

Real Time Static Gesture Recognition using Time of Flight Camera



Netra Lokhande

Abstract: Hand gesture recognition is challenging task in machine vision due to similarity between inter class samples and high amount of variation in intra class samples. The gesture recognition independent of light intensity, independent of color has drawn some attention due to its requirement where system should perform during night time also. This paper provides an insight into dynamic hand gesture recognition using depth data and images collected from time of flight camera. It provides user interface to track down natural gestures. The area of interest and hand area is first segmented out using adaptive thresholding and region labeling. It is assumed that hand is the closet object to camera. A novel algorithm is proposed to segment the hand region only. The noise due to ToF camera measurement is eliminated by preprocessing algorithms. There are two algorithms which we have proposed for extracting the hand gestures features. The first algorithm is based on computing the region distance between the fingers and second one is about computing the shape descriptor of gesture boundary in radial fashion from the centroid of hand gestures. For matching the gesture the distance between two independent regions is computed for every row and column. Same process is repeated across the columns. The number of total region transitions are computed for every row and column. These number of transitions across rows and columns forms the feature vector. The proposed solution is easily able to deal with static and dynamic gestures. In case of second approach we compute the distance between the gesture centroid and shape boundaries at various angles from 0 to 360 degrees. These distances forms the feature vector. Comparison of result shows that this method is very effective in extracting the shape features and competent enough in terms of accuracy and speed. The gesture recognition algorithm mentioned in this paper can be used in automotive infotainment systems, consumer electronics where hardware needs to be cost effective and the response of the system should be fast enough.

Keywords : ToF(Time of flight camera), thresholding, segmentation, hand gesture recognition, static gestures, dynamic gestures, human computer interaction, shape coding, chain coding, fourier descriptors.

I. INTRODUCTION

Human computer interaction is drawing lot of attention recently due to its applications in robotics, human machine interface, gesture controlled activities. The principal components of any gesture recognition system are data set capturing, hand segmentation and tracking, hand feature

identification, feature classification. The classical camera based solution of data acquisition using color cameras have already efficiently employed in gesture recognitions tasks

[1]-[2]. The problem with these solutions are, they are sensitive to lighting conditions, clutter and skin color. The hand motion pose an additional challenge on FPS of camera, to capture detailed motion of gesture. From camera and sensor point of view, the progress of depth sensing devices, like Microsoft Kinect, has greatly promoted the research on HGR. Microsoft Kinect includes a depth camera and a video graphics array (VGA) camera [8]. Both cameras produce image streams at 30 frames per second (fps). The vision based motion capture became barely possible due to Kinect that acquires positional information of individual motion. Depth-based gesture recognition can be categorized into three groups of hand skeleton, spatiotemporal volume of hand, and deep learning-based methods.

From hand localization perspective, various approaches can be employed using depth thresholding either using empirical or automated way [1]. The more empirical and automated approaches are required in case of RGB-D camera such as Kinect. However if depth sensor such as PMD ToF camera is used, it reduces the burden of depth thresholding as a depth which can be obtained using PMD camera is more accurate than other depth sensing techniques[10]. Since there are no publicly available PMD datasets for static and dynamic gestures, we used our own dataset using PMD picoflex ToF camera. Though dataset is recorded with the PMD picoflex camera, the gesture recognition approach presented here is not ToF camera model specific. It is generic approach which is extendible to other methods mentioned in literature. After the hand localization next step is to extract the features which are useful for hand gesture recognition. In past research on human action recognition using video is focused on extracting and using the hand-crafted features [6], [22], [23]. If hand-crafted based features are used it generally have two steps, detecting the features and describing them mathematically. Mostly used and popular feature detection methods are Harris3D [18], Cuboids [19] and Hessian3D [24]. For feature descriptors, popular methods are Cuboids [3], HOG/HOF [4], HOG3D [25] and Extended SURF [24]. Wang et al. [26], used dense trajectories with improved motion based descriptors and other hand-crafted features to achieve results on various datasets. Based on the recent flow of work towards gesture recognition, there would be still efficient and successful methods to achieve better results.

