View-Invariant Human Detection from RGB-D Data of Kinect Using Continuous Hidden Markov Model

Sangheeta Roy and Tanushyam Chattopadhyay

TCS, Innovation Labs,
Kolkata, India
{roy.sangheeta,t.chattopadhyay}@tcs.com

Abstract. In this paper authors have presented a method to detect human from a Kinect captured Gray-Depth (G-D) using Continuous Hidden Markov models (C-HMMs). In our proposed approach, we initially generate multiple gray scale images from a single gray scale image/ video frame based on their depth connectivity. Thus, we initially segment the G image using depth information and then relevant components were extracted. These components were further filtered out and features were extracted from the candidate components only. Here a robust feature named Local gradients histogram(LGH) is used to detect human from G-D video. We have evaluated our system against the data set published by LIRIS in ICPR 2012 and on our own data set captured in our lab. We have observed that our proposed method can detect human from this data-set with a 94.25% accuracy.

1 Introduction

Human activity detection for indoor and outdoor surveillance has a major research interest since last two decades. Activity detection is very effective in human based application like video indexing and retrieval, intelligent human machine interaction, video surveillance, health care, driver assistance, automatic activity detection and predicting person behavior. Some of such applications can be found in literature like [1] (in office), [2], [3], [4] (in retail stores), and in [2] for elderly people monitoring. A significant survey on human activity recognition [22] concludes that the availability of depth information can improve the recognition accuracy. So, the onset of Kinect, a Microsoft gaming platform, with the capability to capture depth along with the color information creates a new area of research on the problem of human activity recognition because of the availability of the depth information along with the color value. On the other hand the results of the HARL competition organized in ICPR 2012 [6] shows that the human activity recognition accuracy increases with the increase in human localization accuracy. Therefore, the human localization approaches on RGB videos need to be modified with the availability of depth information. Traditional back ground modeling based methods didn't work on the RGB-D data

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 325–336, 2014.

when the camera is not static and the lighting condition varies over time. The video frames, on which we have tested our system, contains human leaning over wall, one person occludes another person partially which makes the task of human detection more difficult as color based segmentation didn't work, too. One such example image is shown in Figure 1. In this paper we have concentrated on the problem of human detection from Kinect captured videos by combining RGB and depth information. We have proposed a system that can detect the presence of human in such a G-D video using machine learning technique, namely C-HMM. We have compared our method against other methods like on [6] dataset as well as on our own data set.

Fig. 1. Apparently touching objects in 2D projection plane

2 Related Literature Survey

There have been a large number of methods dealing with recognition of human activity in color image. The recognition problem is very difficult because of large variation involved in human appearances and views. From the literature review it can be seen that, face detection algorithm is often applied [8], [9] for recognizing human in image and video. They addressed the solution in this field either by feature based or image based approach. Bottom-Up analysis is done utilizing feature based approach. Window scanning technique is employed in later case. Muhammad and Atif et al. [10] combines color and motion information to detect face and hence, human. In [11], human detection is achieved by integrating the cascade-of-rejectors concept with the Histogram of Oriented Gradients (HoG) of variable size blocks. They used an AdaBoost training algorithm to learn a cascade of rejecters to eliminate the non human image patches. A new learning method for human detection is proposed in [12] which is based on weak classifier, built from L1-norm minimization learning scheme (LML). The augmentation of edge-based features, texture measures and color information have been used by Schwartz et al. [13]. They handle this high dimensionality resulting from the combination of features, using dimensionality reduction technique Partial Least Squares (PLS). There are several papers [16], [21] that addressed human detecting problem based on body-part. Bhaskar and Jordi [21] presents a technique for view invariant human detection using body-part(head, leg, arm etc) detectors. Human detection is proposed by probabilistic body part assembly in [16]. First, different body parts are detected by Adaboost and after that

