Human Localization at Home Using Kinect

Tanushyam Chattopadhyay
Innovation Lab, Kolkata
Tata Consultancy Services
India
t.chattopadhyay@tcs.com

Sangheeta Roy
Innovation Lab, Kolkata
Tata Consultancy Services
India
roy.sangheeta@tcs.com

UbiComp'13 Adjunct, September 8–12, 2013, Zurich, Switzerland.

10.11452494091.2497323

Abstract
In this paper authors have presented a method to localize and detect human being from Kinect captured sequence of images. The proposed method takes a sequence of gray (G) scale image and the corresponding depth (D) image as input. The gray scale image and the depth information are captured using two different sensors within the same device, Kinect and the processing are executed in the processor attached with Kinect. The proposed method localizes the human by using their motion along x, y direction and then considers all pixels connected with those pixels and over a 3D plane to accomplish the segmentation with an accuracy of 77%. Experimental results demonstrate that our method is robust against existing method for human localization.

Author Keywords
Guides, instructions, author's kit, conference publications Human Activity Detection, Kinect, Video based Localization

ACM Classification Keywords
Experimentation [Human Factors]: Algorithms.

Introduction
Robust indoor surveillance and security system, which has an excellent market traction, also involves research on

human localization. The related survey, prior to 2010 [10, 15, 8],shows that human localization can be accomplished using multiple approaches and also multiple sensors. But the main issue with using multiple sensors is the unobtrusiveness. On the other hand the mobile phone based sensing is much feasible solution but the problem comes if the peerson doesn't carry the mobile phone in bed room while sleeping and forget to carry it after waking up or while the person keeps the phone in a table while watching TV in the living room. Similarly one may not carry the mobile phone into toilet where some causalities may also occur. So it makes clear that some video based approaches are better to localize the human being in bed room and living room. But any video related system always associated with the privacy issue and additional hardware cost for installing camera. In this proposed method we have handled all these issues by using Kinect as the sensor. At the same time the Kinect, with some systems with processing power (like Xbox), are mostly installed in modern homes and thus it can also work as an ubiquitous device to localize human. Microsoft launched Kinect as a gaming platform in 2011 [14] which includes two sensors to sense RGB value as well as the depth. This invention creates a new area of research for indoor human activity recognition because of the availability of the depth information along with the color value. The depth image contains huge information about the person but the person can't be recognized from the depth image which addresses the privacy issue. Microsoft SDK can provide the skeleton points which also hides the person's identity but fails in case the person is occluded by some objects, or not in front view as reported in [17]. We assume that two such systems are installed in every living room and bedroom and kept over TV or mounted on wall. So we are ensuring that our proposed system doesn't incur any additional cost and can yet localize the human (multiple human, too) in bed rooms and living rooms where the person forgets to carry the mobile phone most likely. Moreover we are localizing the human on the edge device (like Xbox) and store the depth image only in the back end so that the privacy issues are also handled.

Existing approaches to human segmentation using video can be roughly divided into three categories mainly (i) Vision based background subtraction (ii) Silhouette based approach and (iii) Cloud point based 3D segmentation. Vision based method [6, 12, 13] is computed on subtracting each new frame from one reference image and thresholding the result. These approaches constantly re-estimate the background model in the video sequence by adaptive Gaussian mixture model, Gaussian distribution, Kalman filter to track the change. The advantage of these method is that it determines the edge of object very limpidly. But it is unable to utilize the crucial depth information for segmentation. Silhouette base methods [2, 7, 11] aim to detect particular silhouette in each frame. In [1], proposed methods can segment object in RGB-D. They suggested segmentation of RGB-D images by mean shift based algorithm. These points are placed all along the edge of object. After that for each frame, position of points is corrected using the gradients. The drawback of this approach is that prior knowledge about the object is unavailable in real time scenario. Now, cloud point based method [3, 9] is emerging as a new technique for 3D segmentation. The problem of these approach is that they have used 3D connectivity information but did not gray scale pixel value or motion for human localization.

