Real-time Disparity Computation and 3D Reprojection for Hand Gesture Recognition

Mohidul Alam Laskar*, Amlan Jyoti Das†, Anjan Kumar Talukdar‡ and Kandarpa Kumar Sarma§
*†‡Department of ECE, Gauhati University, Guwahati-781014, India
Email: mohid.india@gmail.com, amlandas78@gmail.com and anjan.nov@gmail.com
§Department of ECT, Gauhati University, Guwahati-781014, India
Email: kandarpaks@gmail.com

***Abstract*—Human hand gestures as a natural way of interaction and communicating with computers is becoming an emerging and demanding field of research due to its various applications like sign language recognition, human computer interaction, gaming, virtual reality etc. Hand gesture recognition under 2D environment has some limitations as the information about the other dimension (z-axis) is missed. So, hand gesture recognition under 3D environment is becoming a growing field of research. In this paper, we have implemented the realtime disparity computation and proposed a novel technique to detect gesture of Front and Back along with forward and backward movement towards and from the camera respectively. Our technique is based on stereo-vision and we have used the disparity map-based intensity measure of segmented hand and changing of its intensity as feature to classify and recognize the gesture. Stereo calibration followed by rectification is done to get the rectified images and correspondence gives the depth map. Finally three dimensional reprojection is performed. Our technique works well for the gestures and the results are promising.**

Keywords*—Stereo vision, stereo calibration, stereo matching, disparity map, reprojection.

I. Introduction

Vision-based hand gesture recognition is challenging but demanding and have very wide prospects, mainly because of broad range of configurations and aspects [1]. Object recognition or identification, tracking and classification, especially of hands, has been investigated extensively in the research fraternity throughout the world. It has become popular and an influential technology for future world, the study and research of which has various emerging application outlook with marketable importance, hand gestures offer a very fitting and natural way of interacting with an autonomous robot or a computer. Hand gesture is a spatio-temporal pattern [2] and can be static and dynamic or both [3]. The challenges encountered with 2D recognition have encouraged researchers to study 3D gesture recognition where we can extract the depth information also. Mainly two implemented techniques are there by which a 3D gesture can be classified and recognized, one is by Kinect which uses range vision [1] and the other is by stereoscopic vision. Kinect is a combination of a RGB and an IR sensor [4] whereas two RGB sensors are used in a stereoscopic camera. There are applications which could benefit from the stereo vision-based hand gesture recognition such as pervasive computing, augmented reality, compute gaming etc. Stereo vision technique can be also used in industrial automation and surveillance systems. In our approach we have used a stereo camera for 3D hand gesture classification as it is inexpensive and easily available in market. In this paper, the hand gestures of "front" and "back" along with forward and backward movement towards and from the camera are taken for implementation.

Hand gesture recognition systems has been studied extensively and implemented by many enthusiasts and researchers in the past decade throughout the world. In dense stereo algorithms, the disparity at every image point needs to be calculated and has been discussed in [5][6]. As mentioned in [5], for a typical stereo algorithm to compute a disparity map, following are the four steps performed: Matching cost computation, cost aggregation in a support region, optimization and computation of disparity which is followed by disparity refinement. Approaches followed for these algorithms are either local or global. Firstly, local algorithms are window based approach which employs correlation among left and right images where disparity is computed by similarity measurement of pixels or the lowest aggregate of cost matching within a window is calculated [7]. Secondly, in global algorithms, the disparity map is obtained by minimizing a global cost function which is defined in terms of smoothness constraint and image data [8]. Local algorithms are computationally more efficient and robust with respect to global algorithms. Our system prerequisite is to detect the hand from the scene for which we looked for color based segmentation. In [9] color based segmented is utilized in HSV color space. A detailed discussion on color based skin segmentation is there in [10]. Detection of face portion is done in [11][12], where the face is detected using haar cascade classifiers. Stereo vision techniques such as stereo calibration, stereo rectification, stereo correspondence and stereo reconstruction are discussed thoroughly in [13][12]. The disparity computation based on stereo matching algorithm is reported in [14] and is provided in OpenCV. In [15], implementation and evaluation of gesture based on discrete Hidden Markov Models is given. CRF implementation for Indian Sign Language recognition is done in [16]. Sign Language Spotting using Conditional Random Fields is discussed in [17].

