Nurse Care Activity Recognition: A GRU-based Approach with Attention Mechanism

Md. Nazmul Haque
Institute of Information Technology
University of Dhaka
Dhaka, Bangladesh
bsse0635@iit.du.ac.bd

Mahir Mahbub
Institute of Information Technology
University of Dhaka
Dhaka, Bangladesh
bsse0807@iit.du.ac.bd

Md. Hasan Tarek
Institute of Information Technology
University of Dhaka
Dhaka, Bangladesh
bsse0818@iit.du.ac.bd

Lutfun Nahar Lota
Department of Computer Science & Engineering
East West University
Dhaka, Bangladesh
lota.nahar@ewubd.edu

Amin Ahsan Ali
Department of Computer Science & Engineering
Independent University, Bangladesh
Dhaka, Bangladesh
aminali@iub.edu.bd

ABSTRACT

Human activity recognition is a challenging task due to complexity and variations of human movements while performing activities by different subjects. Extracting features to model the temporal evolution of different movements plays an important role in this task. In this paper, we present the approach followed by our team, Dark_Shadow, to recognize complex nurse activities in the "Nurse Care Activity Recognition Challenge" [1]. We present a deep learning method to capture the movements of essential body parts from time series of human activity data collected by sensors and then classify them. Deep learning approaches have provided satisfactory results in various human activity recognition tasks. In this work, we propose a Gated Recurrent Unit (GRU) model with attention mechanism to recognize the nurse activities. We obtain approximately 66.43% accuracy for person-wise one leave out cross validation.

KEYWORDS

Activity recognition, Nurse care activity, GRU, Attention mechanism

ACM Reference Format:
Md. Nazmul Haque, Mahir Mahbub, Md. Hasan Tarek, Lutfun Nahar Lota, and Amin Ahsan Ali. 2019. Nurse Care Activity Recognition: A GRU-based Approach with Attention Mechanism. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct), September 9–13, 2019, London, United Kingdom.* , 5 pages. https://doi.org/10.1145/3341162.3344848

UbiComp/ISWC '19 Adjunct, September 9–13, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6869-8/19/09…$15.00
https://doi.org/10.1145/3341162.3344848

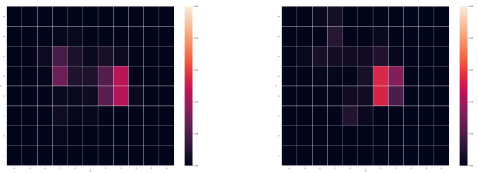
1 INTRODUCTION

Human activity recognition mainly focuses on recognizing different human activities by using data collected from videos or sensors. There are diverse applications of activity recognition in areas such as robotics [17], video analysis [9], gaming, animation, surveillance, [12], and human computer interaction [14]. However, accurate recognition of the human actions from sensor data is challenging because of the complexity of activities, between and within subject variations in performing the activities, and noisy sensor data [10].

In this paper, we focus on nurse care activity recognition for "Nurse Care Activity Challenge" [1]. In this challenge six activities are to be recognized from different sensor data. Inoue et. al prepared a similar dataset and performed Bayesian estimation by marginalizing the conditional probability of estimating the activities for a segment sample attaining an accuracy of only 73.18%. This nurse-care activity recognition task is challenging because the nature of performing the same activities by different nurses vary and some activities contain sub-activities that are similar to other activities [8].

Most existing methods for human activity recognition analyze 3D depth data by constructing mid-level part representations, or using trajectory descriptors of spatial-temporal interest points. Research efforts have also been devoted to motion data analysis, motion detection and recognition, which is widely known as human motion evaluation [2, 4–6]. With the advent of low cost, non-intrusive depth sensors such as Microsoft Kinect [14], research efforts are now devoted to the utilization of 3D skeleton joint positions. Because, the features related to the body part movement extracted from Kinect sensor are more discriminatory as they represent actions properly.

Hossein et. al. [16] use sequences of joint angles and relative positions of joints as features. Among these features they selected most informative sequences of joint angles using entropy of the joint angles. They also determine the most informative relative motions, considering inter/intra-action variations for the differences in 3D positions of each joint pair. After that they calculate Longest Common Sub Sequence within the feature vectors to find the the similarities among different activities. Patrona et. al. [15] propose a framework for online action detection and recognition based on an



(a) Activity 1 (b) Activity 5

Figure 1: The probability distribution of staying in the recording room of activity 1 and 5.

efficient linear search approach. They present local features capturing for both skeleton and kinematics information at the frame level and combine these local features with the moving pose and the angles descriptor with appropriate weighting. After that the model is fed to a binary linear classifier. In [11] authors propose a hierarchical end-to-end convolutional co-occurrence feature learning framework that capture different levels of contextual information gradually. For this, they first independently encode point-level information of each joint and then assemble these information into semantic representation in both spatial and temporal domains.

