Skeleton Tracking Based Complex Human Activity Recognition Using Kinect Camera

Muhammad Latif Anjum[1], Omar Ahmad[1], Stefano Rosa[1],
Jingchun Yin[1], and Basilio Bona[2]

[1] Department of Mechanical and Aerospace Engineering (DIMEAS),
Politecnico di Torino, 10129, Torino, Italy
{muhammad.anjum,omar.ahmad,stefano.rosa,jingchun.yin}@polito.it

[2] Department of Control and Computer Engineering (DAUIN),
Politecnico di Torino, 10129, Torino, Italy
basilio.bona@polito.it

Abstract. This paper presents a new and efficient algorithm for complex human activity recognition using depth videos recorded from a single Microsoft Kinect camera. The algorithm has been implemented on videos recorded from Kinect camera in OpenNI video file format (.oni). OpenNI file format provides a combined video with both RGB and depth information. An OpenNI specific dataset of such videos has been created containing 200 videos of 8 different activities being performed by different individuals. This dataset should serve as a reference for future research involving OpenNI skeleton tracker. The algorithm is based on skeleton tracking using state of the art OpenNI skeleton tracker. Various joints and body parts in human skeleton have been tracked and the selection of these joints is made based on the nature of the activity being performed. The change in position of the selected joints and body parts during the activity has been used to construct feature vectors for each activity. Support vector machine (SVM) multi-class classifier has been used to classify and recognize the activities being performed. Experimental results show the algorithm is able to successfully classify the set of activities irrespective of the individual performing the activities and the position of the individual in front of the camera.

Keywords: OpenNI, Skeleton tracking, Multi-class SVM, Activity recognition, RGBD Dataset.

1 Introduction

With every passing day, the research in robotics is converging to make the robots more and more social. Robots today are jumping out of their once strong field, the industrialized robotics, and are about to invade the human society. The first and the foremost task at their hands is to figure out what is going on around them or to understand what activities the individuals around them are performing. We have presented one such algorithm in this paper which can be implemented

M. Beetz et al. (Eds.): ICSR 2014, LNAI 8755, pp. 23–33, 2014.

with a very low cost single Kinect sensor and enables the robots to recognize the activities being performed in front of them and respond accordingly.

Human activity recognition has been a highly sought after subject in the recent times. After a quick literature review, this research area can be divided into two major groups: (1) the activity recognition research using wearable sensors, and (2) the activity recognition research using RGB and depth cameras. The research based on wearable sensors mainly targets sports activities and athletes. A good insight into this field of activity recognition can obtained from [9], [2] and [4]. The wearable sensors based recognition is not directly applicable to robotics because, in a normal day to day environment, humans are not expected to use wearable sensor for robots' guidance.

Activity recognition using cameras is more relevant to robotics mainly because robots can be easily equipped with a camera. Most of the previous research in this area has been focused on activity recognition using videos and images created by 2D cameras. Huimin et al. [10] used a multi-class SVM classifier for human activity recognition using RGB videos. Their algorithm provides good results both on their own home-brewed dataset and public dataset provided in [11]. Jinhui et al. [5] used motion and structure change estimation in RGB videos to classify different human activities. Similar hand gesture recognition has been implemented on RGB videos by Omar et al. [1]. They have presented results using both SVM and ANN classifiers.

Ever since the emergence of Microsoft Kinect camera that can capture depth images and videos, there is more focus on using depth videos for human activity recognition. Youding et al. [12] used depth image sequences to track human body pose. They have used Bayesian framework and have been successfully able to track 3D human pose. Tayyab et al. [7] used videos from a Kinect camera mounted on a quadrocopter for gesture recognition. Based on the gesture recognition results, the quadrocopter was successfully able to follow the individual performing the gesture. More recent work was presented by Hema et al. [6]. They have used RGB-D videos for activity recognition using both skeleton tracking and object affordances where the training was done using structural support vector machines (SSVM).

The contribution of this paper to this growing field of research is two-fold: (1) provision of the public dataset of RGB-D videos (.oni file format) that can be used for OpenNI skeleton tracker, (2) implementation of an efficient activity recognition algorithm based on skeleton tracking. In summary, we provide the following:

- RGB-D activity dataset purposely built for OpenNI skeleton tracker.
- A modified OpenNI skeleton tracker (as ROS package) for offline recorded videos.
- A ROS package for extracting joints' position based features from videos.
- A ROS package for training and testing using SVM multi-class classifier.

