This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2938295, IEEE Transactions on Neural Systems and Rehabilitation Engineering

A Multi-branch 3D Convolutional Neural Network for EEG-Based Motor Imagery Classification

Xinqiao Zhao, Hongmiao Zhang, *Member, IEEE*, Guilin Zhu, Fengxiang You, Shaolong Kuang, *Member, IEEE*, and Lining Sun

***Abstract*—One of the challenges in motor imagery (MI) classification tasks is finding an easy-handled electroencephalogram (EEG) representation method which can preserve not only temporal features but also spatial ones. To fully utilize the features on various dimensions of EEG, a novel MI classification framework is first introduced in this paper, including a new 3D representation of EEG, a multi-branch 3D convolutional neural network (3D CNN) and the corresponding classification strategy. The 3D representation is generated by transforming EEG signals into a sequence of 2D array which preserves spatial distribution of sampling electrodes. The multi-branch 3D CNN and classification strategy are designed accordingly for the 3D representation. Experimental evaluation reveals that the proposed framework reaches state-of-the-art classification kappa value level and significantly outperforms other algorithms by 50% decrease in standard deviation of different subjects, which shows good performance and excellent robustness on different subjects. The framework also shows great performance with only nine sampling electrodes, which can significantly enhance its practicality. Moreover, the multi-branch structure exhibits its low latency and a strong ability in mitigating overfitting issues which often occur in MI classification because of the small training dataset.**

***Index Terms*—electroencephalogram (EEG), motor imagery (MI), 3D convolutional neural network (3D CNN), multi-branch structure**

I. Introduction

The framework of electroencephalogram (EEG)-based motor imagery (MI) classification is built on the fact that the imagining of body movements is accompanied by a circumscribed event-related synchronization (ERS) or event-related desynchronization (ERD) in EEG [1], [2]. On the basis of ERS/ERD phenomenon, Brain-Computer Interface (BCI) researchers proposed a great number of methods for MI classification. Among the existing feature extraction methods [3]-[6], common spatial pattern (CSP) method [7] is one of the most effective approaches for constructing optimal spatial filters that are sensitive to differences between two kinds imagery [8], [9]. By constructing a few time series whose variances contain the most discriminative information, CSP method can enhance the effect of feature extraction. However, CSP method has issue of searching for the optimal frequency range for each subject which will influence the final performance. To address this issue, the subband CSP (SBCSP) [10] method with different frequency bands as inputs is proposed. This method applies different bandpass filters to separate the raw EEG signal into different frequency bands and then extract MI-related features for each frequency band signal by using CSP. The extracted features are handled through a multiple linear discriminant analysis (MLDA) [11] with fusion algorithm to achieve the final MI-classification results, which shows great uncertainty identification performance.

Inspired by this multiple subbands input idea, a series CSP-based methods are proposed. Filter bank common spatial pattern (FBCSP) [12] first uses a group of band-pass filters and CSP algorithm to extract the optimal spatial features. Then, a feature selection algorithm is used to automatically select discriminative pairs of frequency bands and corresponding CSP features. A classification algorithm is subsequently used to classify the CSP features. Experimental results in [12] show that FBCSP with a particular combination feature selection and classification algorithm, such as Mutual Information based Best Individual Feature (MIBIF) and Naïve Bayesian Parzen Window (NBPW), MIBIF and Fisher Linear Discriminant (FLD) [13], MIBIF and Support Vector Machine (SVM) [14] all yield statistically superior results than SBCSP and CSP. In [15], augmented common spatial pattern (ACSP) features which cover various frequency ranges are generated based on pair-wise projection matrices and classified by convolutional neural networks (CNN). The method yields a slightly better result than FBCSP. A BCI system which combines SBCSP, fuzzy integral, particle swarm optimization (PSO) and multiple linear discriminant analysis (MLDA) classifier is proposed in [16], which exhibits robust performance for offline single-trial classification of MI and real-time control of a robotic arm using MI. Ko et al. [17] propose a BCI framework by using fast fourier transform (FFT) and CSP to extract MI features and then adopting three different classifiers, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and k-nearest

This work was supported in part by the National Nature Science Foundation of China No.U1713218，U1613224， and 61375090. (Corresponding author: Shaolong Kuang)

X. Zhao, H. Zhang, G. Zhu, F. You, S. Kuang, and L. Sun are with the College of Mechanical and Electrical Engineering, Soochow University, Suzhou, 215000, China (e-mail: slkuang@suda.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2938295, IEEE Transactions on Neural Systems and Rehabilitation Engineering

neighbors classification (k-NNC) [18], to classify the transformed MI-based BCI data. Finally, a multimodal fuzzy fusion framework with the Choquet or Sugeno integrals is used to fuse the posterior probability obtained from the results of different classifiers , which significantly enhance the classification accuracy.

In addition to CSP-based methods, several methods combined with other feature extraction algorithms also show great performance on MI-classification tasks. Xie X et al. [19] derive a bilinear sub-manifold learning (BSML) algorithm to reduce the symmetric positive-definite (SPD) matrices space dimensionality in motor imagery BCIs which can be considered as an extension of CSP on covariance matrices in measure of Riemannian distance. Based on this algorithm, BSML-based MI-EEG classification algorithm is proposed which exhibits strong robustness and competitive classification accuracies. In [20], Principal Component Analysis (PCA) transform is applied to decompose and retain the data information of two EEG sampling channels. Then Neural Network Classifier is applied to achieve the final results. This work indicates the feasibility of combining PCA transform with neural network.

