HappyFeet: Recognizing and Assessing Dance on the Floor

Abu Zaher Md Faridee*, Sreenivasan Ramasamy Ramamurthy†, H M Sajjad Hossain, Nirmalya Roy
Department of Information Systems, University of Maryland Baltimore County
{faridee1,rsreeni1,hmsajja1,nroy}@umbc.edu

ABSTRACT

The widespread availability of Internet-of-Thing (IoT) devices, wearable sensors and smart watches have been promoting innovative activity recognition applications in our everyday lives. Recognizing dance steps with fine granularity using wearables is one of those exciting applications. In a typical dance classroom scenario where the instructors are frequently outnumbered by the students, accelerometer sensors can be utilized to automatically compare the performance of the dancers and provide informative feedback to all the stakeholders, for example the instructors and the learners. However, owing to the complexity of the movement kinematics of human body, building a sufficiently accurate and reliable system can be a daunting task. Utilization of multiple sensors can help improve the reliability, however most wearable sensors do not boast sufficient resolution for such tasks and often suffer from various data sampling, device heterogeneity and instability issues. To address these challenges, we introduce *HappyFeet*, a convolutional neural network based deep, self-evolving feature learning model that accurately recognizes the micro steps of various dance activities. We show that our model consistently outperforms feature engineering based shallow learning approaches by a margin ≈7% accuracy on data collected from dance routines (Indian classical) performed by a professional dancer. We also posit a *Body Sensor Network* model and discuss the underpinning challenges and possible solutions associated with multiple sensors' signal variations.

ACM Reference Format:
Abu Zaher Md Faridee, Sreenivasan Ramasamy Ramamurthy, H M Sajjad Hossain, Nirmalya Roy. 2018. HappyFeet: Recognizing and Assessing Dance on the Floor. In *HotMobile'18: 19th International Workshop on Mobile Computing Systems & Applications, February 12–13, 2018, Tempe , AZ, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3177102.3177116

1 INTRODUCTION

With the proliferation of wearable sensors over the last few years, a plethora of exciting new applications is evolving every day. The inbuilt accelerometer, gyroscope, ambient light sensor, altimeter, GPS and heart rate sensors of these wearable devices are being exploited in various application domains ranging from health care, sports, fitness, entertainment etc. While video camera based sensors have also been used widely in various application domains, majority of them suffer from user privacy concerns and overlapped field of view in presence of multiple users such as in case of our proposed dance activity recognition scenario.

Dance activity involves subtle movements of limbs and other body parts in a sequenced fashion. In a professional dance-learning environment, in general there are more students than instructors, which makes it harder for the instructor to dance, teach, and assess the performance of the students simultaneously and divert attention equally to each of the students. Therefore, In this work, we propose to develop "Dance Activity Recognition" (DAR) system which can provide feedback on the steps performed by the students to an instructor. The proposed system helps the instructor to readily identify the mistakes and postulate corrections in mind while the DAR system needs to deal with very fine-grained labeling followed by accurate classification of various dance steps. Recognizing the dance activity is fundamentally different from recognizing and learning the traditional *Activities of Daily Living (ADLs)*. Dancing requires grace and finesse, and involves repetitive movements of the fingers, hands, forearm, elbow, arm, legs, toes, waist, heads etc., in a rhythmic fashion. It also reflects the delicacy and rhythm of different postures along with the cognitive ability and physical fitness of an individual. One step alone may consist of multiple micro steps which span across the various movements of legs, hands, fingers, shoulder, elbow etc. Capturing these movements with a minimal number of accelerometer sensors, recognizing and delimiting these micro steps, and defining a repetitive pattern out of it to recognize the entire dance episode are non-trivial activity recognition problems. This makes a *Dance Activity Recognition (DAR)* system unique in its own context than the traditional *Human Activity Recognition (HAR)* problem.