Vishwas U et.al [33] used random projection and KPCA (kernel PCA) for feature space formation of segmented hand gestures.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

Dr. Netra Lokhande, School of Computer Engineering & Technology, MIT World Peace University, Kothrud, Pune. Country.
Email: netralokhande@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](http://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Their method has achieved near real time performance on time of flight camera. However they did not carry out single blind and double blind testing for hand gesture recognition.

II. GOALS

The end objective of this paper is to develop an, low-complexity, and real-time, platform portable solution for the recognition of both static and dynamic gestures of one hand. The accuracy of gesture recognition come from extracting prominent features for identification of different hand gestures. In particular it is intended to evaluate the algorithms for extracting these features on natural and sign gestures. Also goal here is to compare the proposed method with the recent approaches which are found in literature.

III. RELATED WORK

Due to evolving methods of human machine interaction, events controlling without direct interaction with hardware elements, automotive infotainment, gesture and action recognition is evolving area from researchers. Different temporal models have been proposed. Nowozin and Shotton [27] proposed recognition of simple human actions using a Hidden Markov Model. [6] Wang et al. [28] used a more elaborative inter class separator, hidden-state approach for recognizing human gestures. However, relying on only one layer of hidden states, their model alone is not well efficient to adapt a higher level representation of the data. This is more true when dataset is large.

In the deep learning based approaches Ji et al. [29] proposed using 3D Convolutional Neural Network for automated recognition of human actions in surveillance videos. Their model is based on feature extraction in spatial and the temporal domain by performing window based 3D convolutions. This allows capturing motion and feature information in adjacent frames. Taylor et al. [30] also experimented on using 3D Convolutional Networks for learning spatio-temporal features for gesture data. The methods and work in [31] show that averaging using multiple DNN works better than a single network and larger nets will generally perform better than smaller nets. If enough data is provided to network, averaging multi-column nets [32] applied to action recognition can also further improve the performance.

IV. PROPOSED WORK

The proposed solution for dynamic hand gesture recognition is as shown in Fig.1. Raw image data coming from ToF camera is first captured. Since hand is the nearest object to camera, it is easier to extract the hand boundary using the depth information. A thresholding based on image content is applied to segment the hand region. The unwanted regions or the pixels contributing to noise are eliminated by using region labeling and region filtering as segmented hand is the single contributing region and assumed to be the biggest region. The gradient of the segmented region is computed and then magnitude of gradient is thresholded to obtain the boundary of gesture region (hand region). The centroid of the contours is computed as shown in Fig.1. This centroid is used as a reference point for computation of all the distances of the pixels lying on the contour. The feature which we have used is the distance between the centroid and the contour boundary computed at whole angle rotation 0 to 360 degrees. Same

features are pre computed and used for training and used for matching during runtime. There are two algorithms discussed here for gesture recognition. The first algorithm is for deriving the shape features from the boundary and the other algorithm is based on counting the total number of regions across every row of segmented hand. The accuracy of both algorithms is compared.

