Human Activity Recognition for Domestic Robots

Conference Paper · December 2013

DOI: 10.1007/978-3-319-07488-7_27

CITATIONS
55

READS
436

2 authors:

Lasitha Piyathilaka
Central Queensland University
32 PUBLICATIONS **272** CITATIONS

SEE PROFILE

Sarath Kodagoda
University of Technology Sydney
184 PUBLICATIONS **3,338** CITATIONS

SEE PROFILE

Human Activity Recognition for Domestic Robots

Lasitha Piyathilaka and Sarath Kodagoda

Abstract Capabilities of domestic service robots could be further improved, if the robot is equipped with an ability to recognize activities performed by humans in its sensory range. For example in a simple scenario a floor cleaning robot can vacuum the kitchen floor after recognizing human activity "cooking in the kitchen". Most of the complex human activities can be sub divided into simple activities which can later used for recognize complex activities. Activities like "take meditation" can be sub divided into simple activities like "opening pill container" and "drinking water". However, even recognizing simple activities are highly challenging due to the similarities between some inter activities and dissimilarities of intra activities which are performed by different people, body poses and orientations. Even a simple human activity like "drinking water" can be performed while the subject is in different body poses like sitting, standing or walking. Therefore building machine learning techniques to recognize human activities with such complexities is non trivial. To address this issue, we propose a human activity recognition technique that uses 3D skeleton features produced by a depth camera. The algorithm incorporates importance weights for skeleton 3D joints according to the activity being performed. This allows the algorithm to ignore the confusing or irrelevant features while relying on informative features. Later these joints were ensembled together to train Dynamic Bayesian Networks (DBN), which is then used to infer human activities based on likelihoods. The proposed activity recognition technique is tested on a publicly available dataset and UTS experiments with overall accuracies of 85% and 90%.

Lasitha Piyathilaka
Centre for Autonomous Systems, University of Technology, Sydeny, Australia. e-mail: Jayaweera.M.Piyathilaka@uts.edu.au

Sarath Kodagoda
Centre for Autonomous Systems, University of Technology, Sydney, Australia. e-mail: Sarath.Kodagoda@uts.edu.au

1 INTRODUCTION

Recent advancements in robotics technologies have introduced low cost domestic robots that can vacuum the floor while residents are away or provide company for less mobile or elderly people. It is argued that the success of such domestic service robots can be significantly enhanced by the ability of robots to understand the human activities and to respond them accordingly. Such capabilities will enable robots to make more human like decisions without explicitly being ordered to carry out a certain task. It will also allow the robot to seamlessly integrate with human interactions.

Our research focus is to develop robotic technologies to help and promote independent living for elderly people. It is motivated by the growing number of older people around the world and difficulty of finding enough care staff. In general elderly people gradually lose their cognitive ability to keep track of daily activities. In this context, an assistive robot that can recognize human daily activities will be immensely helpful. For example, an elderly person could be reminded of taking medications in appropriate times and could follow it up until the activity has been completed. In addition, the robot may detect abnormal conditions such as someone laying on the floor or sleeping longer than usual and notify the appropriate personnel.

Detection of human activities is challenging due to several reasons. The first reason is related to noisy sensory inputs, and the second reason is related to the difficulty of modeling highly ambiguous actions. Moreover human activities are performed in different body poses and orientations with inter subject variations. Therefore, video-based human action recognition has unwarranted complexity and limited accuracy.

Recent trend in human activity recognition research is to use low cost RGB-D cameras like Microsoft Kinect^{TM}. These cameras are capable of generating skeleton model of a human with 15 body joints positions and their orientation. In this research our intention is to use these skeleton features to extract relatively unambiguous features to model human activities.

In our previous work [12], we have developed human activity recognition model that used Gaussian mixture based HMM. However, its recognition accuracy is severely compromised, if the actions are performed with different body poses. For example “drinking water” activity can be performed while the person is in different body poses such as sitting, standing or even while walking. This is due to the incorporation of all the features, including non informative and ambiguous ones. However, if we could devise a methodology for identifying the most informative features for a given activity, then it will be better positioned at handling actions done with different body poses.

