15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Real-Time Recognition of Sign Language Gestures and Air-Writing using Leap Motion

Pradeep Kumar and Rajkumar Saini
Deptt of CSE, IIT Roorkee, India
{ pra14,rajkr}.dcs2014@iitr.ac.in

Santosh Kumar Behera and Debi Prosad Dogra
Deptt of Electrical Sciences,
IIT Bhubaneswar, India
{sb29,dpdogra}@iitbbs.ac.in

Partha Pratim Roy
Deptt of CSE, IIT Roorkee, India
proy.fcs@iitr.ac.in

Abstract

A sign language is generally composed of three main parts, namely manual signs that are gestures made by hand or fingers movements, non-manual signs such as facial expressions or body postures, and finger-spelling where words are spelt out using gestures by the signers to convey the meaning. In literature, researchers have proposed various Sign Language Recognition (SLR) systems by focusing only one part of the sign language. However, combination of different parts has not been explored much. In this paper, we present a framework to recognize manual signs and finger spellings using Leap motion sensor. In the first phase, Support Vector Machine (SVM) classifier has been used to differentiate between manual and finger spelling gestures. Next, two BLSTM-NN classifiers are used for the recognition of manual signs and finger-spelling gestures using sequence-classification and sequence-transcription based approaches, respectively. A dataset of 2240 sign gestures consisting of 28 isolated manual signs and 28 finger-spelling words, has been recorded involving 10 users. We have obtained an overall accuracy of 63.57% in real-time recognition of sign gestures.

1 Introduction

Sign Language is a visual language that is used by hearing impaired peoples for communication. Sign language is composed of three features, namely manual signs, non-manual signs, and finger-spelling. Manual signs are the gestures that are represented by hand shapes, motions, and positions, whereas non-manual signs contain facial expression or body postures. They add syntactical information to the gestures that can be the part of a sign or modify the meaning of a manual sign. In finger-spelling gestures, different words are spelt out in local verbal language using fingers during communication [4].

Over the past couple of decades, a handful of SLR systems have been developed by various research groups [20]. However, majority of the existing systems work with one type of data, i.e. manual or non-manual or finger-spelling. Hence their applications are limited. It has been observed that, during gestural communication, the signer starts using finger-spelling to convey the meaning in verbal language. Moreover, finger-spelling is also used while expressing some specific words. Such a scenario is depicted in Figure 1. As shown in Figure 1(a), a signer first performs the sign gesture for the word '*your*' followed by a finger spelling as depicted in Figure 1(b). Therefore, real-time implementation of such a system needs special consideration. Yang et al. [19] have developed a SLR system to recognize manual and non-manual signs of American Sign Language (ASL). The authors have used a set of three video cameras to capture the upper body, side view, and frontal view of the face. Manual signs are discriminated using an hierarchical Conditional Random Field (CRF) and BoostMap embedding, whereas Active Appearance Model (AAM) has been used to extract facial features. To assist physically disabled persons the authors in [5,18] have proposed different rating and choice prediction frameworks using brain signals.

SLR systems have also been proposed by various researchers using different techniques including RGB camera [22], stereo-camera [7], sensor gloves [21], colored gloves [17], 3D depth sensors [20], etc. However, with the development of low cost depth sensing technology such Leap Motion or Microsoft Kinect, it is easy to detect and track hand and finger movements in real-time. These sensors are designed to sense 3D point cloud of the observed scene. Leap motion sensor is specifically designed to track hand and finger movements. The sensor is successfully used by researchers in developing various applications including sign gesture recognition, gaming, rehabilitation, word segmentation [1, 16], etc. Zafrulla et al. [20] have proposed a system suitable for deaf children using Kinect. The authors have used 3D joint position of human skeletal as features and fed them to HMM classifier for recognition. They have tested the system with 60 phrases of ASL with an accuracy of 74.83%. In [15], the authors have investigated the use of Leap motion sensor for recognition of sign gestures. The authors have used 3D finger points as features and Artificial Neural Network (ANN) for the recognition of 26 Australian Sign Language alphabets. In [3], the authors have proposed a system for the recognition of 26 alphabets of ASL using Leap motion. The authors have used velocity, palm normal, and pitch strength as features to recognize sign gestures with the help of k-NN and SVM classifiers. Accuracies of 72.78% and 79.83% have been recorded using k-NN and SVM, respectively.

