Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-020-01979-z

ORIGINAL RESEARCH

Trbaggboost: an ensemble-based transfer learning method applied to Indian Sign Language recognition

S. Sharma[1] · R. Gupta[1] · A. Kumar[2]

Received: 13 December 2019 / Accepted: 9 April 2020

Abstract
An efficient sign language recognition (SLR) system would help speech and hearing-impaired people to communicate with normal people. This work aims to develop a SLR system for Indian sign language using data acquired from multichannel surface electromyogram, tri-axis accelerometers and tri-axis gyroscopes placed on both the forearms of signers. A novel ensemble-based transfer learning algorithm called *Trbaggboost* is proposed, which uses small amount of labeled data from a new subject along with labelled data from other subjects to train an ensemble of learners for predicting unlabeled data from the new subject. Conventional machine learning algorithms such as decision tree, support vector machine and random forest (RF) are used as base learners. The results for classification of signs using Trbaggboost are compared with commonly used transfer learning algorithms such as TrAdaboost, TrResampling, TrBagg, and simple bagging approach such as RF. Average accuracy for classification of signs performed by a new subject is achieved as 69.56% when RF is used without transfer learning. When just two observations of labeled data from a new subject are integrated with training data of an existing SLR system, average classification accuracy for TrAdaboost, TrResampling, TrBagg and RF are 71.07%, 72.92%, 76.10% and 76.79%, respectively. However, for the same number of labelled data from the new subject, Trbaggboost yields an average classification accuracy of 80.44%, indicating the effectiveness of the algorithm. Moreover, the classification accuracy for Trbaggboost improves up to 97.04% as the number of labelled data from the new user increase.

Keywords Ensemble learning · Indian Sign Language · Sign language recognition · Transfer learning

1 Introduction

A substantial number of speech and hearing-impaired people are using sign language as their primary language of communication all over the world (Suharjito et al. 2017). Sign language composes of manual hand gestures and non-manual components such as body language and facial expression. However, there exists a communication gap between sign language users and those who do not understand this language. Use of human interpreters is often costly and not readily accessible. Thus, there is a requirement of developing an efficient sign language recognition (SLR) system which can help the users of sign language to conveniently communicate with non-signers. In Suharjito et al. (2017) the authors reviewed around 20 research papers on SLR and state that continuous development and positive efforts has been made in this area of research. Broadly two sensing technologies are adopted in SLR systems namely, vision based and sensor based. In vision-based approaches, devices such as video camera, Microsoft Kinect and Leap Motion Controller (LMC) have been used to capture motion of hands and finger configuration (Suharjito et al. 2017). For instance, in Raheja et al. (2016), the authors used Microsoft Kinect to classify a sign from a set of 80 signs from the Indian sign language (ISL). In Chong and Lee (2018), the authors used LMC to recognize signs from 26 alphabets and 10 digits in American sign language (ASL) with up to 88.79% accuracy. Vision based approaches often require image segmentation based on skin colour for isolating hand shape or

✉ S. Sharma
sjangid@amity.edu

R. Gupta
rgupta3@amity.edu

A. Kumar
arunkm@care.iitd.ac.in

[1] Electronics and Communication Engineering Department, Amity University Uttar Pradesh, Noida, India

[2] Centre for Applied Research in Electronics, Indian Institute of Technology Delhi, New Delhi, India

Published online: 27 May 2020

tracking hand motion, which makes them difficult to carry out under poor brightness, varying camera depth and view angles, and complex background color and texture (Suharjito et al. 2017). Moreover, wearability of a vision-based system is challenging. With advancements in the field of micro-electromechanical system (MEMS), wearable sensors have gained much attention in recent years. Sensors such as accelerometer, gyroscope, magnetometer, flex and tactile have been reported for the development of SLR systems (Ahmed et al. 2018; Wu et al. 2016). Surface electromyogram (sEMG) sensors have also been used with motion sensors for SLR (Wu et al. 2016; Yang et al. 2016). These sensors may be placed in a glove or on the wrist and forearm region, on one or both hands. Certain unconventional sensors have also been shown to yield high classification accuracies for SLR, such as epidermal-iontronic sensing (Zhu et al. 2018), which is a wearable pressure sensor and WiFi signals (Ma et al. 2018). Sensor-based systems enhance wearability and mobility of SLR system. However, displacement of sensors, perspiration, varying muscular structures of users may affect the performance of the system.

Although SLR has been actively researched for more than a decade, a commercially viable product that may translate sign language to a verbal language is yet to be developed. In recent years, issues required to be addressed for developing a practical SLR system are increasingly being reported. In Wu et al. (2016), the authors achieved an overall accuracy of 96.16% with support vector machine (SVM) classifier for classification of 80 signs from the ASL using an inertial measurement unit (IMU) and four sEMG sensors. This performance was achieved under intra-subject testing when the labeled data used for training the classification model and the unlabeled data used for testing the performance of the model were recorded from a subject in the same session, ensuring no change in placement of sensors. When more unlabeled data recorded in a different session from the same subject was tested, the overall accuracy decreased to 85.24% (Wu et al. 2016). Moreover, for inter-subject testing, that is when the classification model was learned with data recorded from 3 subjects and tested using unlabeled data from a 4th subject new, the overall classification accuracy degraded significantly up to 40%. The authors suggested developing subject-specific SLR system. In Zhu et al. (2018), the authors applied SVM, random forest (RF) and neural network (NN) for the recognition of 26 alphabets and 10 digits from the ASL. The highest intra-subject classification accuracy of 95.2% was achieved with RF, which is an ensemble-based learning method. However, the overall classification accuracy reduced to 65.2% for inter-subject testing. In Yang et al. (2016), the authors reported an optimized tree-structure for classification of 150 subwords from the Chinese sign language using data from one IMU and four sEMG sensors worn on both the hands, on wrist and forearm, respectively. The signs could be classified with overall accuracy of 94.315% under subject-specific testing. However, under subject-independent testing when data from 7 subjects was used for training the classification model and that of an 8th new subject was used for prediction, overall classification accuracy reduced to 87.02%. Inter-subject variability is also known to affect hand gesture recognition in sEMG based prosthetic control devices (Kyranou et al. 2018). Covariate shift, for instance, arising due to shift in sensor placement, varying arm position and additional weight may also cause the test data to change over time (Kyranou et al. 2018).

