Talking Hands: RNN-based Sign Language Recognition

Mohak Kumar Sukhwani
IIIT Hyderabad
mohak.sukhwani@research.iiit.ac.in

Prateek Panwar
University of Ontario Institute of Technology
prateek.panwar@uoit.ca

Abstract

We investigate the utility of mobile accelerometer data to identify the human hand gestures and recognize sign language using Recurrent Neural Networks (RNNs). A set of unsupervised features are learnt, to recognize the phrases from American Sign Language (ASL), using Restricted Boltzman Machines (RBM). We validate the efficacy of our approach by comparing it with the best performing supervised feature set. The proposed unsupervised features outperform the traditional handcrafted features. We use a labelled dataset of 600 accelerometer readings collected from 50 users to validate the proposed approach.

Index Terms: H.5.2 [User Interfaces]: User-centered design; I.5.4 [Applications]: Signal processing; H.5.m. [Information Interfaces and Presentation]: Miscellaneous

1 Introduction and Related Work

The recent surge in the popularity of wearable devices is changing the way we communicate with our surroundings. From healthcare devices to smart assistants these devices have seamlessly permeated into our daily lives. The majority of these devices are physical activity tracking specific – they often track and record steps, sleep patterns, calories burnt etc. They record human activities using various on board sensors such as gyroscope, GPS, IMU and accelerometer. In our proposed work, we explore the use of accelerometer for the fine-grained activity recognition, specifically, sign language recognition.

A variety of algorithms have been applied on the accelerometer data harnessed from mobile devices and other wearables to perform human activity recognition [1, 5, 7]. For action prediction, multiple accelerometer sensors (attached to various body parts) have proven better in terms of accuracy but these methods often fall short in 'out of lab' set-ups due to the discomfort of wearing multiple devices. Widespread availability of the accelerometers in wearables, compact size, low cost and low power requirements justifies our selection. We take in our motivations from the previous human activity recognition experiments [1, 7] and propose the use of accelerometer for 'phrase recognition' in an unsupervised setting. Given a use case, where a deaf person responds to someone who does not know ASL or a deaf person wishes to interact in a foreign country, our solution can come in handy. To the best of our knowledge we haven't witnessed the use of mobile accelerometer data for ASL recognition in past.

In past we have witnessed the use of vision and signal processing based algorithms [5, 7] to recognize hand gestures for interactions. These methods have dealt with recognition of simple 'words' and 'gestures'. We instead use the accelerometer sensor data to generate a description for the actions performed by a person wearing the sensor. We use RBMs to learn an unsupervised feature representation of the accelerometer data. Accelerometer-specific features are learnt by finding the compact low-dimensional representation using RBMs [4, 6]. We use stacked variant of RBM, i.e. DBN (Deep Belief Network), to learn hierarchical organization of explanatory factors in data. DBNs are theoretically more expressive in nature. Being a composition of simple RBM networks they yield better representative features and thus are more suitable for our needs. We explore the utility of stacked RBMs to generate features of unsegmented input accelerometer signals for phrase recognition.

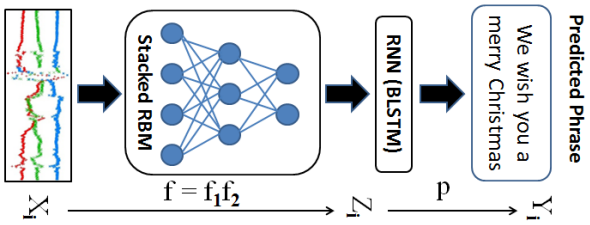

Figure 1: Overview: The stacked RBM network is trained for unsupervised feature generation. The computed features are used for phrase classification using RNN.

2 Approach

Hand-crafted features are often difficult to design and rely on expert knowledge. We use unsupervised features to fill this void. Our proposed framework uses an unsupervised learned representations for phrase recognition instead of handcrafted features [1, 7]. The suggested pipeline is a cascade of two models, namely a stacked RBM for learning representations followed by RNN which is used for classification (Figure 1).