In this paper, we present a real-time framework for manual and finger-spelled gestures using Leap motion sensor. The framework first discriminates between manual and finger-spelling gestures with the help of a SVM classifier. Next, the recognition of discriminated gestures are done using two separate Bidi-

© 2017 MVA Organization

(a) (b)

Figure 1. A user is performs different gestures: (a) manual sign gesture for the word '*your*' (b) finger-spelled Latin language word 'off'.

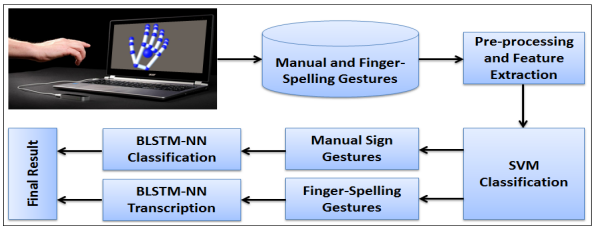


Figure 2. Flow diagram of the proposed framework.

rectional Long Short-Term Memory Neural Networks (BLSTM-NN). One BLSTM-NN has been trained using sequence classification mode, whereas the other one has been trained for sequence transcription for the recognition of finger spelled gestures.

Rest of the paper is organized as follows. Details of the proposed system, pre-processing, feature extraction, and classification are discussed in Section 2. Experiment results are presented in Section 3. Finally, we conclude in Section 4.

2 Proposed System

Leap motion comes with the associated Application Programming Interface (API) that provides an easy access to capture the 3D position of the fingertips with a sampling rate of 120 fps. We have kept the device below the arm of the signer for uninterrupted capturing. We do not put any restriction on gestures type i.e. the signer can perform any type of gesture be it in sign language form or finger-spelled form. A flow diagram of the proposed framework is shown in Figure 2. Raw data captured through the API of the device are then preprocessed and relevant features are extracted. Manual and finger-spelled gestures are distinguished using SVM. Classified gestures are then recognized using two BLSTM-NN classifiers based on sequence modeling and sequence transcription.

2.1 Preprocessing and Feature Extraction

Normalization: Collected 3D raw data may vary in scaling factors. Thus, to make all trajectories of uniform size, a zero-mean (z-score) based normalization has been used [9, 14]. The scheme normalizes the data by computing the standard deviation and mean.

Fingertip Positions: We have extracted instantaneous 3D fingertip positions using the API. Five features vectors (k_1 *to* k_5) have been extracted, where each vector represents a 3D sequence. The fingertip positions (F) for all five fingers of hand can be defined using (1).

$$F = \{k_1, k_2, k_3, k_4, k_5\} \tag{1}$$

Angular Direction: Angular direction (D) features are widely used in various SLR applications and handwriting recognition systems [12]. Angular direction of a 3D point Q is estimated w.r.t. two neighboring points on its either side. To estimate this, we have used the approach of authors defined in [11].

In this work, three angles are considered as angular features. Hence the feature vector D has been composed of five 3D angular feature vectors as given in (2), where each d_k represents a 3D angular feature sequence corresponding to k^{th} fingertip.

$$D = \{d_1, d_2, d_3, d_4, d_5\} \tag{2}$$

2.2 Initial Classification Using SVM

SVM is a kernel based classifier [2, 8, 13]. The basic idea is to map the input data into a high dimensional feature space, where the data can be linearly separable. SVM has the ability to perform both linear and non-linear classification using different kernel functions. For some training data $\{x_i, y_i\}$, where $i = 1, 2, ...n$ and $y_i \in \{-1, 1\}$, two possible constraints as depicted in (3-4) are defined, where w and b denote the hyperplane parameters and offset.

$$wx_i + b \geq +1 \quad for \quad y_i = +1 \tag{3}$$

$$wx_i + b \leq -1 \quad for \quad y_i = -1 \tag{4}$$

It works by finding a decision boundary called margin that maximizes the distance between two hyperplanes. Thus, finding such decision boundary is an optimization problem to minimize $||w||^2$.

