An Novel Hand Gesture System for ASL using Kinet Sensor based Images

[1]Manoj H. M., , [2]Pradeep Kumar B.P., [3]Anil Kumar.C.[4] Rohith.S
, 2nd author's name and surname[2], etc
{manojhm@bmsit.in[1], pradi14cta@gmail.com[2], canilkumarc22@gmail.com[3],
rohithvjp2006@gmail.com[4]}

[1] Assistant Professor, Dept of CSE, BMSIT&M, Bangalore-64., [2] Associate Professor, Dept of ECE, HKBKCE,Bangalore-45, [3] Associate Professor & HoD, Dept of ECE., RLJIT, Doddaballapur, [4]Associate Professor, Dept of ECE.NCET, Bangalore

Abstract. The prime goal of this study is to perform resolution enhancement using experimental analysis. The framework presents hand image resolution enhancement techniques based on multi scale decomposition and edge preservation smoothing. The proposed technique Dual Tree complex wavelet transform (DT-CWT) and edge preservation smoothing (EPS) algorithm images are decomposed into different sub bands and interpolated, after which sub bands are reconstructed to achieve the enhanced image. The study outcome of this phase is studied with respect to PSNR and RMSE for all the letters of ASL.

Keywords: American Sign Language, Camera, Gesture, Wavelet transform, kinet sensor.

1 Introduction

Gestures used for communicating between person and machines also between persons using sign language [5][6]. Gestures can be static (posture / certain pose) or dynamic (series of postures). The static-gestures require less computational complexity whereas dynamic-gestures are more complex. Various techniques have been developed for acquiring necessary information for gestures recognition system [4][6]. Some methods are utilized for external hardware-devices such as data-glove devices and colour-markers which can easily extract the comprehensive-description of Gesture-features. Other methods are based on the appearance of hand which segments the hand and extracts the essential features, these methods are considered as natural, easy and less cost effective than other [7][8].

For most hard of hearing people in the United States, American Sign Language (ASL) is the preferred dialect. For everyday terms, ASL employs approximately 6,000 gestures, with finger spelling for conveying dark words or structured objects, locations, or items. In ASL, communication is often based on available shapes placed in or transferred crosswise over various areas of the endorser's body, despite changes in the head and arm, as well as physical appearance [12]. In any case, proper names and words without a unified sign are spelled in English letter by letter, and ASL understudies often begin their studies by learning the 26 hand shapes that make up the manual letter range [12, 13].

I3CAC 2021, June 07-08, Chennai, India

DOI 10.4108/eai.7-6-2021.2308609

1.1 Hand Gesture Recognition System

The vision is one of the physical senses which computer is instantiated perceptibly during the communication with humans. Thus, the vision based mechanisms are considered more in HGR. The HGR system based on computer vision consists of three different steps and is shown in Figure.1. The typical process includes a) image enhancement to remove noises as a pre-processing step, b) segmenting the palm portion from the captured image of either human body or hand ,c) feature extraction and finally d) classification [14, 15].

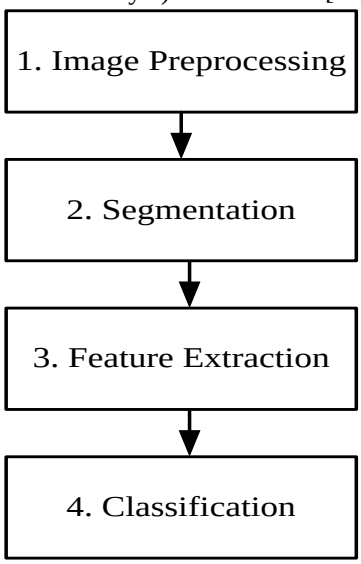


Figure.1. Main Steps of HGR System based on computer vision

1.2. Modeling Hand Gesture

The 2D modelling of hand can be mentioned with motion, shape and deformable templates. The shape based hand modelling can be classified as non-geographic and geographic models [2][3]. The non-geographic models consider the shape based features to model the hand like edges, contour, Eigen vectors etc, which are used for feature extraction and perform analysis too. The flexible/deformable models give an object shape changing flexibility level to pass the little variation of hand shape.