In general, when test data consists of similar instances as in training data, the similar distributions of training and test data result in relatively high classification accuracies. However, for different distributions, where completely new instances occur in test data which were unseen to the training data, the model performance degrades. Consequently, for practical application like SLR system a robust model adaptive to new subject data is important for applicability to the end users. In order to address this issue, one possible solution is to retrain the model from the scratch with target data collected from the new subject. However, it is often impractical and very expensive, because large amount of labelled data is required for training and labelling consumes lot of effort and time. An alternative is to use transfer learning, wherein model learned from a previous dataset (source data) may be adapted to enable classification on a new dataset (target data) (Ali et al. 2019).

In this paper, an ensemble-based transfer learning approach is proposed to enable classification of signs performed by a new subject when a substantial amount of labelled data from other subjects (source data) and only limited amount of labelled data from the new subject (target data) are available. The algorithm is tested on signals recorded using multiple tri-axis accelerometers, tri-axis gyroscopes and sEMG sensors placed on both hands of different signers, while they perform multiple repetitions of 100 commonly used signs in the Indian sign language (ISL). These sensors are placed on the forearm region to make the SLR system easy to wear and mobile. The main contributions of this work are as follows:

1. A novel transfer learning algorithm is proposed named Trbaggboost (transfer bagging and boosting), which combines the ability of two most prominently used ensemble methods, namely bagging and boosting.
2. Application of proposed ensemble-based transfer learning for classification of 100 commonly used isolated signs from ISL, using data from multiple wearable sensors (sEMG, tri-axis accelerometer and tri-axis gyroscope) placed on both the forearms of a signer. Such a

work is being reported for ISL for the first time, to the best of our knowledge.
3. Performance of Trbaggboost with different base learner is compared in terms of classification accuracies for different percentage of labelled data from the new subject as compared to source data available from other subjects.
4. Analysis of minimum amount of labelled data required from a new subject for generating a transfer learning model with acceptable performance.

The remaining paper is organized as follows. Section 2 contains review of algorithms available in literature for transfer learning. The design and implementation of the novel ensemble-based transfer learning approach for the classification of signs performed by a new subject is presented in Sect. 3. Section 4 contains the experimental results for classification of isolated signs from ISL under different scenario. The performance of the proposed algorithm is also compared with simple bagging approach and ensemble transfer learning algorithms namely TrAdaboost (Dai et al. 2007), TrResampling (Liu et al. 2017) and TrBagg (Kamishima et al. 2009). Conclusions are stated in Sect. 5.

2 Related work

2.1 Transfer learning

In transfer learning, the knowledge learned from a previous classification task is used to leverage classification of a different, but related task. Transfer learning algorithms may be categorized based on whether labelled, unlabeled or no data is available from the target domain (Ali et al. 2019). Having a small amount of labelled data from the target domain has shown to significantly improve transferability (Chiang et al. 2017). For instance, in Chiang et al. (2017), the authors proposed a feature-based transfer learning approach to enable classification of human activities with a change in sensors deployed in source and target domain. It is reported that the transfer learning approach outperforms nontransfer-learning models when labelled data from target domain is also available. Moreover, inclusion of ensemble learning in transfer learning algorithms improves stability and accuracy of the algorithms (Liu et al. 2017). Hence, these approaches are explored further in this work. Mainly there are two widely used ensemble learning methods namely, bagging and boosting. Bagging is a method in which number of weak learners are learned on different bootstrap samples using a base learner such as naive Bayes (NB), decision tree (DT) and SVM. The final prediction is determined by carrying out averaging or majority voting on the decision of the individual learners. For instance, in Kamishima et al. (2009), the authors proposed a transfer bagging (TrBagg) algorithm for collaborative tagging of data from two social bookmarking services. In TrBagg, first, the bootstrap samples of training data are created with source and target domain and a number of weak classifiers are learned. Then, weak classifiers that provide more accurate classification of target data are filtered out for taking decision through majority voting. Another bagging-based transfer learning algorithm is reported in Liu et al. (2016). Initially, classifiers are learned on bootstrap samples of source data combined with labelled target data. Then, the classifiers are used to predict labels of unlabeled target data, which are then added to the training data to update the classifiers. Labels of test data are predicted using majority voting on outputs of updated classifiers. The algorithm provided enhanced performance on several publicly available datasets.

In ensemble learning via boosting, the base learners are adapted by focussing more on those samples of the training data that were badly handled by the base learners. Adaptive boosting (Adaboost) is a widely used boosting algorithm in which many weak classifiers are learned on training data (Zhou 2015). Then, weights of misclassified samples from training data are increased, so that its probability to get selected as a training sample for next weak classifier increases. In Dai et al. (2007) and Yang et al. (2016) authors proposed TrAdaBoost method for transfer learning with Adaboost. The TrAdaBoost algorithm iteratively reweights the source data and calculates the error on target data. This will encourage the part of source data most likely to be useful for target data classification, to be used for learning the models. Other than boosting by reweighting, boosting by resampling has also been used in transfer learning. Ensemble of classifiers may be learned on training data formed by resampling useful samples from source data and combining it with labelled target data. For instance, in Liu et al. (2017), TrResampling is presented. In each iteration of TrResampling, weighted resampled source data is used along with labeled target data to form the training data. Weighted resampling gives more emphasis to the source data assigned with higher weights. Afterwards the TrAdaboost algorithm is applied to adjust the impact of source data in developing a model. Transfer learning has been applied in various applications including SLR, as explained in the next subsection.

2.2 Transfer learning applications

Transfer learning has been applied in various fields such as health monitoring, speech signal analysis, face and object recognition, sentiment recognition, text classification and human activity recognition (Weiss et al. 2016). For instance, feature selection based transfer subspace learning was used to perform emotion recognition on one speech corpus by learning from another speech corpus (Song and Zheng 2018).