This paper presents a novel human action recognition approach that uses only 3D skeleton features produced by a depth camera. Each activity was modelled as a Dynamic Bayesian Network (DBN) in which each joint node is probabilistically weighted according to the importance of that joint to the activity being modelled. These joint weights together with their observation probability ensembles, form a

model for each activity. Joint weights are calculated by training HMM for each case of a given activity and estimating the dissimilarity measure between such trained models. The model is firstly evaluated on a publicly available benchmark dataset: Cornell activity Detection Dataset [14]. Then it was tested with our experiments which shows that proposed method is able to achieve higher recognition accuracies even with higher intra-activity variations of 3D skeleton features.

2 RELATED WORKS

Human activity detection is not a new research area that has been looked into by various researchers. In [15] human activities are classified as either normal or aggressive by using a mobile robot and a 3D sensory tracker system. Other researchers have utilized human activity detection to learn and imitate humans activities [2][6]. In [5], audio-based human activity recognition using non-markovian ensemble voting technique is presented. Applicability of this method is limited by the inherent distinguishable sounds associated with activities. Therefore such a system may only be used as a complement to the existing sensory systems.

It is common knowledge that knowing the 3D joint position is helpful for activity recognition. Multi-camera motion capture (MOCap) systems [16] has also been used for activity detection but requires markers attached to joints with a highly calibrated camera system. Therefore, such a system is infeasible to be used in practical robotic scenarios. With the invention of low cost depth cameras, several researchers have used RGB-D skeleton data to recognize activities. In [14] two-layered maximum entropy Markov model with a set of sub-activities is used to detect human activities. There, both the skeleton and 3D point cloud data are used extracting 715 features. However, the algorithm is heavily dependent on a particular sequence of sub activities to form human activities. This can have adverse influence on the generalization aspect due to the individual differences in carrying out activities.

In [16] actionlet ensemble model for human activity detection with depth cameras are proposed. The actionlets are proposed to compensate intra-class variations caused by human activities. This approach mainly differs form ours in many ways. Their actionlets only comprises of different combination of joints, whereas our approach assigns probabilistic weighting for each skeleton joint. Therefore our action ensembles contain more meaningful information than actionlets. Secondly our approach only relies on universal skeleton features whereas actionlet based approach uses depth data associated with each joint position, called Local Occupancy Pattern (LOP). But these LOP features would depend on the objects that the subject interacts with. Therefore it may have difficulty in dealing with a subject performing an activity using different object sizes and shapes.

Use of probabilistic graphical models is one of the most popular techniques that has been used by automatic human activity detection. In [1] researchers used coupled HMM to detect human two hand activities and some others utilized motion template together with HMM to recognize human activities [7]. But these researchers

didn't incorporate all the joint information in their models. However most of the human daily activities are too complex to recognize by only observering few joint features. Therefore those techniques would fail to recognize human daily activities with high intra-activity variations.

The paper is organized as follows. Sect. 3 describes overall activity detection model which is the core of our proposed approach. It details the algorithm we used to calculate joint confidence weights followed by the Dynamic Bayesian Network (DBN) that incorporates joint ensembles. Sect. 4 describes the implementation of the proposed approach and training of DBNs for activity detection. Experimental results are discussed in Sect. 5 followed by the conclusions in Sect. 6.