detected parts are assembled using RANSAC. Mohan et al. [17] have used hierarchical classification architecture using SVM. They also use different components detector at primary level. These method perform poor due to failure of detector in handling of variability of body parts. In [14] authors have presented a graphical model based approach for estimating poses of upper-body parts by fusing depth and RGB color data based on Haar cascade. This method works well for detecting upper-body human pose but not for full human shape. In addition to this, the primary focus of the above methods is identification of front view of human but not view-invariant. Lu et al. [15] uses 2-D head contour model and a 3-D head surface model to detect head of human. Next, segmentation scheme is used to extract the whole contours of human based on head point. The performance of the method extremely depends on the accurate head detection and it uses only depth information not both depth and color. Therefore, from the above discussion, it can be concluded that there are methods to improve human recognition rate but these methods concentrate on color and edge information of images and a little variation in view point but not on color, depth and large view where we can expect much more challenges compared to exiting state of the art methods. Again, the literature also suggests that the segmentation of the input prior to recognition can lead to higher recognition rate [18]. But most of the methods do not take care about proper segmentation. Therefore, improving human detection in Kinect through segmentation, irrespective of view point and background complexity is challenging. In this paper, we present a method of human detection in images captured by Kinect and performs recognition of human using HMM instead of detecting individual body parts. HMM is popular and found robust in printed and handwritten text recognition. It motivates us to use HMM in human recognition. To the best of our knowledge, there is no work on human detection using HMM. This work is motivated by our preliminary work reported in [23]. Our goal is to classify observed feature sequences into human or nonhuman category utilizing depth and color information. Hence, this paper presents view-invariance human detection method using HMM approach. The rest of the paper is organized as follows. The proposed method is described in Section 3 which includes depth based segmentation method, noise cleaning and view-invariant human detection method using HMMs. Section 4 presents the experimental results. Finally, conclusion is drawn in Section 5. In this paper human detection and human localization phrases were some times used interchangeably.

3 Proposed Method

Our proposed method initially attempts to localize the human from the video frame to reduce the time complexity of the method. Kinect based systems have a major limitation that it can work on indoor environment only as the depth sensor uses Infra Red (IR) signal. So the captured videos contain some wall, floor and ceiling parts. We remove noisy components from the candidate regions by removing the floor/ceiling from the image if required. Next we extract the LGH features and train them to classify into human and non-human class. Finally we use C-HMM classifier to separate out human and non human.

3.1 Segmentation Using Depth Connected Operator

Kinect provides two sets of values namely the G image and the D information for each video frame. Any G image is a 2-Dimensional projection of the 3-Dimensional objects located at different distances from the sensor. So it is not possible to segment them on their depth unless we have the depth information exclusively. So we have used the depth information to segment each video frame into number of layers so that each segment contains the pixels those are connected over depth. 8 neighbor connected component analysis is a common method of image processing but we have used depth connectivity instead. The method of depth connectivity is described in details in [23]. The method of is based on the concept of of connected operator for sets as described in [7]. The concept of connected operator on set says that an operator ψ can be said to be a connected operator if the symmetrical difference $P \Delta \psi(D)$is exclusively composed of connected components of D or its compliment D^C. Here we use depth information obtained from the depth sensor of Kinect as the connected operator ψ applying over the gray scale image pixel set G. As per the definition of partition space as stated in [7] this ψ operator partitions the space G into two disjoint subsets G_i and G_j such that $G_i \bigcap G_j = \emptyset \forall i \neq j$. Our proposed method of such segmentation is described below:

- For each video frame/image construct a set (P) by concatenating the gray and the depth information. So P is a two tuple set containing gray value (g) of the pixel and its depth (d) information from the sensor. So $P = (g_i, d_i)$
- Every pixel who are connected over depth are mark with the same depth map label. We have used a threshold to check the depth connectivity.
- For each depth map label we create one image with keeping the gray value of those pixels as it was and mark the rest of the pixels as black.