The 3D model of the hand can be represents and classified into skeletal, geometrical and volumetric models. The geometric models can be utilized in the real-time applications and hand animation. The skeletal models needs less parameters to structure the hand shape. The volumetric models are very complex and needs more parameters to shape the hand [14]. The geometric surfaces significantly performs the simulation of the visual hand image but it needs more parameters and is more time consuming process. The visual shapes like cylinders and ellipsoids are the alternative mechanism of geometric shape [15]. The Figure.2, describes the hand modelling mechanisms

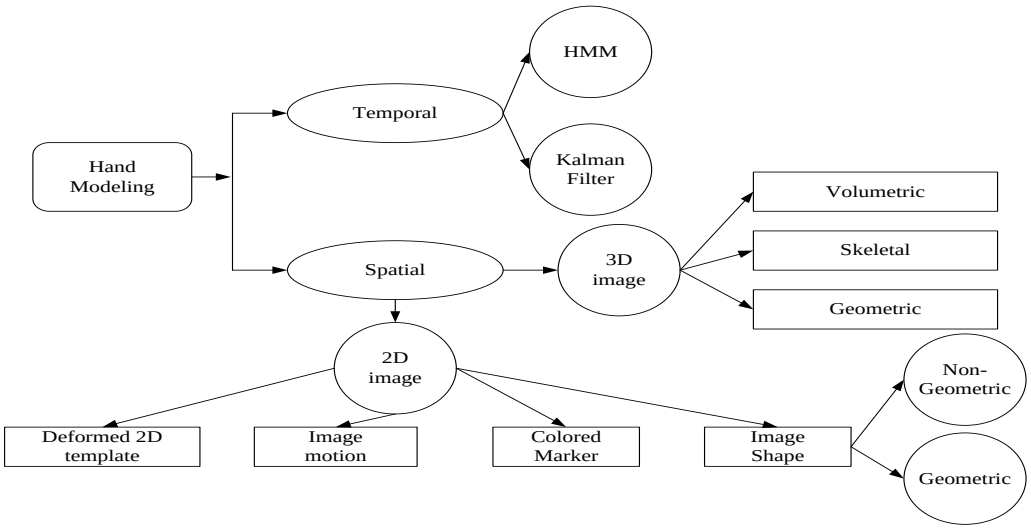


Figure.2. Hand Modelling

1.3 Research Aim And Objectives

A novel system for contact-less Hand gesture recognition using Microsoft Kinect for Xbox is described, and a real-time Hand gesture recognition arrangement is modelled and simulated into numerical computing platform (Matlab). The arrangement permits the user's to choose a situation, and it is clever to notice hand motions prepared by users. To recognize fingers, and to identify the meanings of gestures, and to show the meanings and pictures on screen. The prime aim of the proposed research work is to design a simple framework that can offer enhanced performance of hand gesture recognition system in effective manner considering ASL. In order to accomplish this research goal, following objectives were set:

- ***Preprocessing:*** To design a model that can offer enhanced resolution for input images
- ***Feature Extraction:*** To develop a simple modeling for hand gesture recognition emphasizing on an efficient feature extraction.
- ***Classification:*** To develop a hybridized scheme of hand gesture recognition for increasing recognition performance.
- ***Optimization***: To apply optimization for enhanced performance of hand gesture recognition system in cost effective manner.

2 Resolution Enhancement Of Hand Gesture Images

The hand gesture recognition (HGR) system which mainly emphasizes on the limitations of traditional hand gesture recognition techniques. This section mainly argues two-level architecture for the real-time hand gesture recognition scheme using only one camera as the input device. second describes the resolution enhancement problem and technique for hand gesture images using four different algorithms namely 1) The Nearest Value Algorithm, 2)

The Bilinear Algorithm , 3)The Bi-cubical Algorithm , 4) Dual Tree Complex Wavelet Transformation (DTCWT), further Bilateral Filter is used in order to obtain enhanced performance.

2.1 Skeleton Identification Of Kinect Camera

In this proposed system, "Kinect camera" plays the major role to gather the depth information from the skeleton. The new version of Kinect with its SDK (Software Development Kit) containing the skeleton tracking tool. This unique tool provides the system to collect the 20 joint information of the human body. For each frame, the positions of 20 points are estimated and collected. The 20 joints which is taken as an reference points is as shown in Fig 3

The kinect device provides the both RGB and D-image. This camera utilizes a structured light method to generate the real time depth information which consists of discrete measurements of physical scene. In this study, first creating the depth images of human in different sizes, and shapes, and generate the big dataset. The RGB and D-image are the input images of the system for the recognition of different ASL alphabetic symbols. This skeletal tracker device able to track the skeleton image of one or more persons moving within the kinect area view for gesture driven applications i.e. this tool able to collect 20 joint information about the skeleton. From the skeleton tracked by the Kinect first it extracts the feature of joint positions. Since, each joint has 3 values and also 3 coordinates and the detected skeleton has of 20 joints. So, the feature vector has 60 dimensions. The position of 20 points identified by kinect afterwards it will segment the right hand posture to recognize and stored in the database. From this sets can select the wanted joints of images for representing the postures.