In Hu et al. (2018), a feature incremental RF is proposed when a data from a new sensor is to be included with a model trained on previously available data for activity classification. In Khan et al. (2017), *k*-mean clustering and transfer boost algorithms are used to allow the model to classify new activities, whose limited labeled data may be available, while sufficient amount of labelled data is available for other activities.

Transfer learning has also been applied to the application of sign language when substantial amount of target data for sign language classification is lacking. In Gattupalli et al. (2016) the authors trained a deep learning model on various publicly available video databases for human pose estimation, which they used for initialization in training a deep learning model on ASL database. The resulting models could better classify the position of various body parts such as head, hand, shoulder and elbow from a video sequence as compared to the model trained only using the sign language database. Similarly, when weights learned by training a deep learning architecture on an English speech dataset were used for initialization of deep learning model for classification of British sign language, the model performed better as compared to that learned on the sign language directly (Mocialov et al. 2018). In both the papers, the authors cited the small amount of labelled data in the sign language database as a limiting factor in the performance of the models trained on sign language databases (Gattupalli et al. 2016; Mocialov et al. 2018). In Farhadi et al. (2007), the authors used an animated ASL dictionary signed by an avatar as source data to classify the signs performed by human signers. The attained error rate for frontal view was 99.1% when the word models developed on avatar data were used for classification of signs performed by a human. However, when word models were also developed by using some amount of labeled data from humans, the error rate reduced to 35.8%, showing significant improvement in transferability. In this paper, a novel transfer learning algorithm is proposed that accounts from subject variability and enables classification of signs performed by a subject, when only limited amount of labelled data is available from him.

3 Ensemble based transfer learning system design and implementation

3.1 Database details

A multi-modal and multi-sensor database of signals was recorded for a set of 100 commonly used signs from ISL. These signs were performed according to the ISL dictionary published by Ramakrishna Mission Vivekananda University (Indian Sign Language (ISL) dictionary 2015). Some examples of signs from ISL are shown in Fig. 1 with their description in Table 1. The considered signs consist of postures and motions performed with either dominant hand or with both dominant and non-dominant hands. Ten healthy subjects, 7 females and 3 males, participated voluntarily in this study, out of which 3 were left-hand dominant. Each subject was informed about the objective and process in which the experiment for data acquisition would be carried out. A prior training about performing the sign was also given to make the subject familiar with the experiment. Each subject performed 20 repetitions of each sign, with 5 s of rest between each repetition to avoid muscle fatigue. The signing duration varies between 2 and 4 s for each sign. The subjects were instructed to maintain moderate force level while performing a sign. A total of 20,000 samples of observations, containing around 15 h of usable recording were collected.

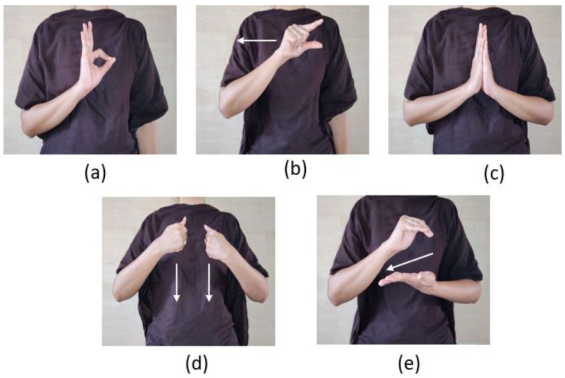

Fig. 1 Examples of signs from ISL

Figure 2 shows the complete experimental setup for data acquisition. Signals were recorded using six Delsys wireless sensors, each of which consists of one sEMG sensor and one IMU containing a tri-axis accelerometer and a tri-axis gyroscope. Signals from sEMG sensors were recorded with a sampling period of 900 μs and those from accelerometer and gyroscopes at 6.7 ms and 16 bits per sample. The sensors wirelessly transmit the signals to a base station, which sends the signals to a computer, where the data is stored and processed. Followed by the basic skin preparation, the sensors were placed on the skin surface using adhesive interface to minimize motion artifacts. As shown in Fig. 2, three sensors were placed on the dominant hand and remaining three were placed on the non-dominant hand in the configuration of a forearm armband.

3.2 Pre-processing and feature extraction

Raw signals of accelerometers and gyroscopes of each sensor were calibrated by removing bias and multiplying it with scale factor. Then, interpolation was performed on calibrated signals to estimate missing sample values. For removing baseline from sEMG signals, the mean of

Table 1 Description of examples of selected signs

Example	Sign	Description
Good	1(a)	Place right **zero** hand, facing out, in front of the chest
Name	1(b)	Place right **U** hand, facing and pointing left, at chest level, move slightly to right
Pray	1(c)	Place both "**Open**" hands, palms touching each other, in front of chest
Chair	1(d)	Place both "**Fist**", facing each other, at chest level and move down slightly
Bank	1(e)	Place left "**Open**" hand, facing up, pointing out, at chest level, right "**Bent**" hand, pointing left, on the left hand and move forward

Basic hand postures given in bold

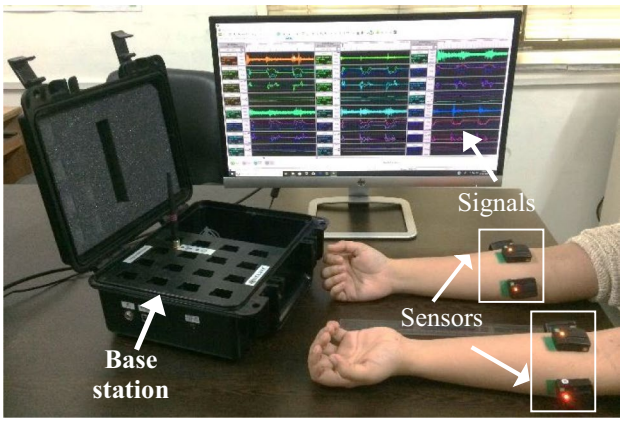

Fig. 2 Experimental setup

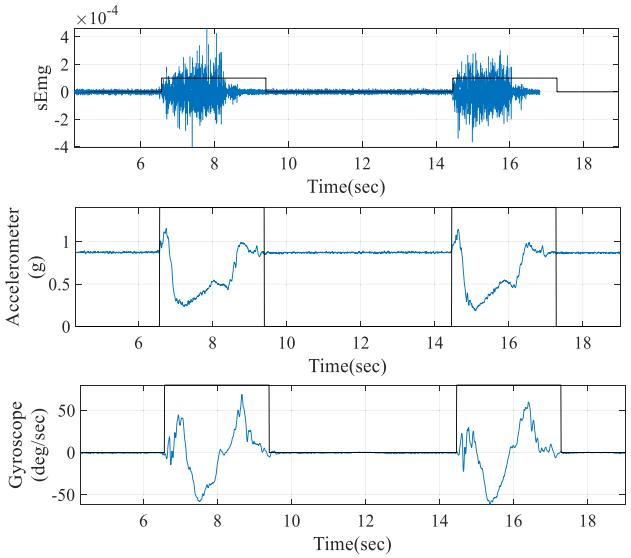

Fig. 3 Plots of signals for the sign "Name"

