The Grouping Of Facial Images Using Agglomerative Hierarchical Clustering To Improve The CBIR Based Face Recognition System

by Muhammad Fachrurrozi

Submission date: 24-Feb-2020 01:43PM (UTC+0700)

Submission ID: 1262955939

File name: Paper_-_1570402260.pdf (369.94K)

Word count: 2481

Character count: 12359

The Grouping Of Facial Images Using Agglomerative Hierarchical Clustering To Improve The CBIR Based Face Recognition System

Muhammad Fachrurrozi
Informatics Engineering Department
Faculty of Computer Science, Universitas Sriwijaya
Palembang, Indonesia
mfachrz@unsri.ac.id

Saparudin
Informatics Engineering Department
Faculty of Computer Science, Universitas Sriwijaya
Palembang, Indonesia
saparudin@unsri.ac.id

Erwin
Computer Engineering Department
Faculty of Computer Science, Universitas Sriwijaya
Palembang, Indonesia
erwin@unsri.ac.id

Clara Fin Badillah
Informatics Engineering Department
Faculty of Computer Science, Universitas Sriwijaya
Palembang, Indonesia
09021181320052@students.ilkom.unsri.ac.id

Junia Erlina
Informatics Engineering Department
Faculty of Computer Science, Universitas Sriwijaya
Palembang, Indonesia
juniaerlinaa@gmail.com

Mardiana
Law Department
Faculty of Law, Universitas Sriwijaya
Palembang, Indonesia
mardiana_rachman@yahoo.com

Auzan Lazuardi
Informatics Engineering Department
Faculty of Computer Science, Universitas Sriwijaya
Palembang, Indonesia
auzanlazuardi@gmail.com

***Abstract*—The facial image grouping runs automatically using the Agglomerative Hierarchical Clustering (AHC) algorithm. The pre-processing performed is feature extraction to get the vector feature of the face image. The AHC algorithm performs a grouping of face image vector features using Manhattan Distance and linkage average, single, and complete linkage methods. Testing cluster validation by comparing the value of Cophenetic Correlation Coefficien (CCC) and comparing the time of face image recognition process using clustering process and without using clustering process. From the test results, it is known that the complete method has a higher CCC value than the other 2 methods, that is equal to 0.904938 with the difference in value of 0.127558 on the single method and the difference of 0.02291 in the average method. Face image recognition system uses clustering process faster than using clustering process.**

Keywords— *Clustering; AHC; Single Linkage; Complete Linkage; Average Linkage*;

I. Introduction

Content Based Image Retrieval (CBIR) is the image process of a database or digital image library in accordance with the visual content of the image [1]. CBIR only focuses on image search using queries on large image databases based on texture, color, shape, and region features. The grouping of facial image can be used to speed up the image search process on facial image recognition system using Image Processing Science. Face recognition is faster and more accurate on CBIR[2] [3][4].

Grouping is divided into two types, namely hierarchy and non hierarchy[5]. Hierarchical Clustering is a grouping algorithm by forming hierarchies of similar data into a tree or dendogram. In Hierarchical Clustering there are two ways of grouping, namely agglomerative and divisive. Agglomerative clustering process based on the amount of data grouped into hierarchy - hierarchy, then hierarchy - the hierarchy becomes a hierarchical unity.

Extraction facial image feature to get vector feature using Local Binary Pattern (LBP). The distance between the vector features is calculated using the manhattan distance which is subsequently grouped using the Agglomerative Hierarchical Clustering (AHC) algorithm. Manhattan Distance provides relatively higher results than Euclidean Distance with high probability [6].

This study focused on comparing the three methods of the AHC algorithm, namely Single Lingkage, Complete Lingkage, and Average Linkage in categorizing facial images and helping to improve the speed of face recognition system.

II. Clustering

A. *Local Binary Pattern* (LBP)

LBP represents a pixel which formed by a 3x3 matrix as a comparison between the center pixel and its surrounding pixel which then converted into binary numbers. The comparison assumes that if the surrounded pixel value is greater than the central pixel value than it will be 1 otherwise 0. After we get 8 binary numbers in each pixel then it will be replaced with the decimal form to get the result.

