Sign Language Recognition in Robot Teleoperation using Centroid Distance Fourier Descriptors

Rayi Yanu Tara[1], Paulus Insap Santosa[2], Teguh Bharata Adji[3]
EE & IT Department, Gadjah Mada University
Yogyakarta, Indonesia

ABSTRACT
Commanding in robot teleoperation system can be done in several ways, including the use of sign language. In this paper, the use of centroid distance Fourier descriptors as hand shape descriptor in sign language recognition from visually captured hand gesture is considered. The sign language adopts the American Sign Language finger spelling. Only static gestures in the sign language are used. To obtain hand images, depth imager is used in this research. Hand image is extracted from depth image by applying threshold operation. Centroid distance signature is constructed from the segmented hand contours as a shape signature. Fourier transformation of the centroid distance signature results in fourier descriptors of the hand shape. The fourier descriptors of hand gesture are then compared with the gesture dictionary to perform gesture recognition. The performance of the gesture recognition using different distance metrics as classifiers is investigated. The test results show that the use of 15 Fourier descriptors and Manhattan distance-based classifier achieves the best recognition rates of 95% with small computation latency about 6.0573 ms. Recognition error is occurred due to the similarity of Fourier descriptors from some gesture.

Keywords
Hand Gesture, Sign Language, Fingerspelling, CeFD, Fourier Descriptor.

1. INTRODUCTION
Hand gesture recognition provides a natural way to communicate with machines (e.g. robot). The use of robot, especially in teleoperation, can reduce the risk factor of task failures and human harms during several activities such as hazardous material handling [1]. Sign language is often used as input method in teleoperation system. Several works that employ visually captured hand gesture in robot teleoperation system has been researched since last decade. Color camera was used to acquire hand image, and the acquired image was processed with Artificial Neural Network (ANN) [2], Fuzzy C-means clustering (FCM) [3], or Support Vector Machine (SVM) [4] to classify the meanings of each hand gesture.

In a sign language, hand shape can give information of hand gestures. Recognizing gestures through hand shape is a challenging process. Some sign language has the similar hand shape, and similar hand shape can be interpreted as different sign because of different viewpoint. Due to its complexity, the research of hand gesture recognition based on hand shape is continuously performed. Recognition of hand posture using two different shape descriptors had been conducted in [5]. Fourier descriptors and hu moments were compared in this research. This experiment used 64 fourier coefficients. Two databases of gestures were used in this experiment (i.e. Triecsch database [6] and self-made database). The result of this experiment showed that fourier descriptors give very good recognition rate rather than hu moments. Performance comparison of fourier descriptors and geometric moment invariants was presented [7]. The comparison was used ASL database to analyze discrimination and feature invariance of hand images. The results showed that both descriptors are unable to differentiate some classes in ASL. Another shape descriptor comparison in hand posture recognition from video was also presented in [8]. The research compares: 1) Hu moments, 2) Zenike moments, 3) Fourier descriptors (common set), and 4) Fourier descriptors (complete set). The recognition also evaluated the use of several classifiers to measure similarity of hand posture with the stored hand posture in the database. Bayesian classifier, support vector machine, k-nearest neighbor, and Euclidean distance were evaluated. Overall result of the presented research showed that the common set fourier descriptors has the highest recognition rate when combined with k-nearest neighbor, reaching 100% in classifiying learning set, and 88% in test set. An application of real time hand gesture recognition system using fourier descriptors was presented in [9]. The presented system uses 56 fourier descriptors from 64 fourier coefficients, and linear combination of Hidden Markov Models (HMM) and Recurrent Neural Network (RNN). With real time processing rates of 22 frames per second and 91.9% correct classification; the presented system achieved good performance. Another approach of hand gesture recognition using fourier descriptors and hidden markov models was also introduced [10]. The system was able to recognize 20 different gestures with average recognition rate of above 90%. All of the research in [5], [7], [8], [9], and [10] were used the same shape signature (i.e. complex coordinate) in fourier descriptors calculation and used hand image from color camera.