Feature Learning using stacked RBM: Let, $\{X_i\}_{i=1}^{N}$ be the unlabelled dataset used for training stacked RBM. X_i is a vector sequence of i^{th} accelerometer signal obtained by sliding window over corresponding input signal. Unsupervised feature learning in stacked continuous RBM is a non-linear projection $f : x \rightarrow z \mid x \in \mathbb{R}^{d_v}, z \in \mathbb{R}^{d_h}$. Z_i is a vector sequence where each vector in the sequence is a learnt compact representation of the input vector. RBM models a distribution over visible variables (**v**) by introducing a set of stochastic hidden features (**h**) and jointly modelling the visible and hidden variables.

Phrase Recognition using RNN: We use RNN-based Bidirectional Long Short Term Memory (BLSTM) network for phrase recognition. The network consists of two bi-directional LSTM networks – one network processes the input from beginning to end while other processes it from end to beginning. The individual output of both the LSTM networks are used to compute the final output. The learnt representation is fed an input to BLSTM network. The Connectionist Temporal Classification (CTC) [3] is used at the output layer of RNN network to label the unsegmented data which uses a forward-backward algorithm. The CTC layer directly outputs the probability distribution of the desired label (phrases in our case). We formulate the problem of phrase recognition as that of a

Name	Contents	Role
OPPORTUNITY	Sampled wrist accelerometer data from the online dataset	RBM weights for feature learning
ASL	600 accelerometer readings 600 for selected phrases	RNN weights for ASL recognition

Table 1: Dataset statistics: Our dataset is a culmination of two datasets. This table describes them in detail, along with the roles they play in our experiments.

Method	Raw Feat.		BoW		Method	Raw Feat.		BoW	
	L_1	L_2	L_1	L_2		L_1	L_2	L_1	L_2
Naive Bayes	0.18	0.19	0.34	0.39	SVM - Linear	0.13	0.13	0.16	0.20
KNN	0.14	0.23	0.25	0.25	SVM - Poly	0.28	0.25	0.325	0.30
Reg. Trees	0.24	0.26	0.285	0.30	SVM - RBF	0.19	0.19	0.45	0.46

Table 2: Phrase classification accuracy using hand crafted features: L_1 and L_2 correspond to normalization scheme. The proposed unsupervised features obtain an accuracy of 0.53 with stacked RBMs.

classification problem where each of the identified 12 phrases are considered as a unique 'class label'.

3 Experiments

An Android application was developed for a mid-range smart phone to record the accelerometer readings from wrist movement while volunteers perform the actions of the corresponding phrase. The application was developed on 'Android Version: 4.2.1' with 'Kernel Version: 3.4.5'. Neural network training and other relevant computations were done on *i*3 processor with 8GB RAM. In all 50 participants, 32 male and 18 female, were recruited from the local university to perform the sign actions. Participant's age ranged from 19 years to 28 years (23 year mean). None of the participants were familiar with ASL which reduces the biases involved with precognition and assists us to establish the robustness of the approach.

Dataset: We effectively use two datasets for our experiments. Table 1 summarises the data used for the experiments. To compensate for the limited availability of the accelerometer data we use an on-body sensor dataset, Opportunity, [2] for our experiments. This dataset [2] is specifically used for stacked RBM training and is composed of readings from the motion sensors which are recorded while users execute typical daily life activities. We isolate the accelerometer data from sensors attached (only) on wrists to train RBM networks.

For the recognition experiments, we select sign language representation of 12 phrases (mostly questions) from ASL– 'Do you go to school?', 'Hey!! What is your name?', 'How many brothers do you have?', 'How many sisters do you have?', 'Is this her/his?', 'Is your dad deaf?', 'Are you a student?', 'Where do you work?', 'Where do you live?', 'We wish you a merry Christmas', 'Do you feel angry?' and 'Do you like your work?'. In all we create a dataset comprising of $50 \times 12 = 600$ accelerometer readings for our experiments. All the selected phrases are predominantly performed using single hand – the phrases are carefully sampled to avoid the complexities involved with the sensor data fusion because to gestures performed using multiple hands. The system is trained with equal probabilities of all the phrases and hence isn't biased to a particular phrase. The experiments are intended to illustrate the utility of the unsupervised features and recurrent neural networks to capture the structure of the accelerometer signal. Youtube videos of the selected phrases were shown to the participants who were then asked to enact the same wearing the mobile device on the wrist. Participants were given time to familiarize themselves with the gestures. The dataset is available online: `www.goo.gl/p9SQjx`

