Human Activity Monitoring System Based on RGB-Depth Sensor

Tin Zar Wint Cho, May Thu Win
Faculty of Information and Communication Technology
University of Technology (Yatanarpon Cyber City), PyinOoLwin
tinzarwintcho@gmail.com, maythu.ycc@gmail.com

Abstract

This paper is related to the domain of human activity recognition in both depth images and skeleton joints. In this paper, for the detection task, a RGB-D sensor (Microsoft Kinect) is used. To obtain discriminative features for action detection, combination of a depth shape features from the 3D space and joints features are investigated. The detection and classification of such features is accomplished by the posture analysis technique, based on K-means and finally, activity recognition are performed by means of HMMs built on the set of known postures to improve performance and accuracy. The proposed system can be evaluated on a new dataset which contains five activities (standing, walking, sit down, lying and bending) and another public dataset MSRDailyActivity3D. The proposed system can be applied to the specific domain of healthcare system including home and hospital to keep older adults functioning at higher levels and living independently.

1. Introduction

Monitoring human activities of daily living is an essential way of describing the functional and health status of a human. Therefore, human activity recognition (HAR) is one of essential components in personalized life-care and healthcare systems; especially for the elderly and disabled [5].To monitor daily activities of the elderly people, video cameras can be deployed in smart environments such as smart homes or smart hospitals to acquire time-series activity video clips [1].

Human activity recognition (HAR) is a hot research topic since it may enable different applications, from the most commercial (gaming or Human Computer Interaction) to the most assistive ones. In this area, HAR can be applied, for example, to detect dangerous events or to monitor people living alone. This task can be accomplished using different sensors, mainly represented by wearable sensors or vision-based devices [12].

As the imaging technique advances, the advent of depth sensors (e.g., Microsoft Kinect) brings great benefits to a variety of visual recognition tasks including object recognition, indoor place segmentation, as well as human activity recognition.

Depth maps, which are a good source of information because they are not affected by environment light variations, can provide body shape, and simplify the problem of human detection and segmentation. Furthermore, the availability of skeleton joints extracted from the depth frames allows having a compact representation of the human body that can be used in many applications.

In this paper, a methodology of HAR based on the recognized body parts features of human depth shapes and skeleton joints is presented and the hidden Markov Model (HMM) is used to recognize human postures.

This paper is organized as follows. Related work is outlined in Section 2. The system architecture is described in Section 3. Section 4 presents the dataset creation. Section 5 presents the MSRDailyActivity3D dataset. Section 6 presents the expected outputs. Conclusions are presented in Section 7.

2. Related Work

Firstly, we review some related activity recognition works based on RGB or RGB-D streams.

In [7], the human body was represented in terms of silhouettes, extracted from RGB images, which were used as input to a framework based on HMM. The general weakness of the methods based on RGB data is that the complexity of the processing chain (e.g., background removal, image normalization), required to obtain adequate silhouette features, limits real-time use. Moreover, such systems are not robust enough to be applied in unconstrained situations, e.g., environments with complex backgrounds or low lighting conditions.

A method to obtain silhouettes from depth information only is presented in [13]. This solution is motivated by the fact that depth images are intensity invariant and then more robust to appearance variations of the human body than RGB ones. The authors trained their system by creating a codebook of body poses so that a new human pose can be represented by its most similar code-word. The major issue of this approach is related to the background removal routine, which needs background images to be known previously, or users to be located away from the background. Such constraints are not always applicable to real contexts.

Here, we review activity recognition approaches based on data provided by the Kinect. In [8], human bodies are modeled as a set of kinematic joints, and actions are defined by the interactions that occur between subsets of these joints. The authors proposed a new feature, called local occupancy feature (LOP), to describe each 3-D joint and introduced the concept of *actionlet* to define a particular conjunction of LOP features. Due to the great number of possible *actionlets*, a data mining technique is used to discover the most discriminative ones and represent an action as an *Actionlet Ensemble*, i.e., a combination of *actionlets*.