In Figure 2 we have shown one such example video frame/image and its four out of five partitions. In this image the backgrounds are marked as black. This image shows that one man is standing in one partition and rest two are residing in an another partition. The main advantage of this method over the simple depth quantization is that human are not separated into two different segments in our approach when the human is between the depth separation.

3.2 Floor and Ceiling Removal

The depth connectivity based segmentation generates multiple images form a input video frame those are connected over depth. Each of these images represents the objects connected over depth by their corresponding gray scale value. The height and width of the images are same as that of height and width of the original video frame/image. The backgrounds are marked as black and an additional tag is added to mark them as background. We have implemented the additional back ground tag information by using a Boolean Flag which is set for background and FALSE for foreground. We shall refer each of these images as a partition of the original video frame. Connected component analysis is used

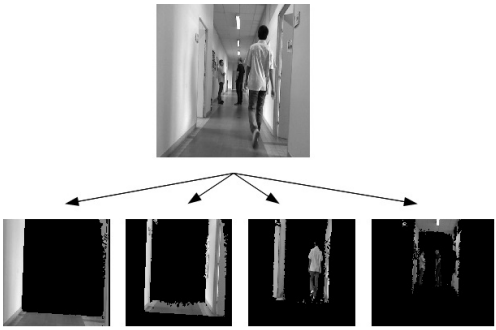


Fig. 2. Objects at different depth from the camera

to mark the different components in each of these partitions. The outcome of connected component analysis shows that some components containing human include some non-human objects through the floor/ceiling. In such cases the component width is almost equal to the width of the input video frame/image. So we formulate a vertical pixel projection based method to eliminate the floor region and thus separate out the human component from the rest part for such components. Here is the description of that proposed method:

- If the width of the component is greater than 75% of the width of input video frame/image execute the following steps. We have used this value as a heuristic obtained from our experiments.
- Count the number of pixels (cnt_i) in a column i for which the background flag is FALSE
- Run a K-Means clustering with K=2 on cnt_i $\forall i \in 0, H$ where H is the height of the image
- The two cluster will represent the columns with higher number of FG pixels (C_1) and lower number of FG pixels (C_2)
- Make the flag for all the pixels from the column those are residing in C_2 as TRUE

This method is explained using the example images. Figure 3. a shows one such partition which contains both human and non human objects. Now the corresponding vertical projection is shown in Figure 3. b. Finally the outcome of our proposed method is shown in Figure 3. c which shows that the connected component is now divided into multiple components. We run connected component analysis as described above on each of these partitions after being modified by the above method. Finally we get some segments containing either human or non human objects. Some outcomes of this method are shown in Figure 4. a and Figure 4. b. Finally some non human objects are removed by effacing of noisy component as described in [23].

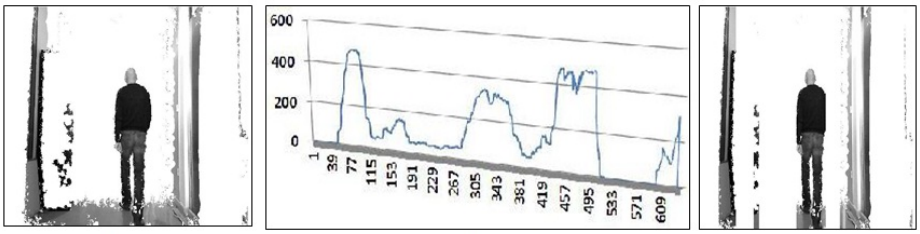


Fig. 3. a) One partition b) Histogram of FG pixels in each column c) Image after floor estimation and removal

(a) (b)

Fig. 4. a) Some segments containing human b) Some segments containing non human

3.3 LGH Feature

Sliding Window is a common technique for many signal processing applications like speech and character recognition. We have used a rectangular sliding window of l pixel width to collect the features of a component. It is shifted from left to right across the normalized gray level segmented image to generate feature vector sequences at each shift position. Adjacent image windows overlap in the vertical direction. This results in a vast amount of features for each frame. The frames are normalized to a pre-defined height before the feature extraction stage. Figure 5 illustrates an example of the sliding window feature extraction process. This feature extraction approach is based on the calculation of the local gradient histogram [24]. Each sub-image is sub-divided into 4 * 4 blocks and from all pixels in each block a histogram of gradient orientations is calculated. Here, we considered 8 orientations. Therefore, the final feature vector concatenation of the 16 histograms results in vector containing 128 features.

