A HYBRID METHOD USING KINECT DEPTH AND COLOR DATA STREAM FOR HAND BLOBS SEGMENTATION

Mostafa Karbasi, Zulkefli Muhammad, Ahmad Waqas, Zeeshan Bhatti, Asadullah Shah, M.Y.Koondhar, Imtiaz Ali Brohi

Department of Computer Science, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia
For Correspondence; Mostafa.karbasi@live.iium.edu.my, Ahmad.waqas@live.iium.edu.my, Zeeshan.bhatti@live.iium.edu.my, Asadullah@iium.edu.my, yaqoobkoondhar@gmail.com, brohiimtiaz@hotmail.com

ABSTRACT: *The recently developed depth sensors such as Kinect have provided new potentials for human-computer interaction (HCI) and hand gestures are one of main parts in recent researches. Hand segmentation procedure is performed to acquire hand gesture from a captured image. In this paper, a method is proposed to segment hand blobs using both depth and color data frames. This method applies a body segmentation and an image thresholding techniques to depth data frame using skeleton data and concurrently it uses SLIC super-pixel segmentation method to extract hand blobs from color data frame with the help of skeleton data. Finally, two segmented blobs are combined to improve the final result by assuming that hands are located in front of the body. The proposed method has low computation time and shows significant result when the basic assumptions are fulfilled.*

Keywords: hand gesture recognition, human computer interaction, simple linear iterative clustering (SLIC), hand detection, posture recognition.

1. INTRODUCTION

Vision based hand gesture needed in hand detection and it could be the most vital things in hand recognition. In some condition hand detection can apply on human body and it works as a boundary and limitation compare to any other things. Now, here mention some obstetrical in this area:

1- Human body has a 3D shape so hand shape as well would give very different lay out according time and position.
2- Light and background make a lot of problem in hand detection. Skin color and background color next to it, also similar color in background n skin color is used as information.

Also depth image is another way to concentrate and find out the fine and right information in RGB image. everything in the background can be discarded with a threshold on the depth values [1]. The TOF camera used in this paper is a Kinect 360. Anyway, this is an inexpensive camera which is IR based. The TOF camera has resolution around 680*480. High resolution can help out the better result with this camera [2].

Kinect and TOF has a depth sensor for gesture recognition. Obviously arm movement and hand shape are using in gesture recognitions. Mostly, contour base on 2D feature shows a hand shape and simulating the hand shape with 2D is the most important and hard process [3] . Furthermore, the arm movement feature is affected by environmental changing, such as individual differences in body size, camera position and so on because the coordinate of centroids of hand region are used as arm movement feature [4].

In this work, we used depth, color data frames and user skeleton data to extract user hand blobs. The user's body extracted using skeleton data adapted to segment depth data. Unnecessary parts are removed by thresholding technique. To extract the user's body, the color data frame is masked by depth data, then SLIC super-pixel segmentation algorithm is employed to cluster color data frame, after that hand blobs are detected based on skeleton data. The main contribution of this paper is developing hand detection methods based on depth and color fames and user skeleton data. It is perfect and robust in hand orientation, motion, position and postures. The proposed method operates accurately and efficiently in uncontrolled environments.

2. LITERATURE REVIEW

There are new technologies such as Kinect and Zcam introduced in the market which has been used by many researchers for different application. Some of researchers worked on hand segmentation, hand counter, color distribution, etc. [5] proposed kd-tree structure for hand and head detection and tracking which is exploited to resolve ambiguities and overlaps. Park et al [2] proposed adaptive hand detection approach by using 3-dimentional information from Kinect and tracks the hand using GHT based method. [6] used RGB stream and depth data for hand detection in their research and contributed in 3D contour model for real time application. In addition, their result shows that proposed method is successful in handling real time interaction in desktop environment with clustering background method. [7] adapted region growing and bilateral filter for depth map enhancement and detection. The proposed method can significantly improve the quality of depth maps and enlarge Kinect's applicant ion fields where high quality depth images are required. [8] used depth data for hand detection based on distances. Their method included background subtraction and shadow removal for removing redundant data. [9] also used depth and depth and color data for hand detection and sign language recognition. They implemented Finger-Earth Mover's Distance (FEMD) as a new approach for sign language recognition. In addition, their methods have been implemented in two applications such as arithmetic computation and rock-paper-scissors game. [1] presented real time system for hand detection and gesture recognition on the base of ToF camera and the RGB. The proposed method not only improved detection rate, but also allows for the hand to overlap with the face, or with hands from other persons in the background. [10] worked on Sign Language Recognition with the help of 3D convolutional neural networks. 3D convolutional neural networks can extract spatio-temporal