The dataset and ROS packages have been made available on the website of our research group, LabRob (`http://www.polito.it/labrob`). We have selected at least three joints or body parts in human skeleton to track using OpenNI skeleton tracker. The change in position in these three joints has been used to construct the feature vectors for each activity. The selection of these joints has been made based on the nature of activity being performed. Once feature vectors for all activities have been constructed, we have used multi-class SVM classifier for training and testing of the algorithm.

2 RGB-D Activity Dataset

Looking at the increasing attention the Kinect sensor is getting from the robotics research community, it is all but necessary that a specific dataset of activities be constructed and made available to the research community. Our dataset contains 200 videos (.oni file format) of 8 different activities being performed by two different individuals. Each video in the dataset starts with a surrender / Psi pose (figure 1) required for calibration in OpenNI skeleton tracker. Before getting to features and statistics of our dataset, let's have a look at already available RGB-D datasets.

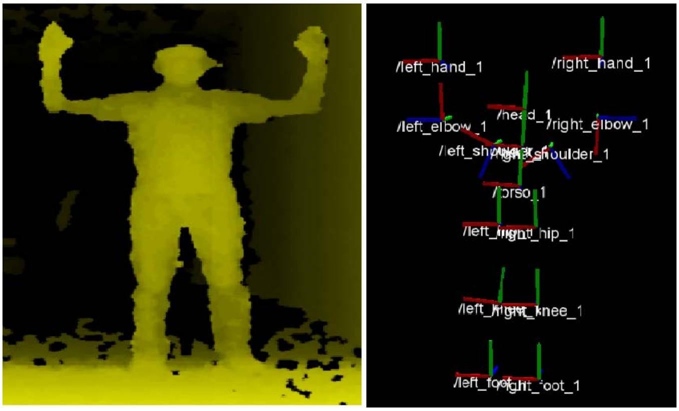

Fig. 1. Surrender / Psi pose in front of Kinect camera and its corresponding skeleton tracking out put shown in rviz

2.1 Available RGB-D Datasets

There are very few datasets publicly available containing RGB-D videos of human activities. Since most of the previous research has been done using RGB videos, most of the datasets contain only RGB videos. A comprehensive list of all available datasets (both RGB and RGB-D) can be found in [8]. Among them there are only two RGB-D datasets available. Cornell Activity Dataset (CAD-120) [6] contains 120 videos of 10 different activities. However they provide data in image format (both RGB and depth images) which require a complex process to convert it into .oni video format for OpenNI tracker. Furthermore, none of their activity starts with a necessary surrender pose. RGBD-HuDaAct [8] is the second available dataset that provides the data in required file format, but since they are not using OpenNI tracker for activity recognition, they do not start the activity with a surrender pose. Our dataset is purposely built to be used for OpenNI skeleton tracker and should serve as a reference for a vibrant ROS OpenNI community.

2.2 Features and Statistics of Our Dataset

We have used Microsoft XBOX 360 Kinect Sensor to record our videos. Each video is recorded using NiViewer[1] and has a resolution of 640x480 with a .oni file format. Each video starts with a surrender pose required for calibration in OpenNI skeleton tracker. The dataset includes 8 different activities with 25 videos recorded for each activity. We have 5 daily life activities while 3 activities consist of umpire's signals from a Cricket match. Table 1 enlists all activities contained in the dataset.

Table 1. Activities performed in dataset

Activity	Activity description	No. of videos
Wave hello	A person waves hello with his right hand	25
Check watch	A person check time from his left hand wrist watch	25
Pick from ground	A person bends and pick something lying on ground and places it on a cupboard	25
Sit stand	A person sits and stands four times in an exercise fashion	25
Sit and drink	A person sits on a chair and drinks water from a bottle	25
Four signal	The cricket umpire gives a four / boundary signal	25
Leg bye signal	The cricket umpires gives a leg bye signal	25
Dead ball signal	The cricket umpires gives a dead ball signal	25

3 Skeleton Tracking

Our algorithm has been implemented using Robot Operating System (ROS)[2]. We use a modified version of ROS wrapper package for OpenNI based skeleton tracking[3]. The original package works with online Kinect camera attached to the PC and publishes user's skeleton positions as a set of transforms (/tf). The available ROS package has been modified to work with offline recorded videos (.oni file format). The tracker can instantly detect user but requires Psi pose for calibration after which it starts tracking the user skeleton. Once calibrated,

[1] The program, NiViewer, is available with OpenNI SDK.