- For the entire image the horizontal and vertical motion components V_x and V_y are determined and a gradient magnitude (m) is computed for each pixel.
- The field vector $\overrightarrow{V}$ is sliced up in an L bin histogram.

– Each bin specifies a particular octant in the angular radian space. Here we consider 8 bins ($360°/45°$) of angular information.
– The concatenation of the 16 histograms of 8 bins provides a 128-dimensional feature vector for each frame.

Let $\overrightarrow{V}=(V_x, V_y)$ and histogram $H = h(1), h(2), ..., h(8)$. The histogram is constructed by quantizing $\theta(x, y) = tan^{-1}\frac{V_y}{V_x}$ and adding up $m(x, y) = \sqrt{(V_x + V_y)}$ to the bin indicated by quantized θ. In mathematical definition,

$$h(i) = \begin{cases} \sum_{x,y} m(x, y) & when\ \theta \in ith\ octant \\ 0 & otherwise \end{cases} \quad (1)$$

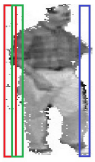


Fig. 5. Path of overlapping sliding window (shown in different colors)

3.4 HMMs Based Recognition

HMMs have been proven to be a powerful stochastic approach and found robust in speech and text (printed and hand written) recognition. It is a special type of dynamic Bayesian networks. We have used HMM in our application because of its ability to cope with variable-length observation sequences obtained from images. Generally, HMM follows the first-order Markov assumption where each state S_t at time t depends only on the state S_{t-1} at time $t - 1$. It contains a fixed number of hidden states. HMM is characterized by 3 matrices: state transition probability matrix A, symbol output probability matrix B, initial state probability matrix π. The parameter A, B and π are determined during learning process. The image is represented as a sequence of feature vectors $X = x_1, x_2, ..., x_T$ also known as sequence of frames. In HMMs, the likelihood of emitting a frame x_t in state i is modelled using a GMM. For a model λ, if O is an observation sequence $O = (O_1, O_2, .., O_T)$ which is assumed to have been generated by a state sequence $Q = (Q_1, Q_2, ., Q_T)$, of length T. We calculate the observations probability or likelihood as follows:

$$P(O, Q|\lambda) = \sum_{Q} \pi_{q1} b_{q1}(O_1) \prod_{T} a_{qT-1} qT b_{qT}(O_T) \quad (2)$$

where π_{q1} is initial probability of state 1, is transition probability from state i to state j and is output probability of state i. The observation likelihoods are computed from a Gaussian Mixture Model (GMM).

$$b_j(x) = \sum_{k=1}^{M_j} c_{jk} \mathcal{N}(x, \mu_{jk}, \Sigma_{jk}) \quad (3)$$

where, M_j is the number of Gaussians assigned to j. and $\mathcal{N}(x, \mu, \sigma)$ denotes a Gaussian with mean μ and covariance matrix σ and c_{jk} is the weight coefficient of the Gaussian component k of state j. Next, the Viterbi decoding searches the subsequence of an observation that matches best to a given HMM. For a classifier of C categories, we choose the model which best matches the observation from C HMMs $\lambda_m = A_m, B_m, \pi_m$, where $m = 1, ..., C$, and $\sum_{m=1}^{c} \lambda_m = 1$. This means when a unknown sequence of unknown category is given, we calculate $P(\lambda_i|O)$ for each HMM $\lambda_m$and select λ_c^* , where