In this paper, we address the problem of localizing human of large variations in gray and depth image obtained from Kinect sensor. Here, we aim to formulate a method that

can localize the human being from the data set [5] published by LIRIS for Human Activity Recognition and Localization (HARL). The set consisted of Kinect captured gray scale and depth images for Human Activity Recognition and Localization. Some such examples are shown in Fig.1. We have also evaluated the method against our own data set which is not a public one yet. We have estimated motion using gray scale value to localize the candidate human and then again used the gray values of pixels to remove the noise and utilized depth to find the 3D connectivity and thus, constructs bounding box containing human.

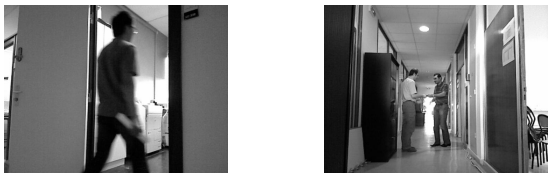

Figure 1: Some Sample Images in the data set

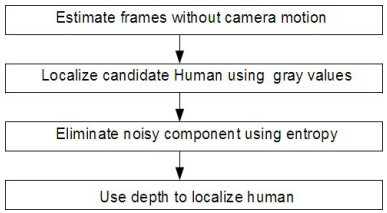

Figure 2: Overview of the proposed method

Proposed Method

In this paper, we have proposed a segmentation method that localizes the candidate objects that may represent a human being in the [5] data set. We finally label the connected components that represent the human object using depth information. One limitation of the motion based approaches is that it can not work well when the capture device is not static (e.g. Kinect placed on a Robot). Hence, first we need to identify the camera motion so that we can process the frames captured at same camera position. The overview of the proposed method has been pictorially described in Figure 2.

Camera motion estimation

In this section we have computed the mean of each frame to identify camera's motion over frame. The method of camera motion detection is described below:

- Dynamic mean $\mu_n(i,j)$ for each pixel $P(i,j)$ for the n^{th} frame of size 640x480 pixels, is calculated by following equation, where i and j specify current position of pixel P in frame along x and y axis, $\forall i \in 1,2,...,Height$ and $\forall j \in 1,2,...,Width$

$$\mu_n(i,j) = ((n-1)\mu_{n-1}(i,j) + P(i,j))/n \quad (1)$$

- Compute the number of pixel (v) satisfying the condition

$$|\mu_n(i,j) - P(i,j)| > \tau_p \quad (2)$$

where τ_p is an heuristically obtained value. We store that frame number n as a candidate for camera motion in an auxiliary buffer (aux) at an index k if $aux(k) - aux(k-1) > 30$ as the video has been capturd at 30 frames per second (FPS).

Candidate human localization

In this step we have localized the candidate human by computing the variation of Gray scale value over time. This method depicts the 2D motion estimation technique. We have proposed this technique based on the following assumptions.

- Only human will have some motion. All the background objects will remain static when there is no camera motion.
- The non-human objets having motion have less variations in gray values.

Based on the above mentioned assumptions if the camera is static from f_p to f_{q-p} where $q > p$, we can conclude that the variation of the G values of the pixels not constituting candidates will be less. So our candidate human localization technique is as follows:

- Compute median ($med(i,j)$) and standard deviation ($\sigma(i,j)$ for each frame along f_p to f_q where $q > p$.
- Now for each frame we compute variation from the median $var_{med}(i,j)$ as $|P(i,j) - med(i,j)|$ for each pixel.
- Mark the pixel (i,j) as a candidate for human object if $var_{med}(i,j) > 3\sigma$

Thus we get a binarized image for each frame where white represents the candidate human and black represents the non human using the following technique:

$$Binary(i,j) = \{1\ if var_{med}(i,j) > 3\sigma; 0 otherwise. \tag{3}$$

The output of the candidate human localization method faces two problems. (i) In case of some small camera motion, motion of the door, some object motion due to airflow, some non-human components are detected as candidate object. (ii) In case of activities with very less human motion like discussion, typing using keyboard, and talking over telephone the parts of human body, mainly the hands are detected as candidate region and rests are not. We shall discuss the approach to get rid of these problems in subsequent subsections.