3 ACTIVITY RECOGNITION MODEL

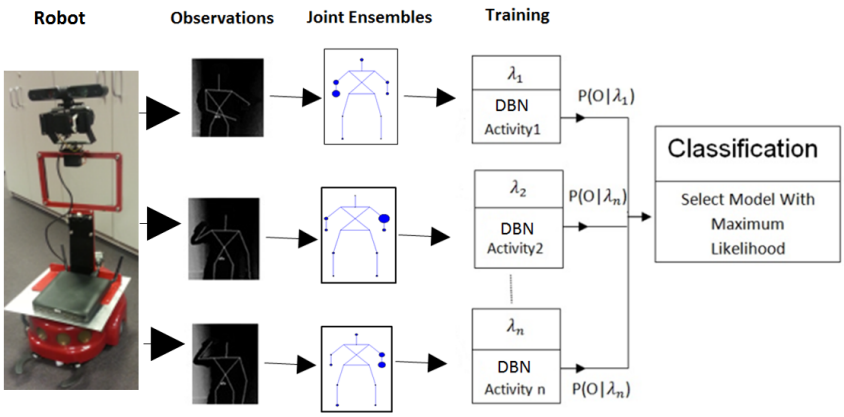

Fig. 1 Block diagram of the recognition process

Fig. 1 shows the overall process which is utilized in the proposed human activity detection method. First we identify joint ensembles and their associated weights for each and every activity in the data-set. Then we train separate Dynamic Bayesian Networks (DBN) by incorporating joints weights for each activity in the data-set. Once a new sequence of skeleton features has been captured, the previously trained models produce likelihood estimation, from which the maximum is selected.

3.1 Learning Action Ensembles

We represent each activity as weighted joint ensembles to better characterize intra class (same activity done with different body poses) variations of human activities. This allows us to identify common joint movements associated with each intra-class activity. The approach can be justified by the fact that all 14 skeleton joints do not contribute equally to a particular human activity. For example, for the activity "drinking water" most descriptive skeleton features would be 3D joint skeleton data of hands and the head. Therefore more weight may be assigned to joint positions of hand and head for the activity "drinking water". Following section describes the learning mechanism that has been utilized to identify joint ensembles and their associated weights for each activity in the data set.

3.1.1 Calculating Joint Confidence Weights

In the proposed algorithm weighted joint ensembles are denoted as W_a^j , where $j\varepsilon J = \{j_1, j_2, ..., j_n\}$ and $a\varepsilon A = \{a_1, a_2, a_3..a_m\}$. Here J is the set of skeleton joints and A is the set of all activities in the dataset and m is the number of activities. In addition the weights of each joint are constrained as in (1) to give probability value for each joint.

$$\sum_{j=1}^{n} W_a^j = 1 \tag{1}$$

We assumed each joint is independent of each other when calculating joint weights for a given activity. For the person p_n , joint j_n and for the activity a_n, we can denote the set of k observation sequence as $O_{J_n}^{p_n} = \{O_1, O_2, O_3, ..., O_k\}$. For each subset of observation sequences $S \subset O_{J_n}^{p_n}$, what we are interested in knowing is the similarity or the likelihood between the observation sequences. When calculating the likelihood of each observation sequence tempo-spatial movement of the joint need to be considered. In order to calculate the likelihood between observation sequences, we should be able to build models that efficiently represent observation sequences. Hidden Markov Models (HMM) have shown a great deal of success to model sequential data [13] and therefore, intra activity likelihood is calculated based on a HMM by training each joint and subsequent testing.

Fig. 2 shows position information (with reference to torso) of the right-hand's wrist joint when "drinking water" activity is performed. It shows few distinguishable clusters. In addition, within each of these clusters, few sub clusters can also be observed. This is due to the variation caused when the subject performs the same activity in different poses. Although unimodal Gaussians are used in HMM to model continuous data, it is not capable of capturing multimodal nature of the joint movements and hence in this research we implemented HMM based on Gaussian Mixture Models (GMM) in order to calculate joint likelihoods.

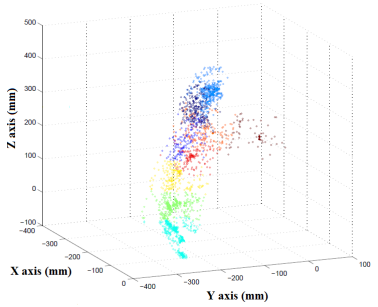

Fig. 2 (x,y,z) positions of the right hand with respect to the torso, when the action "drinking water" is performed

In GMM based HMM, observation probability given states s can be modelled with weighted sum of M component Gaussian densities as,

$$b_s(O) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i) \tag{2}$$

where x is a 3-dimensional continuous-valued joint position vector , $w_i, i = 1,...,M$, are the mixture weights, and $g(x|\mu_i, \Sigma_i), i = 1,...,M$ are the component Gaussian densities. Each component density is a 3-variate Gaussian function with mean of μ_i and covariance matrix of Σ_i.