The fine-grained modifications of the movements needed to enhance the overall performance of a dancer are not always apparent, and therefore assessed appropriately by an instructor. In this context, an autonomous *Dance Activity Recognition (DAR)* system can play an integral role by providing meaningful qualitative feedback to help improve the learning capabilities of the participants. The design of such a system can also help postulate how an individual participant is grasping the dance activity progressively in a group setting compared to a one-to-one learning environment. In order to capture the full extent of movements of the limbs during the dance activities which can vary drastically from subtle to pronounced, a full-fledged Body Sensor Network may be required. However, deploying sensors with different modalities which is prevalent in commercially available devices, introduces heterogeneity, synchronization and sampling instability problems. Considering the complexity of properly capturing the micro steps of dance moves, and

*Equally Contributing authors
†Equally Contributing authors

HotMobile'18, February 12–13, 2018, Tempe , AZ, USA

ACM ISBN 978-1-4503-5630-5/18/02…$15.00
https://doi.org/10.1145/3177102.3177116

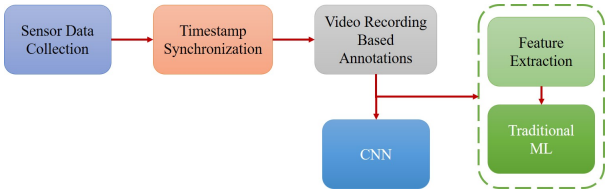

Figure 1: Overall framework of dance activity recognition system

the challenges introduced by employing multiple devices, we make the following contributions in this paper.

- We design an accelerometer-based multi-channel data collection prototype that helps capture the subtle movements of a dancer accurately.
- We propose a multi-channel deep convolutional neural network model that learns the dance moves by automatically and hierarchically learning the features that represent the raw data.
- We investigate empirically the coexistence of multiple heterogeneous devices and the inherent sensor biases, sampling rate heterogeneity, and sampling rate instability.

2 RELATED WORK

It has recently been shown that accelerometer can be used to quantify the performance during simple ballet dance activity [23]. The physical activities of children and adolescents in 7 different dance styles using accelerometer have been investigated and noted that the children were more active than the adolescents [2]. In addition, authors have also hypothesized that there is a requirement for a better teaching method to increase the physical activity. In contrast to a single accelerometer sensor, a sensor network was used as a viable input system for control in a video game involving dance activity called *Dance, Dance, Evolution*; a popular game in Asia [4]. A recurrent and convolutional neural network based models has been used to design and revamp the dynamisms of the game by generating new dance steps [6]. A motion capture and composition system for dance motion was developed by exploiting multiple RGB and depth sensors [12]. Another interesting study in an attempt to preserve the ancient Chinese folk dance, the authors transcribed the motion data into lab annotation which was captured by OptiTrack system [24]. A model that performs classification of Korean pop (K-pop) dances based on human skeletal motion data captured using Kinect sensor in a motion-capture studio environment has been studied [11]. The authors proposed an efficient Rectified Linear Unit (ReLU)-based neural network model without implementing weight learning which is efficient than conventional neural network. [7] proposed a novel technique that helps decompose the dance motions using the Hilbert-Huang transform and compare Waltz and Salsa dance movements with the dance of a Japanese pop group "Perfume". A Kinect-based Thai dance evaluation system was demonstrated in [17] which rates the user's performance and provides helpful and real-time feedback to the user. Recognizing Greek folk dances and their variations using Kinetc II sensor has been proposed in [18]. A mobile based dance education system has been proposed in [8] using wearables. A dance performance evaluation system was developed to decode the performed dance gestures as captured using high-precision motion capture system which

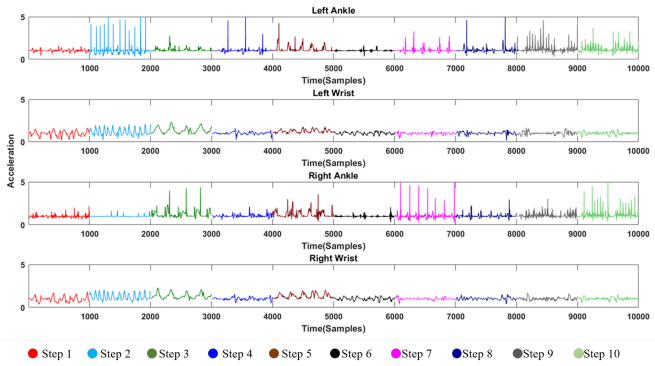

Figure 2: Accelerometer Signals of different activities captured from four *Actigraph* sensors

showed a likelihood value of the recognized gesture in terms of a score [14]. Skeletal pose based dance step recognition system has been proposed in [19] [10]. An image based classical Indian dance norms classification model was demonstrated in [20]. [5] proposed a dance training system using foot mounted inertial sensor which leverages the orientation and position of the foot to evaluate the trainee's movements. [26] depicts that an expert shows a consistent and repeatable pattern of dance activities, an intermediate participant shows a consistent action but is different to that of an expert and the novice shows a pattern which appears disjoint and noisy. Most of the dance recognition related literatures employed video based motion capture system, which requires delicate equipments and fixed installation in a confined setting. In this paper, we propose a Convolutional Neural Network (CNN) based dance activity recognition model using ubiquitous wearable devices focusing on detecting the micro steps of the dancers and mitigating the sensor bias, sampling rate heterogeneity, sampling rate instability, and synchronization problems.