A posture-based approach for action recognition is presented in [15]. The authors represent salient postures as a bag of 3-D points obtained by projecting and sampling the depth maps onto three orthogonal planes. Each posture is then associated with a specific node of an action graph, which is used to model the dynamics of different actions. This method yields better results than those based on 2-D silhouettes; however, 3-D projections obtained from the depth maps are usually quite noisy due to low resolution of the sensor. Thus, further interpolation steps are generally required to repair corrupted projections, and this compromises the overall recognition time.

A histogram-based representation of human postures is presented in [11]. In this representation, the 3-D space is partitioned into n bins using a spherical coordinate system so that each of the 12 considered joints belongs to a bin with a certain level of uncertainty. Linear discriminant analysis (LDA) for C classes is performed to reduce the dimensions of the feature space from n to $C - 1$, and the obtained features are clustered into K visual words. The activities are then represented as sequences of visual words and recognized using discrete HMM classifiers. The main limitations of this approach are the adoption of a complex model for representing the joints and the consequent need for reducing the dimensionality of the feature vectors by means of LDA.

The authors of [6] addressed the problem of reconstructing valid movements from incomplete, i.e., noisy, postures captured by the Kinect. In particular, broken postures are corrected by searching through a motion database for similar postures, which are kinematically valid. Although the method improves wrongly detected postures, it assumes that the motion database always contains postures similar to the ones performed by the user, which is not always true in practical situations.

The authors of [10] proposed an algorithm based on hierarchical maximum entropy Markov model (MEMM) to represent a single activity as a composition of a set of sub-activities. Each sub-activity is initially modeled by analyzing about 700 features extracted from RGB and depth images; then, it is associated with a high-level activity by means of a two-layer MEMM. The framework proposed in [3] aims to demonstrate that using both depth and grayscale data can improve the performance of recognizing complex activities, e.g., users interacting with objects in the environment. Experimental results show that promising recognition and localization accuracies can be obtained, but a computation time analysis is missing. Therefore, suitability for real-time applications is unknown.

The effectiveness of using both color and depth information for activity recognition is also reported in [4]. The authors collected a dataset, called RGBD-HuDaAct, which contains 12 activities performed by 30 different subjects at a distance of about 3 m from a Kinect device. Results obtained by applying multimodal feature representation, i.e., combining color and depth information, are compared to the unimodal counterparts; however, neither an evaluation of time consumption nor a comparison with other approaches is provided.

3. Proposed System and Methodology

The main objective of the proposed system focuses on developing effective feature representation for human activity recognition in both depth sequences and skeleton joints for healthcare system.

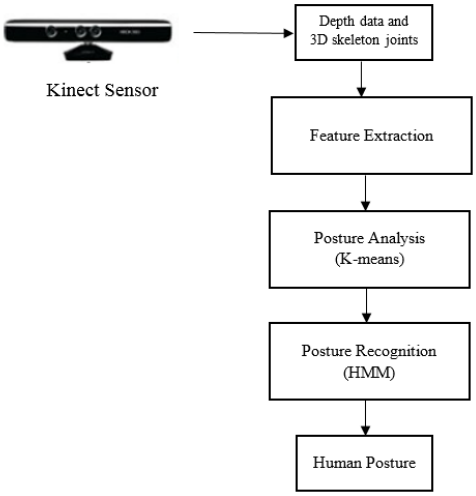
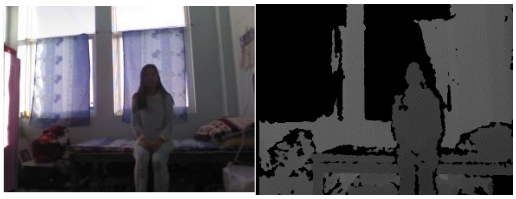


Figure1. System design for proposed system

(a)



(b)

Figure2. Example of color images and depth maps detected by means of Kinect. (a) Standing activity and (b) Sit-down activity

The system design of the proposed system is shown in Figure 1. In this system, depth data and 3D skeleton joints are used as the inputs from the Kinect Sensor and multimodal feature extraction methods such as depth shape features from 3D-space and skeleton joints features are used. The posture analysis technique, based on K-means, is used to detect such features and Hidden Markov Model (HMM), built on the set of known postures, is used to recognize the human activity to improve performance and accuracy.