In this work, we have used the SVM classifier to distinguish the sign language gestures into two class, i.e. either manual signs or finger-spelling gestures. Since, SVM is a non-temporal classifier, three statistical features have been computed for each dimension of the feature vector T, namely, Mean (M), Standard Deviation (SD) and Root Mean Square (RMS). Using this, a new feature vector of 90 dimension has been formed to train the classifier.

2.3 BLSTM-NN classifier based Gesture Recognition

BLSTM-NN is a sequence modeling classifier that has been popularly used in gesture and handwriting recognition problems [6]. The classifier is able to process the input sequences in both directions, i.e. forward as well as backward with the help of two hidden layers. Both the layers are connected to a common output layer. In this work, we have used the classifier for recognition of gestures that belongs to two different categories, namely manual gestures and finger-spelling gestures. Therefore, two different models have been trained using Cross Entropy Error (CEE) based objective function and Connectionist Temporal Classification (CTC) based objective function.

CEE based BLSTM-NN: The network has K output units, one for each class of the gesture sequence [10]. CEE for K classes are defined in (5),

$$CEE = -\sum_{(x,z)\epsilon T} \sum_{i=1}^{K} z_i \ln y_i \tag{5}$$

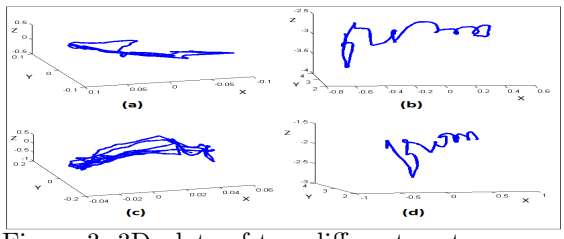


Figure 3. 3D plots of two different gestures performed by two users (column-wise): (a) & (c) sign word '*welcome*' represented in sign language (b) & (d) finger-spelling gestures corresponding to the word '*from*'.

where (x, z) is the input pair with x as the input sequence and z as the target sequence from the training set T. The term y defines the probability such that the input belongs to a particular class.

CTC based BLSTM-NN: It is defined as the negative log probability of correct labelling of the entire training set [11]. The objective function (O) can be computed using (6) for the training set T, where (x, z) represents a pair of input and target sequence.

$$O = -ln(\prod_{(x,z)\epsilon T} p(z|x)) = -\sum_{(x,z)\epsilon T} ln(p(z|x)). \quad (6)$$

O models the label sequence with the given inputs directly.

3 Results

First, we describe the dataset used in our study. The results have been computed using the 4-folds cross validation by dividing the complete dataset into four equal parts, out of which three sets have been used in training and the rest during testing.

3.1 Data Collection

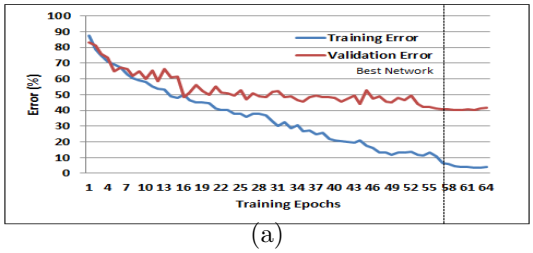
Data collection has been performed involving 10 users. We have considered a total of 56 gestures (28 single-handed-isolated sign gestures of Indian Sign Language (ISL) and 28 different words of Latin language for finger-spelling). All gestures have been performed 4 times by every user using the right hand. Therefore, a total of 2240 gesture have been recorded. Variations between the different gestures can be seen in the Figure 3. Figure 3(a) depicts the 3D plot corresponding to the sign word '*welcome*' performed by two different users, whereas Figure 3(b) represents the finger-spelled gestures for the word '*from*' written by the same two signers.