Nurse care activity recognition is a very challenging and complex task. For example, blood collection activity and blood glucose collection activity are performed approximately in similar way. Identification of discriminatory features and feature selection are necessary to accurate activity recognition. To identify appropriate features, authors in [19, 20, 22] presented a generalized mutual information based feature selection method which can be used for activity recognition. Moreover, the dataset of the activity recognition are usually imbalanced. Different existing methods [3, 13] can be used to solve this problem. However, in this paper, we propose a deep learning based method that implicitly addresses these issues to identify the nurse activities using the data provided by "Nurse Care Activity Recognition Challenge". We make the following major contributions:

- We extract relevant features from two different sensor data and group them.
- A GRU based attention mechanism is used to give more weight to the relevant features.
- A new fusion approach is proposed for aggregating decisions from multiple models.

2 FEATURE EXTRACTION

In the Nurse care activity recognition challenge [1] dataset, Motion capture, Meditag and Accelerometer sensors have been used for collecting data for 6 nurse care activities - 1. Vital signs (pressure, temperature, pulse, respiration) measurement, 2. Blood collection, 3. Blood glucose measurement, 4. Drip retention and connection, 5. Oral care, 6. Diaper exchange and cleaning of area. The data for each activity is divided into 1-minute segments. Human action is considered as a continuous evolution of the spatial configurations of human body segments. To recognise these activities we considered both motion and location sensors data which consists of the position of X, Y, Z planes. However, the features are computed for 2-second windows and each window is classified and then the classifications are combined to make final decisions for each segment. Discussion about feature recognition based on both location and motion analysis is presented in the following sections.

2.1 Location based Features

Human activities are highly correlated to location. Although most of the researches on human activity recognition focused on motion analysis. We combined location sensors data with the motion. Fig. 1 represents the probability distribution of staying in the recording room of a nurse. From the figure we can say that, for different activities, the probability distribution of staying in a place is different. So, location can be an important discriminatory feature for identifying activity. To extract feature we calculated Mean of X and Y, Standard deviation of X and Y, Magnitude, Minimum of X and Y, Maximum of X and Y, Maximum block id, Displacement vector of X and Y and angle.

2.2 Motion based Features

We computed three groups of features from the motion captured data which are: Geometric features, 1st level features and 2nd level features. The extracted geometric features are joint-to-joint distance (JJ_d) , joint-to-joint orientation (JJ_o), line-to-line angle (LL_a), joint-to-line distance (JL_d) which are shown in Fig.3a. Since, the number of combinations will be extremely large if we use all pair of distances, we need to select several important lines and planes in order to reduce the computational cost. As a nurse activity is highly related to the upper body part, we only used the 11 joints in the upper part of the body. The most informative joints for nurse activity are - Top Head (TH), Rear Head (RH), Right offset (RO), Left Shoulder (LS), Left Elbow (LE), Left Wrist (LW), Right Shoulder (RS), Right Elbow (RE), Right Wrist (RW), Left Asis (LA), and Right Asis (RA). We calculated the euclidean distance between a pair of joints and the orientation. We identified some relevant lines connecting two joints and calculated the angle between them. We also computed the distance between a joint to a line.

1st and 2nd level features are extracted same as in [21]. For the 1st level features, we consider the left and right arm of the body parts as the most informative body parts for an action such as the movement of the right hand or left hand. We consider the shoulder joint of each hand as the origin and based on that the barycenter of the hand is calculated which is shown in Equation (1). We calculate the respective range, mean, variance, and skewness of the changes of the barycenter at a time span of t and are used as a set of features.

$$C_2^{(t)} = \frac{P_{L.E}^{(t)} + P_{L.W}^{(t)}}{2} - P_{L.S}^{(t)} \quad (1)$$

For the 2nd level features, the relative position of the end point of each hand (which is the wrist of a hand) from the shoulder is calculated. We also calculated the displacement of the wrist within a time frame.