In most of the existing literature, the gesture recognition deals with the detection of hand part. So initially we detected hand part using color based segmentation techniques. Stereo calibration and rectification is performed to get the camera parameters and hence rectified images of left and right camera. The main objective of our work is to implement a resource limited 3D hand gesture recognition system to classify the Front and Back gestures along with forward and backward movement towards and from the camera respectively by applying stereo vision techniques with the help of a inexpensive stereo camera. Our

978-1-4673-6708-0/15/$31.00 ©2015 IEEE

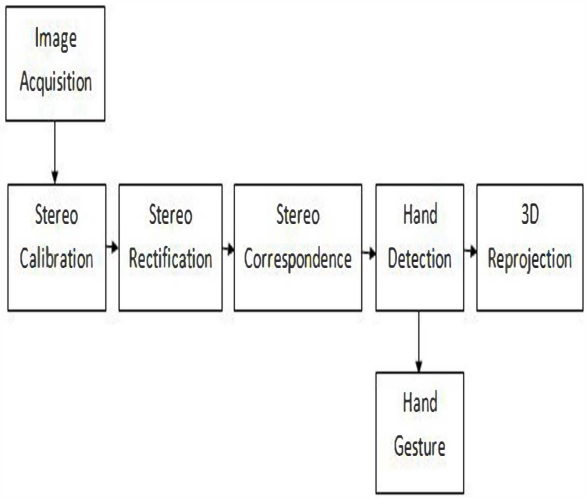


Fig. 1. Block Diagram of proposed method

technique is based on stereo-vision and we have used the disparity map-based intensity measure of segmented hand and changing of its intensity as feature to classify and recognize the gesture. Stereo calibration followed by rectification is done to get the rectified images and correspondence gives the depth map. Finally three dimensional reprojection is performed. Our technique works well for the gestures and experimental results show that the proposed approach is robust and reliable. Section II represents the proposed method. Implementation is given in Section III an results are shown in Section IV.

II. Proposed Method

In a typical stereo vision technique we know the relative positions of the cameras, where we find out the differences in the images observed by the cameras in a similar manner how our own vision mechanism functions with the two eyes. It can either be: (i) Passive, where features viewed by camera are matched under natural lightning conditions, or (ii) Active, by analyzing artificial texture on the scene for improving the feature matching. Triangulation principle is used in both cases to produce the depth map. The main objective is to map the depth data onto an RGB image so that the foreground hand image from background objects can be segmented. Computers do the task of stereo imaging by finding the correspondence between points that are viewed by one camera and the same points as seen by the other camera. With a known baseline separation between cameras and correspondence, 3D location of the points can be computed. The brief methodology of the proposed method is shown in Figure 1. The following steps are implemented for calculating a realtime disparity map from the input left and right images from a stereo webcamera.

1) Stereo Calibration
2) Stereo Rectification
3) Stereo Correspondence
4) Reprojection

A. Stereo Calibration

Stereo Calibration technique is used to calculate the geometrical relationship between the two cameras. It is a onetime process and generates a set of parameters which are applied to correct distortion in images due to lens and camera. Mathematical derivation is available in [12][18]. The rotation matrix (R) and translation (T) vector can be derived from equations:

$$R = R_r(R_l)^T \tag{1}$$

$$T = T_r - RT_l \tag{2}$$

Where, R_l and R_r are rotation matrices whereas, T_l and T_r are translation vectors of left and right cameras respectively.

B. Stereo Rectification

Stereo rectification is a process which deals with correction of the individual images so that they appear as if they are taken by two cameras with row-aligned image planes. Two popular algorithms to carry out stereo rectification are Hartley's algorithm and Bouguet's algorithm. Bouguet's algorithm is preferred over Hartley's algorithm as it minimizes the reprojection error because of its greater accuracy [12].