GMM based HMM was trained for each joint with every observation sequence for a given activity. Given such two HMMs, λ_1 and λ_2, our interest is to find similarities from which the weights can be estimated: for higher similarities higher weights are assigned where as for less similarities lower weights are assigned. This concept of model dissimilarity can be generalized by defining the distance measure $D(\lambda_1, \lambda_2)$, between two HMMs as ,

$$D(\lambda_1, \lambda_2) = \frac{1}{T}[logP(O^{\lambda_2}|\lambda_1) - logP(O^{\lambda_2}|\lambda_2)] \tag{3}$$

where $O^{\lambda_2} = O_1, O_2, O_3..O_T$ is a sequence of observations generated by model λ_2 [13]. Equation (3) is a measure of how well λ_1 matches observations generated by model λ_2, relative to how well model λ_2 matches observations generated by itself. The dissimilarity measure discussed above is none-symmetric. Therefore for better representation (4) can be symmetrized by

$$D_s(\lambda_1, \lambda_2) = \frac{D_s(\lambda_1, \lambda_2) + D_s(\lambda_2, \lambda_1)}{2} \tag{4}$$

Finally, to estimate weights W_a^j associated with a given activity following steps have been followed.

for *activity a=1 to A* **do**
  for *Observation o=1 to O* **do**
    Train GMM based HMM $\lambda_j^a(o)$ for each joint j
  end
  for *joint j=1 to J* **do**
    - For all $S \subset \Lambda = \{\lambda_j^a(1), \lambda_j^a(2).....\lambda_j^a(n)\}$ s.t N(S)=2, calculate dissimilarity measure $D_j^a(n)$ by (4) where $1 \leq n \leq C_3^n$.
    - Calculate total dissimilarity for joint j as $D_{j_{total}}^a = \sum_{n=1}^{C_3^n} D_j^a(n)$
    - Assign weight for the joint as $W_a^j = \frac{1}{D_{j_{total}}^a}$
  end
  Normalize all joint weiights s.t $\sum_{i=1}^{n} W_a^j = 1$ to assign probability value for weights.
end

Algorithm 1: Learning action ensemble joint weights

3.2 DBN for action recognition

Once joint weights are known, we can effectively model each activity by a Dynamic Baysian Network (DBN) as shown in Fig. 3. A DBN is a directed acyclic graph, which represents the conditional independencies and the conditional probability distributions of each node [8]. Shaded nodes represent the observed continuous 3-dimensional joint positions (J_j^t *where* $1 \leq j \leq 14,\ 1 \leq t \leq T$) and transparent squares represent the discrete hidden nodes. We have incorporated joint weights to the observation probability by an exponents as shown in (7). We assumed each human activity is a collection of different poses that evolves over time. Therefore, in the proposed model, top hidden node represents pose class and the middle hidden nodes represent mixture weight components. Pose classes are not directly observed as opposed to the joint positions, which can be directly measured from RGB-D camera's skeleton information.

The proposed DBN can be parameterized by three probabilities A, B and π as follows. First we define individual pose states as $S = \{S_1, S_2, ..., S_N\}$, the state at time t as q_t and K as the number of states. In the proposed model $a_{i,j}$ is the state transition probability from state i to state j and $b_t(i)$ represents the probability of the observation O_t given the i^{th} state of the pose nodes. Then initial state distribution, $\pi = \{\pi_i\}$ can be defined as

$$\pi_i = P(q_i = S_i),\ 1 \leq i \leq N \tag{5}$$

The observation probability distribution can be defined as, $B = \{b_t(i)\}$ where

$$b_t(i) = P(O_t | q_t = S_i),\ 1 \leq i \leq K, 1 \leq t \leq T \tag{6}$$

O_t is the joint observation at time t.