Although CSP-based methods and other BCI frameworks mentioned above have been successfully applied in MI-classification area, the architectures of these frameworks all separate feature extraction and classification into two stages, which causes the features extraction model parameters and followed classifier are trained separately with different objective functions. Accordingly, this training strategy will result in the features extracted by CSP or other features extraction methods are not the most suitable ones for MI classification. In particular, CSP adopts a set of linear transformations (i.e., spatial filters) and it can only be suitable for one-versus one or one versus rest classification strategy. For multi-classes MI classification tasks, a multiple one-versus-rest CSP features extraction strategy have to be applied which will lead to redundant features [21]. These redundant features would not help to improve MI classification performance, but would result in unnecessary computation which affects the real-time performance of the algorithm.

In contrast to CSP based MI classification methods, deep learning methods use only minimal preprocessing and embed all the computational steps in a single network to conduct a fair end-to-end model, which is more suitable for multi-class classification tasks. Due to end-to-end architecture, the parameters of feature extraction model and classifier can be optimized jointly, which shows a significant performance improvement [22]. In addition, applying deep learning methods in MI classification can directly profit from future advances in other areas, such as computer vision and natural language processing, where deep learning methods have already been widely used. With more powerful computing hardware designed for deep learning [23], the real-time performance of MI classification can be enhanced, too.

When it comes to applying deep learning methods in MI classification, an important prerequisite is representing EEG data to a processable form. In order to satisfy this prerequisite, an effective and popular way is representing EEG into a 2D-array with the number of time steps as the array width and the number of electrodes (or other spatial features) as the array height [22]. Another typical approach which is described in [24] is representing EEG signals into a two-dimensional image by short-time Fourier transform (STFT) method. Concretely, the mu and beta frequency band spectral contents are made apparent by preserving the patterns of activation at various locations, times and frequencies.

However, this 2D presentation form will lose most spatial features of EEG and the correlation among nearby sampling electrodes cannot be fully reflected either in the 2D array. As a consequence, the final performance of the MI classification model will be affected. Considering the shortage of 2D representation mentioned above, some other methods with more representation dimensions are proposed to achieve better performance. Bashivan et al. [25] propose a new representation which preserves the spatial, spectral, and temporal structure of the raw EEG. They first estimate the power spectrum of the EEG signal from each electrode and then sum the squared absolute values for three selected frequency bands. After that, the Azimuthal Equidistant Projection (AEP) method is used to map the electrode distribution and construct the input images of the model. This representation shows significant performance improvements over current state-of-the-art approaches in the field of cognitive load classification, which indicates that spatial features are important for EEG-based classification tasks.

Based on the represented EEG, deep learning methods can then be applied to extract related features for MI classification. Luo et al. [26] apply a deep recurrent neural network (RNN) with a cropped training strategy to classify MI. The experiments demonstrate that the spatial-frequency-sequential relationships outperform other spatial-frequency methods and the cropped training strategy can mitigate overfitting which is caused by a small dataset. Among various deep learning methods, convolutional neural networks (CNN) have great advantages in EEG-based MI classification, which can learn MI-related patterns from raw data and exploit hierarchical structure in the natural signals without any priori preprocessing or features selection. Tabor et al. [27] propose a novel architecture which combines CNN and stacked autoencoders (SAEs) for MI classification. This architecture yields 9% improvement over the winner algorithm of BCI competition IV dataset 2b competition.

In the area of CNN appliance in BCI system, several studies show interesting results about the design choice of CNN which is implemented in MI classification. In [22], three CNNs with different depths are used to classify four-class EEG signals. The experiments show that the depth of the CNN significantly affects the classification performance and shallow CNN performs better than the deep one. Results in [21] indicate that when using a single receptive field CNN to classify MI, it is needed to change the parameters of the network for different subjects in order to achieve the best classification performance.

In this study, for EEG-based MI classification tasks, a new 3D representation of EEG is first introduced, which preserves both spatial information and temporal information. Based on this 3D representation, a multi-branch 3D CNN is employed to extract MI related features. The strength of applying 3D CNN

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2938295, IEEE Transactions on Neural Systems and Rehabilitation Engineering

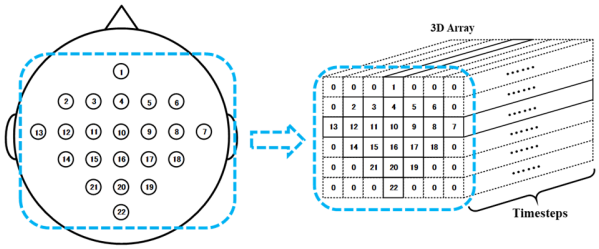

Fig. 1. EEG 3D representation procedure. Left: Electrode montage corresponding to the international 10-20. Right: 3D representation of EEG.

in EEG classification is that the spatial and temporal features in EEG signals can be extracted simultaneously and the relationship between them can be fully utilized. By combining 3D representation of EEG and multi-branch 3D CNN, we achieve state-of-the-art performance on the BCI competition IV-2a data set, significantly enhancing the classification performance with higher average kappa and lower standard deviation of different subjects.