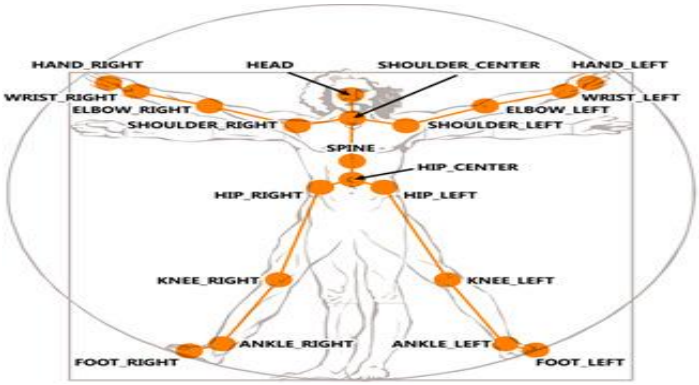


Figure. 3 Human skeleton joints as reference points

2.2 Algorithm: Skeleton identification from Kinect-Sensor.

Input: one or more people image, Output: 20 joint images

a. start
b. capture the image from kinect camera
c. segment the image from skeleton viewer
d. for reference points in body portion finds the depth-information
e. segmented moving body portion is mapped to the skeletal co-ordinates
f. if more than one moving body portion presents
g. calculate skeleton connection map for x,y coordinate
h. display the multiple skeleton moving body
i. end of predefined frames
j. stop.

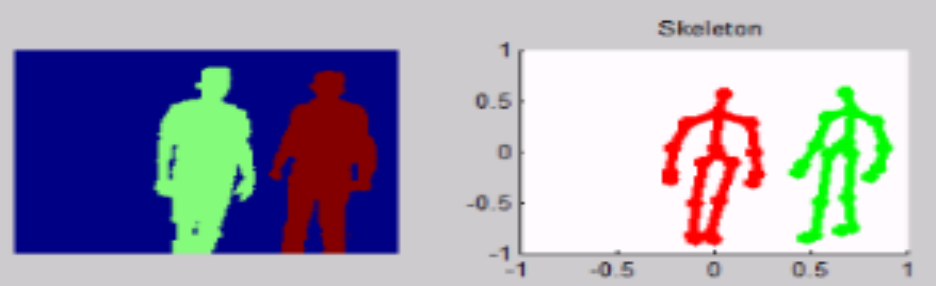


Figure.4. Skeleton-image identification from kinect sensor

Figure 4 shows the outcomes of identifying multiple skeletal signs under different distance using kinet-sensor device which represents the colour and depth image. Figure 5 shows in multiple user environments also our proposed system is identifying the user hand with respect to the distance from camera and extract the hand sign clearly for storing in to database.

Figure. 5. Multiple skeletal sign recognized under different distance

2.3. Multiple Depth Recognition:

(i) To calculate centroid: Here system is considering the three co-ordinates X, Y and Z. Calculating the centroid for each axis independently the mathematical interpretation for centroid is given by

X= sum (x) / length(x)..... .....… 1

Y= sum (y) / length (y)…….....2

Z= sum (z) / length (z)…….....3

(ii) **To calculate mean:** The system will calculate the mean for the entire segmented region calculated by the centroid of the body part by pixel basis. The mathematical interpretation for mean for pixels is given by:

$\bar{x}= \frac{1}{K}\sum_{i=0}^{K} x(i)$ and $\bar{y}= \frac{1}{K}\sum_{i=0}^{K} y(i)$…….....4

The flow graph in figure 6 shows for the multiple depth recognition using Kinect sensor is as follows