3.2 Gesture-Type Recognition using SVM

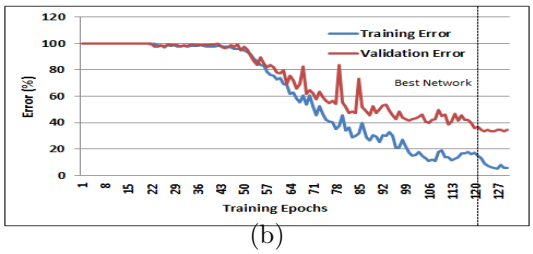
SVM classifier has been trained using a linear kernel. The classification has been performed by varying the regularization parameter (C) from 1 to 100. An accuracy of 100% has been recorded in classification of gestures of manual and finger-spelled type with C=23.

3.3 Gesture Recognition using BLSTM-NN

Two BLSTM-NN classifiers have been trained using the outputs of initial SVM classification with the help of feature vector T. For recognition of manual signs, the network has been trained with maximum error entropy approach with a learning rate of $1e-4$ and a momentum of 0.9. The learning curve of the network is shown in Figure 4(a) that depicts the decrement of training and validation errors during various epochs. It can be verified from the learning curve that, after 57 epochs, there is no change in the validation network. Thus, it has been marked as the Best-Network. An accuracy of 60.35% has been recorded in recognition of all manual sign gestures. Recognition performacne has also been noted against each class of gestures as shown in Figure 5(a), where accuracies vary between of 40% to 100% for different classes.



(a)



(b)

Figure 4. Learning curve of BLSTM-NN for showing variation in training and validation errors: (a) Manual signed gestures (b) Finger-spelled gestures.

The learning curve of CTC network for transcription is shown in Figure 4(b). After 120 number of epochs, the network has been found to be configured as Best-Network because no change in the validation network can be seen. An accuracy of 66.78% has been recorded in recognition of all finger-spelling gestures. Recognition performance has also been recorded for every individual class of gestures as depicted in Figure 5(b), where maximum accuracy of 100% has been recorded for the Latin word '*the*'. It may be noticed from the figure that, lower accuracies are usually recorded for larger words in comparison to smaller words.

A comparison between the recognition rates of manual and finger-spelled gestures along with overall accuracy of the system has been presented in Figure 6, where finger-spelling gestures outperform manual gestures. The overall recognition accuracy of the system has been found to be 63.57% for both types of gestures. The results have been computed using Intel Core i7 processor with 8 GB RAM on Microsoft Windows 7 platform where it took 0.32 second approximately to recognize a test gesture. Hence, the system can be considered as a real-time gesture recognition system.

4 Conclusion

In this paper, we have proposed a new SLR framework for recognition of manual signs and finger-spelling

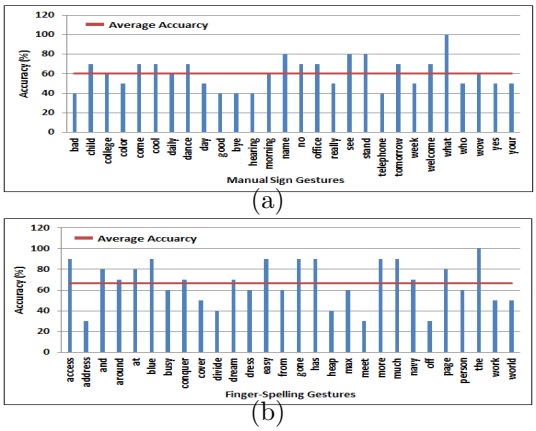


Figure 5. Gesture recognition performance for each gesture class: (a) Manual signed gestures (b) Finger-spelled gestures.

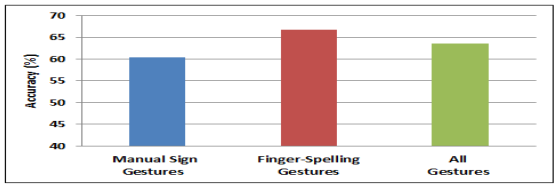


Figure 6. Comparative performance analysis between recognition rates of manual and finger-spelling gestures along with complete system performance.

gestures using Leap motion sensor. The framework facilitates a signer to communicate using modalities in real-time, i.e. manual and finger-spelling. The recognition process has been done in two stages. Firstly, SVM classifier has been used to distinguish input gestures into two classes corresponding to manual and finger-spelling. In the second stage, two BLSTM-NN classifiers have been trained for recognition of distinguished gestures using sequence classification and sequence transcription based approaches. A dataset of 2240 gestures has been prepared using the proposed framework. An accuracy of 100% has been recorded using SVM classifier. An overall accuracy of 63.57% has been recorded by our system for both types of gesture classes. The accuracy of the system is low because it has been tested with a lexicon free approach. Thus, in future, it can be improved by adding lexicon information to the BLSTM. Moreover, other sequential classifiers and their combinations can be tried to enhance the recognition accuracy.