3 PROPOSED MODEL

As discussed above, features are computed for each window within an 1-minute segment. Since, we essentially have sequential data for each segment, we propose to use 2-layer stacked GRU modules

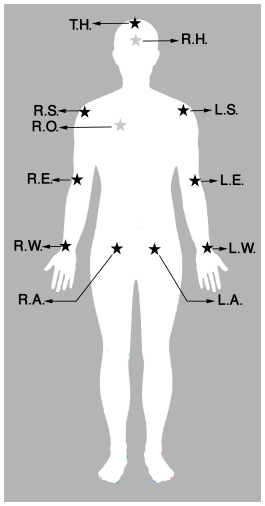
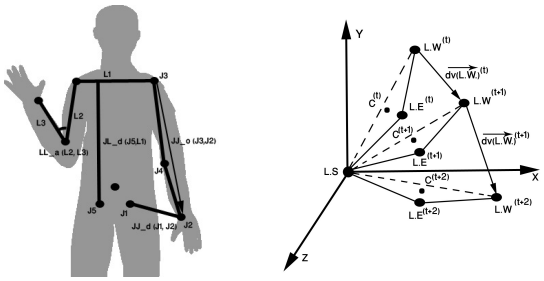


Figure 2: Informative Joints



(a) Geometric Features **(b) Features from Left Arm**

Figure 3: Motion Based Features

with simple and context dependent attention mechanisms for nurse activity recognition. While core components from [7] are adopted, key architectural modifications are needed to solve multivariate nurse activity recognition problems. From the sensor data, 4 sets for features are computed as shown in Fig. 4. These set of features are supplied to four different GRU with attention mechanism models. Each of these models produce class-wise score vectors which are finally combined to output the activity class. The individual GRU with attention mechanism models are presented in Fig. 5. We first describe the individual GRU with attention mechanism models and then describe the procedure to combine the class wise score vectors.

3.1 Gated Recurrent Unit

GRUs are modified versions of standard Recurrent Neural networks (RNN). It captures the long term dependencies by using two kinds of cell state transferring information, update gate ($Z_{(t)}$) and reset gate ($\Gamma_{(t)}$). It also aims to solve the vanishing gradient problem that occurs in standard RNNs. It has fewer parameters compared to Long short term memory (LSTM) which are also modified RNNs that attempts to solve the vanishing gradient problem. The equations for computing $Z_{(t)}$ and $\Gamma_{(t)}$ and the output of the GRU unit, i.e, current memory content ($h_{(t)}$) is given below.

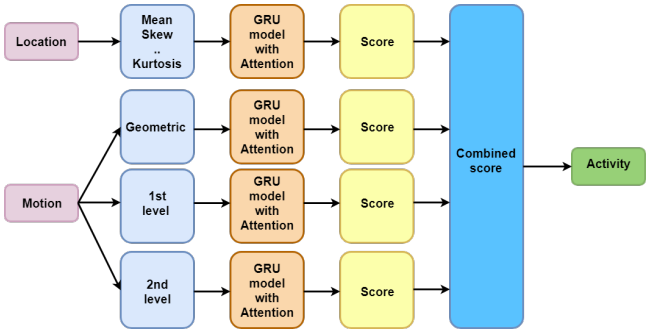


Figure 4: Proposed model full architecture.

$$Z_{(t)} = \sigma(W_{zx} \cdot X_{(t)} + W_{zh} \cdot h_{(t-1)} + b_z) \tag{2}$$

$$\Gamma_t = \sigma(W_{\Gamma x} \cdot X_{(t)} + W_{\Gamma h} \cdot h_{(t-1)} + b_\Gamma) \tag{3}$$

$$h_{(t)} = (1 - Z_{(t)}) \odot h_{(t-1)} + Z_{(t)} \odot \tanh(W_{hx} \cdot X_{(t)} + W_{hh} \cdot (h_{(t-1)} \odot \Gamma_t) + b_h) \tag{4}$$

Here, $\odot$ in (4) represents element-wise multiplication and σ in (2) and (4) represents sigmoid function. W's and b's are the parameters to be learned.