An evaluation of the use of different shape signature in fourier descriptors calculation for shape retrieval was presented [11]. The research compared four shape signatures: 1) complex coordinate, 2) centroid distance, 3) curvature signature, and 4) cumulative angular function. Euclidean distance was used as similarity measurement. To measure the effect of each shape signature in representing a shape, precision and recall ratio were used. The result showed that the use of centroid distance signature in calculating fourier descriptors is significantly better than other shape signature. The centroid distance fourier descriptors is robust and information preserving. This is due to the centroid distance, which captures both local and global features of the shape.

This paper contributes to the use of centroid distance fourier descriptors (CeFD) in sign language recognition system, which will be employed in robot teleoperation system. The sign language is adopted from American Sign Language (ASL) fingerspelling. Only static gesture sign is researched. A

depth imager is utilized to acquire hand image. The centroid distance fourier descriptors are used as the feature vector of inputted gesture. To perform recognition, the fourier descriptors of inputted hand image are compared with the fourier descriptors of each character stored in the gesture dictionary. To obtain the best recognition rate, different classifiers will be evaluated. The following section will clearly explain our research methodology.

2. METHODOLOGY

An overview about the methodology of this research is illustrated in Figure 1.

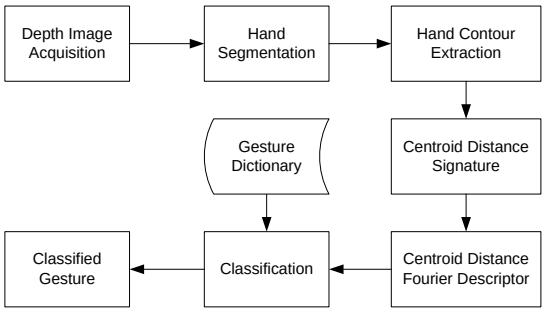


Figure 1. Research methodology

This research employs depth imager to acquire hand image from the human signers. The use of depth imager has benefit in segmenting hand image. Rather than color-based segmentation, segmentation in depth image is more robust since the lighting variation does not affect the image quality. Contour of the segmented hand image is then used to generate centroid distance signature. Hand contour coordinates are arranged into centroid distance signature with equal arc-length point sampling, results in N sampled signature point. Fourier transform of the centroid distance signature yields fourier descriptors, which represent feature of each hand gesture. Generally, this research is separated into two phases: dictionary build phase and classification phase. In the dictionary build phase, the fourier descriptors of each character are stored into a database to develop gesture dictionary. The gesture dictionary and comparison method are derived from [12]. The classification phase has the similar step, except that the fourier descriptors are compared with the dictionary using distance metric as classification methods. The result of classification phase is the meaning of the acquired gesture sign.

3. FEATURES EXTRACTION

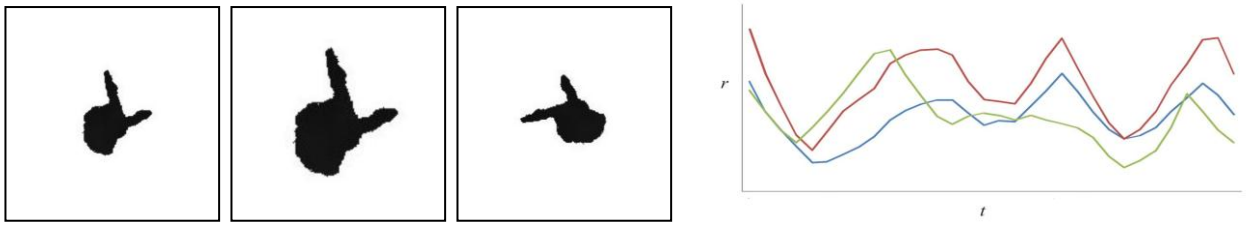
3.1 Hand Segmentation

Microsoft Kinect™ is utilized as depth imager in this research [13]. Human hand is assumed as the closest object in the imager field of view. The closer the object distance to the imager, the lower the voxel (depth pixel) value is. Threshold operation is applied to the image with a threshold value. The threshold value is acquired by summing the hand depth and the closest object distance. This method is adopted from [14]. Figure 2 shows the original image and the segmented hand image.