The LBP algorithm formula can be expressed as the following formula:

$$LBP(x_c, y_c) = \sum_{p=0}^{7} f(g_p - g_c) 2^p \qquad (1)$$

Information:

g_p : central pixel value

g_c : the pixel value around the center

p : number of pixels around the center

And the function f (x) is defined as follows:

$$f(x) = \left\{ \begin{matrix} 1, x \geq 0 \\ 0, x < 0 \end{matrix} \right\}$$

B. *Agglomerative Hierarchical Clustering*

Agglomerative Hierarchical clustering is a clustering algorithm based on the proximity distance between two images into a hierarchy. This process repeats itself until it gets some hierarchy. The hierarchy with the closest distance is combined into one hierarchy. The proximity to the new hierarchy then recalculated and the closest hierarchy is merged again. The process is repeated until all the data (object) clustered into one hierarchy.

Calculating the spacing between two images using the *Manhattan Distance* formulated in formula (2):

$$d = \sum_{i=1}^{n} |u_i - v_i| \qquad (2)$$

where:

d : the distance between the image of u and v.

n : number of variables.

u_i : the value of u on i variable.

v_i : the value of v on i variable

The distance between images is written into a matrix called distance matrix. In order to determine the distance between the two clusters, Agglomerative Hierarchical clustering has 3 methods of grouping data, namely:

1. *Single Linkage*

Single Linkage classifies data based on the closest distance (Min) between the hierarchy. Single Linkage can be formulated in formula (3):

$$d(uv)w = Min[d(uw), d(vw)] \qquad (3)$$

where :

u : the image of u

v : the image of v

w : the image of w

$d(uv)w$: the distance between the hierarchy uv and w.

$d(uw)$: the distance between the hierarchy u and w.

$d(vw)$: the distance between the hierarchy v and w.

2. *Complete Linkage*

Complete Linkage categorizes data by the furthest distance (Max) or the maximum distance between hierarchies. Complete Linkage can be formulated in formula (4):

$$d(uv)w = Max[d(uw), d(vw)] \qquad (4)$$

where :

u : the image of u

v : the image of v

w : the image of w

$d(uv)w$: the distance between the hierarchy uv and w.

$d(uw)$: the distance between the hierarchy u and w.

$d(vw)$: the distance between the hierarchy v and w.

3. *Average Linkage*

Average Linkage classifies data based on the average distance between the hierarchy. Average Linkage can be formulated in formula (5):

$$d(uv)w = \frac{d(uw) + d(vw)}{2} \qquad (5)$$

where :

u : the image of u

v : the image of v

w : the image of w

$d(uv)w$: the distance between the hierarchy uv and w.

$d(uw)$: the distance between the hierarchy u and w.

$d(vw)$: the distance between the hierarchy v and w.

III. Methodology

A. *Data*

The data used as many as 200 images from 20 people, each person taken 10 images with different sides and the same background. With the rules on the image of the face still looks both eyes, nose and mouth. The image is used as training data with dimensions of 150x150 pixels. Examples of face image data can be seen in figure 1.

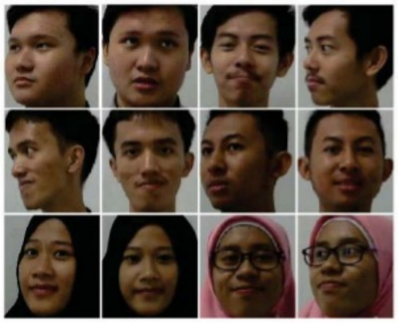

Fig. 1. Example of image data

B. General System

Figure 2 is a block diagram of a clustering system, where there are 5 stages in the grouping of face images.