The rest of this paper is organized as follows: The proposed MI classification framework is described in Sec. II. The experimental results are reported in Sec. III and the discussion is shown in Sec. IV. Sec. V concludes the paper.

II. Methods

In this section, we will illustrate the detailed configurations of the MI classification framework, including 3D representation of EEG, multi-branch 3D CNN and the classification strategy.

A. 3D Representation of EEG

In this study, we design a 3D representation of EEG. The first step of obtaining this 3D representation is transforming each frame of EEG into a 2D array according to the general spatial distribution of sampling electrodes and meanwhile padding the point where there is no electrode with 0. The rectangular shape of this 2D array will make the following feature extraction process implemented much easier. Then, we expand this 2D array to 3D array by using EEG's temporal information. Note that this representation approach is easy-to-use and general enough to be applied in any other EEG-based classification tasks.

For introducing the core idea of 3D representation method, we take twenty-two Ag / AgCl EEG sampling electrodes with 250 Hz sampling rate which have been used in [28] as an example, and design a 3D representation form. Fig. 1 shows the 3D representation procedure according to these twenty-two electrodes. As can be seen, under the premise of ensuring the processability of EEG data, this kind of representation not only preserves the temporal features of EEG (which exist in the sequential data of EEG signals) completely, but also preserves the sampling electrodes spatial features (which exists in the electrodes distribution) to some extent.

B. Multi-branch 3D CNN

For MI classification, a multi-branch 3D CNN is designed based on the 3D representation of EEG. In the following subsections, we will first describe the generic layout of the multi-branch 3D CNN and then detail the specific configurations in each branch network.

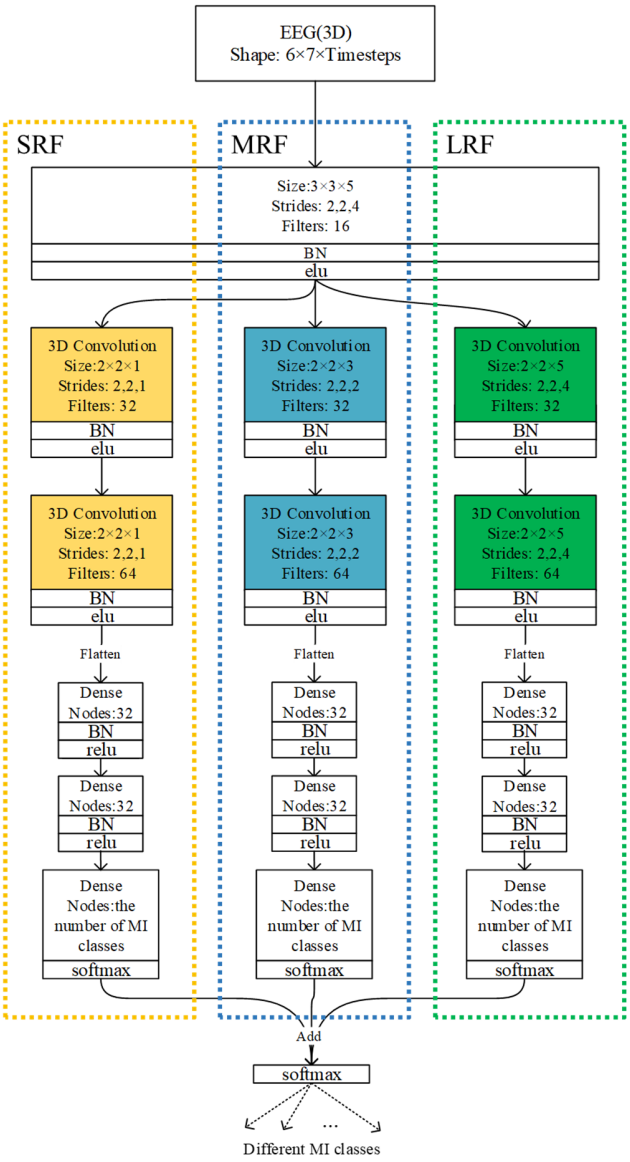

Fig. 2. The multi-branch CNNs architecture. All the convolutional processes adopt zero-padding strategy. 'Shape' means the height, width and depth of the input array. 'Size' means the height, width and depth of the 3D convolution window. 'Strides' means the strides of the convolution along each dimension. 'Filters' means the number of output filters in the convolution. 'BN' means batch normalization layer. 'Nodes' means the dimensionality of the dense layer's output space. Three branch networks are respectively framed by three dashed boxes with different colors.