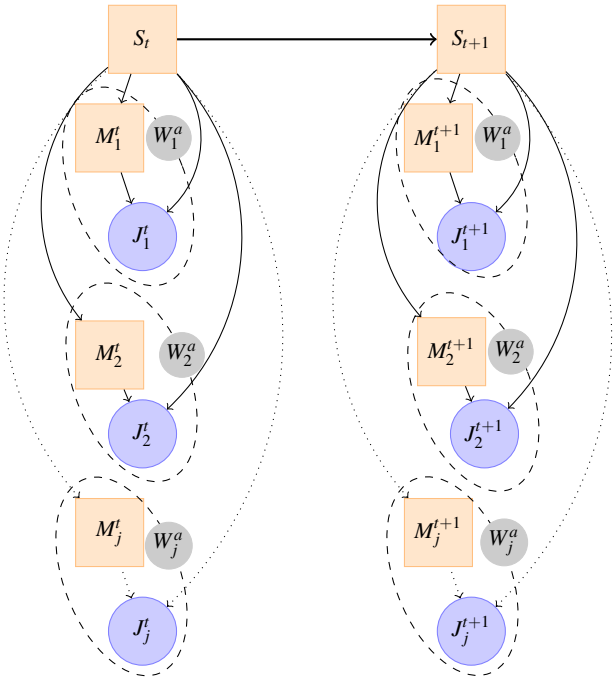

Fig. 3 Graphical representation of the proposed DBN. Square nodes represent discrete hidden nodes and round nodes represent observed continuous 3-dimensional joint positions. Dotted ellipses that encircle observations represent weights associated with each joint.

The observation probability with joint weight W_a^j that represents contribution of that joint to the activity, can be modeled as

$$b_t(i) = \prod_{j=1}^{J} \lceil \sum_{m=1}^{M_i^n} w_{i,m}^j N(O_t^j, \mu_{i,m}^n, \Sigma_{i,m}^j) \rceil^{W_a^j} \tag{7}$$

where J represents the total number of joints, O_t^j the observation vector of the j^{th} node at time t, M_i^j is the number of mixture components in the joint j and state i, and $\mu_{i,m}^j$, $\Sigma_{i,m}^j$, $w_{i,m}^j$ are the mean, covariance matrix, and mixture weight for the j^{th} joint, i^{th} state, and m^{th} Gaussian mixture component, respectively.

Finally the state transition probability distribution can be defined as $A = \{a_{i,j}\}$

$$a_{i,j} = P(q_{(t+1)} = S_j | q_t = S_i), 1 \leq i, j \leq K \tag{8}$$

4 Implementation

The proposed activity recognition model has been implemented using the Bayes Net Toolbox (BNT) for Matlab [9] which is public domain toolkit for modelling Dynamic Bayesian Networks.

4.1 Training Dynamic Bayesian Network

It is a standard practice to use expectation maximization (EM) algorithm to train parameters when a DBN contains any hidden nodes [3]. However it is well known that EM algorithm only converges to a local optimum solution. Therefore initial parameters of the model needed to be carefully chosen in order to get good classification results. In our proposed DBN for activity recognition we used an efficient method to initialize the parameters as explained in our previous research [12].

4.2 Activity recognition

Once a HMM is trained for each action class, we need to select the most likely activity given an observation sequence. Given the observation sequence $O = O_1, O_2, ..., O_t$, and model $\lambda = (A, B, \pi)$ we calculated $P(O/\lambda)$, the probability of the observation sequence once the model is given(likelihood). Then the activity with the maximum likelihood is selected as the most probable activity. The log-likelihood calculation is done using the forward algorithm [13] for HMM that enabled us to recognize activities in real-time.

5 Experiments

First we tested our activity recognition model on the publicly available Cornell Activity Dataset 60 [14] to validate the model. Then we carried out our own experiments on activities with high intra class variations to test the performance of the model to intra activity variations. The empirical results show that proposed framework is capable of recognizing even highly similar activities with reasonable accuracy.

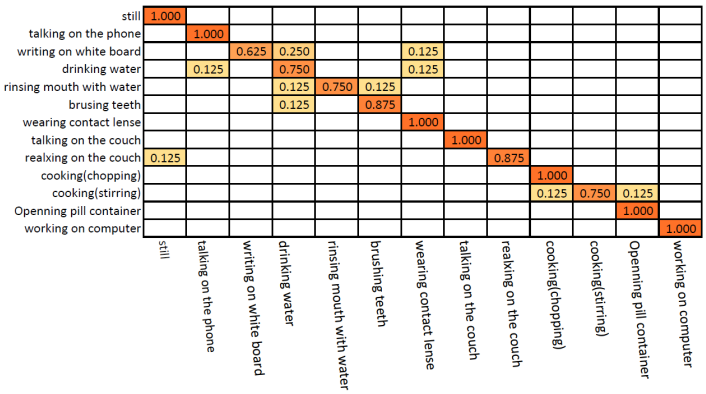

Fig. 4 Confusion matrix for Cornell activity60 dataset .