features from raw data without any prior knowledge. They achieved high accuracy and high speed performance as well. [11] investigated an overview of the main research works based on sign language recognition system and developed system into sign language methods and recognition technique are discussed. Many researchers adapted threshold method for hand detection in different application . [7, 12, 13, 14]used threshold method to identify overlapping of hand of hand-head or hand-hand region. [15] proposed a method for Malaysian Sign Language recognition with the help on image processing but their method was based on RGB with ordinary camera. [16] has done a review on hand detection based on depth data for SLR.

[17] provided set of candidate contours with the help of the foreground segmentation and edge detection . they used foreground segmentation to reduce the region of interest for better selection contour. [18] introduced novel method reduce the complexity of the model by dividing the training set into smaller clusters, and trained PCFs on each of these compact sets. Most of researchers tried to detect hand and head by using Kinect skeleton which can detect and track hand and head easily. For instance, [19, 20] used skeleton model for hand detection. They usually crop hand based on coordinate x, y which obtained from the skeleton frame. [21] proposed predefied version of the ICP(Interactive Closet Point) algorithm to obtain pose estimation.

We build our hand segmentation based on Body extraction, Depth thresholding and Forming hand blob distribution based on depth and skeleton. Then, Simple Linear Iterative Clustering (SLIC) super pixel algorithm was used for detecting hand blobs based on hand skin color.

3. PROPOSED HAND BLOB SEGMENTATION

An overview of the proposed method is shown in Figure 1. The proposed algorithm uses color and data frames with the help of skeleton data provided by Kinect. Firstly, the algorithm employs a body segmentation method based on skeleton data, then a depth thresholding technique is adopted to extract hands by assuming that they are in front of the body. The extracted segments may not be completely accurate based on hands positions; therefore, the wrist hand points of skeleton data are employed to improve hand blob segments. Beside the depth data, the color data frame is segmented using Simple Linear Iterative Clustering (SLIC) algorithm [22] and the proper blobs are selected based on hand points of skeleton data. Finally, a method is proposed to combine the extracted blobs and create more accurate result.

Data Frame Acquisition

Microsoft Kinect provides color, depth, and skeleton data streams. it has 800 – 3500 mm operating range, producing depth image in VGA (640x480) resolution, and capable to capture a depth frame up to 30 frame per second. Each pixel in the depth data frame is a measured distance in millimeter scale.

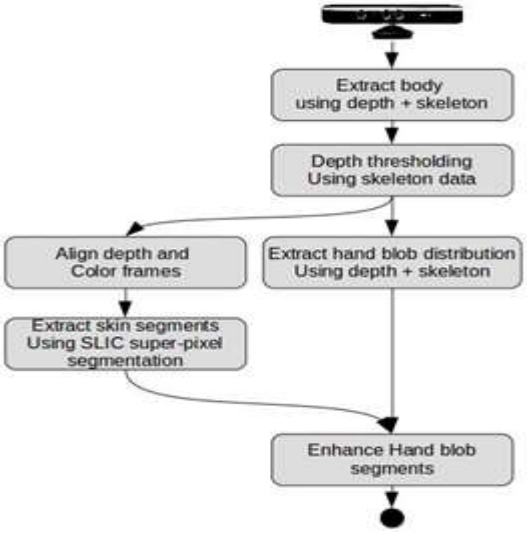
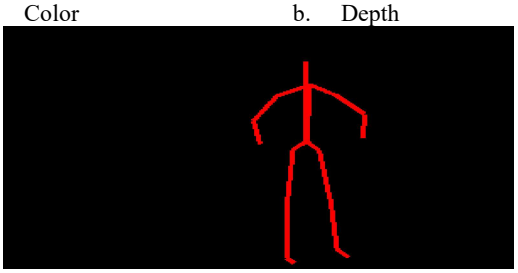