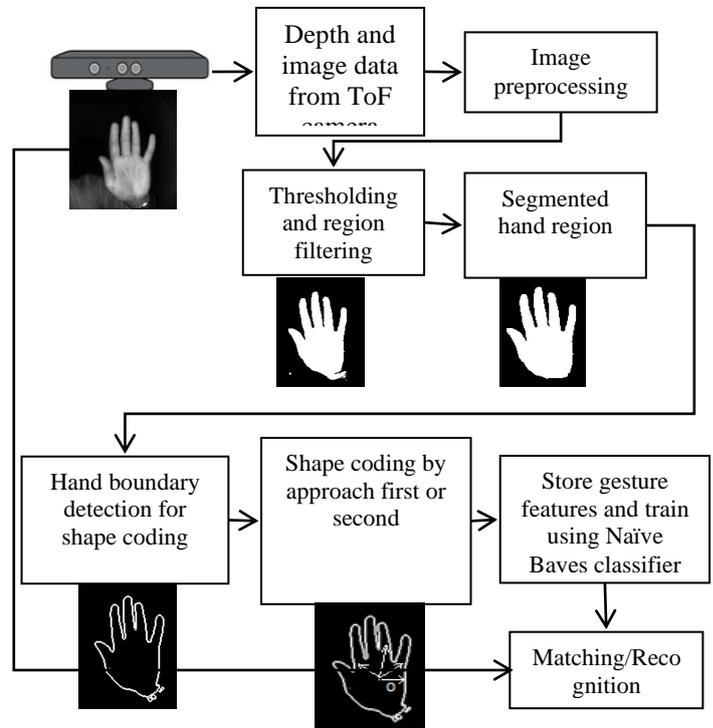


Fig. 1. Proposed method and approach block diagram for hand gesture recognition

A. Hand segmentation

Hand segmentation is carried out by using depth image and thresholding. It is assumed that hand is the closest object to camera and hence single threshold is sufficient to segment the hand. The only problem is that the unwanted noise which is not part of the segmented region needs to be filtered out. This is achieved by using 8 bit region labeling. Here are the details step to carry out the hand segmentation.

B. Hand boundary detection

The shape of an object in image is characterized by the complete object area surrounded by closed contour. The shape can be described by closed contour or region-based shape descriptor [14]. The approaches which we have presented in this paper extracts the shape boundary features. Hence we have relied up on the gesture shape descriptor of hand gesture boundary. Hand boundary is more accurate if segmentation is more accurate. For detecting the boundary of hand, gradient is computed on segmented region. Since segmented hand region is uniform, the gradient inside the boundary region is always zero. The non-zero gradient lies only on hand boundary, we utilize this fact. Hence thresholding is applied to gradient magnitude and hand boundary is retained, as output of gradient thresholding. The result is as shown in Fig. 3.

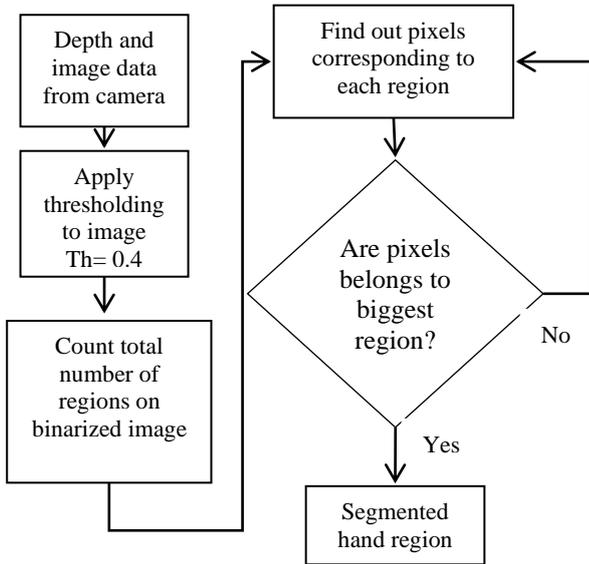


Fig. 2. Steps in hand segmentation

C. Hand centroid computation

For computing the hand centroid we considered the positions of pixels lying on hand boundary.

$$\mu_x = \sum_{i=0}^N X_i \quad \mu_y = \sum_{i=0}^N Y_i$$

Then traversing in 0° to 360° we compute the distances between the centroid and shape boundary. This forms a feature vector for us.

D. Shape feature extractions

There are two major algorithms which we have investigated for extracting the shape boundary features. First algorithm is based on finding the occurrences of multiple regions in each row and column. as shown in Fig. 2. Second algorithm is based on forming a shape code based on distances of every pixel lying on shape boundary from centroid. We compute the distances of nearest occurring pixels in radial way from the centroid at given angle. This is as shown in Fig.3. There are multiple methods which are addressed in the literature, such as chain coding, fourier descriptors, skeleton [14,15]. For representing the shape by code. In our implementation we used the total number of regions found per row and per column. This is shown in following figure. For every row we found total number of independent regions and found length of every independent region. This forms the feature vector for matching.