Figure 2. Depth image (*top*) and the segmented hand image (*bottom*)

3.2 Centroid Distance Signature

Shape signature is used to represent shape contour of an object. The shape signature itself is a one-dimensional function that is derived from shape contour coordinate. Centroid distance signature is one of several types of shape signature. In this research, the segmented hand image is processed with canny edge detection to extract the hand contour. Afterward, the centroid distance signature of hand images is generated from the hand contour. Three hand shapes with same gesture and their centroid distance signature are shown in Figure 3. The centroid distance signature *r(t)* is computed from the coordinates of each contour sequence by applying Equation (1) and Equation (2).



(a) (b) (c) (d)

Figure 3. Three hand images with same gesture and its centroid distance signature (*d*); original (*a-blue*), scaled up 50% (*b-red*), rotated 90°(*c-green*)

$$r(t) = \left(\left[x(t) - x_c \right]^2 + \left[y(t) - y_c \right]^2 \right)^{\frac{1}{2}} \quad (1)$$

$$x_c = \frac{1}{L} x(t), y_c = \frac{1}{L} y(t) \quad (2)$$

where x_c and y_c are the centroid coordinate of hand shape, $x(t)$ and $y(t)$ are the coordinates of each contour, L is the contour length, and t is the contour index. As shown in Figure 3, the centroid distance signature $r(t)$ has the translation invariant property. Rotation of the hand image causes circular shift, and scaling of hand image changes the signature value linearly.

3.3 Centroid Distance Fourier Descriptors

Centroid distance Fourier Descriptors (CeFD) was empirically proven for having higher performance rather than other fourier descriptors [11], [15]. In general, the CeFD is obtained by applying fourier transform on a centroid distance signature. The discrete fourier transform of centroid distance signature $r(t)$ is given in Equation (3).

$$a_n = \frac{1}{N}\sum_{t=0}^{N-1} r(t)\exp\left(\frac{-j2\pi nt}{N}\right), \text{ n=0, 1, …, N-1} \quad (3)$$

N is the total number of sampled points from the signature, and a_n is the fourier descriptor. This research assigns 64 as the N value. The sampling points of centroid distance signature are obtained by applying equal-arc length sampling, which is done by dividing the total contour length L by N. Since the centroid distance signature is real value, there are only N/2 different frequencies in the fourier transform. To make the fourier descriptors invariant to scaling, rotation, and translation, Equation (4) is employed to normalize the fourier descriptors.

$$CeFD = \left|\frac{a_1}{a_0}\right|, \left|\frac{a_2}{a_0}\right|, \left|\frac{a_3}{a_0}\right|, \ldots, \left|\frac{a_{N/2}}{a_0}\right|, \quad (4)$$

where CeFD is the normalized fourier descriptors. The descriptors use only the magnitude values since the phase values are variant to rotation. The dc-component (i.e. a_0) is used to normalize the remaining fourier descriptors to achieve scale invariant.

Figure 4 shows the fourier descriptors of three images from Figure 3.

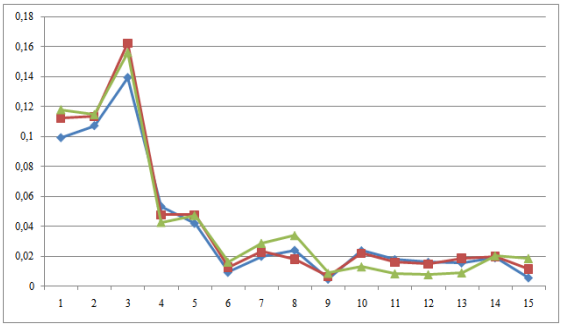


Figure 4. Fourier descriptors of three images in Figure 3

According to the illustration, the normalized fourier descriptors of each character have small deviation. The normalized FDs are proven invariant to translation, rotation, and scaling. When used in shape retrieval, the retrieval precision degrades when using 10 FDs and does not improve significantly when using 15 FDs [16]. Thus, only the first 15 fourier descriptors is employed in this research.