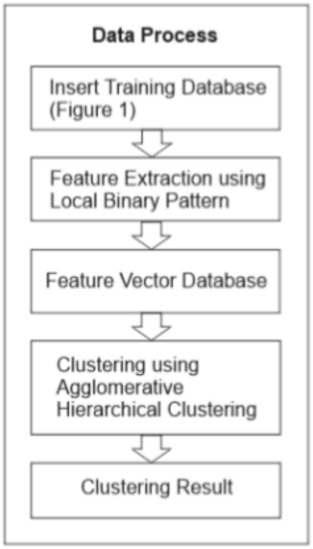
Data Process
Insert Training Database (Figure 1)
Feature Extraction using Local Binary Pattern
Feature Vector Database
Clustering using Agglomerative Hierarchical Clustering
Clustering Result

Fig. 2. General System Diagram

In general, the steps of clustering in this research is as follows:

1. Collect face image.
2. The feature extraction process using LBP method to get the characteristic of face image then transformed into vector feature form
3. Vector feature will be stored in the database.
4. Then do the clustering process using the AHC method on the vektor of the face image in the database.
5. Save clustering result.

IV. IMPLEMENTATION AND RESULT

The grouping process starts after getting the value of the vector feature. Grouping begins by calculating the distance between objects using formula (2), to get all the distance between objects can be calculated and then written into a matrix called distance matrix.

Distance matrix is grouped into several groups according to the three methods. Similarities, the closest distance to the formula (3), the farthest distance by the formula (4), or by the mean distance between the image and the other image by the formula (5) The result of the grouping is the grouped vector feature database.

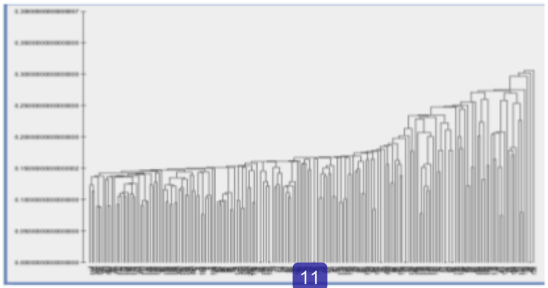

Fig. 3. Result of single linkage

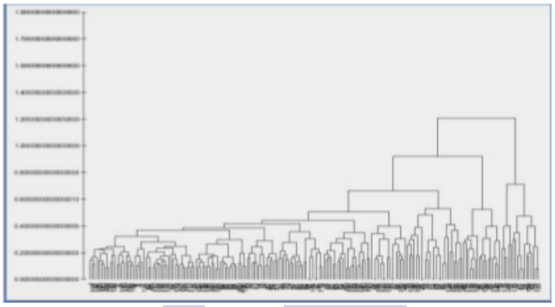

Fig. 4. Result of complete linkage

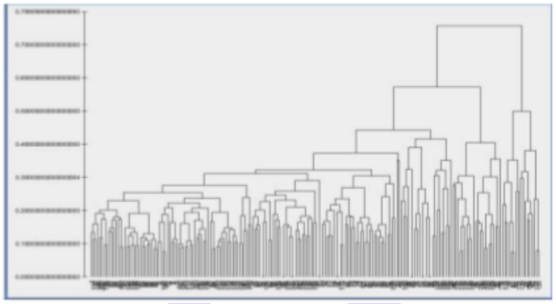

Fig. 5. Result of average linkage

Figure 3, 4, and 5 are the result dendograms of the face image groupings of the three methods of the AHC algorithm. All three methods produce 199 clusters. The validation of the cluster results of each method using Cophenetic Correlation Coefficient (CCC).