References

[1] C. Agarwal, D. P. Dogra, R. Saini, and P. P. Roy. Segmentation and recognition of text written in 3d using leap motion interface. In *ACPR*, pages 539–543, 2015.

[2] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine learning*, 46(1-3):131–159, 2002.

[3] C.-H. Chuan, E. Regina, and C. Guardino. American sign language recognition using leap motion sensor. In *ICMLA*, pages 541–544, 2014.

[4] H. Cooper, B. Holt, and R. Bowden. Sign language recognition. In *Visual Analysis of Humans*, pages 539–562. 2011.

[5] H. Gauba, P. Kumar, P. P. Roy, P. Singh, D. P. Dogra, and B. Raman. Prediction of advertisement preference by fusing eeg response and sentiment analysis. *Neural Networks*, 2017.

[6] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE T-PAMI*, 31(5):855–868, 2009.

[7] R. Grzeszcuk, G. Bradski, M. H. Chu, and J.-Y. Bouguet. Stereo based gesture recognition invariant to 3d pose and lighting. In *CVPR*, volume 1, pages 826–833, 2000.

[8] B. Kaur, D. Singh, and P. P. Roy. A novel framework of eeg-based user identification by analyzing music-listening behavior. *MTAP*, pages 1–22.

[9] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra. Coupled hmm-based multi-sensor data fusion for sign language recognition. *PRL*, 2016.

[10] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra. A multimodal framework for sensor based sign language recognition. *Neurocomputing*, 2017.

[11] P. KUMAR, R. Saini, P. Roy, and D. Dogra. Study of text segmentation and recognition using leap motion sensor. *IEEE Sensors Journal*, 2016.

[12] P. Kumar, R. Saini, P. P. Roy, and D. P. Dogra. 3d text segmentation and recognition using leap motion. *MTAP*, pages 1–20, 2016.

[13] P. Kumar, R. Saini, P. P. Roy, and D. P. Dogra. A bio-signal based framework to secure mobile devices. *JNCA*, 2017.

[14] O. Patsadu, C. Nukoolkit, and B. Watanapa. Human gesture recognition using kinect camera. In *Conference on Computer Science and Software Engineering*, pages 28–32, 2012.

[15] L. E. Potter, J. Araullo, and L. Carter. The leap motion controller: a view on sign language. In *25th Australian computer-human interaction conference: augmentation, application, innovation, collaboration*, pages 175–178, 2013.

[16] K. Vamsikrishna, D. P. Dogra, and M. S. Desarkar. Computer-vision-assisted palm rehabilitation with supervised learning. *IEEE Trans. on Bio. Engg.*, 63(5):991–1001, 2016.

[17] R. Y. Wang and J. Popović. Real-time hand-tracking with a color glove. In *ACM TOG*, volume 28, page 63, 2009.

[18] M. Yadava, P. Kumar, R. Saini, P. P. Roy, and D. P. Dogra. Analysis of EEG signals and its application to neuromarketing. *MTAP*, 2017.

[19] H.-D. Yang and S.-W. Lee. Robust sign language recognition by combining manual and non-manual features based on conditional random field and support vector machine. *PRL*, 34(16):2051–2056, 2013.

[20] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti. American sign language recognition with the kinect. In *13th inte. conf. on multimodal interfaces*, pages 279–286, 2011.

[21] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang, and J. Yang. A framework for hand gesture recognition based on accelerometer and emg sensors. *IEEE Trans. on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(6):1064–1076, 2011.

[22] J. Zieren and K.-F. Kraiss. Robust person-independent visual sign language recognition. In *Iberian Conf. on Pattern Recognition and Image Analysis*, pages 520–528, 2005.