Fig. 3. Proposed region counting across the rows and columns for forming the shape descriptor from the hand shape boundary. Centroid computed for given gesture boundary and the distances taken at different angles

Algorithm 1 Hand Segmentation and Boundary Computation

Input:
Im: Input image
Th: Threshold
Output:
SegOut: Segmented Image
ImOut: Image with hand boundary

```

For each element x of Im
  If x > Th
    SegIm=255;
  Else
    SegIm=0
  End If
End For
nRegions = CountRegions( SegIm)
For each region r =1 To nRegions
  Locations= findRegionLocations(r)
  Cnt_locations=
  CountTotalLocations(Locations)
End for
For each element c of Cnt_location
  Indx=findIndexBiggestCntLocation(c)
End For
SegOut (Indx)=255
  
```

Shape descriptor

Algorithm 2 Computing the radial features

```

for each angle i ∈ 1,..., 360 do
  for r=1:MAX_RAD
    X = r * cos Ø + μx
    Y = r * sin Ø + μy
    image_pixel = im(X,Y)
    If image_pixel == 1
      Feature(r,i) =
      1;
      break;
    end
  endloop
endloop
Feature(r,i)=
Feature(r,i)/max(Feature(r,i))
  
```

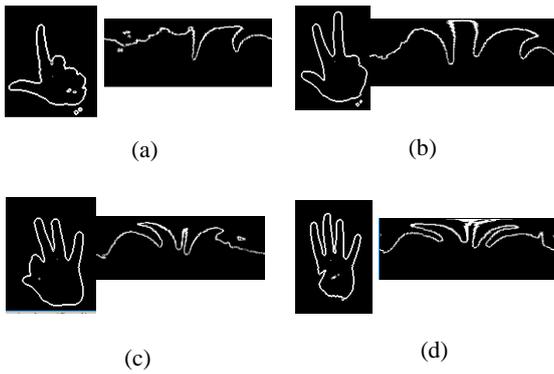


Fig.4. Gestures and its shape descriptors code obtained by using above algorithm for (a) Gesture L, (b) Gesture two, (c) Gesture ok, (d) Gesture number four

V. EXPERIMENTAL RESULTS

Many experiments were done in order to test the presented approach for hand boundary detection and the localization of the palm center. The approach is tested and evaluated over a series of 150 dynamic gestures for which the results are reported in the literature to enable performance comparisons. In our implementation we used few sign numbers and popular gestures for matching, in which the precomputed features were stored and same were used for recognition. For recognition we used naïve bayes classifier. Table I to Table V shows the average accuracy obtained against sign and popular gestures. There are total eight popular gestures and nine sign number gestures, we used for testing our implementation. It was checked on live video. 150 frames of video were given for testing the algorithm. Table I shows the correct number of gestures recognized out of 150 frames. We tested our implementation on Intel i7 5th generation platform with 16GB RAM. The average time taken for recognition is 0 seconds for recognizing particular gesture. In results discussed below accuracy for Victory and OK gestures is lesser than others because, while testing subject slightly tilted the gesture, since our implementation is rotation variant.