3 OVERALL DESIGN AND SYSTEM SETUP

Figure 1 depicts the overall framework of *HappyFeet*, our proposed dance activity recognition system. Since dance involves different movements of limbs to perform distinct steps, it warrants more than one sensor to capture the user's actions with required accuracy. We used four *Actigraphs*, an *Empatica E4* and two *Microsoft Bands* (the details of the placement are described in subsection 3.2). The use of heterogeneous sensors pose challenges associated with multiple sensors data stream synchronization, sensor biases, and sampling rate instability [22], which are discussed in detail in section 4.1. In order to precisely annotate the ground truth, we employed a video recording based labeler called ELAN [15]. Thereafter we fed this annotated data to a hierarchical self-evolving feature-based deep convolutional neural network model to recognize the dance activities. We also posit a feature-based shallow machine learning model as a baseline to compare against the deep learning model. Next, we describe our *HappyFeet* setup and discuss the dance micro steps, sensor data collection process, data stream time synchronization, ground truth annotation and construction of shallow and deep learning models.

Table 1: Description of Activities

Class Label	Description
Step 1	Wave both hands from left to right
Step 2	Stepping right leg forward
Step 3	Clockwise Rotational Movement
Step 4	Walking forward with extended arms
Step 5	Anti-clockwise Rotational Movement
Step 6	Stepping left leg forward
Step 7	Step-by-step slow rotational movement (Clockwise)
Step 8	Step-by-step slow rotational movement (Anti-Clockwise)
Step 9	Step-by-step rapid rotational movement (Clockwise)
Step 10	Step-by-step rapid rotational movement (Anti-Clockwise)

3.1 Defining Dance Steps

In our experiment, we chose to study a classical Indian dance style: *Lasya* which is a subcategory of Manipuri [25] dance form; the dance is noted for its gentle, smooth and subtle limb movements. We collected data from one professional dancer and four learners. We designed a specific dance script for *Lasya* which a beginner would learn during the first few dance sessions. The steps of the dance script are described in Table 1 and depicted in Figure 2. The dance activities described in Table 1 are similar to normal daily activities in many ways, however, these activities involve a large number of minute finer movements. For instance, the first activity which is "Wave both hands from left to right" involves different *mudras* (hand postures) at different positions. Unlike ADLs or other activities, dance activities involves a combination of very short movements which forms the micro level activities. A sequence of such smaller combination of activities helps form a dance routine. In this study, we have recorded 10 such micro level activities. These fine-grained steps have no specific names in the dance literature but the sequence of these steps do have names. This study deals with the micro level dance steps and the sequence of the micro level steps is outside the scope of this study.

3.2 Data Collection

The participants were asked to wear the actigraph (model *wGT3X-BT*) [1] on all the limbs. In addition, to capture the heterogeneity across different devices, the participants were also asked to wear a *Microsoft Band* on the waist and the left hand, and a *Empatica E4* device on the right hand. The *ActiGraph wGT3X-BT* device has much higher sensitivity compared to the Microsoft band. When kept at at rest (Zero g test) [22] Actigraph and Empatica E4 showed constant 1g acceleration, indicating that Actigraph and Empatica have less *Sensor Biases*[22]. However, the Microsoft Band did not pass the Zero g offset test, the resultant acceleration signal was contaminated with noise; hence a low pass filter was used to nullify the sensor bias. Before each data collection session, we also synchronized the clocks of all of the sensors. We collected data for 20 trials out of which the first 10 trials were conducted as such that the participants danced only the specific micro steps repetitively. The remaining 10 trials were recorded as a sequence of all micro steps.