Effacing of noisy component

In this section we shall describe the method to eliminate the non human objects from the candidates. The proposed method is as follows:

- Apply connected component (CC) labeling method to label all candidate pixels.
- Remove all pixels with label l if connectivity of the components belong to that label is too high or too less.
- Based on the assumptions that the chromatic variation of the background objects are less, we remove all the non human components by entropy analysis. The entropy (E) is defined as a measure of histogram dispersion. High entropy is associated with high variation of intensity while low entropy indicates that the intensity variation of pixel values is low, and hence little detail can be derived from them. The number of bins will be 256. For each CC, the local entropy h_i is computed by following equation 4. If entropy is greater than threshold limit, let τ_e (τ_e is 1.8 in our approach), the CC will be processed at next stage. Thus, we remove false positive for better accuracy. The method is:

$$E = \sum_{i=1}^{256} p_i \log(p_i) \tag{4}$$

where p_i is defined as the h_i/N if N is the total number of pixels.

where h_i is value of bin i for CC. We have used a simplified version of Adaboost algorithm to localize connected components that represent human [16]. The entropy values, area and aspect ratio of connected components are considered as features. The weak leaning scheme of our Adaboost uses a weighted histogram based optimal threshold calculator which divides the 3D feature based data into two classes (classes or connected

components representing human and background) with the smallest error. Initially the features are assumed to be distributed uniformly and the cascade of weak classifiers as above is used after boosting the model error in each step. Finally, we have gray level segmentation that will be passed in next level for depth based segmentation. In Figure 3, the example of entropy analysis is given.

(a) (b)

Figure 3: (a)before and (b)after the entropy based noise removal

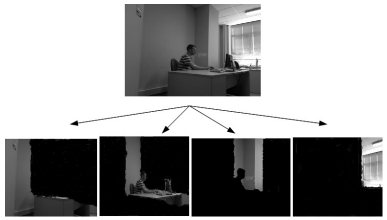

Figure 4: Objects at different depths from the camera

Human Localizing using depth connectivity
The CC representing a potential human is pruned using the depth map values provided by Kinect. It serves two purposes - it either combines neighboring disconnected components (each may be representing different body parts of a person) or it splits connected components if part of the connected component belongs to another object. Every pixel of the connected component has a depth map value and if depth map values of two neighboring pixels are within a threshold δ, the pixels are part of the same connected component else not. We have seen that a simple heuristic in evaluating the threshold δ works adequate. The maximal width of the connected component at the upper half of the connected component is taken as the threshold value δ. For a connected component representing a human being, the maximal width of the connected component at the upper half of the connected component represents the (pixel-level) width of the person in the image. The assumption is that the depth of the person (represented by δ) is maximally the (pixel-level) width of the person. One example image in Fig. 4 shows how different objects are residing on different depths from the camera. The advantage of such a threshold is that varies depending on the size of the connected component or the distance of the camera from the person.

(a) (b)

Figure 5: Example of (a)Over and (b)Under segmentation

Figure 6: Example of comparative results generated by proposed method (first row)and other approach (second row)

Result and Discussion

We have tested our proposed method on G-D data set provided [5]. In addition, we compare our method against against Zivkovic 's[18] on the same dataset. The data contains 107 video sequences as training data and 69 video sequences as test data set. The performance is

