A New CNN Approach for Hand Gesture Classification using sEMG Data

Aysun TUTAK EROZEN

Tofaş Türk Otomobil Fabrikası A.Ş., Bursa, Turkey

Abstract

In this paper, a new CNN architecture is introduced for classification of six different hand gestures using surface electromyography (EMG) data collected from the forearm. At first, two different deep neural networks produced based on Slow Fusion and Inception models separately. Then, the average of accuracy values and standard deviations were calculated for each type of network. The average accuracy was 80.88% and standard deviation was 0.030 for the Slow Fusion based network. For the Inception based network, average accuracy was 82.64% and standard deviation was 0.028. In addition to these two networks, a new CNN architecture is introduced using Slow fusion and Inception models in combination. The architecture has two parallel Inception modules in parallel. Each parallel module is fed by the half of the 3D feature map. The proposed model slowly fuses the information of the parallel modules throughout the network as in Slow-Fusion architecture. The average accuracy achieved with this model was 83.97% and the standard deviation was 0.027. Despite the small data set, the accuracy had increased with the proposed hybrid model. The smaller standard deviation indicates that it is less affected by variations in the training dataset. Our experimental results show that the proposed method gives the best results among the Slow Fusion based and Inception based models.

Keywords: EMG, CNN, Deep Learning, Slow Fusion, Gesture Recognition, Inception.
1. Introduction

For medical motives, sometimes hand amputation is the only choice for patients. After the hand amputation most patients prefer to use prosthetic hands to resume their lives as before. The most important thing is to feel natural when using the prosthetic hands. To advance the prosthetic hands, biomedical researchers prolong to study. There are different methods to control the prosthetic hands movement, but the most favored method is Electromyography (EMG). EMG signals show the electrical activity of the muscles produced during muscle interactions. EMG signals are either collected using needles or the surface electrodes directly from skin. Because of the easiness and being non-invasive, surface electrodes are ideal.

In the past decades, researchers used many approaches and methods for EMG signal classification [1-25]. EMG pattern recognition involves four steps. These steps are signal processing, feature extraction, feature selection and classification. Raw EMG signal has high noise and low amplitude characteristic. In the signal processing stage, unnecessary information is removed from the signal and the low amplitude signal is amplified. During feature extraction, EMG signals can be represented in Time Domain, Frequency Domain or Time-Frequency Domain. [1] In Feature selection step, dimensional reduction is applied. As a last step, pattern classification is performed.

For the hand gestures, the most known feature extraction methods in Time Domain are; Root Mean Square (RMS), Histogram (HIST), Mean Average Value (MAV), Absolute Mean Values Ratio (AMVR), Variance, Zero Crossing (ZC), Waveform Length (WL) [2-6]. Some other features are Integral of Absolute Value (IAV), Wilson Amplitude (WAMP), Skewness, Kurtosis, AR Coefficients [8-10]. Some researchers prefer to use Frequency domain feature extraction methods. The feature extraction methods in frequency domain are Mean Frequency (MNF), Median Frequency (MDF). Other spectral variables that have been applied in the analysis of EMG signal are total power (TTP), mean power (MNP), peak frequency (PKF), the spectral moments (SM), frequency ratio (FR), power spectrum ratio (PSR), and variance of central frequency (VCF) [11-12]. The most favorite Time-Frequency methods are Short Time Fourier Transform (STFT) and Wavelet Transform (WT). [13-17] After feature extraction and feature selection steps there is a classifier at the last stage. For EMG classification, researchers have proposed various classifiers [1-25]. Classifiers can use different algorithms such as LDA[2,4,8], Bayesian [3], k Nearest Neighbor (kNN) [3,4,10], Kernel Regularized Least Squares [4], Random Forest Classifier [5], Support Vector Machines(SVM) [6,18], Artificial Neural Networks [1,7,9,13,14,15,20], Minimum Distance [16] and Fuzzy system[1].