C. Stereo Correspondence

Stereo Correspondence is a process which yields a disparity map. The disparities are actually differences in x-coordinates of the same feature on the image planes viewed in the left and right cameras. Based on the horizontal distance between the locations of a point in the two images and a set of predefined constants, a disparity image is generated. Once the physical coordinates of the cameras or the sizes of objects in the scene are known, depth measurements can be derived from the triangulated disparity measures.
Block Matching Algorithm: The algorithm was proposed and developed by Kurt Konolige [14]. The block matching algorithm was preferred over Graph Cut algorithm in view of the fact that the former is much faster than the later. Here, the strongly matching or high-texture points between the two images are found out. Block matching algorithm is a local algorithm which utilizes small "sum of absolute difference" (SAD) windows to locate matching points between the left and right rectified stereo images. SAD window $C(x, y, \delta)$ is mathematically represented as [19],

$$C(x, y, \delta) = \sum_{y=0}^{wh-1} \sum_{x=0}^{ww-1} |I_R(x, y) - I_L(x + \delta, y)| \tag{3}$$

where, wh is the window height, ww is the window width, δ is the differential value, $I_L(x + \delta, y)$ and $I_R(x, y)$ are the image intensities corresponding to left and right images.

D. Reprojection

Now we have the geometric arrangements of the cameras which is obtained during camera calibration. By applying the triangulation, the disparity map can be turned into physical parameters i.e. in distances where the output is a depth map. Given a disparity d and a two dimensional point (x, y), it can be projected in a three dimensional view by using:

$$Q \begin{bmatrix} x \\ y \\ d \\ 1 \end{bmatrix} = \begin{bmatrix} X \\ Y \\ Z \\ W \end{bmatrix} \tag{4}$$

where,

$$Q = \begin{bmatrix} 1 & 0 & 0 & -c_x \\ 0 & 1 & 0 & -c_y \\ 0 & 0 & 0 & f \\ 0 & 0 & -1/T_x & (c_x - c'_x)/T_x \end{bmatrix} \quad (5)$$

is called the reprojection matrix obtained during the process of camera calibration and is the inverse of the projection matrix. Here, c_x and c_y are the principal points and two dimensional coordinates are $(x/w, y/w)$ in x-coordinate and y-coordinate respectively. T_x is the translation vector used in Q-matrix. The three dimensional coordinates are given by $(X/W, Y/W, Z/W)$ and the depth Z can be calculated by:

$$Z = \frac{fT}{(x_l - x_r)} \quad (6)$$

where, f is the camera focal length, T is the baseline separation between cameras, x_l and x_r are the position of left and right points respectively.

III. Implementation

The detailed implementation steps of the proposed realtime system undergone are illustrated as follows:

A. Pre-processing

Preprocessing steps of any video processing algorithm is essential as we only deal with a particular region of interest. Thus in our system preprocessing is done at the prior and the flow diagram of the preprocessing steps are given in Figure 2. The input video is fed from an inexpensive stereo webcamera with a frame rate of 10 frames per second with a width-height resolution of 320×240. The face part of the operator is detected using haar cascade classifier which is inbuilt in OpenCV as we only need the hand palm which is our region of interest. Color based segmentation is used where the frames are converted to HSV space and thresholding is used to get the binary hand mask. This is followed by morphological closing and opening operations to filter out the noise present and to fill up the holes and thereby we get a precise segmented hand mask. Simultaneously, on the other side, the RGB frames from the stereo camera are converted to Gray frames. The hand mask is then multiplied (logical AND operation) with gray frames to get an image where only gray values of hand are present.