signal recorded during rest duration was calculated and subtracted. Other than baseline removal, sEMG signals was also interpolated as a part of pre-processing steps. Next, the activity duration of each repetition of the sign was detected where the mean and standard deviation of the accelerometer signals change show significant changes in time.

Figure 3 shows two repetitions of pre-processed sEMG, accelerometer and gyroscope signals from one sensor on dominant hand when the signer signs the word "Name", with rest durations on either side. Not much variation in the signal is observed during the rest positions. Whereas, amplitude of signals recorded by all the sensors varies during signing. Hence, the start and end point of each repetitions were determined by detecting the change in accelerometer signals. The activity duration detected for the dominant hand is plotted in solid black line in Fig. 3. Further processing was carried out within the detected activity duration. Within the detected activity duration of the isolated sign, the signals were divided into four windows and features were extracted from each window. Since, sEMG signals are non-stationary random signals consequently they are well studied with statistical features. For sEMG signals, the time-domain and frequency domain features evaluated for each signal window are described below, where x_i represents the sample of sEMG signal in the ith sample and N corresponds to the total number of samples in a signal window.

(1) *Variance (VAR)*: it defines the average power of a sEMG signal (Xi et al. 2017), calculated as,

$$VAR = \frac{1}{N-1}\sum_{i=1}^{N} x_i^2. \tag{1}$$

(2) *Mean absolute value (MAV)*: it is defined as the average of absolute value from sEMG signal amplitude (Phinyomark et al. 2012), defined as,

$$MAV = \frac{1}{N}\sum_{i=1}^{N} |x_i|. \tag{2}$$

(3) *Waveform length (WL)*: a cumulative length of a sEMG signal over the time is known as Waveform length (Phinyomark et al. 2012), calculated as,

$$WL = \sum_{i=1}^{N} |x_{i+1} - x_i|. \tag{3}$$

(4) *Zero crossing rate (ZCR)*: it is a measure of number of times the sEMG signal amplitude crosses zero amplitude value (Phinyomark et al. 2012), given as,

$$ZCR = \sum_{i=1}^{N-1} [sgn(x_i x_{i+1}) \cap |x_i - x_{i+1}| \geq th], \quad (4)$$

where sgn() corresponds to the sign associated with the quantity in the bracket. The threshold *th* was calculated as two times the standard deviation of the signal recorded during rest period.

(5) *Slope sign change (SSC)*: it measures number of times the sign of the slope changes for sEMG signals (Phinyomark et al. 2012).

$$SSC = \sum_{i=2}^{N-1} \left[f\left[(x_i - x_{i-1}) \times (x_i - x_{i+1}) \right] \right], \quad (5)$$

where function $f(x)$=1, when x > *th*, otherwise it is set to zero. The threshold *th* is taken same as that considered in the evaluation of *ZCR*.

(6) *Standard deviation (STD)-temporal and spectral*: it is defined as the amount of variation observed from the mean value. In temporal domain, STD is defined as,

$$STD_t = \left(\frac{1}{N} \sum_{i=1}^{N} (x_i - \eta)^2 \right)^{\frac{1}{2}}, \quad (6)$$

where $\eta = \frac{1}{N} \sum_{i=1}^{N} x_i$ is the mean of the signal segment. Spectral STD is evaluated using the spectral moments (Phinyomark et al. 2012) defined using normalized power spectral density (PSD) of the signal, as,

$$STD_f = \left(\sum_{i=1}^{M} f_i^2 P_i - \left(\sum_{i=1}^{M} f_i P_i \right)^2 \right)^{\frac{1}{2}}, \quad (7)$$

where P_i is the PSD at frequency f_i and $i = 1, \ldots, M$, where M is the number of frequency bins.

(7) *Skewness- temporal and spectral*: it is a measure of symmetry in the shape of distribution. Temporal skewness is defined as,

$$skew_t = \mu_3 / \mu_2^{3/2}, \quad (8)$$

where μ_n is the nth-order central moment evaluated using the probability density function (PDF) of the signal segment. Spectral skewness is defined same as in (8), however the central moments are evaluated using the PSD of the signal (Sharma et al. 2019).

(8) *Kurtosis- temporal and spectral*: it is used to quantify the flatness of signal distribution. Temporal kurtosis is evaluated as

$$kurt_t = \mu_4 / \mu_2^2, \quad (9)$$

where μ_2 and μ_4 are evaluated using PDF of the signal. Spectral kurtosis is also evaluated as given in (9), however, PSD of the signal is used to determine the central moments (Sharma et al. 2019). Spectral skewness and kurtosis have been shown to be effective in classifying finger gestures from forearm sEMG signals.

(9) *Mean frequency (MNF):* IT is defined as the sum of product of PSD of sEMG signal P_i and frequency f_i, divided by total sum of signal spectrum density, given as (Phinyomark et al. 2012).

$$MNF = \sum_{i=1}^{M} f_i P_i / \sum_{i=1}^{M} P_i, \quad (10)$$

where M is the total number of frequency bins.

(10) *Median frequency (MDF)*: it is the frequency where the spectrum divides in two regions having equal amplitude (Phinyomark et al. 2012), such that

$$\sum_{i=1}^{MDF} P_i = \sum_{i=MDF}^{M} P_i = \frac{1}{2} \sum_{i=1}^{M} P_i. \quad (11)$$

where P is the PSD of the signal segment and M is the total number of frequency bins.