5.1 Model Validation through Cornell Activity Dataset

The Cornel activity [14] is consists of 14 activities carried out by four different individuals performing an activity once. Therefore, it is to be noted that the intra-activity complexity is limited to the variation among subjects.They have used the Microsoft Kinect RGBD sensor to record both depth and skeleton data of human daily activities done in a indoor environment. Data has been collected with four different people: two male and two females, recorded for about 45 seconds with each person, without compromising to any occlusion of arms and body. Therefore full skeleton was always observed throughout the activity. With this dataset, 2-fold cross validation testing has been carried out i.e we trained our model on two people and tested on others. Our experiments recorded precision and recall accuracies of 90% and 89% respectively. The confusion matrix is shown in Fig. 4. These results are in general better than the results obtained by [14] as can be seen from the Table 1.

5.2 UTS Experiments

There are few publicly available datasets that include skeleton data, which can be used in activity detection. However, they offer very limited intra-activity variations. The concept of weighted joint ensembles can be better explained and tested with a data set which has higher intra-activity variations.

Therefore, we have collected a dataset (UTS-Skeleton3D) consisting of 3D skeleton data. Fig. 5 shows the hardware set-up of the robot that we developed to aid our experiments. It consists of a RGB-D sensor mounted on a Ambigobot^{TM} mobile robotic platform. RGB-D sensor is mounted on a Pan-Tilt module. The robot, Pan-

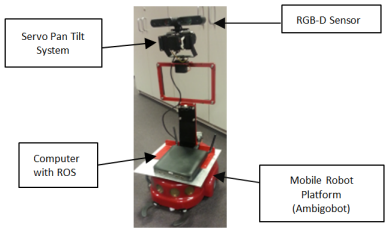

Fig. 5 Hardware set-up of the Robot

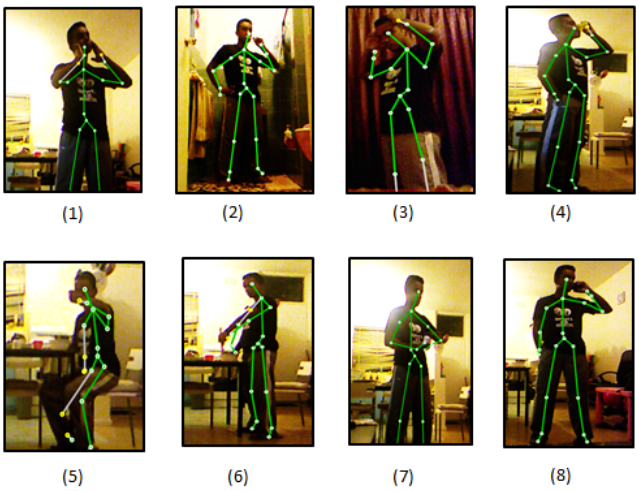

Fig. 6 Samples from our data set. (1) Applying Cream, (2) Brushing Teeth, (3) Combing, (4) Drinking, (5) Eating, (6) Stirring, (7) Opening a pill container, (8) Phone Call

Tilt Module and the RGB-D sensor are interfaced by Robotic Operating System (ROS) and its drivers.