1) Network Overview

3D CNN can adequately extract features from spatial dimension and temporal dimension simultaneously [29] which will fully utilize the relationship between them. In the light of this, three 3D CNNs with different sizes of receptive field are designed to extract MI-related features. These networks are respectively named as small receptive field network (SRF), medium receptive field network (MRF) and large receptive field network (LRF). Then, inspired by ConvNet fusion which has been used in [30], [31] and [32], these three networks are combined together into a multi-branch 3D CNN. This multi-branch structure may improve the performance due to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2938295, IEEE Transactions on Neural Systems and Rehabilitation Engineering

TABLE I
THE RECEPTIVE FIELD SIZE OF SRF, MRF AND LRF

Convolutional Layer num.	Spatial receptive field size	Receptive field size in time domain		
		SRF	MRF	LRF
1st	3×3	5	5	5
2nd	5×5	5	13	21
3rd	6×7	5	29	85

complementarity of the networks with various sizes of receptive field.

Fig. 2 shows the details about the multi-branch 3D CNN. We take the 3D representation with 250 Hz sampling rate described in Sec. II.A as the network's input. In the multi-branch 3D CNN, three branch networks share their first convolutional layer which reduces the number of parameters. In each branch network, a stack of 3D convolutional layers is followed by two dense layers with soft-max activation in the end. These branch networks' respective soft-max results are summed and then input to an additional soft-max function which produces the final classification result.

2) Details about Different Branch Networks

As shown in Fig. 2, there are three convolutional layers (including the shared layer) and three dense layers in each branch network. Particularly, in the light of VGG [33] network's architecture, relatively small convolutional filter size and more convolutional layers are chosen in order to reduce the number of parameters. This approach also increases the non-linear calculation times which makes the decision function more discriminative [33]. The number of the convolutional filters is also incremented layer by layer to make sure that the increasingly richer and richer features are properly extracted. In both convolutional layers and dense layers, batch normalization [34] is implemented to speed up the training process and mitigate overfitting. Due to this architecture setting, each branch network will first extract local features in spatial domain and time domain simultaneously. As the network goes deeper, the network will extract richer features from these local features to achieve a broader receptive field of EEG data. After the third convolutional layer, the network receptive field in spatial domain has covered all sampling electrodes. For different branch networks, the parameters of the filters in the last two convolutional layers are set to different values in order to achieve different receptive field sizes in time domain. These temporal receptive field sizes are correlated with EEG's sampling rate which is 250 Hz in this illustration. In other scenarios, the temporal receptive field size should be adjusted with EEG's sampling rate. The detailed information about the spatial and temporal receptive field sizes of the three branch networks is shown in TABLE I.

C. Motor Imagery Classification Strategy

In the previous section, we described our network architecture in details. In this section, we present the training and testing strategies of multi-branch 3D CNN.

1) Training Strategy

Cropped training is a common procedure to increase training examples. This method has been widely adopted in image recognition areas and significantly improves the training effect [35], [36]. In EEG processing area, Luo et al. [26] and Schirrmeister et al. [22] also applied cropped training to extend EEG training dataset. The experiments in [22] demonstrate that, compared with trial wise training, cropped training leads to a better classification performance.

In this study, a cropped training approach for EEG 3D representation is adopted by sliding a 3D window which covers all electrodes on each EEG data trial along the time dimension with a data stride 1. The 3D window size on time dimension is related to the sampling rate of EEG data and the characteristics of the specific task. Through this approach, the cropped strategy with temporal size of 240 will generate seventy-four cropped data as given an original trial with three hundred and thirteen timesteps. These data and the trials they belong to will have the same labels. We subtract the overall average value of each cropped 3D array from the corresponding 3D array after the cropped approach and shuffle the whole training dataset before each training epoch to improve the training effect [34].

In terms of network optimization, normalized initialization method which is described in [37] is adopted to initialize all weights. The initial value of learning rate is set at 0.01. If the training loss does not decrease during the training process in 1 epoch, the learning rate will be changed to a tenth of what it was before. ADAM with the default parameters values which are described in [38] is adopted as the optimization method and Negative log-likelihood cost is taken as the optimization criterion. During the training process, we keep monitoring the Negative log-likelihood cost on the training dataset. If the cost does not decrease in twenty epochs, we will stop training and restore the network weights from the epoch with the best value of the cost.

2) Testing Strategy

During the test process, a cropped strategy is also used to improve the test accuracy. After calculating the prediction results of all cropped EEG data in each trial, we sum the prediction results together and regard the sum as the final classification result of the corresponding trial. Additionally, considering the real-time capability of 3D CNN network in the test process, we set the cropped strategy data stride parameter at 5 which leads to less computation time.

III. Experiments and Results

A. Experimental Setup

The methods proposed above were evaluated on BCI competition IV 2a dataset. The EEG data in this dataset were recorded from twenty-two Ag/AgCl which corresponded to the international 10-20 system. The recorded signals were sampled with 250 Hz and bandpass-filtered between 0.5 Hz and 100 Hz. An additional 50 Hz notch filter was also enabled to suppress line noise. This dataset contained four kinds of motor imagery from nine subjects. Two sessions on different days were recorded for each subject. Each session was comprised of two hundred and eighty-eight trials. Each trial's timing scheme consisted of a fixation of 2 s, cue time of 1.25 s, followed by a period of a MI of 4 s.