4 Results

For every accelerometer reading we compute the features using the sliding windows of length 30 with an overlap of 75%. The size of the sliding window was empirically selected by computing the validation error. We tried window size of $256, 128, 50, 30, 20$. In our case, window size of 30 performed best and returned minimum error. For a window of 30 frames we have 90 signal points (30 for each x,y,z axis of accelerometer) which are projected to 34 dimension space. The feature dimension for every window is 34 and is computed by the stacked RBM networks as described above. These features are fed into BLSTM network to produce the desired phrase classification. For all the experiments, we use BLSTM size of 10 and number of hidden layers is set at 2. We compared the accuracy of our approach with the handcrafted features. The experiments are baselined by the following two methods:

1. ***Handcrafted features***: Time series signal features for each sliding window are computed [7] capturing various signal properties – Kurtosis, Skewness, Sum of values over a period of time, Signal Power, Log Energy, Average Resultant Acceleration, Energy, Average Absolute Acceleration, Standard Deviation, Coefficients of Variation, Correlation, Mean Acceleration. The feature vector thus computed is of dimension 34, this justifies the reason we choose 34 as our output dimension size for RBMs. The feature set so obtained is stacked along with other other features till full signal is traversed. We performed Bag of Word (BOW) based classification with bag size of 30. Number of clusters (i.e. bags) were empirically selected with experiments being performed over $75, 50, 30, 15$ and 12 bags. The results obtained are compared with the standalone raw features used for classification, Table 2.

2. **RBM *features with* BLSTM**: We obtain a classification accuracy of 0.47 with single layered RBM and 0.53 with stacked RBM. We record a 13% higher accuracy than best performing handcrafted features, Table 2.

The unsupervised RBM-based feature learning is generic and better at capturing signal variations when compared to the hand-crafted features. Stacked RBMs are feature learners which disentangle hidden factors of variations and experimentally perform best among all the baseline methods. The results from best performing handcrafted features [7] are reported for comparison purpose. The framework is segmentation free and does not depend on the hand engineered features, which achieves good (if not better) results than other comparative methods. Our proposed solution is fairly adaptable to the change in sensors which may be sensitive to other human signal modalities.

5 Conclusion

We proposed a framework for 'phrase recognition' from accelerometer data captured from a wearable device. The learned feature representation are obtained by stacked RBM in an unsupervised manner and a sequence classifier, BLSTM, recognizes the desired phrase. Data from much more advance (biological, EEG etc.) sensors capturing better muscular movements would be best suited for our work extension and we expect the system would give better results in such case. We aspire to scale the proposed solution to use sensor data from both hands in future extensions.

References

[1] P. Casale, O. Pujol, and P. Radeva. Human activity recognition from accelerometer data using a wearable device. In *Pattern Recognition and Image Analysis*, 2011.

[2] R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Tröster, J. D. R. Millán, and D. Roggen. The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. In *Pattern Recognition Letters*, 2013.

[3] A. Graves, S. Fernndez, and F. Gomez. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine Learning*, 2006.

[4] G. E. Hinton. Training products of experts by minimizing contrastive divergence. In *Neural Computation*, 2002.

[5] J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Activity recognition using cell phone accelerometers. In *SigKDD Exp. Newsletter*, 2011.

[6] T. Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *International Conference on Machine Learning*, 2008.

[7] Y. Zheng, W. Wong, X. Guan, and S. Trost. Physical activity recognition from accelerometer data using a multi-scale ensemble method. In *Innovative Applications of Artificial Intelligence*, 2013.