The experimental dataset consists of 8 highly similar activities: applying cream on the face, brushing teeth, combing, drinking water, eating cereals, phone call, stirring and opening a pill container in a domestic environment. Four subjects were used to collect the data in which each activity is performed in three different body poses like, “sitting”, “standing” and “walking”. All together there are 96 samples of activities in the experiments. Each subject performed the activity about 45-60 seconds and data is recorded from different camera angles with a Microsoft Kinect sensor. Initially we recorded each joint’s 3-D position and orientation with respect to the sensor. Later we transformed the data w.r.t the torso coordinates to alleviate the effects of the sensor location

First we have calculated joint weights associated with each activity in the dataset by using the algorithm described in the Sect. 7. Fig. 7 shows the joint weights as-

signed for each activity in the dataset. The radius of the dark circle at each joint is proportional to the probabilistic weight assigned by the algorithm. We trained separate models for right-handed people and left-handed people. Therefore each activity is consisted of two models and likelihood calculations were done for each model, once observation sequence is received. From the Fig. 7 it is clear that proposed algorithm is capable of identifying importance of joints for a given activity. For example, the activity " applying cream on the face" has higher probability weights assigned to hand, forearms, and head while relatively low probability values has been assigned to other joints.

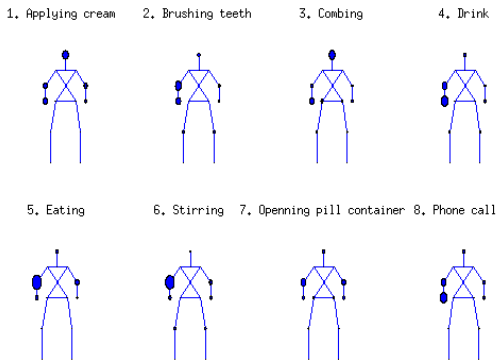

Fig. 7 Learnt weighted joint ensembles for right handed person. The radius of the circle at each joint is proportional to the joint weight.

Once joints weighs have been calculated, the DBN was trained for each activity with their associated joints weights ensembles. K-fold cross validation was used for testing, i.e we left out one sample activity and trained model and weights on others. Left out sample was then used as the activity to be detected. Same procedure was followed for all activity samples in the dataset. Confusion matrix of the test is shown in the Fig. 8. As can be seen, it has a very high rate of activity detection accuracies. It seems the "phone call" activity was slightly confused with "drinking water" activity. This is due to high similarity of the hand movements when these activities are performed and skeleton tracker often fails to track the hand when it is moving very close to the human body. "Stirring" is slightly confused with "Eating cereal" since "Eating cereal" often includes the "Stirring" as a sub activity of it. The proposed method was able to achieve recall and precision accuracies of 85% and 86% respectively. This is a high detection rate given the high intra-class complexity of the dataset. As it can be seen from Table 1 there is a significant improvement of detection rate when joint weighted ensembles are introduced to the DBN, in UTS experiments. This is because UTS experiments contain highly similar activities with very high intra-activity variation.

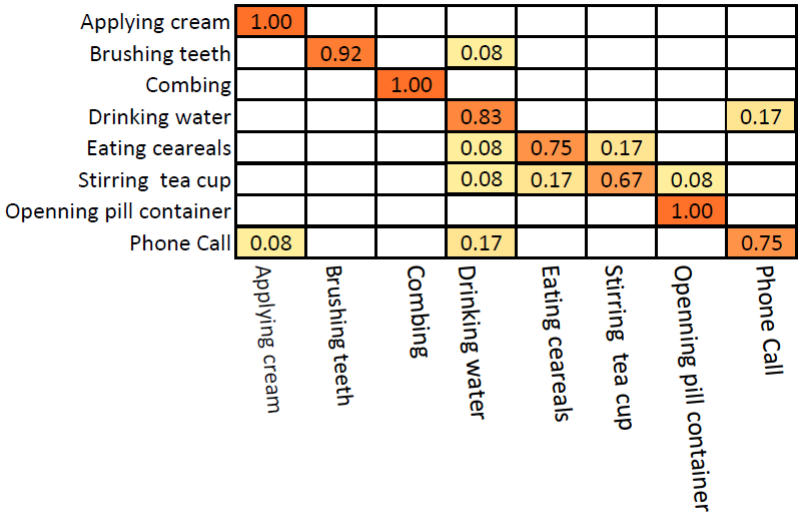

Fig. 8 Confusion matrix for UTS experiments.