B. Disparity Computation

The frames obtained from the pre-processing steps are not row-aligned thus rectification is necessary as it removes the vertical disparity present thereby making both image frames row-aligned. An image of 9×6 chessboard is taken to compute the camera calibration. A total of 30 different pairs of images are taken with chessboard held in different inclinations. For each pair, Rotation and Translation matrices are calculated using *cvStereoCalibrate()* function. The generated matrices are then approximated into one (R,T) pair with minimum error of chessboard corners in left and right camera view. *cvStereoRectify()* function corresponding to Bouguets algorithm is used and output of the function is fed to *cvInitUndistortRectifyMap()* which completes the stereo rectification of left and right images. These calibration matrices are saved as .xml files for further uses as it a one time operation. Stereo Matching is

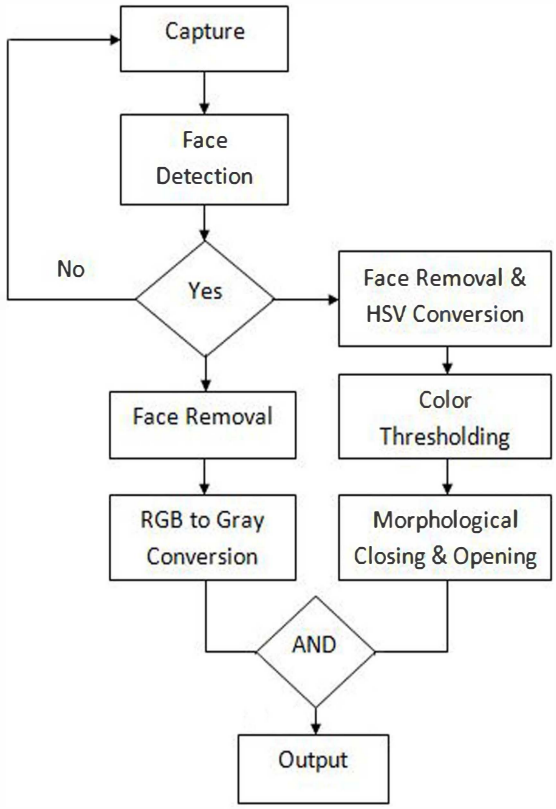


Fig. 2. Flow diagram of pre-processing of stereo images

performed by Block Matching technique to get a disparity map of hand. The function *cvFindStereoCorrespondenceBM()* computes and gives the realtime disparity map. The block matching parameters used in experimentation are SADWindowSize=19, preFilterSize=19, minDisparity=0, numberOfDisparities=128, textureThreshold=12 and uniquenessRatio=3.

C. Post-processing

The post-processing deals with refinement of disparity map. It is first normalized and then smoothed by a 5×5 median filter. Further a morphological grayscale closing operation is done by an rectangular structuring element of size 11×11 to make the disparity map more continuous. The disparity computation is a realtime process and the flow diagram is given in Figure 3. 3D reprojection of disparity is implemented with the help of a point-cloud viewer.

D. Gesture Classification

Front and Back Classification: The output of post-processing steps gives out a disparity map of hand. The "Front" and "Back" gesture is computed by taking the account of change of average pixel value of the disparity map with front and back movements. Two threshold values are set for classification of Front and Back hand gestures. A Front gesture is only classified when the average intensity values are greater than the threshold value $T2$. A Back Gesture is classified only when the average intensity values are less than the threshold value $T3$. Among the two threshold values set the $T2$ is always greater