4. GESTURES CLASSIFICATION

4.1 Gesture Dictionary

Five gestures are employed as gesture vocabulary. These gestures are adopted from ASL fingerspelling. The character similarity graph in [7] is considered when choosing each character, which will be used as the five gestures. The similarity distance of FDs between each character in sign language is essential since it can reduce the possibility of false recognition. Table 1 shows the gesture vocabulary. Fingerspelling illustration in the gesture vocabulary is obtained from [17].

Table 1. Gesture vocabulary

Fingerspelling	Meaning
	Turn Right
	Turn Left
	Forward
	Stop
	Backward

In the gesture dictionary, each character has 15 fourier descriptors as features. To develop reliable gesture dictionary, each character is represented by five signers. Each of signer gives two hand poses with different conditions. The total training dataset is 50 gestures, 10 gestures for each character. Having 10 variations of gesture data for each character is enough since the fourier descriptors themselves have the invariant property. The dictionary GD is represented in a matrix, as shown in Equation (5).

$$GD = \begin{bmatrix} C_1 & FD_{11} & FD_{12} & \ldots & FD_{1j} \\ C_2 & FD_{21} & \ldots & \ldots & \ldots \\ C_3 & FD_{31} & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ C_i & FD_{i1} & \ldots & \ldots & FD_{ij} \end{bmatrix}, \quad (5)$$

where FD_{ij} is the j^{th} fourier descriptors of i^{th} fingerspelling gestures, and C_i is the character of i^{th} fingerspelling gesture.

Having 50 gestures as dictionary is not the most efficient representation and increases the computational load. Similarity between each gesture that has the same character is considered to overcome the inefficiency of dictionary size. The similarity measurement employs Euclidean distance, which shown in Equation (6). Distance between two gestures that has value less than 0.05 can be merged. This technique reduces the dictionary size to 80% from its original size, from 50 gestures into 40 gestures. Performance comparison

between the complete gesture dictionary and the reduced gesture dictionary will be covered in Section 5.

4.2 Gesture Classifier

Classifying hand gesture is performed using similarity degree matching. Euclidean distance, Manhattan distance, and Canberra distance are evaluated in this research. In similarity degree matching with distance metric, the smallest distance is considered as a match. Thus, the smallest FDs distance between a character in gesture dictionary and inputted gesture is taken as a classified gesture sign. Equation (6) can be used to calculate both Euclidean distance and Manhattan distance.

$$D_j = \left(\sum_i \left| FD_i - GD_{ij} \right|^y \right)^{\frac{1}{y}}, \quad (6)$$

where D_j is the distance between inputted gesture and j^{th} character in dictionary. Euclidean distance calculation is performed by using 2 as y, and using 1 as y for Manhattan distance. For the Canberra distance calculation, Equation (7) is used. Canberra distance CD_j is very sensitive to small variation in input value.

$$CD_j = \sum_i \frac{\left| FD_i - GD_{ij} \right|}{\left| FD_i \right| + \left| GD_{ij} \right|}, \quad (7)$$

5. EXPERIMENT AND RESULT

Several experiments are conducted to measure overall performance of the classifier. Testing dataset is created by acquiring 90 gestures for each character from five different signers, having a total of 450 gestures as testing dataset. Both training dataset and testing dataset are used in this experiment. When using the training dataset and complete gesture dictionary, the recognition rate achieves 100% with three different distance metrics (i.e. Euclidean distance, Manhattan distance, and Canberra distance). Nonetheless, the use of reduced gesture dictionary gives lower recognition rates of: 1) 95.7% with Euclidean distance; 2) 97.2% with Manhattan distance; and 3) 92.8% with Canberra distance. Good recognition result with training dataset is very common, since the training dataset is used to build the dictionary. Further test using the testing dataset is conducted to validate the performance of each classifier with the gesture dictionary. Recognition results of three classifiers using both datasets are shown in Table 2.