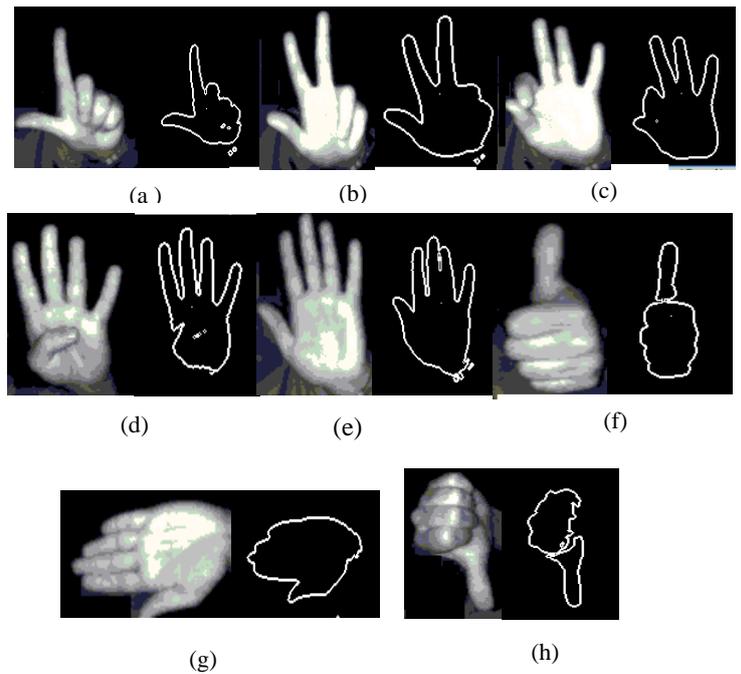


Fig.5. Different natural gestures we used in our implementation and their centroid of (a) Gesture L. (b) Gesture Victory. (c) Gesture ok. (d) Gesture number 4. (e) Gesture full hand. (f) Gesture thumbs up. (g) Gesture left. (h) Gesture thumbs down.

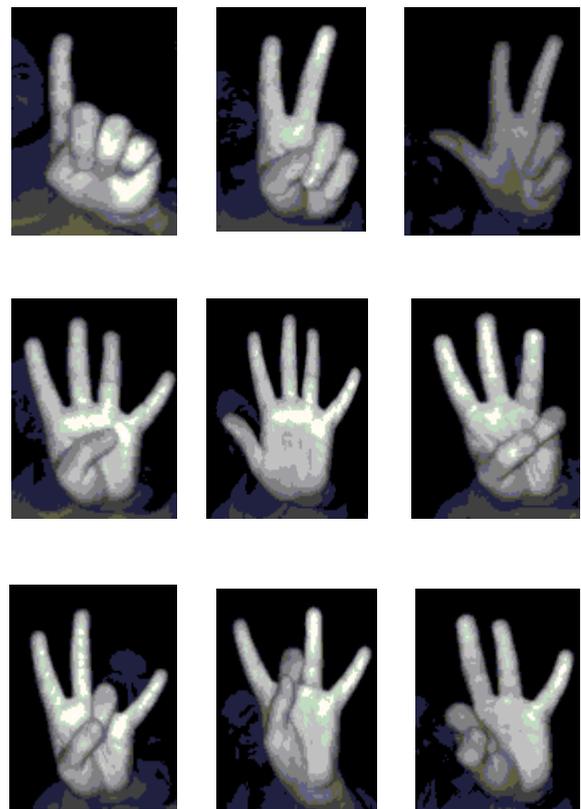


Fig. 6. Different sign number gestures we used in our implementation

Table- I: Recognition rates for popular gestures using radial shape descriptors [single blind]

Gesture	No Of correctly matched	Accuracy	Time(s)
Open hand	150	100%	0.1864
Victory	149	99.33%	0.1877
OK letter	121	80.66%	0.1855
Letter L	150	100%	0.1901
Letter four	147	98%	0.1797
Left	150	100%	0.1712
Thumbs up	149	99.33%	0.1855
Thumbs down	124	82.67%	0.1677
Average	142.5	95%	0.1817

Table – II: Recognition rates for popular gestures using radial shape descriptors [double blind]

Gesture	No Of correctly matched	Accuracy	Time(s)
Open hand	146	97.33%	0.1864
Victory	149	99.33%	0.1877
OK letter	138	92%	0.1855
Letter L	145	96.66%	0.1901
Letter four	131	87.33%	0.1797
Left	150	100%	0.1712
Thumbs up	142	94.66%	0.1855
Thumbs down	139	92.66%	0.1677
Average	142.5	94.97%	0.1817