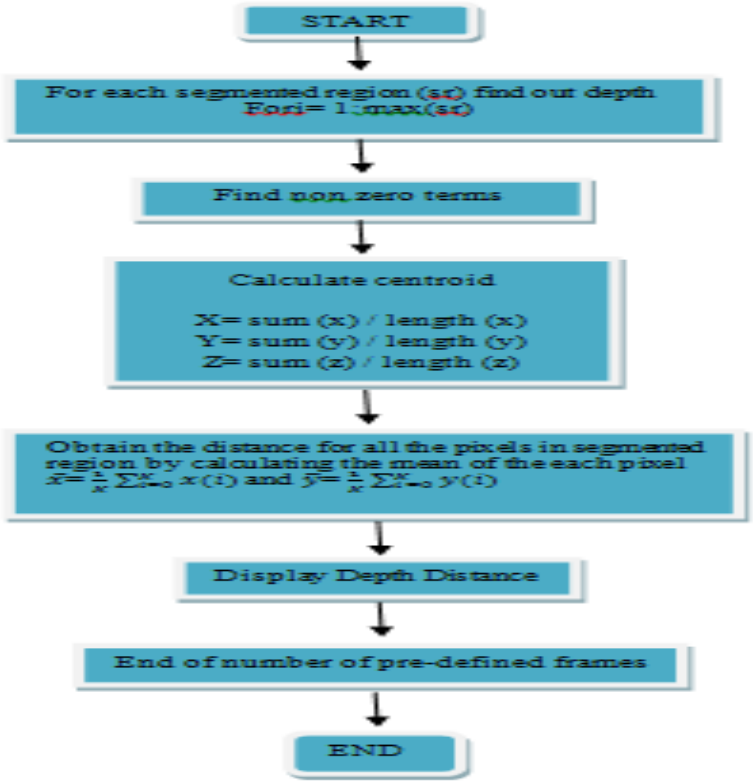


Figure. 6 The flow graph for the multiple depth recognition

2.4. Data Acquisition and Pre-Processing

The raw information acquired from the Kinect sensor via the Natural User Interface (NUI) contained 512×424 depth data, 1920×1080 RGB data, and 26-joint body skeleton data. The hand region in the depth image was illustrated using spatial thresholds in X-axis direction [Tx_min, $Tx\_$], Y-axis direction [Ty_min, Ty_max] and Z (depth)-axis direction [$TDepth\_min$, $TDepth\_max$]. As demonstrated in Figure 6, the Kinect depth sensor placed at position $\boldsymbol{S}$ has angles of view α (horizontal) and β (vertical). The declaration of the depth image is $RxbyRy$ pixels. The position of the "hand" joint (x, y, D) in the depth image can be attained from the Kinect skeleton data (Figure 8a). Thus, the spatial thresholds are illustrated as:

$$[T_{x\_min}, T_{x\_max}] = [x - \frac{d_x}{2}\frac{\frac{R_x}{2}}{Dtan\frac{\alpha}{2}}, x + \frac{d_x}{2}\frac{R_x/2}{Dtan\frac{\alpha}{2}}] \quad \text{.....5}$$

$$[T_{y\_min}, T_{y\_max}] = [y - \frac{d_y}{2}\frac{R_y/2}{Dtan\frac{\beta}{2}}, y + \frac{d_y}{2}\frac{R_y/2}{Dtan\frac{\beta}{2}}] \quad \text{.....6}$$

$$[T_{Depth\_min}, T_{Depth\_max}] = [D - \frac{d_z}{2}, D + \frac{d_z}{2}] \quad \text{.....7}$$

where dx , dy and dz are stable dimensions (in millimetres) of the hand's region. The hand's region in the depth image is revealed in Figure 7. The hand's region in the color image can also be gained by mapping the hand's region on top of the color image (Figure 8 a).

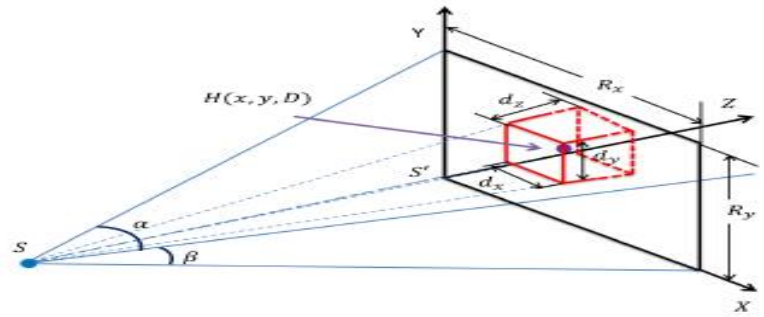


Figure. 7.. Illustration of the hand region segmentation: the$dx \times dy \times dz$hand region at
(x, y, D) was segmented from the$Rx \times Ry$depth image obtained using a depth sensor located at the position S.

(a) (b)

Figure. 8. Illustration of data obtained using Kinect.
(a) RGB Color image of the hand region. (b) Depth image of the hand region.

The following table illustrates the result analysis of proposed system can measure the recognition accuracy in different distance ranges. As sample we are experimenting skeletal signs as 10 times with different distance like 850mm to 1000mm …3000mm to 3500 mm and evaluating the recognition accuracy with time. The recognition accuracy is calculated in terms of percentage like for first experiment we are considering the (skeletal and camera) distance as 850mm to 1000mm and getting the 70% of recognition accuracy. Like this from the experiment analysis results, we can get the following results, which is shown is following table 1.