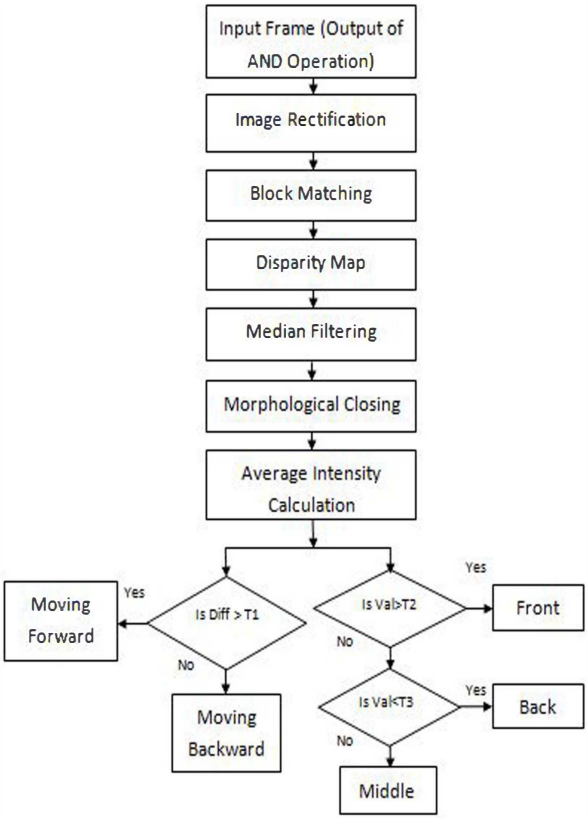


Fig. 3. Flow diagram of real-time disparity computation and gesture classification

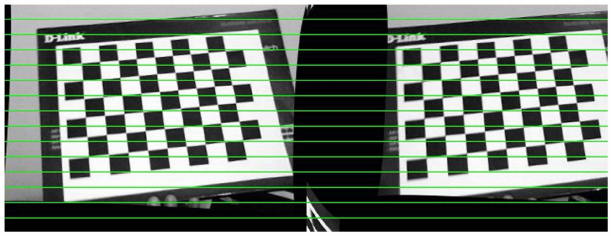
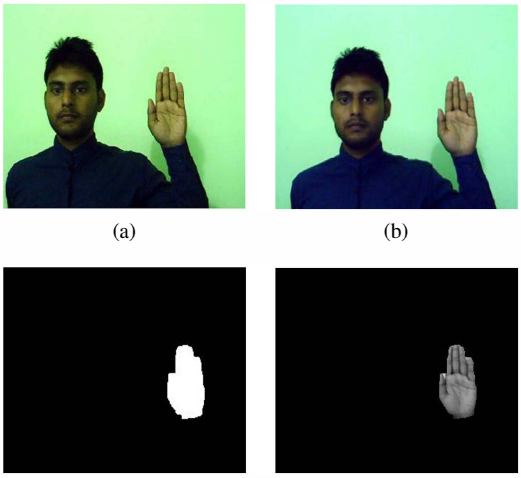


Fig. 4. Rectified left and right chessboard images



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

Fig. 5. (a) RGB image of left camera, (b) RGB image of right camera, (c) Color based segmented image, (d) Skin tone Gray image

than $T3$. There is a difference between the thresholds which is called a t and given by:

$$t = T2 - T3 \tag{7}$$

This difference threshold t, is created and kept deliberately to extract out distinct Front and Back gestures.

Moving Forward and Moving Backward Classification: Moving Forward and Moving Backward Classification is calculated by frame by frame substraction of average intensity values in front and back movements of hand. A threshold value $T1$ is set for the decision making. The gesture classification in realtime is given in the flow diagram Figure 3.

IV. Results

The input frames from left and right camera are taken as RGB images and shown in Figure 5(a) and Figure 5(b) respectively. These images are not row aligned, thus calibration and rectification is done to make them row aligned. The rectified left and right chessboard image is shown in Figure 4. Figure 5(c) shows color based thresholded image which corresponds to a hand mask. The resulting image of AND operation which corresponds to a gray skin tone of hand is shown is Figure 5(d). Realtime disparity computation results along with the respective left and right images are shown in Figure 6.

Results of "Front" gesture in "Moving Forward" is shown in Figure 7, while for "Back" gesture in "Moving Backward" condition is shown in Figure 8. We have tested the forward and backward movements of hand for 10 sample videos and tabulated the number of counts. There are two possible cases, in first case we have taken the test videos where the hand movements starts from back to front. In second case, the test videos are taken where the hand movements starts from front to back. The number counts of the classified results obtained after testing for the forward movement case is given in Table I and for the backward movement case is given in Table II.

The change in intensity value with the distance between the hand and camera for both "Front" gesture and "Back" gesture for 10 samples is shown in Figure 9(a) and Figure 11(a) with an average signature is shown in Figure 9(b) and Figure 11(b) respectively. The reprojection of the disparity map is shown in Figure 10.

