information and the level of accuracy of the method decreases.

2.3.2. Watershed Segmentation

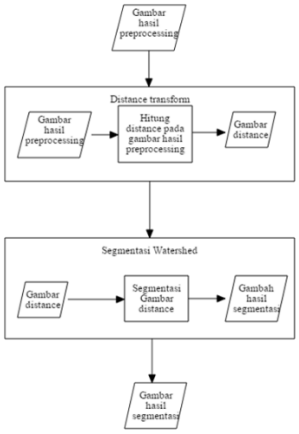


Figure 4 Flowchart Watershed

After a *median* value of image *thresholding filter* and pass through the stages and grades have diproleh, a process to search for watershed line that represents the object to be segmented. The first step in to operate the watershed to identify parts of an image is the minimum intensity.

First calculate *the distance* value exist in the image matrix. With *Cityblock* method.

$$d_{ij} = \sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|$$

0	0	1	1	0
0	1	2	2	1
1	2	3	2	1
1	2	3	3	2
1	2	3	4	3

Figure 5 Matrix *Distance* Based Sample Results Figure 3

he first step to implement a *watershed* method is to identify the minimum value of a picture 3:13, after it carried out a comparison value with eight neighbors using the structure element 3x3. Where the center pixel is a value of 1, because the value of 1 is the minimum value of all other values.

0	0	1	1	0	0	0
0	1	2	2	1	0	0
1	2	3	2	1	0	0
1	2	3	3	2	1	0
1	2	3	4	3	2	1
1	2	3	4	4	3	2

Figure 6 Formation Process Dam Watershed

The next step in the pixel center decided to find the minimum value of eight neighbors, for example, to find the minimum value of the eight first neighbors connected will be compared with the pixel center, if the condition of the center pixel is lower than the value of another image, then do a search value image that almost approaching the center pixel value, to determine the minimum value and the watershed line.

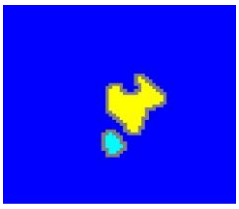
Once found minimunnya value, there will be a line of grades 1 to 1 neighboring. The concept that occurs in this watershed are:

1. Determine the area (region) R1, R2, .., Rn, by representing the object that has a minimum value.
2. Do a comparison point of the center pixel with eight neighbors using the structure element 3x3.
3. Form the dam (line) at the position where the two regions are connected.
4. Repeat step three so that all areas are connected, and the obtained result

The results obtained after processing the image threshold, implement distance and using watershed segmentation.



(a) (b)



(c)

Figure 7 (a) Figure Threshold (b) After Applying Distance (c) Segmentation Watershed

2.3.3. extraction features

The next stage after determining the next image segmentation in determining the values of the features in the image. Values of these features is the Area, Perimeter, and compactness workflow *k-means clustering* process is

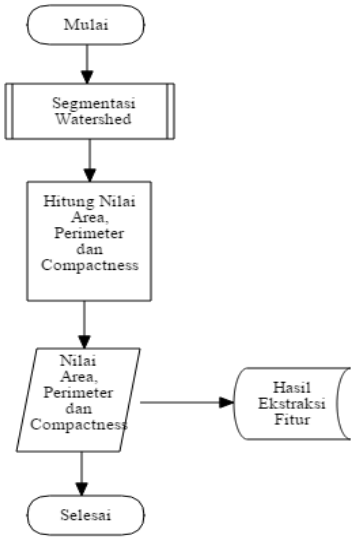


Figure 8 Workflows Feature Extraction

2.3.3.1. Area

O notation in the equations are a point or pixel of an object in a binary image, while the index of *x* and *y* is the location of that point in the matrix of size *m x n.*

$$A = \sum_{x=1}^{m} \sum_{y=1}^{n} O(x, y)$$

2.3.3.2. Perimeter

Perimeter is the outermost part of an object adjacent to the pixels of the background.