[2] http://www.ros.org/

[3] http://wiki.ros.org/openni_tracker

it starts publishing 3D positions and rotations quaternions of 15 joints or body parts with respect to a fixed frame of reference. The published skeleton joints include both feet, knees, shoulders, hands, elbows, hips, head, neck, and torso. The fact that it can track virtually every joint and part of the body signifies the potential of OpenNI tracker for human activity recognition. With a proper combination of different joints, we can recognize any movement or activity.

4 Constructing Feature Vectors

Constructing feature vectors is the most crucial step of this work, where we decide which joints or body parts are to be tracked for each activity and how to arrange them in a mathematical model for the construction of training and testing data for SVM.

4.1 Selection of Joints to be Tracked

The OpenNI tracker publishes 3D positions and rotation quaternions of 15 joints. If we track all available joints for the construction of feature vectors, the algorithm will become computationally heavy and might not work in real time. Besides, not all joints or body parts are undergoing change in position in any given activity. So, it is all but natural to select the joints and body parts to be tracked for the construction of feature vectors for each activity. The OpenNI tracker publishes positions of all joints relative to a fixed frame of reference. If we use the position relative to a fixed frame of reference, the same activity being pepformed at different positions in front of the camera may give different results. We have, therefore, used joint position relative to another joint position to account for different positions of the user in front of the camera. Table 2 summarizes the joints and their references tracked for each activity to construct the feature vectors.

The joint or body part undergoing the most distinct motion during an activity has been tracked to construct the feature vectors. For example, waving involves a continuous and distinct motion of the right hand and right elbow. Similarly it would be the most useful to track the position of the left hand and left elbow during the check watch activity. Bending down to pick something can be distinguished by tracking position of head and right hand while sit stand activity involves distinct motion of head, hip and torso. Although the selection of joints to be tracked is manual and based on the intuitive understanding of the motion scenarios, this selection is only for the training purposes. Once trained, the algorithm will not require manual selection of joints to be tracked. Figure 2 shows the sample depth images during first four activities in the dataset.

Sitting in a chair and drinking from a bottle is a fairly complex activity. We have tracked the position of right hip, right hand and left hand to construct its feature vector. Making a four signal can be distinguished by tracking right hand and right elbow along with the position of left hand. We have purposely included the leg bye signal in the dataset because it includes movement of the lower part

Table 2. Summary of the joints tracked along with their references for each activity in dataset

Activity	Joints tracked	Reference points
Wave hello	/right_hand /right_elbow /right_elbow	/head /torso /neck
Check watch	/left_hand /left_elbow /left_hand	/head /torso /torso
Pick something from ground	/right_hand /head /right_shoulder	/right_foot /right_foot /right_foot
Sit stand	/head /right_hip /torso	/right_foot /right_foot /right_foot
Sit on a chair and drink from bottle	/right_hip /right_hand /left_hand	/right_foot /head /head
Four signal	/right_hand /right_elbow /right_elbow	/head /head /neck
Leg bye signal	/right_hand /right_knee /right_knee	/head /head /left_foot
Dead ball signal	/right_hand /left_hand /head	/head /head /right_foot

of the body i.e. foot and knee. We have tracked the position of right hand, right knee and right foot for this activity. The dead ball signal involves movement of the two hands along with slight bending down of head, so we have tracked these three part to construct its feature vector. Figure 3 shows the depth images of the next 4 activities in our dataset.