Figure1. An Overview of Proposed Method

a. Color b. Depth



c. Skeleton

Figure2. Color, Depth, Skeleton

Depth frame is required to be converted into grayscale image for visualization purposes. Conversion of the depth frame into 8-bits grayscale image is done using Equation 1.

$$P_N = 255 \times \left(\frac{D_N - 800}{3500 - 800}\right) \quad (1)$$

Where is the depth data, is the new pixel data, and N is the pixel number? Some pixel in the depth frame which has been captured Kinect are not in the operating range value – considered as unknown pixel – because of the occlusion effect. Occlusion effect has been solved by employing median filter [23, 24].

Hand segmentation using depth data

Hand segmentation is performed by following the three steps, (a) body extraction, (b) depth thresholding, (c) and finally forming hand blob distribution based on depth and skeleton data frames for each hand. The highlighted steps result a probability distribution demonstrating the probability of each pixel belonging to the hand blob.

User's body is extracted by the underlying techniques provided by Kinect. The sensor adopts the depth data frame to determine user location and estimate skeleton points. After body segmentation, depth thresholding is performed on the resulted depth data frame. Depth thresholding simply any depth points lower than a predefined value and keep the rest of the depth points. The threshold value is defined as mentioned in Equation 2 and Equation 3.

$$T_x = max\left(0.5\times\left(P_x^{leftelbow}+P_x^{rightelbow}\right), 0.5\times\left(P_x^{righthand}+P_x^{lefthand}\right)\right) \quad (2)$$

$$T_y = max\left(0.5\times\left(P_y^{leftelbow}+P_y^{rightelbow}\right), 0.5\times\left(P_y^{righthand}+P_y^{lefthand}\right)\right) \quad (3)$$

Where P is location of a skeleton point and T is predefined depth threshold. After depth thresholding, the hand blob distribution is formed for each hand. The probability distribution is a multivariate normal distribution as described in Equation 4.

$$N(x,\mu,\Sigma) = \left(\sqrt{(2\pi)^k|\Sigma|}\right)^{-1}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \quad (4)$$

Where x is the depth point including the depth pixel location and depth value, μ is location of each extracted hand blob, and ∑ is a diagonal matrix as shown in Equation 5.

$$\Sigma = I_{k\times k}.\sigma \quad (5)$$

Where I is an identity matrix and σ is a predefined value. The formed distribution determines all possible depth points and their probability measures.

Hand segmentation using color data

Color data is masked using depth data to extract the user's body, then hand blobs are detected based on skin color using simple linear iterative clustering (SLIC) super-pixel algorithm. The color data frame is clustered using SLIC super-pixel algorithm, then clusters within similar color to skin color are extracted and are merged together as Region-Of-Interests.

Extracted skin pixels are employed to form a mask which all skin pixels are set to one and others set to zero. In next step, all regions that do not include the hand points determined by skeleton data have been removed from the mask. In result, the extracted color frame and the mask represents the hand blobs.

Simple Linear Iterative Clustering (SLIC)

SLIC (Simple linear iterative clustering) is a very generic technique which is easy to implement. This technique only uses a *k* value parameter in its algorithm, where k is the desired number of super-pixels, which are also, approximately, sized equally. In the CIELAB color space having color images, the procedure of clustering initiates with an initialization step, where the 'k' parameters for initial cluster centers is represented as Equation 6.

$$C_i = (l_i, a_i, b_i, x_i, y_i)^T \quad (6)$$

And is sampled with S pixels apart on a regular grid space. Further, the grid interval is computed using Equation 7.