Activity	Video	Frame	Recall	Precision	Fscore	Recall	Precision	Fscore
		Our method				Zivkovic		
1	21	11,087	.51	.51	.50	.30	.22	.25
2	12	457	.58	.40	.43	.35	.24	.28
3	30	1843	.71	.49	.56	.25	.20	.22
4	44	2386	.76	.50	.58	.19	.15	.16
5	16	2349	.93	.65	.75	.15	.10	.12
6	14	2102	.81	.67	.72	.27	.20	.22
7	13	747	.59	.30	.44	.24	.19	.21
8	17	709	.75	.49	.57	.30	.21	.24
9	12	4357	.40	.41	.40	.24	.20	.21
10	15	3722	.48	.44	.45	.21	.15	.17

Table 1: Activity wise average RPF on training data-set

Activity	Video	Frame	Recall	Precision	Fscore	Recall	Precision	Fscore
		Our method				Zivkovic		
1	12	5786	.73	.64	.67	.32	.26	.28
2	9	353	.72	.42	.46	.24	.21	.22
3	17	923	.71	.47	.55	.17	.13	.14
4	27	1054	.75	.45	.54	.27	.23	.24
5	9	1059	.97	.67	.79	.17	.15	.15
6	9	1202	.91	.68	.77	.32	.23	.26
7	8	439	.76	.47	.56	.24	.23	.23
8	14	607	.77	.42	.52	.21	.16	.18
9	6	4402	.37	.43	.39	.16	.13	.14
10	7	1072	.63	.49	.55	.17	.12	.14

Table 2: Activity wise average RPF on test data-set

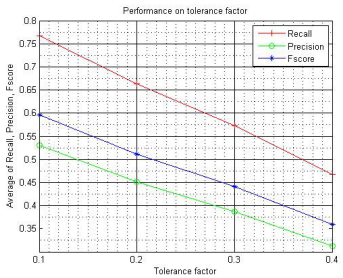

Figure 7: RPF for different tolerance factor (training dataset).

measured using the metric Recall, Precision [5], and F-score [4]. Recall (R) is defined by the number of detected action locations matched with the ground truth action locations, Precision (P) is defined as the number of true detected actions against total number of actions. The F-score (F) is computed as a function of Recall and Precision. Localization accuracy can be defined as a function of intersection of the area of ground truth and test data [5]. The activity wise RPF result on training and test data is respectively shown in Table 1 and Table 2. Figure 7 shows the effect of different tolerance values on localization accuracy on R,P, F on training data set. The comparative study of RPF at different tolerance factor on test data has been listed in Table 3. So once we add a tolerance factor with the ground truth bounding box, the corresponding area also increases and thus the R and P value reduce when the localization is quite perfect. Figure 6 shows the localization result of some images by our approach. The proposed method generates two types of errors namely under segmentation and over segmentation. We have observed that the method under segments when the human is very close (within 0.5 c.m) to the door/wall as shown in Figure 5. The proposed method also fails by performing over segmentation when depth value is noisy. The noise in depth usually comes as a system constraint of the depth sensor used in Kinect. Another reason of error in our proposed method comes when camera motion is not estimated properly.

Conclusions

In this paper, we have presented a simple yet novel approach that aims to localize human from Kinect captured data taken in real time environment. We observe that the proposed method is generic enough and can work up to a good accuracy level for any type of activity. The algorithm, at first, extracts the moving objects from the background by detecting the change of camera motion and then segments the object by depth connectivity. We have tested our method on standard data set. Overall qualitative result is over 77% shows that the method is promising for real world application for human localization. The system can also quantify movement levels and perhaps track multiple humans around and can also get the human's identity from facial and/or skeletal recognition in future without hampering the privacy issue.