For each channel of tri-axes accelerometer and gyroscope, the mean and standard deviation were evaluated for each window. The features from accelerometer and gyroscope are well known to determine position and orientation respectively (Wu et al. 2016). Next, the extracted features were normalized and principal component analysis (PCA) was applied to reduce the dimension of feature vector while 99% variance is retained. PCA also reduces the complexity of the learning model by preserving only relevant features for training. The proposed transfer learning algorithm for classification of signs is explained in the following subsection.

3.3 Trbaggboost algorithm

Let us consider that sufficient amount of labelled data from multiple subjects is available as source data for training a model to classify the set of 100 signs from ISL. It is shown in the next section that the model can classify signs performed by the known subjects with high accuracy. However, the classification accuracy degrades significantly when tested on signs performed by a new subject (target data), whose labelled data was not used for training the model. This is due to the difference in the characteristics of source and target data. For training a model using only the data from the new subject, as suggested in Wu et al. (2016), will require a large amount of labelled data from the new subject. To reduce the cost and effort for labelling new subject data, only a few labelled samples of each activity performed by the new subject, called as auxiliary

data (D_{Aux}) are included with the source data (D_S) to train the model. To further enhance the classification performance, we propose *Trbaggboost* algorithm, which is a two-stage process for transferring the knowledge gained by source data to new subject data.

Trbaggboost consists of two stages, one utilizes the bagging concept while the other uses boosting. In the first stage, using the bagging approach, an ensemble of base learners is learned with K number of bootstrap samples of source data, D_S^k, $k = 1, 2, \ldots K$. Hence, each learner in stage 1 (L_k^1, $k = 1, 2, \ldots K$) learns only from a segment of source data, whereas prediction is made on the auxiliary data (D_{Aux}). The second stage is the boosting stage which is based on the concept of boosting by resampling (Seiffert et al. 2008). For each learner in stage 1, the misclassified samples of auxiliary data ($\overline{D}_{Aux}^k$, $k = 1, 2, \ldots K$) are appended with the corresponding bootstrap sample of source data for that learner (D_S^k). The appended data [$\overline{D}_{Aux}^k$, D_S^k] are used for training the learners of stage 2, L_k^2, $k = 1, 2, \ldots K$. In this way, the samples from the auxiliary data that are different from the source data samples and hence, misclassified are included in training data for L_k^2, giving them more weightage than those samples that were accurately classified by L_k^1. Finally, the target data D_T, is tested with each model trained in the boosting stage, L_k^2, $k = 1, 2, \ldots K$. PredictionsP_k, obtained from ensemble learners are combined using majority voting method to determine the labels of D_T. The steps in the proposed Trbaggboost algorithm are summarized in Algorithm 1.

Algorithm 1: **Trbaggboost**

Input- Labelled data D_S and D_{Aux}, unlabelled data D_T, a base learning algorithm and maximum number of bootstrap samples K.

Processing- Initialize empty vector of predictions on test data, $P = []$
For $k = 1, 2, \ldots, K$
1. Learn model L_k^1 with k^{th} bootstrap samples from source data D_S^k
2. Predict label for D_{Aux} using model L_k^1
3. Find $\overline{D}_{Aux}^k$, the misclassified samples of D_{Aux}
4. Learn model L_k^2 with [$\overline{D}_{Aux}^k$, D_s^k] as training data
5. Predict labels of D_T with model L_k^2 and save as P_k
6. Append P_k in vector P

End

Output- Use majority voting on P to determine the labels for D_T

The performance of the proposed algorithm is determined with different base learners, such as DT, SVM and RF. It may be noted that in ensemble learning, the base learners need not necessarily be weak. In fact, strong learners more often give better performance (Zhou 2015). The base learner with which the proposed algorithm yields best performance will be selected for further analysis. Results for selection of suitable base learner and performance of the proposed algorithm are given in the next section.

4 Results

For SLR, the signals recorded from multiple sEMG and IMU sensors on both hands for 100 ISL signs performed by 10 subjects were processed to extract features as described in Section III B. Three different testing scenarios are considered for SLR. First, the accuracy with which a model can be trained to classify the considered 100 ISL signs when the test data was derived from the same distribution as the training data was determined. The average classification accuracy achieved for 100 signs with five-fold cross validation for different classifiers is shown in Table 2. In DT, the nodes are split according to Gini index till there are at least 2 samples in a node. SVM classifier performed the best with radial basis kernel function and RF was initialized with 50 trees. SVM could classify the signs with highest average classification accuracy of 97.08% and standard deviation 0.01% over five-fold cross validation. In the second testing scenario, out of data from 10 subjects, data of 7 subjects were used for training the classifier and the remaining data from other 3 subjects were used for testing to determine the classification accuracy. Hence, the classification accuracies were determined for all 120 unique subject combinations for forming training and testing data. The average of classification accuracies over all subject combinations is given in second row of Table 2. The classification accuracies degrade for all classifiers by around 26%. Nevertheless, RF yielded the best performance under different distribution of training and testing data with average classification accuracy of 69.56% and standard deviation 2.70%.

The degradation in classification accuracies is due to the difference in distribution of training and testing data. To illustrate the similarity and difference between the training and testing data for the above two scenarios, receiver operating curves (ROC) are plotted for multiple data distributions in Fig. 4. Here, the training data are assigned label 1, irrespective of the sign performed and the subject who performed it. The testing data is assigned a label 0. The training and testing data are concatenated to form a single dataset, which is segmented into new training and testing data according to the aforementioned binary labels. A stratified five-fold split was used to ensure the preservation of each class and to cover entire data. An RF classifier was learned on the training data and its ROC was plotted. The

Table 2 Classification accuracies (average ± standard distribution) without transfer learning

	DT (%)	SVM (%)	RF (%)
Similar distribution	63.19 ± 0.03	97.08 ± 0.01	95.69 ± 0.01
Different distribution	39.49 ± 3.13	68.66 ± 3.93	69.56 ± 2.70