Table- III: Recognition rates for popular gestures using row and column region features [single blind]

Gesture	No Of correctly matched	Accuracy	Time(s)
Open hand	127	84.67	0.1857
Victory	133	88.67	0.1952
OK letter	110	73.33	0.1859
Letter L	122	81.33	0.1881
Letter four	95	63.33	0.1823
Left	110	73.33	0.1904
Thumbs up	124	82.67	0.20
Thumbs down	112	74.67	0.1840
Average	116.62	77.75	0.1889

For comparison we studied the various approaches which uses depth sensor, Microsoft Kinect sensor and checked for the accuracy we obtained on sign language gestures. We found that the average accuracy obtained was better than the approaches we compared. We tested the proposed solution in case of single blind (The person knows which gesture is recognized by system) and double blind (The person doesn't know which gesture is recognized by the system). The comparison is done against standard approaches mentioned in the literature. The accuracy results are summarized in Table I and Table II. We also tested our implementation on popular gestures mentioned in Table III and Table IV and compared the results with the standard methods mentioned in literature.

Table -4 : Recognition rates comparison for popular gestures using radial shape descriptors [single blind]

Gesture	[7]	[10]	[11]**	[12]	[13]	Proposed method
Open hand	99	-	-	-	-	100%
Victory	89.2	87	83	90	91	99.33%
OK letter	84.8	35	94	97	94	80.66%
Letter L	91	87	95	96	94	100%
Letter four	90.5	77	94	96	97	98%
Left	85.2	-	-	-	-	100%
Thumbs up	89.5	-	-	-	-	99.33%
Thumbs down	88.2	-	-	-	-	82.67%
Star Trek	84	-	-	-	-	-
Average	89.1	71.5	91.5	94.8	94	95

Table- V: Recognition rates comparison for sign number gestures using radial shape descriptors [single blind]

Gesture	[7]	[8]	[8]**	[9]	[2]	Proposed method
One	91.75	92	96	84	100	100
Two	89.25	82	86	85	100	100
Three	90	91	94	75	80	90
Four	100	86	88	71	100	100
Five	100	85	92	82	80	95
Six	77	93	96	84	90	98
Seven	84.25	91	96	82	95	98
Eight	76.25	89	93	80	100	89
Nine	74	97	98	80	100	100
Average	86.9	89.6	93.2	80.3	93.9	96.66

Real Time Static Gesture Recognition using Time of Flight Camera

In the final test we tested our system from the point of view of naturally controlling the software using gestures. This has been shown in Fig.7. We used thumbs down gesture to close the application. Gesture 'L' is trained to open the excel application and Gesture victory is used for opening windows media player and play the song.