The entire software is implemented using OpenCV 2.4.9 library and PCL 1.6.0 library on C/C++ programming in Visual Studio 2010 on a PC with Intel Pentium B970 2.30 GHz CPU and 4GB RAM with pre-installed Windows 7 Ultimate operating system. The stereo video frames are captured using Minoru 3D webcam.

V. Conclusion

In this paper, we have proposed a real-time disparity computation method for hand gesture in a 3D environment for "Front" and "Back" gestures with moving forward and backward conditions. Color based segmentation as a preprocessing is implemented to get the hand image out of the moving video inputs. Stereo vision techniques are applied which resulted in a disparity map. The disparity is then reprojected in a three dimensional view with the help of a point cloud viewer. The experimentation with the gesture classification are tested

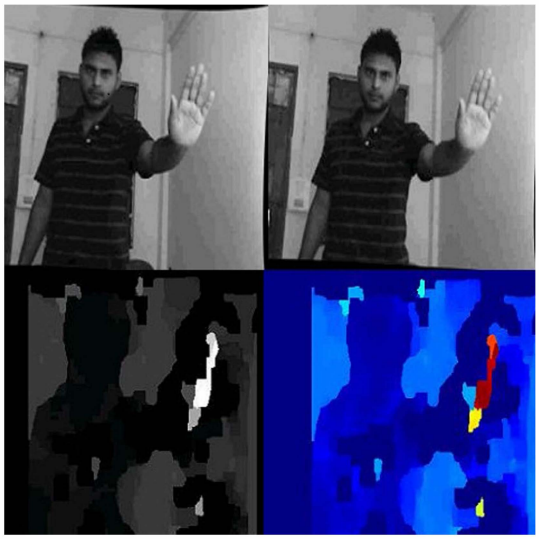


Fig. 6. Realtime disparity computation of left and right images with a colormap of depth map