Additional to these conventional approaches, in recent years computer processing speeds have increased and researchers have used GPU and TPU to perform experiments. With the widespread use of high-speed processors, studies in deep learning area have become much more common. In 2016, Alard et al. [21] proposed two-stage CNN for classification to control a robotic arm. The network was fed with 2D images of spectrogram. A gesture was considered a success if no more than two false consecutive or no more than four non-consecutive miss-classifications occurred during a 10 s period.
With this assumption the average success rate was 93.14% over all participants. In 2017, Alard et al. [22] proposed to use Slow Fusion model CNN. To prevent overfitting, they applied transfer learning. The input utilized in this work (Time x Spatial x Frequency) is that they can both represent non-stationary information. The target network yielded an average accuracy of 97.81% on the evaluation dataset for 17 participants. Robotic arm controlled with CNN behaves
similar with joystick. In 2018, Hu et al. [23] presented a CNN-RNN architecture and made a comparison for different benchmark datasets. Attention-based hybrid CNN-RNN architecture has the best results with accuracy %87 for NinaProDB1 dataset. Shen et. al. [24] developed a classification method based on CNN and Stacking Ensemble Learning. Chen et.al.[25] used a compact CNN architecture called EMGNet. It consists of four convolutional layers and a max pooling layer without using the full connection layer as the final output. The main idea was to have smaller number of parameters. It was validated on the Myo Dataset that the average recognition accuracy of EMGNet can achieve 98.81%

In this paper, a hybrid CNN model is proposed. The proposed model is a combination of Slow-Fusion and Inception [26] models. Inception architecture is one of the state-of-the-art architecture for image classification. Slow-Fusion model is used for video classification. It slowly fuses the temporal information throughout the network such that higher layers get access to progressively more global information in both spatial and temporal dimensions [27]. In this proposed model, there are two Inception modules in parallel. First Inception module fed by the information of 4 electrodes and the second Inception module fed by the rest of the electrodes. To fed parallel CNN’s with split data is the common behavior in Slow-Fusion architecture. The input feature map has 3 dimensions like [22]. But In [22] 2 CNNs fed by same electrodes but half of the complete feature map. The proposed model performance is higher than both Inception architecture and Slow-Fusion architecture alone. The rest of the paper is organized as follows: In the second section, the dataset, feature extraction method and the introduced networks are described in detail. The experimental results are discussed in Section 3. The Section 4 provides a conclusion.

2. Materials Methods

EMG pattern recognition involves four steps. These steps are signal processing, feature extraction, feature selection and classification. The biggest opportunity to use CNN is that it does not require feature selection step. Feature selection is completed automatically in deep learning algorithms. In this section, the dataset, preprocessing and designed deep learning networks are summarized.

2.1. The Dataset

In this paper, the dataset “EMG Data for Gestures Data Set” is obtained from UC Irvine Machine Learning Repository [28]. The dataset contains EMG data of 36 people for 6 different type gestures. The data was collected by MYO Thalmic bracelet worn on user’s forearm. This bracelet has 8 sensors that simultaneously get myographic signals. The signals are sent through a Bluetooth interface to a PC. The EMG data is time-series data and the sampling frequency is 1 kHz. The existing gestures in the dataset are shown in Figure 1. These gestures are; hand at rest, hand clenched in a fist, wrist flexion, wrist extension, radial deviations and ulnar deviations.
In the dataset the subjects perform two series of 6 gestures. Using these time-series data, 864 feature maps were generated after preprocessing. %80 of the feature maps were benefited for training, and the remaining were handled for testing purpose.

2.2. Preprocessing of Raw Data and Feature Extraction

EMG signals are acquired from muscles. EMG amplitude is the sum of the electric potential differences within a muscle related to all the active motor units near the electrodes on the skin. It is a function of time and it is defined in terms of its amplitude, frequency, and phase. EMG signal is modelled as in Eq. (1). In this equation, \( x(n) \) is the modelled EMG signal, \( e(n) \) is the point processed representing the firing impulse. \( h(r) \) represents the MUAP (motor unit action potentials) and \( w(n) \) shows the zero-mean additive Gaussian noise and \( N \) is the number of the motor firings.

The amplitude range of EMG signal is 0-10 mV (+5 to -5) prior to amplification. It has high noise levels and its frequency range is between ±50 Hz and ±500 Hz.

\[
x(n) = \sum_{r=0}^{N-1} h(r)e(n - r) + w(n)
\]  

The dataset contains time-series raw data. However, preprocessing should be applied to the raw data to extract features before feeding the networks.