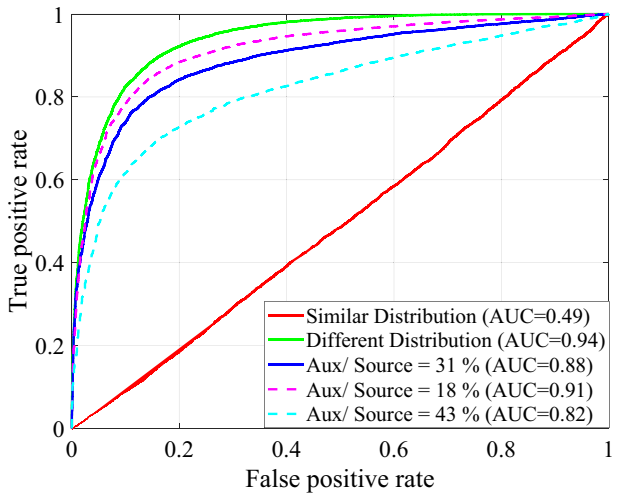

Fig. 4 ROC-AUC curve tested on different distribution sets

corresponding area under curve (AUC) indicates the level to which the classifier is able to identify the training and testing data as distinct. Such an approach for testing similarity and difference between the training and testing data is suggested in Bickel et al. (2009). As seen in Fig. 4, when the testing and training data are derived from same subjects as explained in the first testing scenario, the AUC is minimum (AUC = 0.49) indicating that the training and testing data have similar distribution. Whereas, when training data and test data are derived from different subjects as done in scenario 2, the AUC is maximum at 0.94, indicating that the test and training data have different distribution.

In the third testing scenario, small amount of auxiliary data from new subjects was included with the source data and used for training the classifier. The classifier was then tested with target data from new subjects to assess its performance. Under the third scenario, the classifier learned to identify difference between the training and the test data distributions had ROC in between the ROCs plotted for the first two scenarios (see Fig. 4). As the percentage of auxiliary data to the source data in the training data increases from 18 to 31% and 43%, the AUC reduces. This indicates that the difference between the testing and training data progressively decreases as more labeled data from new subjects is included in the training data. With reducing difference between the training and testing data, the SLR is expected to better classify signs performed by a new subject. Next, the performance of the proposed algorithm with different base learners is determined to select a base learner for further analysis.

4.1 Base learner for Trbaggboost

The performance of proposed algorithm evaluated with three base learners is given in Table 3. The base learners, DT, SVM and RF were designed same as stated earlier in this section. Moreover, the classification accuracies are evaluated when different percentages of auxiliary data to source data are used in training data. In Table 3, the percentage of auxiliary data to the source data is given in first column, along with the number of repetitions of each sign required to be performed by a new user for training the transfer learning model. The classification accuracies given in the remaining three columns are the mean of accuracies achieved for all possible subject combinations in target data. When just two repetitions of each sign are performed by the new user, the mean classification accuracies achieved using Trbaggboost with DT, SVM and RF as base learners is 55.27%, 80.51% and 80.44%, respectively. Hence, with 6% auxiliary data, the achieved classification accuracies are significantly higher as compared to those achieved without transfer learning, which are 39.49%, 68.66% and 69.56% for DT, SVM and RF, respectively, as given in Table 2. Also, as the percentage of auxiliary data to the source data increases, the classification accuracies of Trbaggboost with all the base learners also

Table 3 Classification accuracy of different base learners in Trbaggboost

Percentage of auxiliary data/source data (Repetitions by new user)	Trbaggboost DT	Trbaggboost SVM	Trbaggboost RF
6% (2)	55.27	**80.51**	80.44
9% (3)	61.56	**84.94**	**84.83**
12% (4)	64.62	86.85	**87.97**
15% (5)	67.82	87.46	**90.17**
18% (6)	71.04	89.57	**91.50**
24% (8)	76.88	90.00	**93.62**
31% (10)	79.93	91.13	**94.95**
43% (14)	83.11	91.38	**96.08**
55% (18)	85.00	91.66	**97.04**
Average	71.69	88.17	**90.73**

Best performance shown in bold

increases. When the percentage of auxiliary data to source data is 43%, the amount of labeled data from the new user is similar to that available in the first testing scenario. However, the mean classification accuracies achieved with Trbaggboost algorithm are higher than those achieved with only single stage of classification, whose results are reported in first row of Table 2. On an average, the classification accuracies of Trbaggboost with DT, SVM and RF as base learners were 71.69%, 88.17% and 90.73%, respectively. Since, RF outperforms DT and SVM as base learner in Trbaggboost, it is considered as a suitable base learner for the Trbaggboost algorithm.

4.2 Performance evaluation of Trbaggboost

The performance of the proposed algorithm is now compared with a baseline bagging approach, such as RF, and transfer learning algorithms, namely TrAdaboost, TrResampling and TrBagg. Out of 10 subject data, just like in second and third testing scenarios, data from 7 subjects is treated as labelled source data, while those of the remaining 3 subjects is further split into labelled auxiliary data and target data for testing. The classification accuracies of the considered algorithms are evaluated for different subject combinations.

In Fig. 5, the mean of the classification accuracies over all possible subject combinations is plotted with respect to the percentage of auxiliary data to the source data in the training data. It is observed that when the percentage of auxiliary to source data was 6%, 9%, 12% and 15%, Trbaggboost achieved average classification accuracies of 80.44%, 84.83%, 87.97% and 90.17% respectively. Here, the transfer capability achieved with Trbaggboost is much better than that achieved with TrAdaboost, TrResampling, TrBagg and RF. When larger number of labelled instances from new subject enter the training data, the SLR system performs better. Above 43% of auxiliary to source data, the performances of all the five algorithms become comparable. Hence, when labelled data from target domain is available in abundance, transfer learning is not as effective. However, when only few instances of auxiliary data are available, it is advantageous to use transfer learning approaches. This helps to reduce the manual labelling effort for all the new subject data.