3.2 Simplified Attention

We add simplified attention mechanism to the basic GRU units to give more weight to more relevant features. These weights ($\alpha_{(t_i)}$s) are also learned. Using the weights we compute a score vector $c_{(i)}$. The following equations shows the computations that are performed in this module.

$$e_{(t_i)} = \tanh\left(W_{as} \cdot h_{(t_i)} + b_{as}\right) \tag{5}$$

$$\alpha_{(t_i)} = \frac{\exp(e_{(t_i)})}{\sum_t \exp(e_{(t_i)})} \tag{6}$$

$$c_{(i)} = \sum_t \alpha_{(t_i)} h_{(t_i)} \tag{7}$$

3.3 Context Dependent Attention

We also use context dependent attention that exploits spatial as well as temporal contexts of the data. The equations for capturing context dependent scores using attention are given bellow.

$$e_{(t_i)} = \tanh\left(W_{ac} \cdot h_{(t_i)} + b_{ac}\right) \tag{8}$$

$$\alpha_{(t_i)} = \frac{\exp(e_{(t_i)}^T \cdot e_s)}{\sum_t \exp(e_{(t_i)} e_s)} \tag{9}$$

$$c_{(i)} = \sum_t \alpha_{(t_i)} h_{(t_i)} \tag{10}$$

W_{ac} and b_{ac} in (8) are parameters to be learned. e_s in (9) allows the preservation of context information and is learned jointly when training the network. A summation of the relative weights of the time steps is generated as the context dependent score vector in (10).

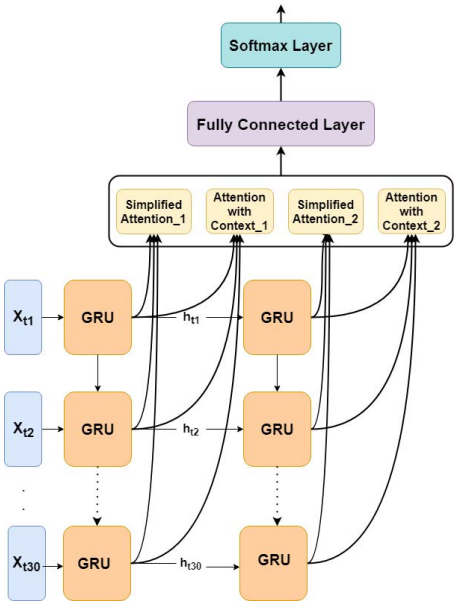


Figure 5: Stacked 2 layer GRU model with simplified and context sensitive attention mechanism

3.4 Fully Connected and Softmax layers

Here the attention score vectors are used to produce class-wise scores. First, the attention score vectors produced by the stacked GRU modules are concatenated and then batch normalization to put them in the same scale. This vector is provided as input to two fully connected layers. We use Rectified Linear Unit (ReLU) in these layers. Also dropout is used to increase the robustness of the model as well as remove any simple dependencies between the neurons preventing over fitting. Finally, we use softmax to compute class-wise scores. We train the model with learning rate α and decay factor λ.

3.5 Combining model outputs

As four different models give four different class-wise scores for six different activities, we need to combine each of model's scores to take final decision. First, we use class-wise scores as a matrix g which is given bellow. As we have six classes and four different models, the matrix would be $6 * 4$.

$$g = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{14} \\ s_{21} & s_{22} & \dots & s_{24} \\ \vdots & \vdots & \ddots & \vdots \\ s_{61} & s_{62} & \dots & s_{64} \end{bmatrix}$$

We add the entries of the rows of g to produce the final combined class-wise scores. Here, S represents the class-wise combined scores.

$$S = \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_6 \end{bmatrix}$$

Then final decision is calculated using equation (11).

$$\hat{y} = arg\max_i \sum_{i=1}^{6} S_i \tag{11}$$

4 EXPERIMENTAL RESULTS

We tested our algorithm on a challenging dataset named "Nurse Care Activity Recognition Challenge" created for a open lab activity recognition challenge[1].

4.1 Dataset

The dataset contains of 6 activities with 8 subjects, 5 repetitions for each, yielding about 240 activity sequences and 407 recorded minutes. The six activities are Vital signs measurements, Blood Collection, Blood Glucose Measurement, Indwelling drip retention and connection, Oral care, Diaper exchange and cleaning of area. They recorded each activity using accelerometers, meditag data and motion capture data including 29 body markers. The Data were divided in 1-minute segments. Train data contains 282 segments from 6 subjects. Test data contain 116 segments from 2 other subjects. For our experiment we took 2 second window size, resulting 7781 windows in training data.