3.3 Synchronization and Annotations

The *ActiGraph wGT3X-BT* has a tri-axis accelerometer sensor, that gives us acceleration data for x, y and z axis at the desired sampling frequency (in this case 100 Hz) along with the UNIX time-stamp of each of the readings. The *Lasya* dance form in our experiment contains ten separate micro steps. The granularity of the steps can be varied if desired, for finer granular step identification the annotation, training and inference would warrant extensively more data collection over more dance sessions. In our case, the dance routine lasted for roughly one minute and each of the dance steps taking up between six to fourteen seconds (they are not of equal lengths). We recorded each dance session using a video camera and annotated each micro step of the dance session by synchronizing the video with the accelerometer data stream. We synchronized the signals from each sensor, the video and the timestamps associated with them using *ELAN* software [15]. At the start of each dance routine, the participants were asked to jump thrice as high as possible. These three jumps showed a peak in the resultant acceleration signal. The annotator used the peaks to synchronize the sensor data stream and the video feed, all at the same time. We deduced the starting and ending frame for each of the micro level dance moves and labeled them accordingly. Annotation is done with respect to the video feed not the accelerometer data as ELAN is originally a linguistic annotation tool. When doing the alignment, we also noted the video and accelerometer synchronization offset. With this information, we were able to derive the standardized time stamps across devices and crop the data that is of interest using the equation $T_{acc.} = T_{video} - O_{video} + O_{acc}$ (T_{acc} denotes the accelerometer time, T_{video} is the video time, O_{video} is synchronization offset of the video and O_{acc} is the synchronization offset of the accelerometer). Because of the initial clock synchronization, all sensor samples are also properly aligned with each other in the end.

3.4 Model Building

We design a deep learning model for recognizing the dance steps and compare its performance against several shallow learning approaches after the extraction, synchronization and annotation of the data, we carefully split the data between train and test set while also ensuring that both the sets have similar label distribution. This early separation of training and test samples ensures absolute zero overlapping between the samples. First, we applied filtering to nullify the effect of noise. We applied both *Kalman* filter and *Median* filter and noted that the frequency response of the filtered signals were similar. Therefore, we chose Median filter instead of *Kalman* filter as it has lower computational complexity. After filtering the noise, we divided the accelerometer data into 50 sample window with a sliding window approach (90% overlap) which helps prevent the model from being dependent on initial positioning of the windows. For shallow learning models, we extracted a total of 46 time and frequency domain features [3]. We calculated *Pearson's correlation coefficient* between the features and used a threshold t to remove the highly correlated features. We optimized the hyper-parameters of the classifiers with 5-fold cross validated (with stratified sampling) *Randomized Search* and performed the final evaluation on the held-out test data using the accuracy, recall and the F1-scores.

The CNN architecture employs three *convolution* layers which is the main building block for the self-evolving feature learning. These layers are followed by the two fully connected layers that take care of the actual classification by working on the features

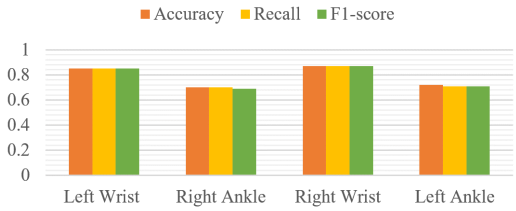

Figure 3: Accuracy, Recall and F1-score for each sensor in identifying all activities using Random Forest

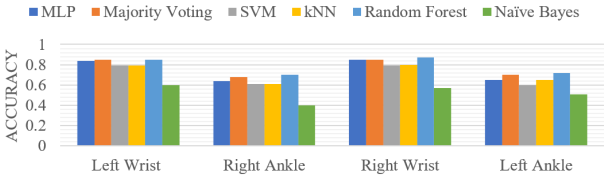

Figure 4: Comparison of Accuracy for each sensor in identifying all activities using all shallow learning algorithms

learned by the previous layers. Each of these layers actually consists of the following fundamental components:

- The convolution layer [13] consists of the following operations in sequence - convolution, batch normalization [9], rectilinear activation function and average pooling.
- The first fully connected layer also has a rectilinear activation function.
- We introduce a dropout [21] layer between the two fully connected layers which helps to make the network robust against over-fitting issue.
- We employ *soft-max* layer to get the final class label from the fully connected layers.