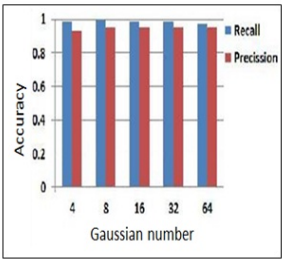
$$c^* = argmax_m P(\lambda_m|O) \tag{4}$$

An HMM should be learned for each class. For our application 2 HMMs have been used to model human and non human. The 128 dimensional LGH features extracted from each sliding window of image were used to represent sequence of local feature vectors. The extracted feature of each window is arranged row-wise to form complete vector set. The task of the learning algorithm is to find the best set of state transitions and observation probabilities. The Baum-Welch recursive algorithm is used to obtain the final parameters of HMMs. For classifying an observed symbol sequence O, classifier choose the model whose likelihood is highest as the recognition result. The recognition is performed using the Viterbi algorithm.

4 Result and Discussion

It is well known fact that when the training sample size is small, the recognition rate becomes low. The performance of any recognition system depends not only size number but also the well variation of samples used in training phase as these both are very crucial to estimate HMM parameters. To construct a robust recognition system, appropriate training patterns are important. This means training pattern should capture the maximum test pattern variation. In this experiment training and test samples were completely segregated to make this evaluation more reliable. We have developed a working prototype of human detection using x86 PC system. C/C++ and OpenCV library were used for segmentation and feature extraction on a windows environment. We have used the popular HTK toolkit for HMMs training and evaluation [20]. Our data set includes G frames of the Kinect module are encoded as lossy JPEG images and D frames of the Kinect module are encoded in lossy 16bit JPEG2000 images with a compression factor of 20, resulting in 30KB per frame. We validated our proposed system on two data sets. One of these data sets is our own. This data set includes the Kinect captured RGB-D images recorded at our lab which includes more than 30 videos each having more than 1000 frames. The other data set used in our experiment is the [6] data set published by LIRIS for Human Activity Recognition and Localization (HARL) in ICPR 2012. This database contains 107 training and 69 test videos with gray and depth information, captured in Kinect sensor in indoor scenes. These videos contain different views of human with a vast range of poses like standing, walking, siting while performing action. These are

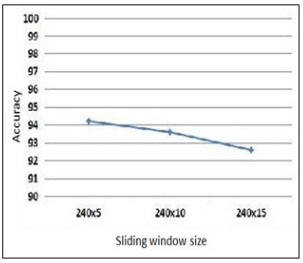
taken under different illumination condition with camera movement and scale variation by different person against complex background. We have used 35,434 segments during training phase. Among them 14,867 segments are positive and 20,567 segments are negatives. Positive consists of human containing image. On the other hand, representation of negative images are floor, ceiling and wall component (shown in Figure 4. a, 4. b). We have applied our method for all those test images for human detection. For human detection, a set of human images is used in the training of HMM. The images in the training set represent different views of different persons taken from segmentation results generated by previous step. Once the models have been trained for human and non human, the Viterbi algorithm was applied to find the most likely state sequence and its likelihood in the recognition process. The observation sequences for a image are formed from image window or block that are extracted by scanning the image from left-to-right. The observation vectors consist of 128 features. In training set, we have measured the performance of system using a 10-fold cross validation. Recall, describes how many object have been correctly detected, with respect to the total number of objects in the dataset and Precision evaluate how many detected objects are matched with respect to the total number of detected objects. We define recall (R) as $R = \frac{c}{c+m}$ and precision(P) as $P = \frac{c}{c+fp}$ where c indicates the correct recognition, m means misses and fp is the false positive. If we can't detect a human we define this error as miss as our intention of the research is to localize the human. On the other hand if our proposed method detects one component as human though it is actually a non human one, we define it as false positive. We have observed that our P is always less than R. The main reason behind that error comes from the lack in training of different instances of non human objects. In Figure 6. a we have shown the recall and precision against Gaussian number. We have evaluated our algorithm performance in terms of *variation of Gaussian number, variation of states and different sliding window size*. Experiments show that learned HMM classifiers have good performance for detecting human. Results show that most of the humans have been correctly recognized. The HMM system was tested with different number of states and Gaussian numbers. In Figure 6. b, the recognition accuracies are given in terms of both of these. Next, we inspected that increasing the number of states up to 8 states improves the performance of the HMM recognizer, but a larger number and small number states decreases its accuracy. The best average accuracy was obtained for an HMM with 7 states and a mixed output probability of 16 Gaussians, 94.25% with Local gradient Histogram(LGH) on unseen samples. We have shown the effect of sliding window size on recognition accuracy in Figure 6. c. We observed that the increase in sliding window size after a limit reduces the recognition accuracy. The advantage of the sliding window based-HMM is that the detection of human is very robust. Some qualitative images detected from our approach are shown in Figure 7. a and c. Our proposed method can detect the human in case of improper segmentation and even when the human is blended while sitting and partially occluded by other object and human as shown in Figure 7. b. We have compared the performance of our proposed