In this study, 1.25 s period of EEG data after the visual cue in each trial was chosen as the experimental data. These data were then represented into 3D representation without any further preprocessing. With regard to each cropped EEG data, we presented the corresponding label in a one-hot-vector format.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2938295, IEEE Transactions on Neural Systems and Rehabilitation Engineering

TABLE II
COMPARISON OF CROSS-VALIDATION RESULTS OF NETWORKS WITH DIFFERENT ARCHITECTURES

Subject	Multi-branch 3D CNN	SRF 3D CNN	MRF 3D CNN	LRF 3D CNN
1	77.397	73.738	76.112	74.743
2	60.140	57.079	58.048	54.936
3	82.927	81.172	82.559	80.405
4	72.288	70.439	70.748	66.484
5	75.836	73.990	75.117	72.173
6	68.988	66.355	67.402	62.711
7	76.036	75.265	73.637	70.754
8	76.855	76.425	75.226	74.369
9	84.665	81.996	82.647	82.188
Mean	75.015	72.940	73.500	70.974
Standard deviation	7.344	7.667	7.597	8.570
p-value (Multi-branch)	-	3.43E-04	1.14E-04	6.18E-05
p-value (SRF)	3.43E-04	-	2.19E-01	1.38E-02
p-value (MRF)	1.14E-04	2.19E-01	-	8.08E-04
p-value (LRF)	6.18E-05	1.38E-02	8.08E-04	-

p-value ("NETWORK") means the significant difference (p-value) of network performance between 3D CNN network with "NETWORK" architecture and each other 3D CNN network with different architecture.

The p-values shown in each experimental results were calculated from two-tailed paired t-test. For the evaluations using 10-fold cross-validation, the combination of training and testing dataset of BCI IV dataset 2a was first randomly divided into ten subsets of equal size. In each run, nine subsets were used as training data and a single subset was used as the validation data. The final accuracies reported in the following evaluations were all obtained by averaging fifty results with different model initializations.

B. 3D CNN Architecture Evaluation

1) Four-classes MI Classification Experiments

SRF, MRF, LRF and Multi-branch 3D CNN proposed above were evaluated in TABLE II through 10-fold cross-validation.

Comparing SRF, MRF and LRF with each other, it can be found that networks with different receptive field sizes always perform differently on the same subject. The result also indicates that one branch network cannot always perform well on all subjects. For example, SRF was the best-performing network for subject 7, but it was the worst-performing network for subject 1. This means, the most suitable receptive field size is a highly subject dependent factor. For single-branch networks such as SRF, MRF and LRF, we need to change the network's parameters for different subjects to improve the performance. As the network receptive field size in time domain grow wider, the classification accuracy starts significantly dropping down, such as LRF network which is 2.526% lower than MRF network on the mean value with p-value < 0.01.

Comparing three single-branch networks with the multi-branch network, we found that, unlike single-branch networks, the multi-branch network can always significantly outperform other methods on all subjects. From an overall perspective, the multi-branch network also achieved higher mean and lower standard deviation of different subjects with p-values < 0.01. This means, the multi-branch network is more robust and more effective than other three single-branch networks.

2) Learning Progress Visualization

In order to get an inside view behind the learning processes of different networks, four different 3D CNN networks (SRF network, MRF network, LRF network and Multi-branch network) losses (the Negative log-likelihood cost) and accuracies of all subjects on the experimental testing dataset were monitored for thirty training epochs. The results are shown in Fig. 3.

For subject 3 and subject 9, it can be found that three branch networks all encountered slight overfitting issues since the 7th epoch by observing the loss value. In contrast, the classification loss of the multi-branch network first decreases a little and then remains at an almost fixed level during the training process which do not show any obvious signs of overfitting.

For the rest subjects, unlike subject 3 and subject 9, the classification loss of each branch network increases sharply since around the 4th epoch and reaches a level which is significantly higher than the level in the beginning of the training process. This indicates a severe overfitting issue which is more serious than the issues happened on subject 3 and subject 9. However, the multi-branch network still performs well during the training process of without any obvious signs of overfitting. It is worth to mention that, after overfitting, the prediction accuracy will not significantly drop as the loss rises a lot, but will stabilize around its historic peak. According to this phenomenon, we can adjust the training strategy such as abandoning early stopping strategy, etc.

Overall, the results in this part of experiments indicate that, for single-branch 3D CNN, such as SRF, MRF and LRF, the overfitting issue does exist and the severity of the issue is highly subject dependent. However, the multi-branch structure shows its resistibility to overfitting on different subjects, which leads to an improved performance.

C. The Influence of Filtering

In the experiments mentioned earlier, we used the raw EEG data which have only been filtered with a band-pass filter (0.5-100 Hz) and a notch filter at 50 Hz. To analyze filtering influence on the MI classification framework proposed in this study, we evaluated the multi-branch 3D CNN on a dataset which had been further low-pass filtered at 38 Hz and an another dataset which had been further band-pass filtered between 4 Hz and 38 Hz. The performance of multi-branch 3D CNN networks with different kinds of filtering range were analyzed in TABLE III through 10-fold cross-validation.