Table 1 Recognition Accuracy Comparison for Different Datasets

Dataset	**DBN only**		**Proposed Method**	
UTS Experiments	Recall 66%	Precision 69%	Recall 85%	Precision 86%
Dataset	**Maximum Entrophy Markov Model [14]**		**Proposed Method**	
Cornel activity 60	Recall 57%	Precision 69%	Recall 90%	Precision 89%

6 CONCLUSIONS

In this paper, we presented weighted joint ensembles based human activity recognition system using skeleton features generated from an inexpensive RGB-D sensor. In the proposed technique, joint weights model the importance of that particular joint to the activity. Then we trained a DBN for each activity in the datasets and maximum log-likelihood estimation is calculated in-order to select the most probable model for a given sequence of observations. The proposed algorithm was tested with a challenging publicly available dataset and through UTS experiments with very promising accuracies. More importantly, it is shown that the proposed model is robust to intra-activity variations when people perform the same activity in different body poses.

In a real situation, the humans perform activities in a continuous way. Therefore future work involves detecting end of the activity to improve the model to a long term activity recognition system. In addition currently we are using supervised learning techniques to recognize activities that were previously seen. The reliability of the system can be further improved if the system can detect the difference between a new activity and a previously trained activity.

References

1. Brand, M., Oliver, N., Pentland, A.: Coupled hidden Markov models for complex action recognition. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition **0**, 994–999 (1997). DOI 10.1109/CVPR.1997.609450
2. Demiris, Y., Meltzoff, A.: The robot in the crib: A developmental analysis of imitation skills in infants and robots. Infant Child Dev **17**(1), 43–53 (2008)
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B **39**(1), 1–38 (1977)
4. Hartigan, J.A., Wong, M.A.: A K-means clustering algorithm. Applied Statistics **28**, 100–108 (1979)
5. J. Stork L. Spinello, J.S.K.O.A.: Audio-based human activity recognition using non-markovian ensemble voting. In: Proc. of IEEE International Symposium on Robot and Human Interactive Communication (RoMan) (2012)
6. Lopes, M., Melo, F.S., Montesano, L.: Affordance-based imitation learning in robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1015–1021. USA (2007)
7. Martinez-Contreras, F., Orrite-Urunuela, C., Herrero-Jaraba, E., Ragheb, H., Velastin, S.A.: Recognizing Human Actions Using Silhouette-based HMM. 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance pp. 43–48 (2009)
8. Murphy, K.: Dynamic bayesian networks: Representation, inference and learning. Ph.D. thesis, UC Berkeley, Computer Science Division (2002)
9. Murphy, K.P.: The bayes net toolbox for matlab. Computing Science and Statistics **33**, 2001 (2001)
10. Nefian, A.V., Liang, L., Pi, X., Liu, X., Murphy, K.: Dynamic Bayesian Networks for Audio-Visual Speech Recognition. EURASIP Journal on Advances in Signal Processing **2002**(11), 1274–1288 (2002). DOI 10.1155/S1110865702206083
11. Nefian, A.V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., Murphy, K.: A coupled hmm for audio-visual speech recognition. In: in International Conference on Acoustics, Speech and Signal Processing (CASSP?02, pp. 2013–2016 (2002)
12. Piyathilaka, L., Kodagoda, S.: Gaussian mixture based hmm for human activity recognition uisng 3d skeleton features. 8th IEEE Conference on Industrial Electronics and Applications (2013)
13. Rabiner, L.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257 –286 (1989). DOI 10.1109/5.18626
14. Sung, J., Ponce, C., Selman, B., Saxena, A.: Human activity detection from rgbd images. In: Plan, Activity, and Intent Recognition, vol. WS-11-16. AAAI (2011)
15. Theodoridis, T., Agapitos, A., Hu, H., Lucas, S.M.: Ubiquitous robotics in physical human action recognition: A comparison between dynamic anns and gp. In: ICRA, pp. 3064–3069. IEEE (2008)
16. Wu, Y., Yuan, J., Liu, Z., Wang, J.: Mining actionlet ensemble for action recognition with depth cameras. 2012 IEEE Conference on Computer Vision and Pattern Recognition **0**, 1290–1297 (2012). DOI http://doi.ieeecomputersociety.org/10.1109/CVPR.2012.6247813