2.3.3.3. Compactness

Compactness *(compactness)* of an object can be measured using the equation.

$$C = \frac{4 . \pi . A}{P^2}$$

2.3.4. Training K-Means Clustering

Results of the extraction feature that has been calculated to be trained so

that the data stored training data. Here is a k-means training process:



Figure 9 Flowchart Training K-Mean

2.3.5. Examination

Stages of testing is the last stage of that test when the picture or image included to recognize the results of image classification either already trained or untrained image. The steps in the testing phase is as follows:



Figure 10 Flowchart Tests

2.4. examination

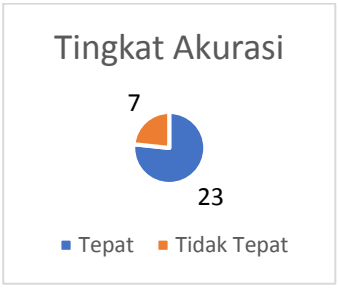
esting method of *k-means clustering* is to test the accuracy of the method to the image of a CT scan of the lungs. This study used 100 data, which will be divided into two, namely A1 = 50 training data and test data A2 = 30. Testing method of *k-means clustering* as follows.

Tabel 1 Testing Methods *K-Means Clustering*

Data	Cluster	
	Normal	Abnormal
50 Data Training	30 data	20 data

The test data to training data to determine the CT scan image of normal and abnormal *bronchus*.

Tabel 2 Level Accuracy



Gambar 4. 1 Diagram Accuracy

8 of 30 Data are data that does not match, then the accuracy rate *of k-means clustering* method was 76.6%.

3. Closing

3.1. Conclusion

From the research, the analysis of image processing, training and testing methods *watershed* and *k-means clustering* to identify patterns CT scan image of *the bronchus* to obtain information*bronchus* state of normal and abnormal. As for the accuracy of the application of the method was 76.6% and can recognize the image of CT scan properly.

Application of the method can be implemented properly and may also be applied to other cases related to the health of the lungs and other body parts.

3.2. Suggestion

From the research, analysis, image processing, training and testing are

suggestions that may be useful if there is going to conduct similar research, that ensure validation of the input image is in conformity with the image of penelititan goal. Picture a good image and accordingly is required to obtain a higher degree of accuracy again.

BIBLIOGRAPHY

[1] Pino P., Kiel dan Suharto., 2013. Pengaruh Lama Waktu Kematian Terhadap Kemampuan Pergerakan Silia Bronkus Hewan Coba Post Mortem Yang Diperiksa Pada Suhu Kamar Dan Suhu Dingin. thesis. Malang : Faculty of Medicine Diponegoro University.

[2] Turnip, Alfian., 2015. Analisis Penggunaan Teknik Post Processing Volume Rendering (Vr) Pada Ct Angiografi Arteri Abdominalis Dengan MSCT. Skripsi. Medan : Universitas Sumatra Utara.

[3] Nalwan, A., 1997. Pengolahan Gambar Secara Digital. Jakarta: Elex Media Komputindo

[4] Afrizal, Syafitri., 2013. Segmentasi Bronchus dan Bronchiolus pada Citra CT Scan Paru – Paru Menggunakan Watershed Filter, Applied Business and Engineering Conference (ABEC), Politeknik Caltex Riau, Riau, hal. 1-7.

[5] Nugroho, Setiawan Qodri., 2015. Pengelompokan Kayu Kelapa Menggunakan Algoritma K-Means Berdasarkan Tekstur Citra Kayu Kelapa Dua Dimensi (2D). Semarang : Universitas Dian Nuswantoro.

[6] Muis, S., 2006. Teknik Jaringan Saraf Tiruan. Yogyakarta: Graha Ilmu

[7] Faradila, 2007. Penyelesaian System Persamaan non Linier dengan metode Jaringan Saraf Tiruan Hopfield. Skripsi. Jakarta : Universitas Negri Hidayatullah.

[8] MacQueen, J. B., 1967. Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1: 281-297.