CCC is the correlation coefficient between the original matrix elements of distance matrix and the resulting elements of the dendogram (Cophenetic matrix)[7]. CCC can be formulated in formula (6):

$$coph = \frac{\sum_{i<k}(d_{ik} - \bar{d})(d_{cik} - \bar{d}_c)}{\sqrt{\left[\sum_{i<k}(d_{ik} - \bar{d})^2\right]\left[\sum_{i<k}(d_{cik} - \bar{d}_c)^2\right]}} \quad (6)$$

Keterangan :

$coph$: cophenetic correlation coefficient

d_{ik} : the distance matrix between object i and k

$\bar{d}$: average of d_{ik}

d_{cik} : distance of cophenetic object i and k

$\bar{d}_{c}$: average of d_{cik}

The resulting value of cophenetic correlation coefficient ranges between -1 and 1. The closer to the value of 1 means the quality resulting from the clustering process is said to be good, whereas if the value of CCC approaching value -1 means that the resulting quality of the clustering process is not good.

TABLE I. COPHENETIC CORRELATION COEFFICIENT VALUES

Single Linkage	**Complete Linkage**	**Average Linkage**
0.777381	0.904938	0.882028

In Table I the value of *Cophenetic Correlation Coefficient* obtained is single linkage is 0.777381, *complete* circle is 0.904938, and the *average linkage* is 0.882028. All three methods have a CCC value close to 1, which means the three methods are able to group well on the research data (200 face images). However, from the three methods it is known that the *Complete Linkage* method has a higher CCC value than the other two methods with a difference of value of 0.127558 on *single linkage* method and the difference of 0.02291 on the *average linkage* method. This shows the complete linkage method is better in grouping research data of 12.76% of the *single linkage* method and by 2.29% of the *Average Linkage* method.

The CBIR speed time test using AHC is calculated using the start time and stop time of program execution. Computation time is done to get the computation time of each method process in assisting face recognition system in recognizing face image. CBIR speed computation time test without clustering and CBIR with clustering using AHC is done with the formula in equation (7), that is:

$$R_{time} = \frac{\sum W}{n} \qquad (7)$$

Keterangan :

R_{time} : the average of time

n : number of variables

W : computing's time

Tables II and III are the results of calculating the computational time of the face recognition system using clustering and without clustering. The average value of time calculation result is calculated using formula 7. In Table II experiment introduction on 1 object (1 face image) the same without using clustering process as much as 10 times introduction experiment and using clustering process 10 times introduction experiment on each method. While in table III experiment introduction on 6 objects (6 images face) the same without using the clustering process as much as 10 times the introduction experiment and using clustering process 10 times the introduction of experiments on each method. Testing the computing time of this facial recognition system proves whether the three proposed methods can help reduce the time of face recognition system in recognizing face image.

TABLE II. TABLE II. INTRODUCTION OF FACE IMAGE USING 1 OBJECT IMAGE

Recognition to -	Computation time without clustering process (s)	Computation time uses clustering process (s)		
		Single Linkage	Complete Linkage	Average Linkage
1	0.781	0.759	0.607	0.779
2	0.781	0.764	0.754	0.765
3	0.797	0.754	0.075	0.766
4	0.829	0.755	0.754	0.766
5	0.797	0.749	0.074	0.765
6	0.766	0.764	0.772	0.750
7	0.766	0.771	0.751	0.765
8	0.766	0.779	0.754	0.765
9	0.781	0.754	0.750	0.765
10	0.813	0.753	0.764	0.766
Rata – rata	**0.7877**	**0.7602**	**0.6055**	**0.7652**

TABLE III. TABLE II. INTRODUCTION OF FACE IMAGE USING 6 OBJECT IMAGE

Recognition to -	Computation time without clustering (s)	Computational time using clustering process (s)		
		Single Linkage	Complete Linkage	Average Linkage
1	0.779	0.722	0.766	0.757
2	0.897	0.751	0.750	0.758
3	0.766	0.765	0.750	0.772
4	0.875	0.754	0.766	0.755
5	0.781	0.757	0.765	0.757
6	0.766	0.755	0.751	0.758
7	0.780	0.753	0.797	0.752
8	0.766	0.776	0.750	0.756
9	0.782	0.751	0.750	0.776
10	0.766	0.770	0.750	0.753
Rata – rata	**0.7958**	**0.7554**	**0.7595**	**0.7594**

In table II test for face recognition not real-time on 1 object, it is known that the three methods are able to increase the speed of face recognition system with single linkage method by 3.49%, complete linkage method is 23.13%, and the average circumference method is 2.25%.