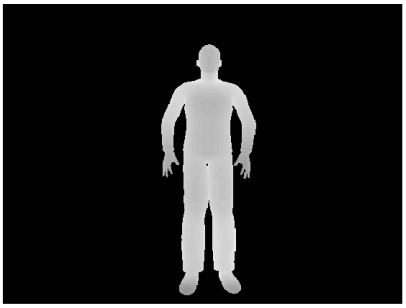


Figure3. Example of Depth silhouette detected by means of Kinect

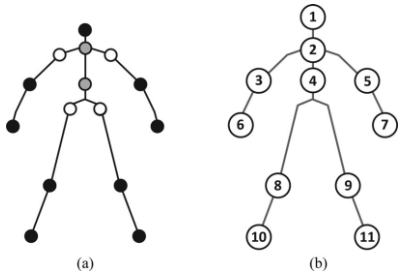


Figure4. Examples of Skeleton joints detected by means of Kinect. (a) Fifteen joints. Reference joints(gray): *neck, torso*. Selected joints (black): *head, elbows, hands, knees, feet*.Discarded joints (white): *shoulders, hips*. (b) Eleven joints of the feature set.

Figure 2 shows some example of color images and depth maps detected by means of Kinect. Example of depth silhouettes detected by means of Kinect are shown in Figure 3 and examples of skeleton joints detected by means of Kinect are shown in Figure 4. In Figure 4, some noisy joints that are not relevant at all for activity recognition (i.e., hip and shoulders) have been discarded. The final set of joints we chose as features is shown in Figure 4 (b), while the joints we discarded are white shown in Figure 4(a).

3.1. Pre-processing

The main task of pre-processing is to find the depth human silhouettes from the noisy background using depth camera. Depth video frames include an unrestricted environment having a number of uncertain objects, obstacles, noisy background and freely movement of a human in the scenes.

Therefore, we propose a floor removal technique based on a least squares method [9] to remove these noisy effects in depth images. The pixels of the floors can be calculated as:

$$m_1(x + y + z) = C_f \cdot \cdots \cdot m_n(x + y + z) = C_f \quad (1)$$

whereC_f defines the depth pixel values of floor in each frame, m_1 and $m_n$are the constant values of all three coordinate axis of each floor pixel. In order to eliminate the floor from the scene, *y p*arameters correspond to a given pair of *x* and *z* axis. Usually, the depth value *y* in a spaced grid having least value (i.e., equal to zero) is used to ignore floor from background.

While, to segment all objects and human silhouettes from the video scene, we can localize all expected objects from the scene based on connected component labeling (CCL) method to label all candidates pixels separately. In CCL, the variation of pixel intensity in an image is observed using raster scanning. Thus, we can detect all the non-object components acting as background by entropy analysis.

In addition, this system can measure every depth pixel of the connect component and differentiate the depth values of corresponding neighboring pixels within a specific threshold values. Thus, the depth center map values help to divide the 3D feature-based data into two proper classes (i.e., connected component representing foreground objects as human or materials and background). Furthermore, temporal continuity constraints between frames are applied to easily extract depth human silhouettes from the scenes.

3.2. Feature extraction from depth shape

Firstly, depth image history features are obtained from the overall pixel intensity information of human body shape in sequential activities. Secondly, temporal motion features are extracted by capturing the intra/inter motion variation among different body parts. Thirdly, optical flow features are examined by considering the directional angular values among consecutive frames[2].

3.3. Feature extraction from skeleton joints

To analyze the joints information, we utilized body part model [14] that includes human skeleton focusing on significant parts such as head, neck, torso, arms, legs, hands, and feet as shown in Fig. 3.