The classification accuracies plotted in Fig. 5 for different percentage of auxiliary to source data and five algorithms RF, TrAdaboost, Trbaggboost, TrBagg and TrResampling were analyzed using two-way ANOVA to determine if the difference in the classification accuracies is statistically significant. The p-values for different percentages of auxiliary to source data and the five algorithms were both zero. These values indicate that the percentage of auxiliary to source data and considered algorithms significantly affect the mean classification accuracies taken over 120 subject combinations. Figure 6 shows the results of post-hoc analysis of ANOVA test conducted using Tukey's honestly significant difference post hoc test. Here, horizontal axis represents the average classification accuracies with 100 isolated signs from 120 subject combinations and vertical axis represents the different percentage of auxiliary to source data. The bold circles represent the mean accuracy and horizontal lines indicate the 95% confidence interval. The mean and confidence intervals of classification accuracies obtained with RF, TrAdaboost, Trbaggboost, TrBagg, TrReampling algorithm are plotted in different colors in Fig. 6. As seen in Fig. 6, the vertical grey dotted line representing the 95% confidence interval around

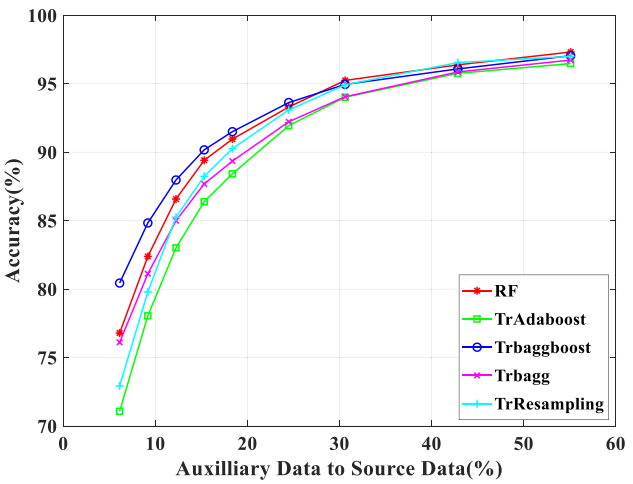

Fig. 5 Comparison of average classification accuracies with different percentage of auxiliary to source data

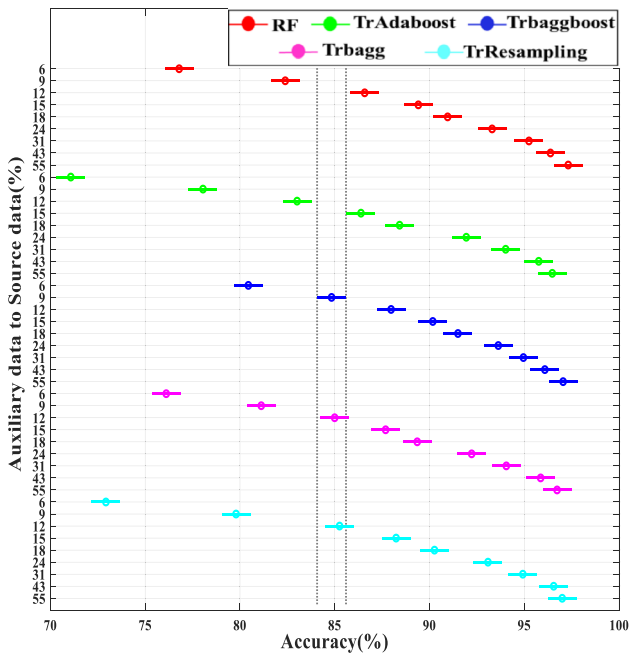

Fig. 6 Post hoc analysis of two-way ANOVA

the mean accuracy for Trbaggboost with 9% of auxiliary to the source data does not overlap with the confidence intervals of the remaining four algorithms for similar training data. This indicates that there exists a significant difference between classification accuracies of Trbaggboost algorithm as compared to RF, TrAdaboost, TrBagg and TrResampling when the percentage of auxiliary to the source data is 9%. Similar observation may be made when the percentage of auxiliary to the source data is 6%. Hence, Trbaggboost outperforms the other four algorithms when just 2 or 3 repetitions of signs are available from the new subject as auxiliary data. However, with 12% auxiliary to the source data, TrBagg and TrResampling yield classification accuracies similar to that obtained with Trbaggboost at 9% auxiliary to the source data. Above 24% of auxiliary to the source data there was an overlapping between the confidence interval of all the five algorithms, which indicates that all the five algorithms provide comparable transfer learning performance. Hence, the proposed transfer learning approach is particularly effective when the labelled data from the new subject is very limited.

5 Conclusion

In this paper, a practical issue in SLR is addressed. A database of 100 isolated signs from ISL was acquired using wearable sensors namely sEMG, accelerometer and gyroscope placed on both the forearms of the signers. When classification of signs is tested on data from signers whose labelled data was also used in training the SLR model, an average classification accuracy of 97.08% was achieved. However, when the SLR is used for classifying signs performed by a new subject whose data was not used in training the SLR system, the average classification accuracy degrades by around 25%. Hence, a novel transfer learning algorithm, Trbaggboost is proposed that combines the utility of two ensemble methods namely bagging and boosting. The performance of the algorithms was tested when different percentage of auxiliary to source data are used as labelled training data in the transfer learning algorithms. It was observed that when small amount of labelled data, such as 2 or 3 observations of each sign is available from the new subject, Trbaggboost performs significantly better than other four algorithms, RF, TrAdaboost, TrBagg, TrReampling. When just 2 observations of each sign performed by a new subject were included with the labelled data from other subjects to train the SLR, its performance improved to 71.07%, 72.92%, 76.10%, 76.79% and 80.44% with TrAdaboost, TrResampling, TrBagg, RF and Trbaggboost. When there is sufficient amount of labelled data from the new subject, the performance of all the approaches improves and become comparable. Also, RF is more suitable as a base learner in comparison with SVM and DT for the proposed approach. As a future scope this frame work can be extended with deep learning approach to enhance the performance of SLR systems.

Acknowledgements The author thankfully acknowledges the support and funds provided by Science and Engineering Research Board (SERB), a statutory body from the Department of Science and Technology (DST), (ECR/2016/000637) Government of India. The author also thanks for the support and patience of all the volunteers in recording the data.