In testing table III for face recognition not real-time on 6 objects, it is known that the three methods are able to increase the speed of face recognition system with single linkage method 5.08%, complete linkage method is 4.56%, and the average linkage method is 4.57%.

V. CONCLUSION

From the results of testing new AHC method with single method, complete, and average of good association of facial image grouping. Among the three methods, it is known that the Complete Linkage method is better than the other two methods (single and average linkage).

Facial image recognition using 3 Agglomerative Hierarchical Clustering methods can help improve the computing time speed of face recognition system.

REFERENCE

[1] R. Chaudhari and A. M. Patil, "Content Based Image Retrieval Using Color and Shape Features," *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.*, vol. 1, no. 5, pp. 386–392, 2012.

[2] A. K. V. M. N. Anbazhagan, "Image Clustering and Retrieval using Image Mining Techniques," 2010.

[3] V. S. V. S. Murthy, E. Vamsidhar, J. N. V. R. S. Kumar, and P. Sankara Rao, "Content Based Image Retrieval using Hierarchical and K-Means Clustering Techniques," *Int. J. Eng. Sci. Technol.*, vol. 2, no. 3, pp. 209–212, 2010.

[4] S. Pandey, P. Khanna, and H. Yokota, "Clustering of hierarchical image database to reduce inter-and intra-semantic gaps in visual space for finding specific image semantics," *J. Vis. Commun. Image Represent.*, vol. 38, pp. 704–720, 2016.

[5] R. Nainggolan, "Algoritma Modified K-Means Clustering pada penentuan Cluster Centre Berbasis Sum of Squared Error (SSE)," 2014.

[6] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," *Database Theory – ICDT 2001*, pp. 420–434, 2001.

[7] A. Rodrigo and C. Tadeu, "A cophenetic correlation coefficient for Tocher ' s method," no. 1, pp. 589–596, 2013.

.

The Grouping Of Facial Images Using Agglomerative Hierarchical Clustering To Improve The CBIR Based Face Recognition System

ORIGINALITY REPORT

10%	6%	7%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Binus University International Student Paper	1%
2	Nurkhairany Amyra Mokhtar, Yong Zulina Zubairi, Abdul Ghapor Hussin. "Detecting multiple outliers in linear functional relationship model for circular variables using clustering technique", AIP Publishing, 2017 Publication	1%
3	Submitted to Rochester Institute of Technology Student Paper	1%
4	doaj.org Internet Source	1%
5	Willem Boshoff, Rossouw Jaarsveld. "Market Definition Using Consumer Characteristics and Cluster Analysis", South African Journal of Economics, 2019 Publication	1%
	www.mdpi.com	

6	Internet Source	1%
7	haralick.org Internet Source	<1%
8	Submitted to Stefan cel Mare University of Suceava Student Paper	<1%
9	Submitted to University of Stellenbosch, South Africa Student Paper	<1%
10	www.omicsonline.org Internet Source	<1%
11	Submitted to National University of Singapore Student Paper	<1%
12	Efendi Nasıbov, Cagin Kandemır-Cavas. "OWA-based linkage method in hierarchical clustering: Application on phylogenetic trees", Expert Systems with Applications, 2011 Publication	<1%
13	journal.portalgaruda.org Internet Source	<1%
14	"Text, Speech and Dialogue", Springer Science and Business Media LLC, 2006 Publication	<1%
15	Ing-Guey Jiang, Li-Chin Yeh, Wen-Liang Hung,	

	Miin-Shen Yang. "Data analysis on the extrasolar planets using robust clustering", Monthly Notices of the Royal Astronomical Society, 2006 Publication	<1%
16	"Encyclopedia of Optimization", Springer Science and Business Media LLC, 2009 Publication	<1%
17	iopscience.iop.org Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		