In joint information features, firstly, let $\mathbf{J_i}$be one of the 11 joints detected by means of the Kinect, the feature vector **f** is defined as:

$$\mathbf{f} = [\ \mathbf{j}_1, \mathbf{j}_2, \mathbf{j}_3, \mathbf{j}_4, \mathbf{j}_5, \mathbf{j}_6, \mathbf{j}_7, \mathbf{j}_8, \mathbf{j}_9, \mathbf{j}_{10}, \mathbf{j}_{11}\ ] \qquad (2)$$

where each $\mathbf{j_i}$is the vector containing the 3-D normalized coordinates of the i^{th} joint $\mathbf{J_i}$detected by the Kinect. Thus,

$$\mathbf{j}i = \frac{\mathbf{Ji}}{s} + \mathbf{T}, \quad 1 \le i \le 11 \qquad (3)$$

where*s i*s the scale factor which normalizes the skeleton according to the distance, *h*, between the neck and the torso joints of a reference skeleton:

$$\boldsymbol{s} = \frac{||\mathbf{J4}-\mathbf{J2}||}{\mathbf{h}} \qquad (4)$$

and *T* the translation matrix needed to set the origin of the coordinate system to the torso.

3.4. Posture Analysis

Now, these depth and joint features vectors are further normalized based on K-means clustering algorithm. And then, the classification is done according to a “max wins” voting strategy. The process of classifying the detected features into *k* classescan be viewed as building a *k*-words vocabulary. Each posture can be represented as a single word of the vocabulary, and therefore, each activity can be considered asan ordered sequence of vocabulary words.

3.5. Posture Recognition

A Hidden Markov Model is defined by the tuple $\lambda \equiv (S, M, A, B, \Pi)$ where:

$S = \{S_0, \ldots, S_N\}$is the set of hidden states of the model. The state at time *t* is denoted by q_t.

M is the set of observation symbols of the model.

$A = \{a_{ij}\}$is the state transition probability distribution:

$$a_{ij} = p\ (q_{t+1} = S_j | q_t = S_i) \qquad (5)$$

$B = \{b_j\}$ is the observation symbol probability distribution in state *j*:

$$b_j = p\ (v\ | q_t = j) \qquad (6)$$

$\Pi = \{\pi_j\}$is the initial state probability distribution:

$$p_{ij} = p\ (q_0 = j) \qquad (7)$$

One of the most powerful features of HMMs is that it can model both large duration and small duration activities.

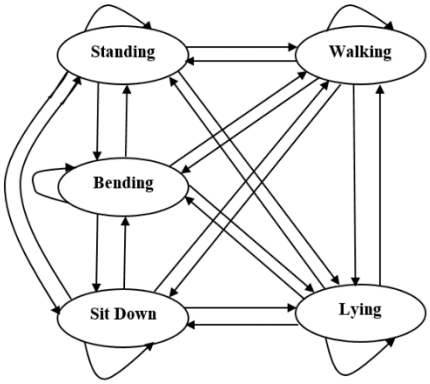


Figure 5. Example structure of HMM for the proposed system

Figure. 5 show the example structure of the HMM for the proposed system. To train and recognize different activities, we can apply the feature clusters to the Hidden Markov models (HMM). The idea is to encode each activity in terms of postures and build the corresponding HMM. Each HMM is trained on the posture sequences of each activity and classified according to the largest posterior probability.

4. Dataset Creation

Our dataset is created by using the Kinect. Both the depth and RGB image were recorded at a rate of 30 frames per second. The depth images were saved as 11 bit images. The resolution of the depth and RGB images were 640 × 480. We placed the Kinect at a fixed height from the floor so as to capture the subject's entire body. All actions were performed at a fixed distance from the Kinect. Our dataset consists of the following five actions: standing, walking, sit down, lying and bending. All actions are performed by facing the camera.