The results, presented in TABLE III, indicates that the classification performance of the multi-branch 3D CNN network is not changed by further low-pass filtering at 38 Hz, but changed significantly by further low-pass filtering at 4 Hz with p-value<0.01. This means, the proposed framework has a slack demand for low-pass filtering and the corresponding cutoff frequency should be above 38 Hz to ensure the performance. However, the framework classification accuracy is significantly degraded by 12.141% with p-value < 0.01 when implementing band-pass filtering between 4Hz and 38Hz. This means, high-pass filtering should be used with great caution when applying the proposed 3D CNN framework for MI

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2938295, IEEE Transactions on Neural Systems and Rehabilitation Engineering

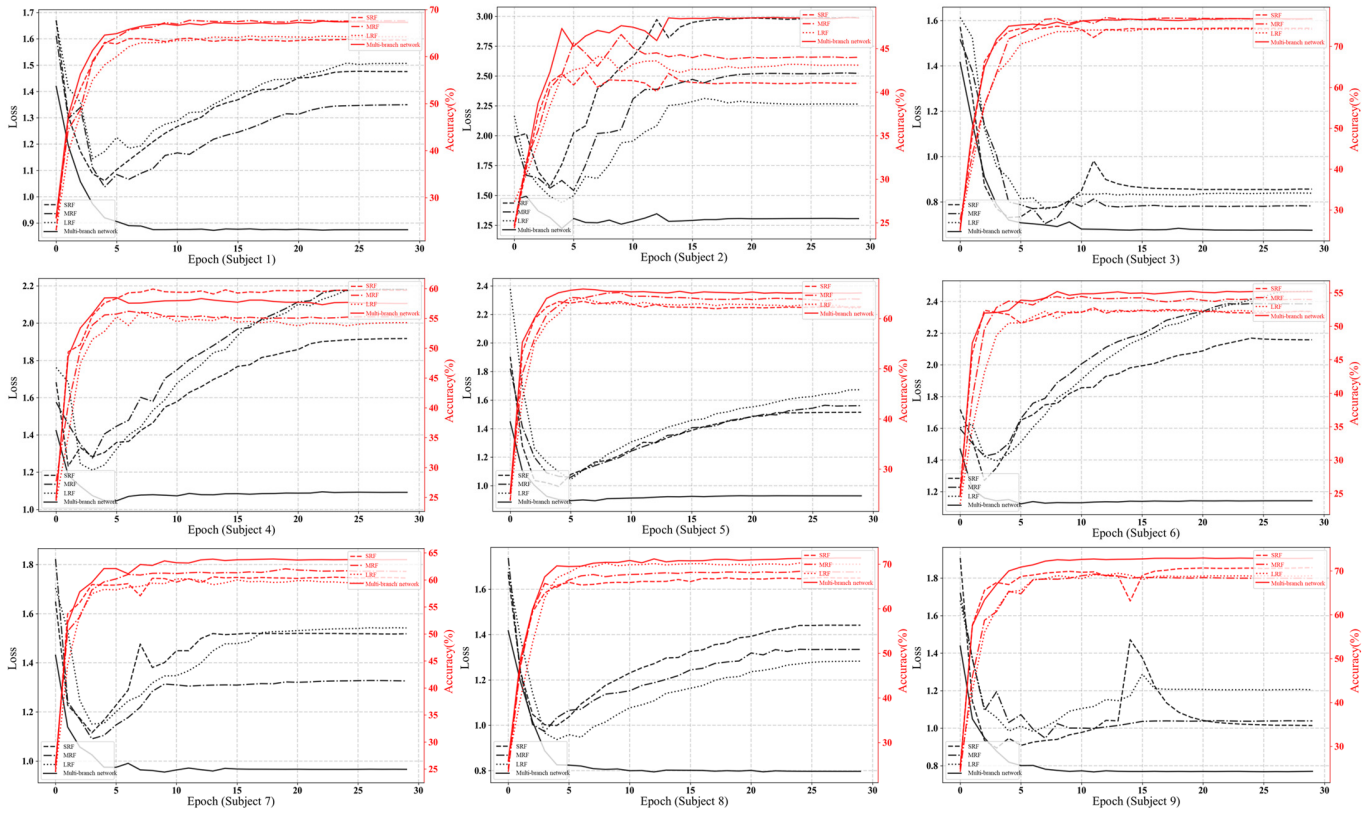

Fig. 3. The test losses and test accuracies of all nine subjects during thirty epochs.

TABLE III
COMPARISON OF CROSS-VALIDATION RESULTS OF DIFFERENT FILTERING METHODS

Subject	Frequency range (Hz)			
	0.5-38	4-38	0.5-4	Raw (0.5-100)
1	77.122	64.475	68.592	77.397
2	59.961	39.770	60.318	60.140
3	82.835	73.486	73.196	82.927
4	71.729	61.251	66.668	72.288
5	75.851	55.121	75.354	75.836
6	68.528	55.948	63.538	68.988
7	76.265	69.257	73.341	76.036
8	77.434	67.576	66.294	76.855
9	84.440	78.008	71.968	84.665
Mean	74.907	62.766	68.808	75.015
Standard deviation	7.432	11.465	5.043	7.344
p-value (0.5-38)	-	1.07E-04	4.12E-03	3.82E-01
p-value (4-38)	1.07E-04	-	8.14E-02	1.08E-04
p-value (0.5-4)	4.12E-03	8.14E-02	-	3.56E-03
p-value (0.5-100)	3.82E-01	1.08E-04	3.56E-03	-

p-value ("RANGE") means the significant difference (p-value) of network performance between dataset filtered with "RANGE" range and each other dataset filtered with different range.

classification, but an adequate high-pass filtering such as filtering with cutoff frequency greater than 0.5 Hz can make the performance of the framework reach an acceptable level.