a

Number of states	Number of Gaussians					
		4	8	16	32	64
	4	81.63	83.67	85.38	87.79	88.82
	5	80.12	82.38	87.34	90.39	91.56
	6	84.21	87.32	90.12	91.12	92.34
	7	91.69	94.25	94.21	93.93	92.97
	8	83.12	85.23	88.38	90.23	89.12

b



c

Fig. 6. a)Recall and precision accuracy on testing set b) Recognition accuracy for different number of states and Gaussians on testing set c) Recognition accuracy vs Sliding window size

a b c

Fig. 7. a) Example of view-invariant human recognition generated by proposed method b) Recognition of partial segmented human part c) Detection results of the proposed method

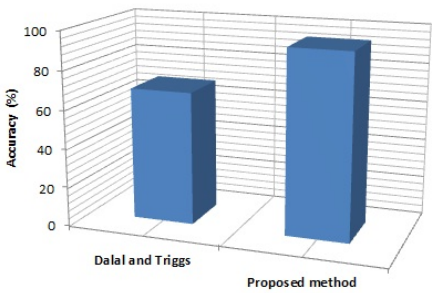


Fig. 8. Comparative result

method against a state of the art method described in Figure 8. We observe that our algorithm outperforms the [19] method. [19] works only when the human is in upright position but in the real world human can be found in any orientation.

5 Conclusions

In this paper we have presented a method that combines color and depth information in the pre-processing phase to localize the candidate segments containing human being and thus the proposed method overcomes the limitations of 2D based methods. The use of depth with robust machine learning framework makes the system robust against variations in viewpoint. So the performance of the proposed system outperforms the state of the art methods. We have shown that using locally normalized histogram of gradient orientations features descriptors in a overlapping window with HMM framework gives very good results for person detection. These results show that our method is promising to recognize human for numerous applications such as video indexing and retrieval, intelligent human machine interaction, video surveillance, health care, driver assistance, automatic activity detection and predicting person behavior. Performance of our proposed method partly depends on the candidate human localization and all errors are mostly coming from the failure in proper localization. This method can precisely identify whether the candidate segment contains human being or not and thus if the area of the candidate region is much bigger than the actual human region, the accuracy falls. So we are currently working to find a better localization method.

References

1. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: RGBD Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments. In: Proc. ISER 2010 (2010)
2. Trinh, H., Fan, Q., Pankanti, S., et al.: Detecting Human Activities in Retail Surveillance Using Hierarchical Finite State Machine. In: Proc. ICASSP 2011, pp. 1337–1340 (2011)
3. Trinh, H., Fan, Q., Gabbur, P., Pankanti, S.: Hand tracking by binary quadratic programming and its application to retail activity recognition. In: Proc. CVPR 2012, pp. 1902–1909 (2012)
4. Gabbur, P., Pankanti, S., Fan, Q., Trinh, H.: A pattern discovery approach to retail fraud detection. In: Proc. KDD 2011, pp. 307–315 (2011)
5. Sinha, A., Chattopadhyay, T., Mallik, A.: Segmentation of Kinect Captured Images using Grid Based 3D Connected Component Labeling. In: Proc. VISAPP 2013, pp. 327–332 (2013)
6. Wolf, C., Mille, J., Lombardi, L.E., Celiktutan, O., Jiu, M., Baccouche, M., Dellandrea, E., Bichot, C.-E., Garcia, C., Sankur, B.: The LIRIS Human activities dataset and the ICPR 2012 human activities recognition and localization competition, Technical Report RR-LIRIS-2012-004. In: Proc. ICPR 2012 (2012)