Average Pooling at the end of each *convolution layer* automatically helps smoothen out the data as it picks the average value and provides a low pass filtered version of the signal along with dimensionality reduction that nullifies the need to use filtering beforehand. For the CNN based approach we use the same sliding window technique mentioned before. After this preprocessing step, each sample consists of a 50x3 data points as we consider 3 axes in the accelerometer data.

4 EXPERIMENT SETUP AND EVALUATIONS

The experiments were conducted on a Windows platform consisting of Intel i7 6700HQ Quad Core Hyper-threaded CPU, 16GB DDR4 RAM and Nvidia Quadro M1000M workstation class GPU with 2GB of DDR3 RAM. For the signal processing, filtering and shallow learning tasks we used MATLAB R2017a [16] and python (scikit-learn). For the deep learning task, we performed the processing with GPU optimized version of Tensor-flow which cut down training time by 3.5 times compared to running the same code on CPU. The size of the training and testing raw samples for the experiments were of 51148x150 and 25192x150 respectively (66% vs 34% ratio). For this particular experiment, we annotated the dataset with 10 output labels. Table 1 describes the nature of these labels and Figure 2 shows the time series (time vs absolute magnitude) plot of labels. Due to the variable lengths between dance steps, we ended up with a little bit imbalanced distribution of the class labels (Table 2). This

Table 2: Class distribution

Class Label	Percentage	Class Label	Percentage
Step 1	7.96%	Step 6	5.16%
Step 2	3.55%	Step 7	12.19%
Step 3	11.79%	Step 8	13.02%
Step 4	12.74%	Step 9	12.65%
Step 5	9.11%	Step 10	11.77%

kind of class imbalance can make it very difficult to achieve high accuracy with supervised learners. This shows a great point that real world data is often riddled with less than desirable characteristics compared to synthetic datasets.

4.1 Device Heterogeneity

Deploying multiple sensors simultaneously poses some challenging *heterogeneity* issues such as sensor bias, non-uniform sampling rate, sampling rate instability etc. *Sensor bias* is a type of heterogeneity that is caused due to the differences in the precision, resolution, and range values of the devices [22]. *Sampling Rate Heterogeneity* occurs when two different devices starts recording the data at two different sampling rates. For instance, Device X records at 100 Hz and Device Y records at 75 Hz. This heterogeneity is undesirable as the proposed learning framework required equal number of data points from all the devices. A more challenging heterogeneity is *Sampling Rate Instability (SRI)* which is the irregularity between successive timestamps of consequent data points. To control and test the effect of heterogeneity, we maintained redundancy when collecting data with heterogeneous sensors. As described earlier, *Actigraph* was worn on all the limbs, Empatica E4 and Microsoft Band were worn on both of the hands. However, the *Actigraph* collected data at 100 Hz, the *Empatica E4* at 32 Hz and the Microsoft band at 128 Hz. We noticed that the Microsoft band data was suffering from SRI as it was missing certain data points. At the end we categorized these BSN heterogeneity issues into three groups:

(1) All the devices collect data at a constant sampling rate.
(2) Device X collects data at a constant sampling rate and Device Y collects data at a different constant sampling rate.
(3) Devices collect data at a constant rate, however one of the devices collects data at a varying sampling rate.

In this paper, we only focus on building the dance activity recognition system on the first case (multiple sensors collecting data at the same constant rate) and discuss the possible solutions of the other issues in Section 5.

4.2 Results

In this section, we discuss the preliminary results by analyzing the data from the four *Actigraph* sensors (placed on each of the limbs) for the professional dancer. First, we compare the accuracy between the baseline (shallow learning techniques) and CNN model using individual sensors and then we compare the performance of both using multiple sensors referred as BSN. We extracted 46 features from the time windows and used a Pearson's correlation co-efficient threshold of 0.75 to select only 24 of the features. We then train *Naive Bayes*, *K-nearest neighbor* (k=5), *Linear SVM*, *Multilayer Perceptron* and *Random Forest* as the baseline to compare against our CNN model. The classification accuracy is reported in Figure 4. *Random Forest* (with 2000 tress) performed the best (shown in Figure 3) among the shallow learning classifiers so we