D. MI Classification with Fewer Sampling Electrodes

Full-set EEG sampling electrodes are difficult to achieve in the practical application of BCI system. The research shows that BCI system with fewer electrodes which cover the motor cortex and sensorimotor cortex can also achieve good performance [17]. Considering this, the classification performance of proposed framework with fewer sampling electrodes was studied in this section.

We first designed four kinds of sampling electrodes selection mode which named A-type, B-type, C-type and D-type. A-type, B-type, C-type all had nine sampling electrodes, but the electrodes distribution of A-type was sparser than the ones of B-type and C-type. Although B-type and C-type both had almost the same electrodes distribution shape, B-type focused more on the electrodes at the front and C-type focused more on the ones at the back. D-type was designed as a complementary research for C-type to study if the performance of proposed framework can be improved on account of a denser electrodes distribution. The input representation form and the corresponding parameters of 3D CNN were both adjusted for different kinds of electrodes selection mode. With regard to the input representation, following the rules described in Sec.II.A, each frame of EEG is transformed into a 2D array according to the general spatial distribution of sampling electrodes and meanwhile padded where there is no electrode with 0. The detailed information is shown in Fig. 4 and TABLE IV. The performance of proposed framework with different sub-set sampling electrodes selection mode was evaluated through 10-fold cross-validation in TABLE V.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2938295, IEEE Transactions on Neural Systems and Rehabilitation Engineering

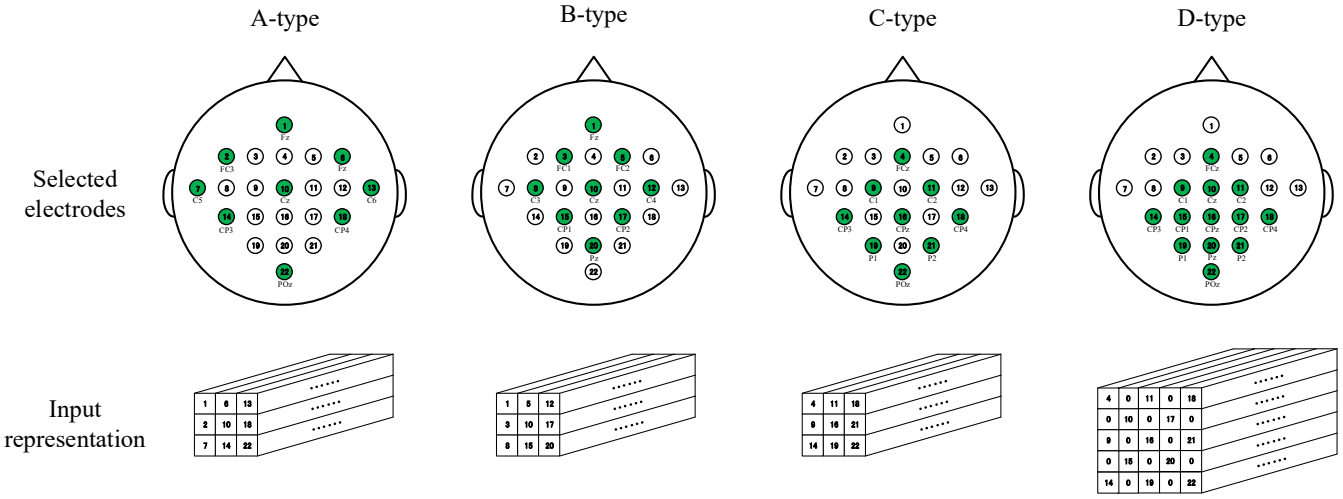

Fig. 4. Four kinds of sampling electrodes selection mode and the corresponding input representation methods. Circles filled with green represent the electrodes which are selected.

TABLE IV
THE MULTI-BRANCH 3D CNN NETWORK PARAMETERS WITH REGARD TO DIFFERENT SAMPLING ELECTRODES SELECTION MODE

Convolutional Layer num.	Branch Type	A-type& B-type& C-type		D-type	
		Shape	Stride	Shape	Stride
1st	Shared layer	2×2×5	1,1,4	3×3×5	2,2,4
2nd	SRF	2×2×1	2,2,1	2×2×1	2,2,1
	MRF	2×2×3	2,2,2	2×2×3	2,2,2
	LRF	2×2×5	2,2,4	2×2×5	2,2,4
3rd	SRF	2×2×1	2,2,1	2×2×1	2,2,1
	MRF	2×2×3	2,2,2	2×2×3	2,2,2
	LRF	2×2×5	2,2,4	2×2×5	2,2,4

Comparing full-set with different electrodes selection modes, it can be found that the mean classification accuracy of the full-set is significantly higher than the ones of A-type and B-type with p-values < 0.01, but the mean accuracies of full-set, C-type, and D-type have no obvious differences. The p-value between full-set and C-type and the p-value between full-set and D-type are both greater than 0.05, too. This means, the proposed framework with fewer sampling electrodes such as C-type and D-type can also reach a competitive MI classification accuracy, which can be employed in the practical applications of BCI system. Note that, comparing with the proposed framework with C-type input, a denser sampling electrodes such as D-type will significantly improve the MI classification performance with p-value < 0.01, but will bring more calculation. This also indicates that a more radical electrodes selection mode of which electrodes are less than C-type within this range should be applied with more caution. Furthermore, for practical application, only classification accuracy is not enough for deciding which method is better, time consuming also need to be considered. We will discuss the electrodes selection strategy further in the next section.