7. Salembier, P., Serra, J.: Flat Zone Filtering, Connected Operator, and Filters by Reconstruction. Proc. IEEE Transactions on Image Processing 1995, 1153–1160 (1995)
8. Jin, R., Hauptmann, A.G.: Learning to Identify Video Shots With People Based on Face Detection. In: Proc. ICME 2003, pp. 6–9 (2003)
9. Low&, B.K., Hjelmas, E.: Face Detection: A Survey, Computer Vision and Image Understanding 2001 (2001)
10. Khan, M.U.G., Saeed, A.: Human Detecion in Videos. Journal of Theoretical and Applied Information Technology (2009)
11. Zhu, Q., Yeh, M., Cheng, K., Avidan, S.: Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In: Proc. CVPR 2006, pp. 1491–1498 (2006)
12. Xu, R., Zhang, B., Ye, Q., Jiao, J.: Cascaded L1-norm Minimization Learning (CLML) classifier for human detection. In: Proc. CVPR 2010, pp. 89–96 (2010)
13. Schwartz, W.R., Kembhavi, A., Harwood, D., Davis, L.S.: Human detection using partial least squares analysis. In: Proc. ICCV 2009, pp. 24–31 (2009)
14. Jain, H.P., Subramanian, A., Das, S., Mittal, A.: Real-time upper-body human pose estimation using a depth camera. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2011. LNCS, vol. 6930, pp. 227–238. Springer, Heidelberg (2011)
15. Xia, L., Chen, C., Aggarwal, J.K.: Human Detection Using Depth Information by Kinect. In: Proc. CVPRW 2011, pp. 15–22 (2011)
16. Micilotta, A., Ong, E., Bowden, R.: Detection and tracking of humans by probabilistic body part assembly. In: Proc. British Machine Vision Conference 2005, pp. 429–438 (2005)
17. Mohan, A., Papageorgiou, C., Poggio, T.: Example-based object detection in images by components. Proc. IEEE Transaction on Pattern Analysis and Machine Intelligence 2001, 349–361 (2001)
18. Alon, J., Athitsos, V., Yuan, Q., Sclaroff, S.: A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation. Proc. IEEE Transaction on Pattern Analysis and Machine Intelligence 2009, 1685–1699 (2009)
19. Dalal, N., Triggs, B., Schmid, C.: Human Detection Using Oriented Histograms of Flow and Appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)
20. Young, S.J., Jansen, J., Odell, J.J., Ollason, D., Woodland, P.C.: The HTK Hidden Markov Model Toolkit Book. Entropic Cambridge Research Laboratory (1995)
21. Chakraborty, B., Rudovic, O.N., Gonzlez, J.: View-invariant human-body detection with extension to human action recognition using component-wise HMM of body parts. In: Proc. FG 2008, pp. 1–6 (2008)
22. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Computing Surveys (CSUR) 2011 43(3) (2011)
23. Chattopadhyay, T., Roy, S.: Human Localization at Home Using Kinect. In: Proc. HomeSys, UbiComp (Adjunct Publication) 2013, pp. 821–828 (2013)
24. Jos, A., Serrano, R., Perronnin, F.: Handwritten word-spotting using hidden Markov models and universal vocabularies. In: Proc. Pattern Recognition 2009, pp. 2106–2116 (2009)