Table. 1 analysis results for recognition accuracy calculation

Distances from Kinect In mm	Number of times checked	Number of times recognised	Recognition Accuracy in %
850-1000	10	7	70 %
1000-1500	10	10	100 %
1500-2000	10	10	100 %
2000-2500	10	10	100 %
2500-3000	10	10	100 %
3000-3500	10	8	80 %

3. Framework For Resolution Enhancement

In this research methodology using experiential analysis image enhancement is executed as shown in fig 9. The techniques used here are Edge preservation smoothing and multi-scale decomposition. To obtain the enhanced image, the method of DT-CWT and EPS algorithm images are decomposed to a different sub band. At a regular interval of 5 minutes, samples of sign language are recorded from Kinect camera at a distance of 1500-2000mm. Values of PSNR, RMSE, are used to further proceed with quantitative analysis. The output of this phase is an enlarged image.

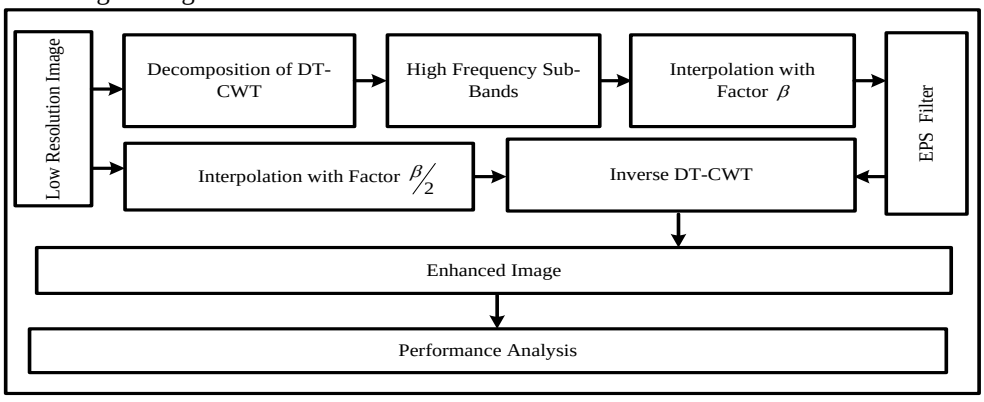


Figure.9. Block diagram representing framework for resolution enhancement

3.1. The Nearest Value Algorithm

Nearest Neighbour interpolation or proximal interpolation is way by which multiple dimensions can be interpolated. For a random point in space surrounding the nearest point (Neighbouring) would have its value approximated leading to the interpolation problem. Hence, the algorithm of the nearest neighbour selects the nearest point and neglects the value of points around it.

The above mentioned is the nearest neighbour algorithm which selects the nearest point eliminating the value associated to the random points around it. The original disk on file (I) is initialized and the size of the resized image is computed (I_{new}).Now, the size of the original file on disk is calculated. Further to check a condition that size of I and I_{new} match or not the number of rows and columns are compared respectively. If the resized image has higher number of rows then its new value will be ranging from 1-to-rows of original image considering first row and first column. If the original image has higher number of rows then its new value will be ranging from 1-to-rows of resized image considering first row and first column. Same procedure is applied to check the condition for the columns. The value of PSNR is calculated to be 23.3591, RMSE is 17.3214 for the input image as shown in fig 10.

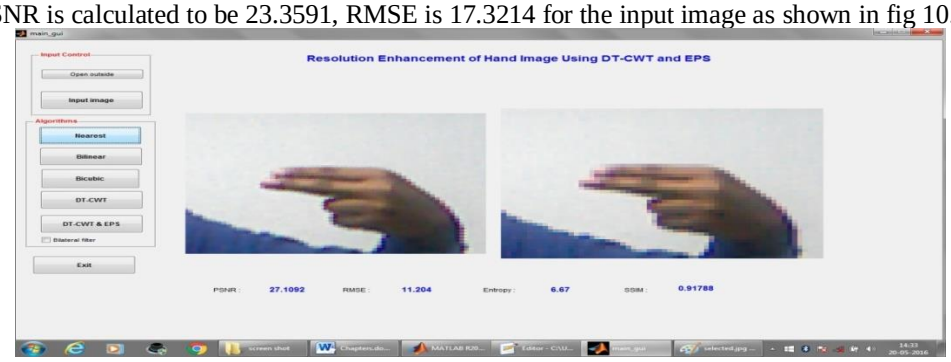


Figure.10. Application of Nearest Neighbor Transformation Function

3.2. Bilinear Interpolation

An image transformation process used in cases where pixel matching is impossible is called as bilinear interpolation. When compared to other methods of transformation bilinear interpolation considers closet 2x2 neighbourhood of known pixel values surrounding the unknown pixel's computed location.