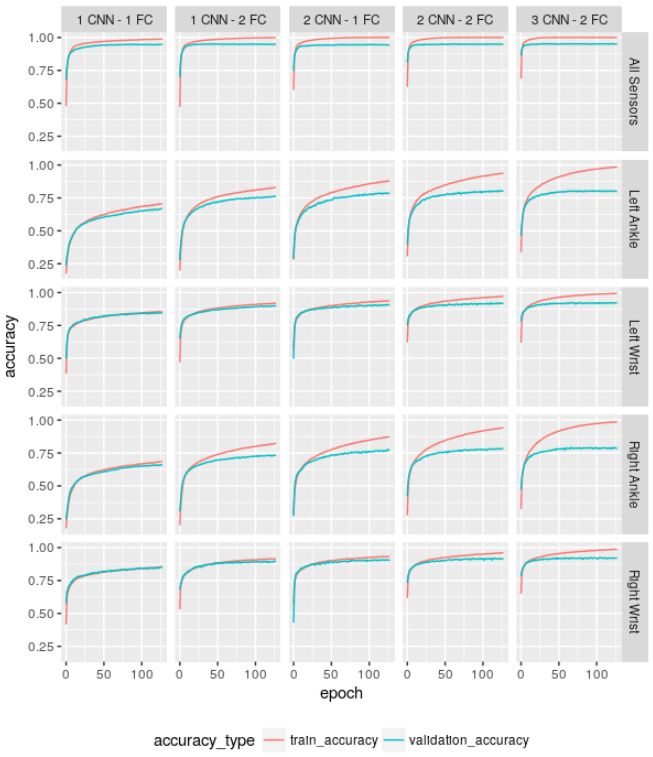

Figure 5: Accuracy comparison across sensors and CNN architectures

refer to it as our baseline henceforth; we then compare the baseline's performance with that of the deep learning model. We run our deep learning model on each of the sensor data separately and noted that the CNN model is consistently performing better than the baseline (*Random Forest*) which is shown in Table 3.

These results encouraged us to move to the next logical step, combining multiple sensors streams to develop CNN based *Body Sensor Network* and investigate whether it improved the discriminative power of the model. For example, Step 7 described in Table 1 is an activity involving step-by-step slow rotational movement in clock wise direction where both the legs should capture similar patterns for the activity. In Figure 2 we see that the patterns of Step 7 for right ankle is more distinctive and repetitive in nature when compared to that of left ankle. For left ankle the baseline classifier accuracy was 69% whereas for the right ankle it was 84% (figure not included due to space constraint). We hypothesized that combining data from multiple sensors would help mitigate the scarcity of discriminative data for certain classes. Our deep learning architecture was fluid in terms of employing multiple sensors requiring no complicated and computationally expensive modification of the current network structure. We have 3 channels per sensor data stream on which the CNN performs 3D convolution operations. With the added homogeneous sensors, the CNN would then be performing *nx3*-dimensional convolution operations (n being the total number of sensors). We already had the sensor data aligned (section 3.3) so this was a fairly straightforward process and without any special change in the current CNN structure we were able to get 93.81% hold-out test accuracy. This is not surprising; as we have stated previously, the intuition behind using CNN for dance activity recognition was to generate self developing features that are able to capture the truer representation of the data compared to the hand-crafted heuristics. In contrast to 3 channel single sensor data stream, the 12 channel 4 sensors (two Actigraphs, one Empatica, one Microsoft band) data stream gives CNN more patterns to holistically analyze and yield better accuracy. We experimented with different CNN architectures for both the individual sensors and BSN model to find the optimal hyper-parameters and the network structure. The details of the hyper-parameters are described in Table 4. In Figure 5, we compare the training and validation set accuracy for the combined BSN and each of the sensors (placed row-wise) and five network structures of increasing complexity (more convolutional and fully connected layers, placed column-wise). Table 5 shows the accuracy on held-out test dataset for each model. It can be noted that for each of the individual sensors, adding more layers improves the accuracy; adding the 2nd CNN layer over base 2 layer network seems to provide the higher boost ($\approx$ 8%) compared to just adding 2nd fully connected layer. Adding another CNN and fully connected layer each provide $\approx$1% more accuracy gain. Surprisingly, the BSN model achieved high accuracy with even the simplest network model (1 CNN layer and 1 Fully Connected layer), the difference between this model and the most complex model was just $\approx$0.67%, hinting that a simpler CNN architectures become as effective as complex models when larger supplementary sensor data streams are available.