$$s = \sqrt{N/k} \quad (7)$$

to generate super-pixels with approximately equal size. In a 3x3 neighborhood for the lowest gradient position, the centers of the super-pixels are moved to their corresponding seed locations. The reason for applying this technique is to evade centering of a super-pixel on an edge, and consequently decreasing the chances of seeding a super-pixel with a noisy pixel. Further, in the subsequent step each pixel *i* is connected with the nearest cluster center, for which the location of search region is overlapping. This technique is significant and efficient for increasing the execution time by minimizing the search region size to a small area, which is then achieved a result by reducing the calculations for the distance of the region. This approach has an advantage over the k-means clustering technique, in which all cluster centers are compared with each pixels. The size SxS is an approximate expected range of the super-pixel, hence the search around the super pixel center for the similar pixels is done within a region of 2Sx2S. As the nearest cluster center is linked with each pixel, the cluster centers are updated in next step to be the mean vector of all the pixels belonging to the cluster. The new and previous locations of the cluster centers are computed using L_2 norm with a residual error E. These steps of assignment and update are iteratively repeated until the error converges and generally 10 iterations would be enough for most of images. Finally, the nearby super-pixels are re-assigned to disjoint pixels in a post-processing step that imposes connectivity.

Enhance hand segmentation using both depth and color data

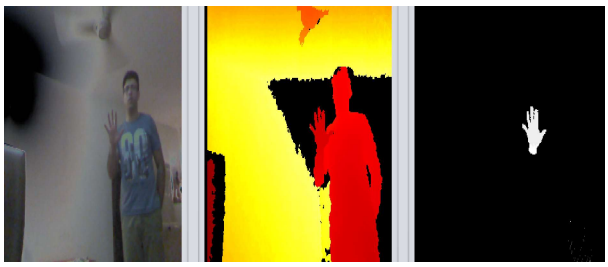
Hand segmentation in depth (D) and color (C) data frame is performed concurrently to extract two independent hand blobs for each hand. Depth-based hand blobs (D) are probability distribution and Color-based hand blobs (C) are binary masks setting the hand pixels to one and the rest to zero. Extracted hand blobs are adopted to generate enhanced results. To do so, it is formulated as described in Equation 8.

$$P_{enhance} = P_D \cap P_C \quad (8)$$

Where depth-based hand is blob and is color-based hand blob mask. Due to the fact that depth-based and color-based distributions are independent, Equation 9 can be simplified as,

$$P_{enhance} = P_D \times P_C \quad (9)$$

Then, the resulted enhanced distributions filtered with a simple threshold value. All values lower than predefined threshold are set to zero. Figure 3 (a, b and c) below, shows color, depth and enhanced frames respectively.

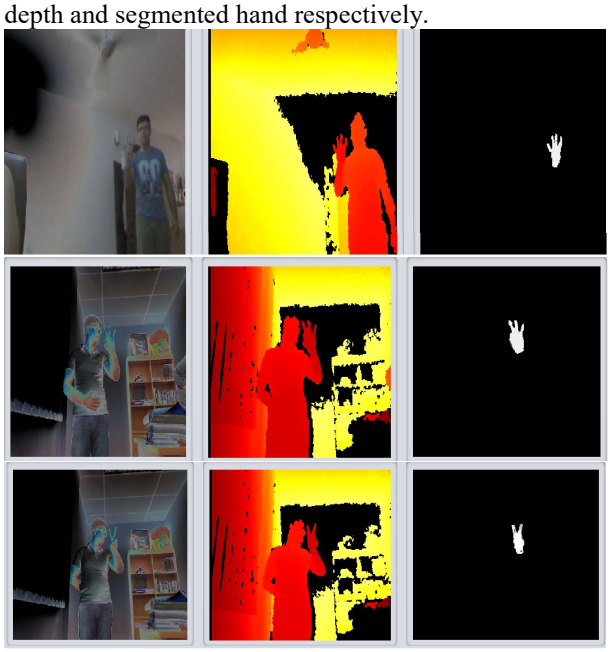


a. Color b. Depth c.Enhanced

Figure 3

4. EXPERIMENTS AND DISCUSSIONS

In order to measure the performance of this method, the color and depth data frames in different environments are collected in figure 4. Below figures have shown segmented hand in different environments. Figure 4(a),(b) and (c) shown color, depth and segmented hand respectively.



a. Color b. Depth c. Segmented

Figure4: Hand Segmented in Different Environments.