4.2 Mathematical Formulation of Feature Vectors

Let us consider $j1$, $j2$ and $j3$ be the three joints we are tracking in a given activity. The 3D position of each of these joints is published in successive frames. Equation 1 gives the formulation of feature vector for activity A1.

$$\begin{aligned} FV = \{A1, &\{(j1_{x,0}, j1_{y,0}, j1_{z,0}), (j2_{x,0}, j2_{y,0}, j2_{z,0}), (j3_{x,0}, j3_{y,0}, j3_{z,0})\}, \\ &\{(j1_{x,1}, j1_{y,1}, j1_{z,1}), (j2_{x,1}, j2_{y,1}, j2_{z,1}), (j3_{x,1}, j3_{y,1}, j3_{z,1})\}, \\ &..., \{(j1_{x,n}, j1_{y,n}, j1_{z,n}), (j2_{x,n}, j2_{y,n}, j2_{z,n}), (j3_{x,n}, j3_{y,n}, j3_{z,n})\}\} \end{aligned} \quad (1)$$

where $j1_{x,0}$ indicates the x position of the joint $j1$ in frame number 0 relative to the reference joint while $j1_{x,1}$ indicates the x position of the same joint in

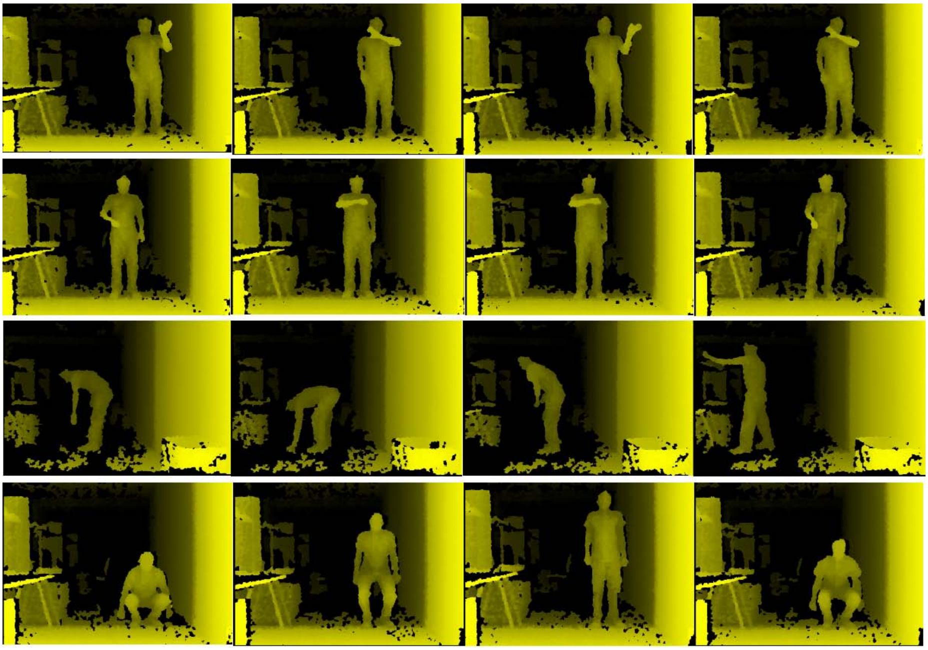

Fig. 2. Depth images at four different positions during the waving, checking watch, picking something from ground and sit stand activities respectively

frame number 1 relative to the same reference joint. The first element in each feature vector is the label of the activity, indicated as $A1$ in equation 1. Based on the length of the longest activity, we have tracked joints' positions for 2260 consecutive frames in each activity. So the dimension of each feature vector is 1x2260.

5 Training and Testing with Multi-class SVM

We now have 200 feature vectors constructed using the mathematical model presented in equation 1. All these feature vectors are put into a matrix to construct feature data as given in equation 2.

$$featureData = \begin{cases} (A_i, f_i) & i = 1, ..., n \\ & A_i \in \{1, 2, ..., 8\} \\ & f_i \in R^{\{1X2260\}} \end{cases} \tag{2}$$

where A_i represents the activity label and f_i represents the feature sets of the activity. The number of activity videos is represented by n which is 200 in our case. Our goal now can be stated as: Given the feature vector of any activity f_i from the n videos, we have to successfully predict its label A_i.