The Fourier Transform (FT) is the one of the well-known methods to extract features. FT assumes that the frequency characteristic of the signal is same for an infinite time. It does not contain information of frequency changing over time. The FT of a continuous time signal \( x(t) \) is defined as below Eq. (2). \( X(\omega) \) is the frequency domain of signal. \( \omega \) is angular frequency, \( t \) is time and \( i \) is imaginary number.

\[
X(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x(t) e^{-iat} dt
\]
On the other hand, in nature, most signals have frequency contents changing over time. The frequency of the EMG is changing over time as well. The FT is not appropriate for these kinds of signals. Therefore, the Short Time Fourier Transform (STFT) is developed for these types of signals.

In STFT, the signal is cut into blocks of finite length, and then the FT of each block is computed. To improve the result, blocks are overlapped. Each block is multiplied by a window that is tapered at its end points. [36] The graphical representation of the magnitude of the Short Time Frequency Transform $X(\tau, \Omega)$, is called the spectrogram of the signal. STFT of a continuous signal is given in Eq. (3). In this equation $w$ stands for the window function. $\tau, \Omega$ shows the time and frequency information.

$$X(\tau, \Omega) = \int_{-\infty}^{\infty} x(t)w(t - \tau) e^{-j\omega t} dt$$  \hspace{1cm} (3)

There are different window functions that are used as a window for STFT. In this paper, Hann window was selected. The Hann window is defined as in Eq. (3). Here, $\omega(n)$ shows the window function, $M$ shows the number of points in Hann window.

$$\omega(n) = 0.5 - 0.5 \cos \left( \frac{2\pi n}{M-1} \right) \quad 0 \leq n \leq M - 1$$  \hspace{1cm} (4)

In the dataset, the collected EMG data samples have different lengths in time. During feature extraction, initial 1024 sequential points are split from each sample at first. The time-series of 1024 points long data was used to find the spectrogram characteristic. For this purpose, STFT was applied to these split data samples. The 1024 points long data is separated into 64 points long samples. The sequential samples had overlapping 32 points. Then for each block 64-point FFT was applied using 64-point Hann window. As a result, 33x33 matrix was formed for each electrode.

There were 8 electrodes measuring EMG signal, so that, the feature matrix of a gesture was 33x33x8. The last step was L2 normalization of feature matrix before feeding the neural networks (Figure 2).

![Figure 2](image-url)
2.3. Deep Learning Networks

In this paper three different 3D-CNNs were developed. The first network was built based on Slow Fusion model. In this model, input data was separated into two pieces. Each piece fed the disconnected parallel layers. These layers were connected throughout the network. Figure 3. shows the Slow Fusion model in this study.

Slow Fusion model starts with two halves of the whole feature map. Each half is 33x33x4 in size. This means that each parallel layer carries the information of 4 electrodes. Firstly, each data piece pass through two CNN layers with rectified linear unit (ReLU) and batch normalization (BN). The first CNN layer has 11 filters with size of 7x7. The second CNN layer has 11 filters with size of 5x5. Each parallel branch produces a map.

![Slow fusion model](image-url)

**Figure 3.** Slow fusion model
The two maps that produced from two branches are concatenated. The merged map pass through another two CNN layers with rectified linear unit (ReLU) and batch normalization (BN). These layers have same amount and same size of filters. Later, the output map is flattened and connected to a Full Connected (FC) layer. There are two FC layers having 100 neurons sequentially. After each FC layer, there are BN and dropout stages. Dropouts are applied to prevent overfitting. The latest step is to apply SoftMax activation function to classify the gestures. The values of SoftMax function defines the class.