Computation of PSNR $(\beta) = 10\log_{10}\left(\frac{255}{\alpha}\right)$; Computation of RMSE $(\gamma) = \sqrt{MSE}$

The above mentioned is the bilinear interpolation method which considers 2x2 neighboring pixels. The original disk on file (I) is initialized and the size of the resized image is computed (I_{new}).Now, the size of the original file on disk is calculated. Further to check a condition that size of I and I_{new}match or not the number of rows and columns are compared respectively. If the resized image has higher number of rows then its new value will be ranging from 1-to-rows of original image considering first row and first column. If the original image has higher number of rows then its new value will be ranging from 1-to-rows of resized image considering first row and first column. Same procedure is applied to check the condition for the columns. The value of PSNR is calculated to be 23.591, RMSE is 17.3214 for the input image as shown in fig 11.

Figure. 11. Application of bilinear Transformation Function

3.3. Bicubic Interpolation.

This technique is implemented using Lagrange polynomials cubic splines or cubic convolution algorithm. Bicubic interpolation can be chosen over other methods if speed is not a constraint. Smoother images are obtained as output with lesser interpolation artifacts. The above mentioned is the bicubic interpolation method which is based on cubic convolution algorithm, Lagrange polynomials. The original disk on file (I) is initialized and the size of the resized image is computed (I_{new}).Now, the size of the original file on disk is calculated. Further to check a condition that size of I and I_{new}match or not the number of rows and columns are compared respectively. If the resized image has higher number of rows then its new value will be ranging from 1-to-rows of original image considering first row and first column.If the original image has higher number of rows then its new value will be ranging from 1-to-rows of resized image considering first row and first column. Same procedure is applied to check the condition for the columns. The value of PSNR is calculated to be 24.953, RMSE is 17.3214 for the input image as shown in fig 12.

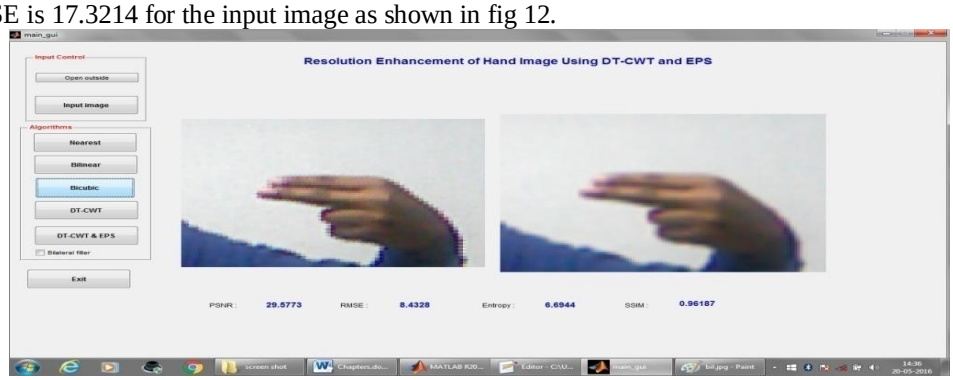


Figure.12. Application of bicubic Transformation Function

3.4. Bilateral Filtering Process

Bilateral filter is a noise reducing filter for images with the property of non-linearity and preserving the edge. Each pixel has an intensity value a picture restored by weighted average from nearby pixels. Weight can be based on Gaussian distribution and the sample output is shown in fig 13.

The bilateral filter is defined as:

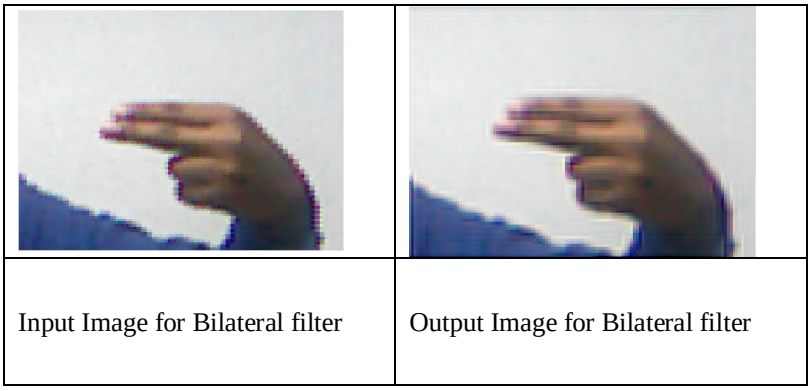
$$I^{filtered}(x) = \frac{1}{W_p} \sum_{xi \in \Omega} I(x_i)\, fr(\| I(x_i) - I(x) \|) g_s(\|x_i - x\|) \quad \ldots\ldots 8$$