5 LIMITATIONS AND FUTURE DIRECTIONS

This paper only discusses the preliminary findings of our BSN architecture, a lot of areas still remain unexplored and we would like to keep investigating and improve the model. Although we did capture the dance activity data from four *Actigraph* sensors, two *Microsoft Bands* and one *Empatica E4* sensor, in this paper we limited the BSN model to use just the *Actigraphs*. As mentioned in Section 3.2, we faced issues related with failed zero-g test, mismatched sampling rates, missing data points (either due to sensor fault or sampling rate instability) etc. The next challenge is to deal with the heterogeneity issues both experimentally and analytically (Section 4.1). We plan to investigate the effectiveness of recent deep learning architectures such as *Restricted Boltzmann Machine* and *Generative adversarial networks (GANs)* in reconstruction of these erroneous data samples. As stated previously, we collected redundant data with multiple sensors on the same limb; with large quantities of such unlabeled data from a high precision sensors, we can train an RBM to learn the intrinsic probability distribution of the dance activities and use that to reconstruct the missing data points. During this preliminary study, we limited ourselves to detecting 10 beginner level dance steps, but we would like to detect more advanced dance steps in future that consists of more complex micro-level gestures with a more advanced data collection and annotation setup. We also had to restrict ourselves to using the data collected on the dance steps performed by the professional dancer as the data collected from the students had extreme and unpredictable variances (due to their unfamiliarity with the dance routines). This complicated the process of building a consistent dance activity recognition model. In order to provide feedback to the instructor about the dance routines

Table 3: Comparison between CNN and Random Forest for each sensor

	CNN			Random Forest		
Sensor Position	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Left Ankle	79.45%	79.66%	79.36%	71.87%	72.08%	71.87%
Left Wrist	91.90%	92.44%	91.32%	84.34%	84.26%	84.34%
Right Ankle	76.75%	76.91%	76.66%	69.40%	69.88%	69.40%
Right Wrist	91.84%	91.79%	91.30%	87.95%	88.13%	97.96%

Table 4: Hyper-parameters of CNN model

Hyper-parameters	Values
No. of maximum convolution layers	3
No. of filters in convolution layers	32, 48, 64
Convolution filter dimensions	21x1, 15x1, 7x1
No. of maximum fully connected layers	2
No. of neurons in fully connected layers	64, 10
Batch size	64
Dropout rate	1.0
Max number of epochs	128

Table 5: Final Test Accuracy across sensors and CNN architectures

Sensor Position	1 CNN 1 FC	1 CNN 2 FC	2 CNN 1 FC	2 CNN 2 FC	3 CNN 2 FC
All Sensors	93.53%	93.44%	94.20%	93.56%	93.81%
Left Ankle	65.67%	73.42%	77.13%	78.65%	79.45%
Left Wrist	85.23%	88.29%	90.50%	90.96%	91.90%
Right Ankle	64.17%	72.05%	74.94%	75.15%	76.75%
Right Wrist	83.77%	88.21%	89.29%	90.69%	91.84%

performed by the students and track their progress, we will need to extend the model; it will also be interesting to investigate the qualitative improvements of the model and whether the adoption of the model improve the overall teaching and learning experience of both the students and the instructors. We also want to investigate whether the model can be adopted to a mobile/cloud based settings and the possibility of providing real time feedback to the users. Finally, the efficacy of a multi-modal combination of the BSN based approach with vision based approaches could also be investigated.

6 CONCLUSION

In this paper, we have presented the initial stages of a deep convolutional neural network based dance activity recognition model that can automatically learn the morphologically distinct features from multi-channel sensor data and then use those features to correctly identify the dance steps. We achieved ≈7% better accuracy than existing popular hand-crafted feature engineered machine learning approaches. The fact that the CNN architecture, apart from a few hyper-parameter tuning steps does not require time consuming feature engineering, without which the traditional models perform poorly. This supports our initiative of using deep learning architectures for daily activity recognition tasks. We also described the problem of heterogeneity and sampling rate instability for different sensor modalities in body sensor network and discussed how to overcome this adversity. We demonstrated a video recording based ground truth data annotation synchronizer for accurate labeling of large amount of dance activity data. Finally, we demonstrated the flexibility of the CNN approach, and showed that our Dance Activity Recognition system, *HappyFeet* can be easily extended reliably to heterogeneous body sensor network.