Table 2. Recognition results with complete data set

Gesture	Complete Dictionary			Reduced Dictionary		
	ED	MD	CD	ED	MD	CD
B	97	97	68	97	97	64
L	87	92	84	86	92	90
5	99	98	99	99	98	99
S	100	100	96	100	100	96
1	91	88	82	90	88	83
Average	94,8	95	85,8	94,4	95	86,4

From the results, the Manhattan distance achieves the highest average recognition rate of 95% in classifying the complete dataset, whilst the Canberra distance has the lowest average recognition rate. The Euclidean distance has the lowest recognition rate in recognizing gesture "L". Meanwhile, the Canberra distance has the same problem for gesture "3" and "5". Overall, the Manhattan distance slightly outperforms Euclidean and Canberra distances. The use of different gesture dictionary gives no significant effect to the recognition rate. For three different distance metrics, slight variation under 1% occurs when the reduced dictionary is employed.

Table 3 shows the confusion matrix for recognition using Manhattan distance that explicitly details the recognition rate result. Gesture "S" obtains the best recognition rate of 100%. Recognition of gesture "1" gets the lowest recognition rate of 88%. Thus, the recognition is often misclassified as gestures "B" and "1". Gesture "L" has the most ambiguous gesture sign over another gesture even though having 92% recognition rate. The "L" gesture is often recognized as gestures "5", "S", and "1".

Table 3. Confusion matrix for recognition using Manhattan distance and complete dictionary

	B	L	5	S	1
B	97	2	0	0	1
L	0	92	4	2	2
5	0	0	98	2	0
S	0	0	0	100	0
1	9	3	0	0	88

This experiment also evaluates the computation latency. Different image resolutions are used to approximate the actual computation loads. This test uses square images from Figure 3.a with resolutions of: 1) 400 pixels; 2) 600 pixels; and 3) 800 pixels. This experiment utilizes a PC with Ubuntu OS, 2.0 GHz dual core processors, and 3 GB RAM. Table 4 shows the computation latency test. It explains that the total computation load is significantly raised when the image resolution increases in the process of contour extraction and CeFD calculation. Segmentation and classification are less affected with the variation of image sizes. Classification process is the fastest process among the processes done in this research. This occurs since it only calculates the distance between feature set of inputted gesture with feature set in gesture dictionary. With computation time of 6.0573 ms, this method has the capability to process an 800 x 800 pixels image of 165 frames per second.

Table 4. Computation latency test result

Process	Image 1 (ms)	Image 2 (ms)	Image 3 (ms)
Segmentation	2,7841	2,8043	2,8869
Contour extraction	0,6972	1,5398	2,8752
CeFD	0,0874	0,1684	0,2917
Classification (dictionary)	0,0045	0,0034	0,0035
Total	**3,5732**	**4,5159**	**6,0573**

6. CONCLUSION AND FUTURE WORK

In this paper, sign language recognition by comparing Centroid distance Fourier Descriptors (CeFD) of inputted gesture with gesture dictionary is considered. The sign language is used as mobile robot teleoperation commands. In this research, five gesture signs are adopted from ASL fingerspelling. Acquiring hand gesture is done using depth imager. A gesture dictionary