Where the normalization term;

$$W_{p=} \sum_{xi \in \Omega} fr(\| I(xi) - I(x) \|) gs (\| xi - x \|) \quad \text{.....9.}$$

Ensures that the filter preserves image energy and

- $I^{filtered}$ is the image after filtration: Iis the original input image
- xare the directs of the current pixel to be filtered: Ω is the window centered in x
- f_r is the range kernel for smoothing differences in intensities. This utility can be a Gaussian function
- $g_s$is the spatial kernel for smoothing differences in coordinates. This function can be a Gaussian function.



Input Image for Bilateral filter	Output Image for Bilateral filter

Figure 13.Application of bilateral filtering process

3.5. Dual Tree Complex Wavelet Decomposition Transformation (DTCWT)

Using the dual tree complex wavelet transformation method the image is decomposed. This happens with respect to Discrete Continuous Wavelet Transformation (DWT) and Continuous Wavelet Transformation (CWT) .In DWT the basis function used is symlet mother wavelet. Image will be decomposed into two parts, the approximation coefficients and detailed coefficients. Only approximation coefficients are considered and similar mechanism is implemented for CWT. The algorithm of DTCWT is being performed above for the resolution enhancement for a hand gesture by splitting the image mainly into real and imaginary parts. To evaluate the analysis and synthesis parameter taken into consideration, the dual filter function is worked on. The dual cell structure is determined via cplx dual 2D function. Frequency values are divided into lower and higher components. The higher frequency components are normalized to nullify the effect of frequencies lying outside the desired range of detection. The lower frequencies have their highest value used in the algorithm among all of them. Further lower frequency image is converted into original image with the inclusion of the new dual tree cell structure. for the input image as shown in fig 14.

Figure.14. Application of DTCWT and EPS algorithm

Table.2. Comparison of algorithms for sign H

Algorithms	PSNR	RMSE
Nearest	26.24	12.37
Bilinear	28.08	10.01
Bicubic	28.11	10.11
DT-CWT	28.50	13.8
DT-CWT & EPS	29.07	10.006

Table 3 showing the detailed tabulated parametric values for all the signs using DT-CWT conditions respectively considering performance parameters of PSNR, RMSE

Table.3 Numerical Outcome of Performance parameters of DT-CWT

ALPHABET	A	B	C	D	E	F	G	H	I
PSNR	28.5	29.61	30.35	28.96	29.02	29.14	28.84	26.24	30.77
RMSE	13.8	8.66	8.63	10.49	10.9	10	12.38	12.37	8.44

ALPHABET	J	K	L	M	N	O	P	Q	R
PSNR	28.82	29.45	28.45	28.33	29.43	29.77	30.12	29.25	29.43
RMSE	12.15	12.15	11.15	9.6	11.56	10.32	7.85	11.9	8.43

ALPHABET	S	T	U	V	W	X	Y	Z
PSNR	29.59	29.89	30.45	27.78	27.32	27.55	28.56	26.72
RMSE	8.8	8.99	9.2	12.73	9.34	13.92	12.44	14.6

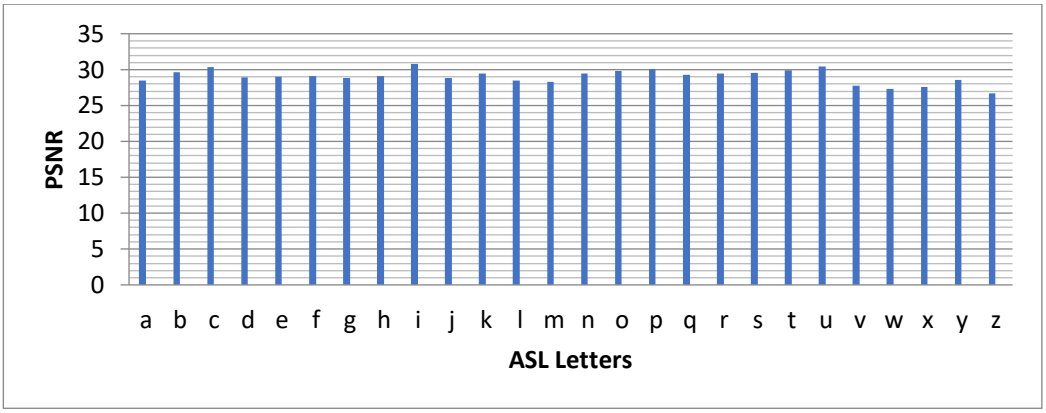


Figure.15. Illustration of PSNR values for the algorithm DT-CWT and EPS.