Tolerance factor	Recall	Precision	Fscore
.1	0.7778	0.5207	0.5962
.2	0.6858	0.4455	0.5162
.3	0.5905	0.3809	0.4436
.4	0.5028	0.3224	0.3753

Table 3: Variation of RPF against tolerance on test dataset

References

[1] A. Bleiweiss and M. Werman, "Fusing Time-of-Flight Depth and Color for Real-Time Segmentation and Tracking", *Proceedings of the DAGM 2009 Workshop on Dynamic 3D Imaging Pages 58 - 69*

[2] A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov, "Bilayer segmentation of live video," *Proceedings of the Conference on Computer Vision and Pattern Recognition, CVPR 2006*

[3] A. Sinha, T. Chattopadhyay, A. Mallik. "Segmentation of Kinect Captured Images using Grid Based 3D Connected Component Labeling," *Proceedings of the 9th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP 2013*

[4] C.J.van Rijsbergen. 'Information Retrieval,' *Butterworths, London, 2nd edition,1979.*

[5] C. Wolf, J. Mille, L.E Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E Dellandrea, C.-E. Bichot, C. Garcia, B. Sankur, "The LIRIS Human activities dataset and the ICPR 2012 human activities recognition and localization competition," *Technical Report RR-LIRIS-2012-004, LIRIS Laboratory, March 28th, 2012.*

[6] D. Koller , J. Weber,T. Huang,J. Malik, G. Ogasawara, B. Rao, S. Russell, "Towards robust automatic traffic scene analysis in realtime," *Proceedings of the 33rd IEEE Conference on Decision and Control, 1994, vol.4, pp.3776,3781, 14-16 Dec 1994*

[7] D. Terzopoulos and R. Szeliski, "Tracking with kalman snakes," *In Active vision, Andrew Blake and Alan Yuille (Eds.). MIT Press, Cambridge, MA, USA 3-20.*

[8] E. Munguia-Tapia , S. S. Intille and K. Larson, "Activity Recognition in the Home Using Simple and Ubiquitous Sensors," *Proc. 2nd Intĺ Conf. Pervasive Computing (Pervasive 04), pp.158 -175 2004*

[9] F. Hegger, N. Hochgeschwender, K. Gerhard,Kraetzschmar and P. G. Ploeger. 'People Detection in 3d Point Clouds using Local Surface Normals.' *RoboCup 2012: Robot Soccer World Cup XVI, Lecture Notes in Computer Science Volume 7500, 2013, pp 154-165, Mexico, 2012*

[10] J. Ben-Arie, Z. Wang, P. Pandit, S. Rajaram, "Human activity recognition using multidimensional indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol.24, no.8, pp. 1091-1104, Aug 2002*

[11] M. Isard and A. Blake, "Condensation conditional density propagation for visual tracking," *International Journal of Computer Vision, vol. 29, no. 1, pp. 528, 1998*

[12] N. Friedman, S. Russell, "Image Segmentation in Video Sequences: A Probabilistic Approach," *Inc., San Francisco, 1997*

[13] P. Kaew, T. Pong, and R. Bowden, "An Improved Adaptive Background Mixture Model for Realtime Tracking with Shadow Detection," *In Proc. 2nd European Workshop on Advanced Video Based Surveillance Systems, AVBS01. Sept 2001*

[14] The teardown. (2011), *Engineering Technology, vol. 6, no.3, pp. 94-95, April 2011.*

[15] U. Maurer, "Activity Recognition and Monitoring Using Multiple Sensors on Different Body Positions," *Proc. Intĺ Workshop on Wearable and Implantable Body Sensor Networks, pp.99 -102 2006*

[16] Y. Freund, R. E. Schapire, 'A Short Introduction to Boosting," *Journal of Japanese Society for Artificial Intelligence,14(5):771-780, September, 1999.*

[17] Yang Zhao; Zicheng Liu; Lu Yang; Hong Cheng, ''Combing RGB and Depth Map Features for human activity recognition," *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific , pp.1,4, 2012*

[18] Z.Zivkovic, F. van der Heijden, 'Efficient Adaptive Density Estimapion per Image Pixel for the Task of Background Subtraction,' *Pattern Recognition Letters, vol. 27, no. 7, pages 773-780, 2006.*