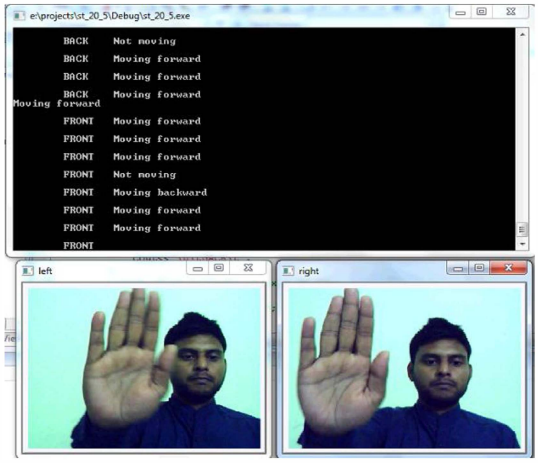


Fig. 7. Front gesture with moving forward condition

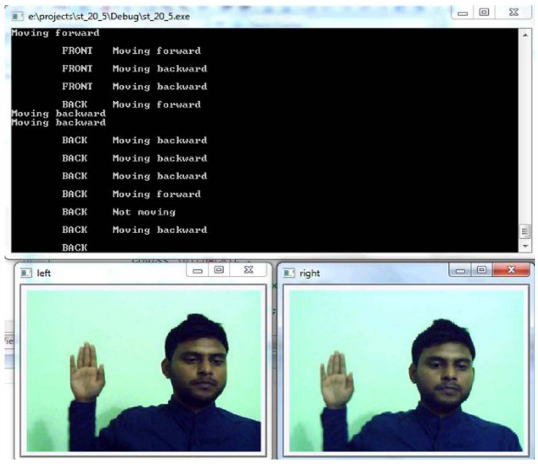


Fig. 8. Back gesture with moving backward condition

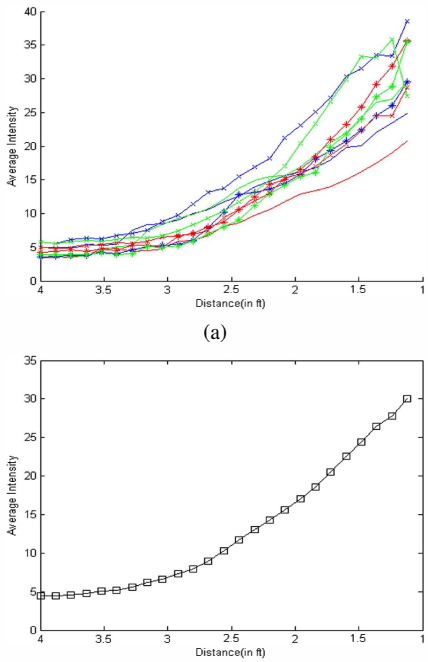


Fig. 9. (a) Sample signature for gesture "front", (b) Average signature for gesture "front"

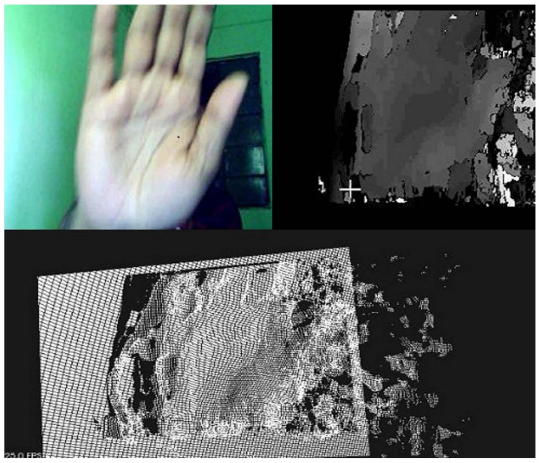
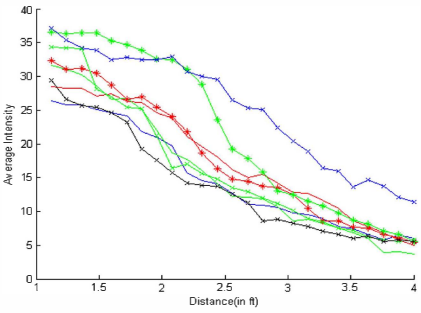


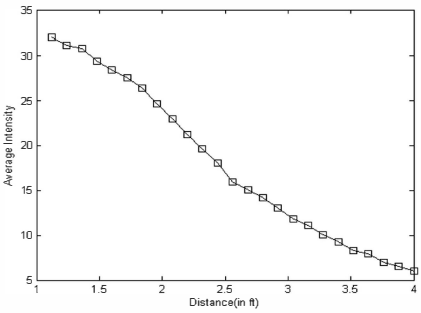
Fig. 10. Realtime 3D reprojection of disparity in a point-cloud viewer

Test no	Moving Forward	Moving Backward	No Movement
test-1	20	1	4
test-2	17	4	4
test-3	23	3	5
test-4	24	2	2
test-5	19	4	4
test-6	24	3	2
test-7	22	4	2
test-8	25	4	1
test-9	21	4	4
test-10	28	2	5

TABLE I. TESTING RESULTS FOR MOVING FORWARD CASE



(a)



(b)

Fig. 11. (a) Sample signature for gesture "back", (b) Average signature for gesture "back"

Test no	Moving Forward	Moving Backward	No Movement
test-1	4	16	8
test-2	4	13	5
test-3	4	14	7
test-4	3	14	9
test-5	2	15	6
test-6	1	18	6
test-7	3	14	6
test-8	3	14	3
test-9	3	15	3
test-10	2	17	6

TABLE II. TESTING RESULTS FOR MOVING BACKWARD CASE

several times and with the results obtained as of now it can be concluded the proposed method is robust and can be used for a gesture recognition system.

ACKNOWLEDGMENT

The authors are thankful to the Ministry of Communication and Information Technology, Govt. of India for facilitating the research.