4.2 Result Analysis

Table 1: Accuracy of single(1) and 2-Layer Model

	Model	M1	M2	M3	M4	Combined
F-1	1-layer	54.17	54.17	20.83	33.33	41.67
	2-layer	58.33	58.33	20.83	62.50	59.33
F-2	1-layer	40.00	40.00	72.00	64.00	60.00
	2-layer	44.00	44.00	80.00	72.00	88.00
F-3	1-layer	46.88	46.88	62.50	43.75	68.75
	2-layer	46.88	46.88	56.25	71.88	61.13
F-4	1_layer	35.14	35.14	64.84	43.24	51.35
	2-layer	67.57	67.57	72.97	56.76	56.76
F-5	1-layer	36.36	36.36	40.91	40.91	50.00
	2-layer	65.91	65.91	59.09	63.64	61.36
F-6	1-layer	40.00	40.00	65.00	60.00	75.00
	2-layer	55.00	55.00	65.00	65.00	72.50
avg.	1-layer	42.09	42.09	54.35	47.54	57.79
	2-layer	56.28	56.28	59.02	65.29	**66.43**

For classification with machine learning algorithm, generally 10 fold cross validation is used. We did not use 10 fold cross validation. This is because, if we use 10 fold cross validation then subject specific feature may influence the classifier and we will obtain biased results. To identify a particular nurse activity, it is thus required to use person independent training and test set. We use leave one subject out cross-validation. Windows of a single segment is feed into the model in each iteration during the training of the model. To run the experiments, we used Google Colaboratory.

We have shown our experimented results for two different approaches namely 1-layer GRU with attention model and 2-layer GRU with attention model shown in Table 1. It can be observed from the table that by adding one layer, the average accuracy (for

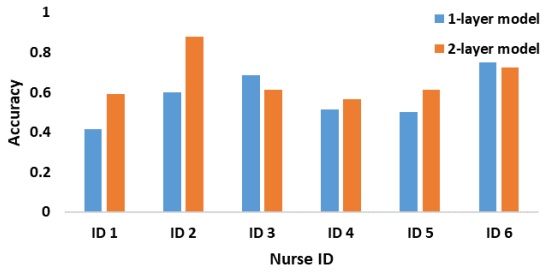


Figure 6: Accuracy comparison of two different model

Table 2: Confusion matrix of leave one subject out CV

	C_1	C_2	C_3	C_4	C_5	C_6
C_1	23	10	1	2	3	9
C_2	2	56	0	0	0	5
C_3	1	16	8	1	4	6
C_4	2	3	0	19	1	7
C_5	2	1	2	0	20	8
C_6	1	3	1	0	2	58

leave one subject out CV) of every model along with combined model increase. Again, we also see that the performance of combined model is better than that of the individual models. This shows that by combining all the models improves the ability of identifying complex nurse activities through utilizing all the extracted features. The comparison of average accuracy of two different combined models is shown in Fig. 6.

We have also performed class specific accuracy as shown in Table 2. It is observed that our model can correctly identify C_1,C_2, C_5 and C_6 as these are distinguishable activities. But for C_3, our model confuses with C_2 as these activities have similar movement of hands of the nurse near the patient's hands.

5 CONCLUSION

In this paper, we use a deep learning based method with attention mechanism which achieves reasonable result for leave one subject out cross validation. This is because stacked GRU based models with attention mechanism are capable to capture the sequential nature of the data. We extracted a number of features proposed in literature. More relevant features for the identification of nurse activity may improve the overall performance which we will address in future. The recognition result for the testing dataset will be presented in the summary paper of the challenge[18].

ACKNOWLEDGMENTS

This research is partially supported by a grant from Independent University, Bangladesh.