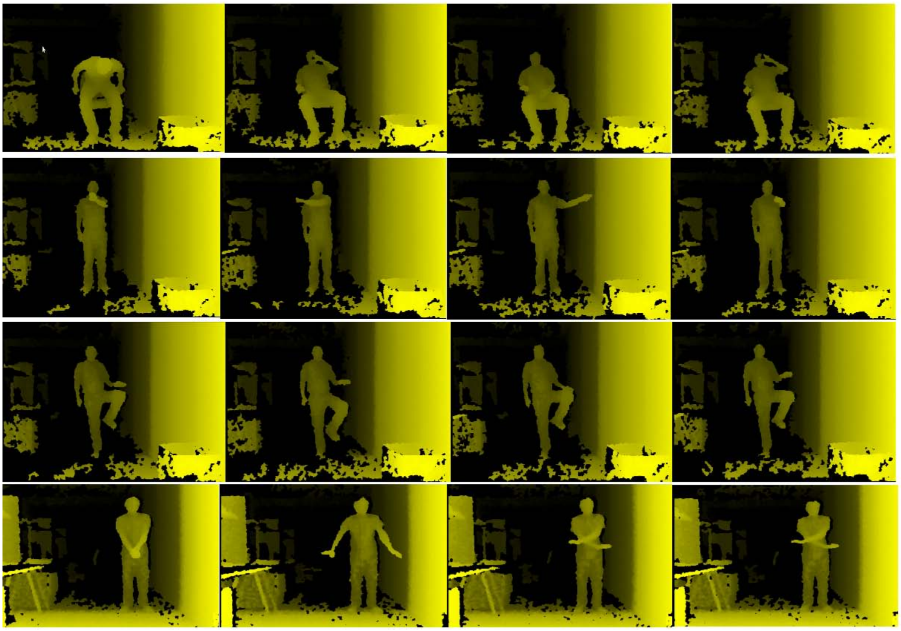

Fig. 3. Depth images at four different positions during the sitting and drinking, four signal, leg bye signal and dead ball signal activities respectively

Support Vector Machines (SVM) is an increasingly becoming a popular tool to solve this kind of classification problem. SVM has produced accurate results in many areas of machine learning including text categorization, gesture recognition and face detection. We are using multi-class SVM classifier which can classify more than two categories of classes. The classification strategy is based on one-against-one approach where each feature set is matched against all samples in the training data. A voting strategy is used for testing where label with maximum positive votes is assigned. Readers interested in learning SVM and other kernel based learning methods are directed towards [3].

Our dataset contains 200 videos in total with 25 videos of each activity. We have trained our SVM based algorithm on 120 videos (15 videos of each activity) while the testing of the algorithm is done on remaining 80 videos (10 videos of each activity) . OpenCV library LibSVM has been used for training and testing of data. We have used linear Gaussian kernel for training the SVM.

6 Experimental Results

It was initially assumed that tracking only two joints or body parts in any activity would be sufficient for high accuracy activity recognition. The results with two joints tracking were satisfactory but not to our expectation. We then added a third joint to be tracked in each activity. The results for both the cases are presented separately.

6.1 Recognition Results While Tracking Two Joints

The two joints undergoing continuous and distinct motion were selected for this experiment. The first two joints in table 2 against each activity were selected for tracking in this experiment. The recognition accuracy turned out to be 92.5% (74 correct recognitions against 80 test videos). The confusion matrix for this experiment is shown in table 3.

Table 3. Confusion matrix for activity recognition while tracking two joints

No.	Activity	1	2	3	4	5	6	7	8
1	Wave	9	0	0	0	0	1	0	0
2	Check watch	0	10	0	0	0	0	0	0
3	Pick from ground	0	0	10	0	0	0	0	0
4	Sit stand	0	0	0	10	0	0	0	0
5	Sit and drink	0	0	0	1	8	0	1	0
6	Four signal	2	0	0	0	0	8	0	0
7	Leg bye signal	0	0	0	0	0	0	10	0
8	Dead ball signal	0	0	1	0	0	0	0	9

Two notable activities that are overlapping and causing a confusion are waving hello and four signal. Both the activities involve similar movements of the right hand with the difference being the position of the hand. Three other activities have one incorrect label prediction each in table 3. These confusions have been avoided with the inclusion of a third joint in tracking algorithm.

6.2 Recognition Results While Tracking Three Joints

A third joint (the third entry in table 2 for each activity) was included in tracking algorithm keeping in mind the activities being confused with each other. For example, the position of right elbow was tracked with reference to neck in four signal and waving activities to further enhance the distinguishing features. With the addition of one more joints for tracking, we were able to obtain 98.75% result with only one video in 80 test videos being confused with another. The confusion matrix for this experiment is given in table 4.