is built as the reference of each sign character. Comparison is performed by measuring fourier descriptor similarities between inputted gesture and the gesture dictionary which results in recognized gesture sign. Recognition of gesture sign using three different distance metrics (i.e. Euclidean, Manhattan, and Canberra) and two dictionary sets (i.e. complete, and reduced) is evaluated. Five persons are involved to develop gesture dataset of 100 images per gesture signs. Hence, the total images are 500 images. The dataset is separated into two datasets (i.e. 50 for training dataset and 450 for testing dataset). From the experimental results, the classification of inputted gesture achieves the best recognition with Manhattan distance metric. With average recognition rate of 95%, the accuracy of Manhattan distance metric in classifying gesture sign based on its feature vector (i.e. centroid distance fourier descriptors) is considered high in this area of research. Total computation latency in processing square image of 800 pixels in each side is 6.0573 ms, which is equal to 165 processes per second. Moreover, with computation latency of less than 10 ms, this method is fast enough to be employed in real time application because some imager only has acquisition rate of less than 100 frames per second (e.g. Kinect sensor has 30 FPS). Future work will aim at improving the recognition rate by applying other classification methods and developing more reliable gesture dataset. Another issue will be the stability evaluation of the current method when applied in real time application.

7. REFERENCES

[1] Kwok, K. 1999. Research on the use of robotics in hazardous environments at Sandia National Laboratories. 8th International Topical Meeting on Robotics and Remote System, Pittsburgh.

[2] Ariyanto, G. 2007. Hand gesture recognition using Neural Networks for robotic arm control. National Conference on Computer Science & Information Technology, Indonesia.

[3] Wachs, J. 2007. Real-time hand gesture telerobotic system using the Fuzzy C-Means clustering, Fifth Biannual World Automation Congress.

[4] Liu, Y., Gan, Z., and Sun, Y. 2008. Static hand gesture recognition and its application based on Support Vector Machines. Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel / Distributed Computing.

[5] Conseil, S., Bourennane, S., and Martin, L. 2007. Comparison of fourier descriptors and hu moments for hand posture recognition. European Signal processing Conference (EUSIPCO).

[6] Triesch, J. and Malsburg, C.v.d. 1996. Robust classification of hand postures against complex backgrounds. IEEE International Conference on Automatic Face and Gesture Recognition, 1996, pp. 170-175.

[7] Barczak, A.L.C., Gilman, A., Reyes, N. H., and Susnjak, T. 2011. Analysis of feature invariance and discrimination for hand images: Fourier descriptors versus moment invariants. International Conference Image and Vision Computing New Zealand IVCNZ2011.

[8] Bourennane, S., and Fossati, C. 2010. "Comparison of shape descriptors for hand posture recognition in video." Signal Image and Video Processing, 2010: 1-11.

[9] Wah Ng, C. 2002. "Real-time gesture recognition system and application." Image and Vision Computing 20.13-14, 2002: 993-1007.

[10] Chen, F. 2003. "Hand gesture recognition using a real-time tracking method and hidden Markov models." Image and Vision Computing 21.8 2003: 745-758.

[11] Zhang, D., and Lu, G. 2002. A comparative study of fourier descriptors for shape representation and retrieval. Proceedings of Fifth Asian Conference on Computer Vision. 2002: 1-6.

[12] Santosa, P.I., and Tara, R.Y. 2012. Dictionary of basic hand gesture sign language. 2012 International Conference on Future Information Technology and Management Science & Engineering (FITMSE 2012), Hongkong.

[13] Microsoft Kinect. http://en.wikipedia.org/wiki/Kinect. 2012.

[14] Tara, R.Y., Santosa, P.I., and Adji, T.B. 2012. "Hand Segmentation from Depth Image using Anthropometric Approach in Natural Interface Development", International Journal of Scientific and Engineering Research Vol: 3-5, May 2012.

[15] Shih, F.Y. 2008. Image Processing and Pattern Recognition: Fundamentals and Techniques, Wiley and Sons, Canada.

[16] Zhang, D. 2002. Image Retrieval Based on Shape. PhD Thesis, Monash University.

[17] ASL Fingerspelling. http://www.lifeprint.com/asl101/fingerspelling/images/abc1280x960.png. 2011.