REFERENCES

[1] 2019. Nurse Care Activity Recognition Challenge. https://doi.org/10.21227/2cvj-bs21
[2] Mohamed Abdur Rahman, Ahmad M Qamar, Mohamed A Ahmed, M Ataur Rahman, and Saleh Basalamah. 2013. Multimedia interactive therapy environment for children having physical disabilities. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, 313–314.
[3] Pritom Saha Akash, Md. Eusha Kadir, Amin Ahsan Ali, and Mohammad Shoyaib. 2019. Inter-node Hellinger Distance based Decision Tree. In *IJCAI*.
[4] Antonio Padilha Lanari Bo, Mitsuhiro Hayashibe, and Philippe Poignet. 2011. Joint angle estimation in rehabilitation with inertial sensors and its integration with Kinect. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 3479–3483.
[5] Chien-Yen Chang, Belinda Lange, Mi Zhang, Sebastian Koenig, Phil Requejo, Noom Somboon, Alexander A Sawchuk, and Albert A Rizzo. 2012. Towards pervasive physical rehabilitation using Microsoft Kinect. In *2012 6th international conference on pervasive computing technologies for healthcare (PervasiveHealth) and workshops*. IEEE, 159–162.
[6] Yao-Jen Chang, Shu-Fang Chen, and Jun-Da Huang. 2011. A Kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities. *Research in developmental disabilities* 32, 6 (2011), 2566–2570.
[7] M. N. Haque, M.T.H. Tonmoy, S. Mahmud, A. A. Ali, M.A.H. Khan, and M. Shoyaib. May 3-5, 2019. GRU-based Attention Mechanism for Human Activity Recognition. In *in the proceedings of 1st International Conference on Advances in Science, Engineering and Robotics Technology*. ICASERT.
[8] Sozo Inoue, Naonori Ueda, Yasunobu Nohara, and Naoki Nakashima. 2015. Mobile activity recognition for a whole day: recognizing real nursing activities with big dataset. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1269–1280.
[9] Ahmad Jalal, Yeon-Ho Kim, Yong-Joong Kim, Shaharyar Kamal, and Daijin Kim. 2017. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern recognition* 61 (2017), 295–308.
[10] Sonia Jubil. 2015. Applications and Challenges of Human Activity Recognition using Sensors in a Smart Environment. *International Journal for Innovative Research in Science Technology* 2, 04 (2015).
[11] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. 2018. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055* (2018).
[12] Xiaoqiang Li, Yi Zhang, and Dong Liao. 2017. Mining key skeleton poses with latent svm for action recognition. *Applied Computational Intelligence and Soft Computing* 2017 (2017).
[13] Wei Liu, Sanjay Chawla, David A Cieslak, and Nitesh V Chawla. 2010. A robust decision tree algorithm for imbalanced data sets. In *Proceedings of the 2010 SIAM International Conference on Data Mining*. SIAM, 766–777.
[14] Roanna Lun and Wenbing Zhao. 2015. A survey of applications and human motion recognition with microsoft kinect. *International Journal of Pattern Recognition and Artificial Intelligence* 29, 05 (2015), 1555008.
[15] Fotini Patrona, Anargyros Chatzitofis, Dimitrios Zarpalas, and Petros Daras. 2018. Motion analysis: Action detection, recognition and evaluation based on motion capture data. *Pattern Recognition* 76 (2018), 612–622.
[16] Hossein Pazhoumand-Dar, Chiou-Peng Lam, and Martin Masek. 2015. Joint movement similarities for robust 3D action recognition using skeletal data. *Journal of Visual Communication and Image Representation* 30 (2015), 10–21.
[17] Lasitha Piyathilaka and Sarath Kodagoda. 2015. Human activity recognition for domestic robots. In *Field and Service Robotics*. Springer, 395–408.
[18] Alia Sayeda Shamma, Paula Lago, Shingo Takeda, Tittaya Mairittha, Nattaya Mairittha, Farina Faiz, Yusuke Nishimura, Kohei Adachi, Tsuyoshi Okita, Sozo Inoue, and Francois Charppillet. 2019. Nurse Care Activity Recognition Challenge: Summary and Results. In *Proc. HASCA*.
[19] Sadia Sharmin, Amin Ahsan Ali, Muhammad Asif Hossain Khan, and Mohammad Shoyaib. 2017. Feature selection and discretization based on mutual information. In *2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, 1–6.
[20] Sadia Sharmin, Mohammad Shoyaib, Amin Ahsan Ali, Muhammad Asif Hossain Khan, and Oksam Chae. 2019. Simultaneous feature selection and discretization based on mutual information. *Pattern Recognition* 91 (2019), 162–174.
[21] Benyue Su, Huang Wu, Min Sheng, and Chuansheng Shen. 2019. Accurate Hierarchical Human Actions Recognition From Kinect Skeleton Data. *IEEE Access* 7 (2019), 52532–52541.
[22] Nguyen Xuan Vinh, Shuo Zhou, Jeffrey Chan, and James Bailey. 2016. Can high-order dependencies improve mutual information based feature selection? *Pattern Recognition* 53 (2016), 46–58.