The illustration shows that, the proposed method is satisfactory enough efficient in performing hand segmentation. The segmentation is failed when the distance between hand and camera is equal or more than the distance between body and camera, resulted in no segmented object in the image.

Computation latency is also evaluated for proposed method. The experiment is performed by measuring computation latency time in processing three images. Another three computation time data is also obtained from live captured image of human pose. The experiment is conducted using a PC with Microsoft Windows 8 with Intel Core i5 1.70GHz x 4 processors and 4 GB RAM. The result of this experiment is shown in Table 1 which elaborates that the proposed method is capable to process a VGA sized image up to 255 frames per second and still much higher than the sensor frequency rate.

The results in Table 1 show that the hand detection using depth data frame has minimum computation time. Hand detection using color data frame requires more time to perform than depth-based hand detection because it needs to perform SLIC algorithm to extract hand blobs. The results demonstrate that enhanced hand detection proposed in this paper requires more time, because it performs hand detections based on color and depth and then it combined the two results to enhance the hand blob.

Table1: Hand Segmentation using Different Frames

Method	Computation time (millisecond)
Hand blob segmentation using depth data frame	245.4
Hand blob segmentation using color data frame	395.2
Enhance hand blob segmentation using color and data data frames	413.6

5. CONCLUSION

In this paper, a novel method is proposed to hand blob segmentation algorithm using depth and color data frame. The hand blob segmentation is achieved using Body extraction, Depth thresholding and Forming hand blob distribution based on depth and color using skeleton data. Simple Linear Iterative Clustering (SLIC) super pixel algorithm is adopted to segment hand blobs based on hand skin color. The experiment shows significant results and it prove that it is more accurate, efficient, and robust for various states of hand articulation, distortion with orientation or scale changes. Moreover, it can perform in an uncontrolled random environment and it has minimum runtime effort using depth data as compared to using color data frames for hand blob segmentation. The presented work is planned to be employed in a Malaysian sign language recognition system.

6. REFERENCES

[1] Van den Bergh, Michael, & Van Gool, Luc. (Combining RGB and ToF cameras for real-time 3D hand gesture interaction. Paper presented at the Applications of Computer Vision (WACV), 2011 IEEE Workshop on 2011).

[2] Keller, Maik, Lefloch, Damien, Lambers, Martin, Izadi, Shahram, Weyrich, Tim, & Kolb, Andreas. Real-time 3D reconstruction in dynamic scenes using point-based fusion. Paper presented at the 2013 International Conference on 3D Vision-3DV (2013).

[3] Pavlovic, Vladimir I, Sharma, Rajeev, & Huang, Thomas S. Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Transactions on pattern analysis and machine intelligence, 19(7), 677-695 (1997).

[4] Jaemin, Lee, Takimoto, Hironori, Yamauchi, Hitoshi, Kanazawa, Akihiro, & Mitsukura, Yasue. A robust gesture recognition based on depth data. Paper presented at the Frontiers of Computer Vision,(FCV), 19th Korea-Japan Joint Workshop on (2013).

[5] Suau, Xavier, Ruiz-Hidalgo, Javier, & Casas, Josep R. (Real-time head and hand tracking based on 2.5 D data. IEEE transactions on multimedia, 14(3), 575-585 2012).

[6] Yao, Yuan, & Fu, Yun. Real-time hand pose estimation from RGB-D sensor. Paper presented at the 2012 IEEE International Conference on Multimedia and Expo (2012).

[7] Chen, Li, Lin, Hui, & Li, Shutao. Depth image enhancement for Kinect using region growing and bilateral filter. Paper presented at the Pattern Recognition (ICPR), 21st International Conference on (2012).