REFERENCES

[1] ActiGraph. 2017. (2017). http://www.actigraphcorp.com/
[2] Kelli L Cain, Kavita A Gavand, and Terry L Conway et al. 2015. Physical activity in youth dance classes. *Pediatrics* (2015), peds–2014.
[3] Diane J Cook and Narayanan C Krishnan. 2015. *Activity learning: discovering, recognizing, and predicting human behavior from sensor data.*
[4] Nick Crampton and Kaitlyn Fox et al. 2007. Dance, dance evolution: Accelerometer sensor networks as input to video games. In *Haptic, Audio and Visual Environments and Games, 2007. HAVE 2007. IEEE International Workshop on.* IEEE, 107–112.
[5] Quoc Khanh Dang, Duy Duong Pham, and Young Soo Suh. 2015. Dance training system using foot mounted sensors. In *SICE.* 732–737.
[6] Chris Donahue, Zachary C. Lipton, and Julian McAuley. 2017. Dance Convolution. In *ICML.* 1039–1048.
[7] Ran Dong, DongSheng Cai, and Nobuyoshi Asai. 2017. Nonlinear dance motion analysis and motion editing using Hilbert-Huang transform. In *CGI.* 35:1–35:6.
[8] Augusto Dias Pereira dos Santos. 2017. Smart Technology for Supporting Dance Education. In *UMAP.* 335–338.
[9] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning.* 448–456.
[10] Sotiris Karavarsamis and Dimitrios Ververidis et al. 2016. Classifying Salsa dance steps from skeletal poses. In *CBMI.* 1–6.
[11] Dohyung Kim, Donghyeon Kim, and Keun-Chang Kwak. 2017. Classification of K-Pop Dance Movements Based on Skeleton Information Obtained by a Kinect Sensor. *Sensors* 17, 6 (2017), 1261.
[12] Yejin Kim. 2017. Dance motion capture and composition using multiple RGB and depth sensors. *IJDSN* 13, 2 (2017).
[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems.* 1097–1105.
[14] Sohaib Laraba and Joëlle Tilmanne. 2016. Dance performance evaluation using hidden Markov models. *Journal of Visualization and Computer Animation* 27, 3-4 (2016), 321–329.
[15] Hedda Lausberg and Han Sloetjes. 2009. Coding gestural behavior with the NEUROGES-ELAN system. *Behavior research methods* 41, 3 (2009), 841–849.
[16] MATLAB. 2017. *version 9.2.0 (R2017a).* The MathWorks Inc., Natick, Massachusetts.
[17] Ob-orm Muangmoon, Pradorn Sureephong, and Karim Tabia. 2017. Dance Training Tool Using Kinect-Based Skeleton Tracking and Evaluating Dancer's Performance. In *IEA/AIE 2017, Part II.* 27–32.
[18] Eftychios Protopapadakis and Athanasios Voulodimos et al. 2017. A Study on the Use of Kinect Sensor in Traditional Folk Dances Recognition via Posture Analysis. In *PETRA.* 305–310.
[19] Sriparna Saha and Rimita Lahiri et al. 2016. Human skeleton matching for e-learning of dance using a probabilistic neural network. In *IJCNN.* 1754–1761.
[20] Shubhangi and Uma Shanker Tiwary. 2016. Classification of Indian Classical Dance Forms. In *IHCI.* 67–80.
[21] Nitish Srivastava and Geoffrey E Hinton et al. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR* (2014).
[22] Allan Stisen and Henrik Blunck et al. 2015. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proc of the 13th ACM Conference on Embedded Networked Sensor Systems.* 127–140.
[23] Thiel, David V and Quandt, Julian and Carter, Sarah JL and Moyle, Gene. 2014. Accelerometer based performance assessment of basic routines in classical ballet. *Procedia Engineering* 72 (2014), 14–19.
[24] Jiaji Wang, Zhenjiang Miao, and et al. 2017. Using automatic generation of Labanotation to protect folk dance. *J. Electronic Imaging* 26, 1 (2017).
[25] Wikipedia. 2017. Manipuri Dance — Wikipedia, The Free Encyclopedia. (2017). https://en.wikipedia.org/wiki/Manipuri_dance
[26] Miguel Xochicale, Chris Baber, and Mourad Oussalah. 2017. Analysis of the Movement Variability in Dance Activities Using Wearable Sensors. In *Wearable Robotics: Challenges and Trends.* Springer, 149–154.