5. MSRDailyActivity3D dataset

The MSRDailyActivity3D dataset consists of daily activities captured by Microsoft Research using a Kinect device. There are sixteen activities which include drink, eat, read book, call on cell phone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lie down on sofa, walk, play guitar, stand up, and sit down. The total number of activity samples, in which ten subjects are involved, is 320. This dataset has been designed to cover human daily activities in a living room.

6. Expected Outputs

During training/testing phase, the proposed system collected the raw depth data, and extracted human silhouettes and the joints information. Then, random input activities were trained and recognized using HMM.

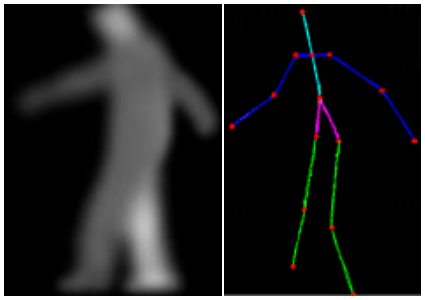


Figure6. Some example of experimental results of the proposed system

The proposed system here can provide the expected outputs as shown in Figure 6. These are some of expected outputs in this system. Our approach is intended to improve the performance and accuracy of recognition.

7. Conclusion

In this system, a methodology to detect and recognize the human activity in both depth images and skeleton joints is proposed. In order to obtain a suitable representation of the human body, spatial depth shape features and temporal joints features are used to improve classification performance.

Both of these features are fused together to recognize different activities using the hidden Markov model (HMM). Our proposed system can be applied to the specific domain of healthcare system including home and hospital for the elderly, disabled or children to support their health in place via continuous monitoring of their daily activities.

References

[1] Ahmad Jalal, NaehaSarif, Jeong Tai Kim, Tae-Seong Kim, "Human Activity Recognition via Recognized Body Parts of Human Depth Silhouettes for Residents Monitoring Services at Smart Home", Indoor Built Environ 2013;22;1:271–279

[2] A. Jalal, J. Kim and T. Kim, "Human activity recognition using the labeled depth body parts information of depth silhouettes," *in Proceedings of SHB symposium*, Korea, pp.1-8, Oct. 2012.

[3] B. Ni, Y. Pei, P. Moulin, and S. Yan, "Multilevel depth and image fusion for human activity detection," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1383–1394, Oct. 2013.

[4] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: A color-depth video database for human daily activity recognition," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, Nov. 2011, pp. 1147–1153.

[5] Chan M, Esteve D, Escriba C, Campo E, "A review of smart homes-Present state and future challenges", Compute Methods Programs Biomed 2008;91:51–81.

[6] H. Shum, E.Ho,Y. Jiang, and S. Takagi, "Real-time posture reconstruction for Microsoft kinect," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1357–1369, Oct. 2013.

[7] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog. Proc.*, 1992, pp. 379–385.

[8] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 914–927, May 2014.

[9]Jalal, A.; Kim, Y.: Dense depth maps-based human pose tracking and recognition in dynamic scenes using ridge data. In: Proceedings of the IEEE conference on advanced video and signal-based surveillance, pp. 119–124 (2014)

[10] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from RGBD images," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2012, pp. 842–849.

[11] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog. Workshops*, 2012, pp. 20–27.

[12] O. C. Ann and L. B. Theng, "Human activity recognition: a review," in *Proceedings of the IEEE International Conference on Control System, Computing and Engineering (ICCSCE '14)*, pp.389–393, IEEE, Batu Ferringhi, Malaysia, November 2014.

[13] R. Gupta, A. Y.-S. Chia, and D. Rajan, "Human activities recognition using depth images," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 283–292.

[14] Salvatore Gaglio, *Member, IEEE*, Giuseppe Lo Re, *Senior Member, IEEE*, and Marco Morana. "Human Activity Recognition Process Using 3-D Posture Data" IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. 45, NO. 5, OCTOBER 2015

[15] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog. Workshops*, 2010, pp. 9–14.