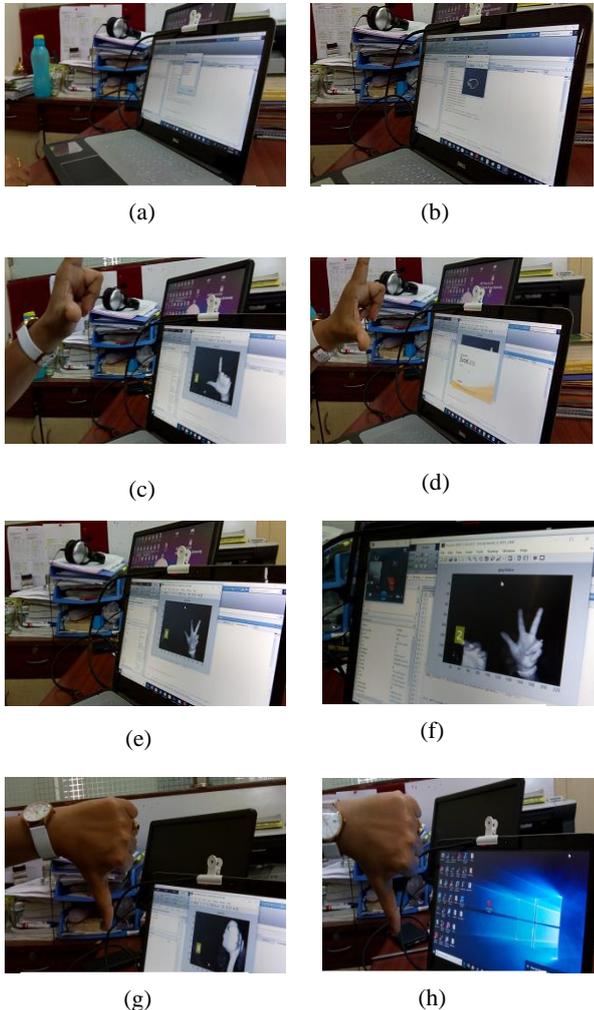


Fig.7. Software application control using gesture recognition (a) Application starts (b) Training features. (c) Gesture L is performed. (d) Opening excel application. (e) Gesture victory is performed (f) Opening windows media player. (g) Thumbs down gesture for closing the application. (h) Application closes.

VI. CONCLUSION

A method for dynamic recognition of hand gestures using Time of Flight camera is presented in this paper. There are two algorithms we proposed, first one was based on computing row wise and column wise features and second one is based on traversing at different angles from 0 to 360 degrees it finds the radial distance between centroid and hand boundary. We formed the features vectors using radial distances for these angles. Our system is able to perform with average accuracy 95% with single hand dynamic gestures with eight natural gestures and nine sign language gestures. Since we have used Time of Flight camera in the implementation, lighting condition, clothing, skin color have little impact on the results obtained using current implementation. In the proposed implementation first

algorithm is dependent on hand size and distance from the camera whereas second approach is independent on the hand size of the user. This is one of the major advantages of our method as compared to [1]. The proposed solution has many applications in human machine interaction, augmented reality, automotive infotainment and natural control of a software application. In future the results can be more improved by using large dataset for training. In the first algorithm we discussed, it can be made size and distance independent from camera by normalizing the row wise and column wise region lengths. This can be normalized by dividing the length of maximum region found in given gestures shape. Multiple gestures can also be used to widen the application areas of our implementation.

REFERENCES

- Guillaume Plouffe and Ana-Maria, , Static and Dynamic Hand Gesture Recognition in Depth Data Using Dynamic Time Warping, IEEE Trans. Instrumentation and Measurement, 2015
- A. R. Várkonyi-Kóczy and B. Tusor, "Human-computer interaction for smart environment applications using fuzzy hand posture and gesture models," IEEE Trans. Instrum. Meas., vol. 60, no. 5, pp. 1505-1514, May 2011.
- P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in International Conference on Multimedia. ACM, 2007.
- I. Laptev, "On space-time interest points," International Journal of Computer Vision, 2005.
- Mu-Chun Su, A Fuzzy Rule based Approach to Spatio-Temporal Hand Gesture Recognition, IEEE Trans. On Systems, Man and Cybernetics, vol. 30, No. 2, May 2000.
- Di Wu, Lionel Pigou, Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition, IEEE Trans. Patt. Anal and Machine Intell., Jan-2016
- Daniilo Avola, Marco Bernadi, Exploiting Recurrent Neural Networks and Lep controller for the Recognition of Sign Language, IEEE Trans on Multimedia, DOI 10.1109/TMM.2018.2856094
- Reza Azad*, Maryam Asadi, Dynamic 3D Hand Gesture Recognition by Learning Weighted Depth Motion Maps, IEEE Trans. On circuits and systems, DOI 10.1109/TCSVT.2018.2855416
- Tom M. Mitchell, Machine learning, 2010, McGraw Hill
- Fabrizio Pece, Jan Kautz, and Tim Weyrich. Three depth camera technologies compared <http://www.cs.ucl.ac.uk/staff/F.Pece/page9/files/abstract.pdf>, 2011.
- Y. Li, "Multi-scenario gesture recognition using Kinect," in Proc. IEEE Int. Conf. Comput. Games, Jul./Aug. 2012, pp. 126-130.
- Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in Proc. ACM Int. Conf. Multimedia, 2011, pp. 1093-1096.
- Z. Ren, J. Meng, and J. Yuan, "Depth camera based hand gesture recognition and its applications in human-computer-interaction," in Proc. IEEE Conf. Inf., Commun. Signal Process., Dec. 2011, pp. 1-5.
- N. Pugeault and R. Bowden, "Spelling it out: Real-time ASL fingerspelling recognition" in Proc. IEEE Intl. Conf. Comput. Vis. Workshops (ICCV Workshops), Barcelona, Spain, Nov. 2011, pp. 1114-1119
- I. Oikonomidis, N. Kyriazis, and A. A. Argyros, , "Efficient model-based 3D tracking of hand articulations using Kinect," in Proc. Brit. Mach. Vis. Conf., 2011, pp. 101.1-101.11.
- K. Otiniano-Rodríguez and G. Cámara-Chávez, "Finger spelling recognition from RGB-D information using kernel descriptor," in Proc. IEEE Conf. Graph., Patterns Images, Arequipa, Peru, Aug. 2013, pp 1-7
- F. Pedersoli, S. Benini, N. Adami, and R. Leonardi, "XKin: An open source framework for hand pose and gesture recognition using Kinect," Vis. Comput., vol. 30, no. 10, pp. 1107-1122, Oct. 2014.
- D. Wu and L. Shao, "Silhouette analysis-based action recognition via exploiting human poses," IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 2, pp. 236-243, 2013.