References

Ahmed MA, Zaidan BB, Zaidan AA, Salih MM, Lakulu MM (2018) A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017. Sensors 18(7):2208. https://doi.org/10.3390/s18072208

Ali SM, Augusto JC, Windridge D (2019) A survey of user-centred approaches for smart home transfer learning and new user home automation adaptation. Appl Artif Intell 33(8):747–774. https://doi.org/10.1080/08839514.2019.1603784

Bickel S, Brückner M, Scheffer T (2009) Discriminative learning under covariate shift. J Mach Learn Res 10:2137–2155

Chiang YT, Lu CH, Hsu JY (2017) A feature-based knowledge transfer framework for cross-environment activity recognition toward smart home applications. IEEE Trans Hum Mach Syst 47(3):310–322. https://doi.org/10.1109/THMS.2016.2641679

Chong TW, Lee BG (2018) American sign language recognition using leap motion controller with machine learning approach. Sensors 18(10):3554. https://doi.org/10.3390/s18103554

Dai W, Yang Q, Xue GR, Yu Y (2007) Boosting for transfer learning. In Proceedings of the 24th international conference on Machine learning ACM, pp. 193–200. doi: 10.1145/1273496.1273521

Farhadi A, Forsyth D, White R (2007) Transfer learning in sign language. IEEE conference on computer vision and pattern recognition, pp. 1–8. doi: 10.1109/CVPR.2007.383346

Gattupalli S, Ghaderi A, Athitsos V (2016) Evaluation of deep learning based pose estimation for sign language recognition. In Proceedings of the 9th ACM international conference on pervasive technologies related to assistive environments, 12, pp. 1–7. doi: 10.1145/2910674.2910716

Hu C, Chen Y, Peng X, Yu H, Gao C, Hu L (2018) A Novel feature incremental learning method for sensor-based activity recognition. IEEE Trans Knowl Data Eng 31(6):1038–1050. https://doi.org/10.1109/TKDE.2018.2855159

Indian Sign Language Dictionary (2015) Ramakrishna mission Vivekananda University, Coimbatore. http://indiansignlanguage.org/dictionary/. Accessed 25 May 2020

Kamishima T, Hamasaki M, Akaho S (2009) TrBagg: a simple transfer learning method and its application to personalization in collaborative tagging. Ninth IEEE international conference on data mining, 6, pp. 219-228. doi: 10.1109/ICDM.2009.9

Khan MA, Roy N (2017) Transact: transfer learning enabled activity recognition. IEEE international conference on pervasive computing and communications workshops, pp. 545–550. doi: 10.1109/PERCOMW.2017.7917621

Kyranou I, Vijayakumar S, Erden MS (2018) Causes of performance degradation in electromyographic pattern recognition in upper limb prostheses. Front Neurorobotics 12:58. https://doi.org/10.3389/fnbot.2018.00058

Liu X, Wang G, Cai Z, Zhang H (2016) Bagging based ensemble transfer learning. J Ambient Intell Hum Comput 7(1):29–36. https://doi.org/10.1007/s12652-015-0296-5

Liu X, Liu Z, Wang G, Cai Z, Zhang H (2017) Ensemble transfer learning algorithm. IEEE Access 6:2389–2396. https://doi.org/10.1109/THMS.2016.2641679

Ma Y, Zhou G, Wang S, Zhao H, Jung W (2018) SignFi: sign language recognition using WiFi. Proc ACM Interac Mob Wearable Ubiquitous Technol 2(1):23. https://doi.org/10.1145/3191755

Mocialov B, Hastie H, Turner G (2018) Transfer learning for British Sign Language Modelling. In Proceedings of the fifth workshop on NLP for similar languages, varieties and dialects, pp. 101–110

Phinyomark A, Phukpattaranont P, Limsakul C (2012) Feature reduction and selection for EMG signal classification. Expert Syst Appl 39(8):7420–7431. https://doi.org/10.1016/j.eswa.2012.01.102

Raheja JL, Mishra A, Chaudhary A (2016) Indian sign language recognition using SVM. Pattern Recognit Image Anal 26(2):434–441. https://doi.org/10.1134/S1054661816020164

Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A (2008) Resampling or reweighting: a comparison of boosting implementations. 20th IEEE international conference on tools with artificial intelligence, 1, pp. 445–451. doi: 10.1109/ICTAI.2008.59

Sharma S, Gupta R, Kumar A (2019) On the use of multi-modal sensing in sign language classification. 6th IEEE international conference on signal processing and integrated networks (SPIN), pp. 495–500. doi: 10.1109/SPIN.2019.8711702

Song P, Zheng W (2018) Feature selection based transfer subspace learning for speech emotion recognition. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2018.2800046

Suharjito, Anderson R, Wiryana F, Ariesta MC, Kusuma GP (2017) Sign language recognition application systems for deaf-mute people: a review based on input-process-output. Proced Comput Sci 116:441–448. https://doi.org/10.1016/j.procs.2017.10.028

Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. J Big Data 3(1):9. https://doi.org/10.1186/s40537-016-0043-6

Wu J, Sun L, Jafari R (2016) A wearable system for recognizing American sign language in real-time using IMU and surface EMG sensors. IEEE J Biomed Health Inform 20(5):1281–1290. https://doi.org/10.1109/JBHI.2016.2598302

Xi X, Tang M, Miran SM, Luo Z (2017) Evaluation of feature extraction and recognition for activity monitoring and fall detection based on wearable sEMG sensors. Sensors 17(6):1229. https://doi.org/10.3390/s17061229

Yang X, Chen X, Cao X, Wei S, Zhang X (2016) Chinese sign language recognition based on an optimized tree-structure framework. IEEE J Biomed Health Inform 21(4):994–1004. https://doi.org/10.1109/JBHI.2016.2560907

Zhou ZH (2015) Ensemble learning. Encycl Biom. https://doi.org/10.1007/978-1-4899-7488-4

Zhu Z, Wang X, Kapoor A, Zhang Z, Pan T, Yu Z (2018) EIS: a wearable device for epidermal American Sign Language recognition. Proc ACM Interact Mob Wearable Ubiquitous Technol 2(4):202. https://doi.org/10.1145/3287080