Secondly, the Inception based model was conducted. In this model, the parallel branches are fed by whole feature map. The input has size 33x33x8. At each branch, different filters are applied in CNN layers. Figure 4 shows this model. At first branch, there are two CNN layers. First CNN layer is formed by 11 filters with size 1x1. Second CNN layer is formed by 11 filters with size 3x3. There are also two CNN layers at second branch. Here, first CNN layer is formed by 11 filters with size 1x1 and second CNN layer is formed by 11 filters with size 5x5. At third branch, there is a maximum pooling layer with a stride of 3x3 followed by a CNN layer formed by 11 filters with size 1x1. Fourth branch has only one CNN layer composed of 11 filters with size 1x1. At each CNN layer the ReLU activation function and BN are applied. After merging of four branches, there are FC layers which are same with previous model.

![Inception model](image-url)
Thirdly, hybrid model was developed as a novel approach. This model is a combination of Slow Fusion and Inception models. The input data separated into two data pieces like the Slow Fusion model. Each piece feeds an Inception module. This means that there are two inception modules parallel to each other. At the end of each Inception module, the data maps are merged. The layers after this stage are the same as other models. Figure 5. displays this model.

After the network structures were created, all networks were trained from scratch using following parameters.

- Optimization: SGD
- Batch Size: 100
- Epochs: 150
- Activation function: ReLU
- Output function: SoftMax

![Hybrid model](image)

**Figure 5.** Hybrid model
3. Experimental Results and Discussion

In this paper three different 3D-CNNs were developed for hand gesture classification using sEMG data gathered from the forearm. For the experiments, the dataset “EMG Data for Gestures Data Set” is obtained from UC Irvine Machine Learning Repository. This dataset contains raw EMG data of 36 people for 6 different type gestures. Using the time-series data, 864 feature maps were generated after preprocessing. There were 8 electrodes measuring EMG signal. The features maps were produced using STFT and they were 33x33x8 in size. The dataset totally contains 864 feature maps, so that there are 144 feature maps for each gesture. The whole dataset is balanced. The train and validation data sets are split randomly, they are not balanced in own self. To measure the performance, the train set and validation set were created in 100 diverse ways. In all cases, the train dataset has 688 data and the validation dataset has 176 data. It should be noted that the constructed networks were trained from scratch for 150 epochs with a batch size of 100. As a performance metric, accuracy was calculated with validation data. Accuracy is the percent ratio of the number of true classified gestures over the number of all classified gestures in validation dataset. The accuracy was calculated for each round and the average accuracy of 100 turns were evaluated at the end. Table 1. shows comparison of the average accuracy and the standard deviation of accuracy for each developed network.

The visualization of the results is indicated in the Figure 6. It figures out that minimum average accuracy was obtained with Slow Fusion model. The average accuracy was 80.88% and standard deviation was 0.030 for the Slow Fusion based network. With Inception model the average accuracy is higher and standard deviation is smaller. This means that the Inception model is more stable network. For the Inception based network, average accuracy was 82.64% and standard deviation was 0.0028.

As it seen from the results, the proposed network achieved the highest accuracy and lowest standard deviation. The average accuracy achieved with this model was 83.97% and the standard deviation was 0.027. Despite the small data set, the accuracy had increased with the hybrid model. The smaller standard deviation indicates that it is less affected by variations in the training dataset.

<table>
<thead>
<tr>
<th>Models</th>
<th>Average Accuracy</th>
<th>Std. Dev. of accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slow Fusion Model</td>
<td>%80.88</td>
<td>0.030</td>
</tr>
<tr>
<td>Inception Model</td>
<td>%82.64</td>
<td>0.028</td>
</tr>
<tr>
<td>Proposed (Hybrid) Model</td>
<td>%83.97</td>
<td>0.027</td>
</tr>
</tbody>
</table>
In this paper, a hybrid CNN model is introduced for hand gesture classification using sEMG data. The proposed model is a combination of Slow-Fusion and Inception models. There are two Inception modules in parallel at the beginning. First Inception module fed by the information of 4 electrodes and the second Inception module fed by the rest of the electrodes. To fed parallel CNN’s with split data is the common behavior in Slow-Fusion architecture. Then the outputs of the inception models are fused throughout the network. The 3D input feature maps are constructed by STFT of all electrodes time-series raw data. The experimental results demonstrate that the proposed method yields 83.97% accuracy and 0.027 standard deviation. Moreover, it was denoted that the introduced model is more successful than Slow-Fusion and Inception models using alone.

References