REFERENCES

[1] C. Zhang, X. Yang and Y. Tian "Histogram of 3D Facets: A Characteristic Descriptor for Hand Gesture Recognition", *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1 - 8, Shanghai, China, April 2013.

[2] M. Elmezain, A. A. Hamadi and B. Michaelis "Improving Hand Gesture Recognition Using 3D Combined Features", *Second International Conference on Machine Vision* , pp.128 - 132, Dubai , UAE, Dec. 2009.

[3] A. Just, O. Bernier and S. Marcel "Recognition of Isolated Complex Mono- and Bi-Manual 3D Hand Gestures", *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 571 - 576, May 2004.

[4] J. Han, L. Shao, D. Xu and J. Shotton "Enhanced Computer Vision with Microsoft Kinect Sensor: A Review", *IEEE Transactions on Cybernatics*, vol. 43, issue. 5, pp. 1318 - 1334, Oct 2013.

[5] D. Scharstein and R. Szeliski "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms", *International Journal of Computer Vision*, vol. 47, issue. 1-3, pp. 7-42, April 2002.

[6] F. Tombari, S. Mattoccia and L. D. Stefano "Classification and evaluation of cost aggregation methods for stereo correspondence", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, Anchorage, AK, June 2008.

[7] S. A. Adhyapak, N. Kehtarnavaz and M. Nadin "Stereo matching via selective multiple windows", *Journal of Electronic Imaging*, vol. 16, issue. 1, pp. 013012 1-14, Jan 2007.

[8] V. Kolmogorov and R. Zabih "Computing Visual Correspondence with Occlusions via Graph Cuts", *In Proceedings of the Eighth IEEE International Conference on Computer Vision*, vol. 2, pp. 508-515, 2001

[9] S. S. Rautaray and A. Agrawal "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, Springer Netherlands, Nov 2012.

[10] S. L. Phung, A. Bouzerdoum and D. Chai "Skin Segmentation Using Color Pixel Classification: Analysis and Comparison", *IEEE Transactions on Pattern Analysis and Machine Intellegence*, vol. 27, issue. 1, pp. 148-154, Jan 2005.

[11] P. Viola, M. Jones "Rapid object detection using a boosted cascade of simple features", *In Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1, pp. 511 - 518, 2001.

[12] G. Bradski and A. Kaehler "Learning OpenCV- Computer Vision with OpenCV library", OReilly Media Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[13] B. Cyganek and J. P. Seibert "An Introduction to 3D Computer Vision Techniques and Algorithms", John Wiley & Sons Ltd, New Jersey, 2009.

[14] K. Konolige "Projected Texture Stereo", *IEEE International Conference on Robotics an Automotion(ICRA)*, pp. 148 - 155, Anchorage, AK, May 2010.

[15] Y. Dennemont, G. Bouyer, S. Otmane and M. Mallem "A Discrete Hidden Markov Models Recognition Module for Temporal Series: Application to Real-Time 3D Hand Gestures", *3rd International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 299 - 304, Istanbul, Oct. 2012.

[16] A. Choudhury, A. K. Talukdar and K. K. Sarma "A Conditional Random Field based Indian Sign Language Recognition System under Complex Background" *Fourth International Conference on Communication Systems and Network Technologies(CSNT)*, pp. 900 - 904, Bhopal, April 2014.

[17] H. D. Yang, S. Sclaroff and S. W. Lee "Sign Language Spotting with a Threshold Model Based on Conditional Random Fields", *IEEE Transactions on Pattern Analysis and Machine Intellegence*, vol. 31, issue. 7, pp. 1264 - 1277, July 2009.

[18] D. Forsyth and J. Ponce "Computer Vision: A Modern Approach", Englewood Cliffs, NJ: Prentice-Hall, 2003.

[19] V. Shenoy H, P. Bongale, V. Roy and S. David "Stereovision based 3D Hand Gesture Recognition for Pervasive Computing Applications", *9th International Conference on Information, Communications and Signal Processing (ICICS)*, pp. 1 - 5, Tainan, Dec. 2013.