7 Results Analysis and Future Works

The accuracy of results signifies the potential of skeleton tracking for activity recognition. A depth image based skeleton tracker is even better because it makes the image background, lighting conditions etc irrelevant producing robust algorithm. A high accuracy in results can also be attributed to fairly distinct activities in our dataset. Future works should test the algorithm on slightly similar activities (for example head rotation when saying yes versus head rotation when saying no). We are also working on making our dataset robust to include

Table 4. Confusion matrix for activity recognition while tracking three joints

No.	Activity	1	2	3	4	5	6	7	8
1	Wave	10	0	0	0	0	0	0	0
2	Check watch	0	10	0	0	0	0	0	0
3	Pick from ground	0	0	10	0	0	0	0	0
4	Sit stand	0	0	0	10	0	0	0	0
5	Sit and drink	0	0	0	0	9	0	1	0
6	Four signal	0	0	0	0	0	10	0	0
7	Leg bye signal	0	0	0	0	0	0	10	0
8	Dead ball signal	0	0	0	0	0	0	0	10

more activities and more users performing the activities. It can also be a good challenge to include activities involving multiple people (for example two people shaking hands). Another goal would be to integrate object recognition and tracking with skeleton tracker to distinguish between actually drinking something and making drinking like hand movement.

8 Conclusion

We have presented an RGB-D activity dataset and an OpenNI tracker based activity recognition algorithm. The use of skeleton tracker is especially significant because with proper selection of joints to be tracked, we can recognize very complex and long activities. We have tested the algorithm for activities involving movement of arm, hand, head, torso and knees. Feature vectors for each activity have been constructed using relative position of three joints with respect to different reference joints for each activity. The joints undergoing continuous and distinct movement have been selected for feature vector construction in each activity. SVM multi-class classifier has been used for training and testing of data. The experimental results show 98.75% accuracy when tracking three joints in each activity.

References

1. Ahmad, O., Bona, B., Anjum, M.L., Khosa, I.: Using time proportionate intensity images with non-linear classifiers for hand gesture recognition. In: Sakim, H.A.M., Mustaffa, M.T. (eds.) The 8th International Conference on Robotic, Vision, Signal Processing & Power Applications. LNEE, vol. 291, pp. 343–354. Springer, Heidelberg (2013)
2. Choudhury, T., Consolvo, S., Harrison, B., Hightower, J., LaMarca, A., Legrand, L., Rahimi, A., Rea, A., Bordello, G., Hemingway, B., Klasnja, P., Koscher, K., Landay, J., Lester, J., Wyatt, D., Haehnel, D.: The mobile sensing platform: An embedded activity recognition system. IEEE Pervasive Computing 7(2), 32–41 (2008)
3. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press (2000)

4. Ermes, M., Parkka, J., Mantyjarvi, J., Korhonen, I.: Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. IEEE Transactions on Information Technology in Biomedicine 12(1), 20–26 (2008)
5. Hu, J., Boulgouris, N.V.: Fast human activity recognition based on structure and motion. Pattern Recognition Letters 32(14), 1814–1821 (2011)
6. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. The International Journal of Robotics Research 32(8), 951–970 (2013)
7. Naseer, T., Sturm, J., Cremers, D.: Followme: Person following and gesture recognition with a quadrocopter. In: IEEE RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 624–630 (2013)
8. Ni, B., Wang, G., Moulin, P.: Rgbd-hudaact: A color-depth video database for human daily activity recognition. In: Consumer Depth Cameras for Computer Vision Advances in Computer Vision and Pattern Recognition, pp. 193–208 (2013)
9. Parkka, J., Ermes, M., Korpipaa, P., Mantyjarvi, J., Peltola, J., Korhonen, I.: Activity classification using realistic data from wearable sensors. IEEE Transactions on Information Technology in Biomedicine 10(1), 119–128 (2006)
10. Qian, H., Mao, Y., Xiang, W., Wang, Z.: Recognition of human activities using svm multi-class classifier. Pattern Recognition Letters 31(2), 100–111 (2010)
11. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: 17th International Conference on Pattern Recognition (ICPR 2004), pp. 32–36 (2004)
12. Zhu, Y., Fujimura, K.: A bayesian framework for human body pose tracking from depth image sequences. Sensors 10(5), 5280–5293 (2010)