19. I. Laptev, "On space-time interest points," International Journal of Computer Vision, 2005.[5] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in Visual Surveillance and Performance Evaluation of Tracking and Surveillance. IEEE, 2005.
20. Qingxiao Niu, Hua Zhang, "A Shape Contour Description Method based on Chain Code and Fast Fourier Transform", 2011 Seventh International Conference on Natural Computation
21. S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, 2002, pp. 509–522
22. L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal laplacian pyramid coding for action recognition," IEEE Transactions on Cybernetics, vol. 44, no. 6, pp. 817-827, 2014.
23. L. Liu, L. Shao, F. Zheng, and X. Li, "Realistic action recognition via sparsely-constructed gaussian processes," Pattern Recognition, doi: 10.1016/j.patcog.2014.07.006., 2014.
24. G. Willems, T. Tuytelaars, and L. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in European Conference on Computer Vision. Springer, 2008.
25. A. Klaser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in British Machine Vision Conference, 2008.
26. H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," International Journal of Computer Vision, 2013.
27. S. Nowozin and J. Shotton, "Action points: A representation for low-latency online human action recognition," Tech. Rep., 2012.
28. S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 2. IEEE, 2006, pp. 1521–1527.
29. S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013.
30. G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in European Conference on Computer Vision. Springer, 2010.
31. D. Wu and L. Shao, "Deep dynamic neural networks for gesture segmentation and recognition," European Conference on Computer Vision and Pattern Recognition Workshops, 2014.
32. D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in IEEE Conference on Computer Vision and Pattern Recognition, 2012.
33. Vishwas Udpikar, Jitesh Butala, "Static gesture recognition using PMD ToF camera," 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)

AUTHORS PROFILE



Dr. Netra Lokhande, received her B.E degree from Karnatak University. M.E from Govt. College of Engineering (COEP), Pune University, and her Ph.D. from JJTU, Rajasthan. She has 22 years of teaching experience from various well known institutes from Pune University. Presently

she is working as an Associate Professor in School of Computer Engineering and Technology, MIT World Peace University, Pune, Maharashtra, India. She has worked as Head of Dept. and Dean Academics in previous institute. She is also UG and PG guide for students. She has published more than 30 research papers in International/National conferences and Journals. She is the member of Editorial Board of number of International Journals.