[8] Karbasi, Mostafa, Bhatti, Zeeshan, Nooralishahi, Parham, Shah, Asadullah, & Mazloomnezhad, Seyed Mohammad Reza. Real-Time Hands Detection in Depth Image by Using Distance with Kinect Camera. International Journal of Internet of Things, 4(1A), 1-6 (2015).

[9] Ren, Zhou, Meng, Jingjing, Yuan, Junsong, & Zhang, Zhengyou. Robust hand gesture recognition with kinect sensor. Paper presented at the Proceedings of the 19th ACM international conference on Multimedia (2011).

[10] Huang, Jie, Zhou, Wengang, Li, Houqiang, & Li, Weiping. Sign language recognition using 3D convolutional neural networks. Paper presented at the 2015 IEEE International Conference on Multimedia and Expo (ICME) (2015).

[11] Bhoir, Pooja P, Itkarkar, Rajashri R, & Bhople, Shilpa. Hand Gesture Recognition Based on Hidden Markov Models.

[12] Breuer, Pia, Eckes, Christian, & Müller, Stefan. Hand gesture recognition with a novel IR time-of-flight range camera–a pilot study. Paper presented at the International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications (2007).

[13] Liu, Xia, & Fujimura, Kikuo. Hand gesture recognition using depth data. Paper presented at the Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on (2004).

[14] Biswas, Kanad K, & Basu, Saurav Kumar. Gesture recognition using microsoft kinect®. Paper presented at the Automation, Robotics and Applications (ICARA), 2011 5th International Conference on (2011).

[15] Karabasi, Mostafa, Bhatti, Zeeshan, & Shah, Asadullah. A model for real-time recognition and textual representation of malaysian sign language through image processing. Paper presented at the Advanced Computer Science Applications and Technologies (ACSAT), International Conference on (2013).

[16] Karbasi, Mostafa, Bhatti, Zeeshan, Aghababaeyan, Reza, Bilal, Sara, Rad, Abdolvahab Ehsani, Shah, Asadullah, & Waqas, Ahmad. REAL-TIME HAND DETECTION BY DEPTH IMAGES: A SURVEY. Jurnal Teknologi, 78(2) (2016).

[17] Hamester, Dennis, Jirak, Doreen, & Wermter, Stefan. Improved estimation of hand postures using depth images. Paper presented at the Advanced Robotics (ICAR), 16th International Conference on (2013).

[18] Keskin, Cem, Kıraç, Furkan, Kara, Yunus Emre, & Akarun, Lale. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. Paper presented at the European Conference on Computer Vision (2012).

[19] Xiao, Zheng, Mengyin, Fu, Yi, Yang, & Ningyi, Lv. 3D human postures recognition using kinect. Paper presented at the Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2012 4th International Conference on (2012).

[20] Zainordin, Faeznor Diana, Lee, Hwea Yee, Sani, Noor Atikah, Wong, Yong Min, & Chan, Chee Seng. Human pose recognition using Kinect and rule-based system. Paper presented at the World Automation Congress (WAC), (2012).

[21] Coscia, Pasquale, Palmieri, Francesco AN, Castaldo, Francesco, & Cavallo, Alberto.. 3-D Hand Pose Estimation from Kinect's Point Cloud Using Appearance Matching. arXiv preprint arXiv:1604.02032 (2016).

[22] Achanta, Radhakrishna, Shaji, Appu, Smith, Kevin, Lucchi, Aurelien, Fua, Pascal, & Süsstrunk, Sabine. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence, 34(11), 2274-2282 (2012).

[23] Matyunin, Sergey, Vatolin, Dmitriy, Berdnikov, Yury, & Smirnov, Maxim. Temporal filtering for depth maps generated by kinect depth camera. Paper presented at the 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), (2011).

[24] Xia, Lu, Chen, Chia-Chih, & Aggarwal, Jake K. Human detection using depth information by kinect. Paper presented at the CVPR 2011 WORKSHOPS (2011).