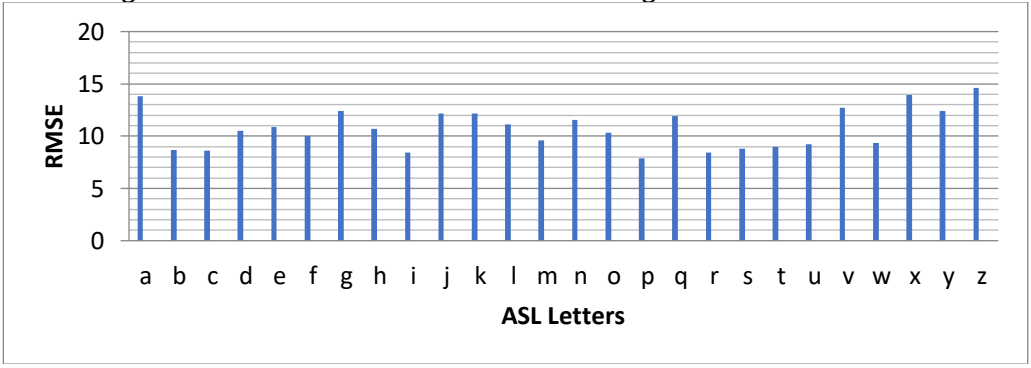


Figure.16. Illustration of RMSE, values for the algorithm DT-CWT and EPS.

4. Conclusion.

Thus, this study quickly abridges about every thing of the calculations being executed for the upgrade of a hand signal acknowledgment framework alongside the examination procedures required behind it. A novel picture determination improvement procedure in view of DT-CWT and EPS channel. The method breaks down the LR input picture utilizing DT-CWT. EPS (Bilateral) sifting is utilized to safeguard the edges and de-noising the picture and to additionally improve the execution of the proposed method as far as RMSE, PSNR

References

[1] A.R. Sarkar, G. Sanyal, and S. Majumder, "Hand gesture recognition systems: a survey", International Journal of Computer Applications, 71.15, 2013

[2] N. Neverova, Natalia, "A multi-scale approach to gesture detection and recognition", Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013

[3] S. Yang, P. Premaratne, and P. Vial, "Hand gesture recognition: An overview", Broadband Network & Multimedia Technology (IC-BNMT), 5th IEEE International Conference, 2013

[4] T. Osunkoya and J-C. Chern, "Gesture-based human-computer-interaction using Kinect for Windows mouse control and Powerpoint presentation", Department of Mathematics and Computer Science, Chicago State University, Chicago, IL 60628, 2013

[5] J. Katkar, Jayshree, "Hand Gesture Recognition and Device Control", Hand, 2017
[6] J.G. Kyle and B. Woll,“Sign language: The study of deaf people and their language”, Cambridge University Press, 1988
[7] M.C. Thomas and A. P. M. S. Pradeepa,“A comprehensive review on vision based hand gesture recognition technology”, International Journal 2.1, 2014
[8] H. Zhouand H. Hu,“A survey-human movement tracking and stroke rehabilitation”, University of Essex, Colchester United Kingdom, 2004
[9] M. Turkand M.Kölsch,“Perceptual user interfaces”, Emerging Topics in Computer Vision, Prentice Hall, 2004
[10] K.K. Vyas, A. Pareek, and S. Tiwari, "Gesture Recognition and Control", International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, Retrieved on 16th August, 2017
[11] D.H. Stefanov, Z. Bien, and W-C. Bang,“The smart house for older persons and persons with physical disabilities: structure, technology arrangements, and perspectives”, IEEE transactions on neural systems and rehabilitation engineering, vol.12.2, pp.228-250, 2004
[12] D.C. L-Martin, “Universal grammar and American sign language”, Universal Grammar and American Sign Language, pp.1-48, 1991
[13] M.L. McIntire, “The acquisition of American Sign Language hand configurations”, Sign Language Studies, 16.1, pp.247-266, 1977
[14] M. Alsheakhali, “Hand gesture recognition system”, Computer Engineering Department, The Islamic University of Gaza, Gaza Strip, Palestine, 2011
[15] S.D. Badgujar, “Hand Gesture Recognition System”, International Journal of Scientific and Research Publications, 4.2, 2014