E. Time Consuming Analysis

For evaluating the practical performance of the BCI system which is based on the proposed framework, the training time and testing time of several frameworks with different settings (i.e., SRF network, MRF network, LRF network, multi-branch 3D CNN network with C-type sampling electrodes, multi-branch 3D CNN network with D-type sampling electrodes and multi-branch 3D CNN network with full-set sampling electrodes) were analyzed in Fig. 5. In details, the time consuming of fifteen training epochs on one subject whole training dataset (about twenty-one thousand samples) was recorded as the training time. Similarly, the time consuming of predicting MI classes through the described testing strategy on one subject whole testing dataset (about twenty thousand samples) was recorded as the testing time.

TABLE V
COMPARISON OF CROSS-VALIDATION RESULTS OF MULTI-BRANCH 3D CNN NETWORKS WITH DIFFERENT SAMPLING ELECTRODES SELECTION MODES

Subject	Full-set	A-type	B-type	C-type	D-type
1	77.397	66.344	55.865	74.691	78.848
2	60.140	54.766	48.654	61.367	61.659
3	82.927	67.690	68.602	80.274	81.418
4	72.288	67.612	57.992	71.421	71.834
5	75.836	72.004	65.920	68.486	71.549
6	68.988	61.554	54.734	68.185	68.455
7	76.036	69.157	68.542	74.436	76.332
8	76.855	72.884	62.260	77.422	78.417
9	84.665	78.752	78.037	84.341	85.875
Mean	75.015	67.862	62.289	73.403	74.932
Standard deviation	7.344	6.859	8.964	6.945	7.335
p-value (Full-set)	-	4.46E-04	2.88E-05	9.15E-02	9E-01
p-value (A)	4.46E-04	-	5.56E-03	4.6E-03	9.87E-04
p-value (B)	2.88E-05	5.56E-03	-	1.96E-04	8.65E-05
p-value (C)	9.15E-02	4.6E-03	1.96E-04	-	8.77E-03
p-value (D)	9E-01	9.87E-04	8.65E-05	8.77E-03	-

p-value ("MODE") means the significant difference (p-value) of network performance between multi-branch network with "MODE" sampling electrodes and each other multi-branch network with different electrodes selection mode.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2938295, IEEE Transactions on Neural Systems and Rehabilitation Engineering

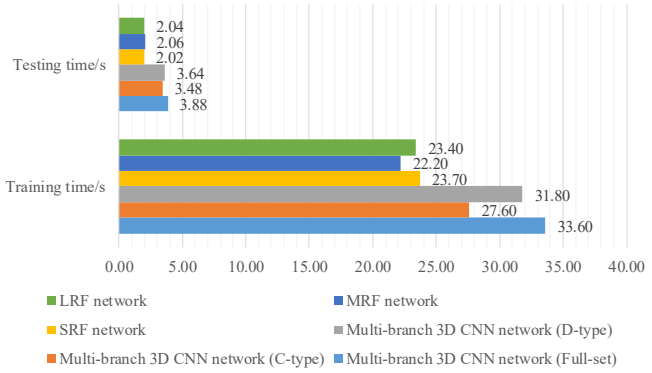

Fig. 5. Training time and testing time of different networks

As shown in the experimental results, the time consuming of multi-branch network with full-set sampling electrodes is the longest one, but we think 33.6 s training time and nearly 1.94E-04 s testing time per sample are both in an acceptable range, which can ensure the quick configuration and the real-time performance of BCI system. Due to more computation, the multi-branch network consumes much more time than single-branch networks during both training and testing process. Besides, considering in conjunction with the results in Sec.III.D, for the requirement of classification accuracy, the multi-branch networks with full-set electrodes is the optimal choice. For the requirement of quick set-up, real-time performance and relatively loose classification precision, the multi-branch networks with C-type and D-type sampling electrodes selection modes can both be good choices. Between them, C-type input can provide shorter training time and be more real-time but will lead to a slightly worse performance than D-type.

F. Comparison with the state of the art

We introduced three state of the art MI classification methods in the literature and compared these methods with the multi-branch 3D CNN in TABLE VI. Cohen's kappa coefficient [40] value which was a common indicator for BCI IV dataset 2a was used to evaluate the performance of different networks. The kappa values reported in TABLE VI which were in decimal forms were all obtained by averaging fifty results with different model initializations. The classification results were calculated on the testing dataset of each subject respectively using the weights trained on the training dataset of the corresponding subject. Kappa value is defined as (1) where P_o is the proportion of observed agreement and P_e is the probability that agreement is due to chance.

$$K_{appa} = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

Here we briefly describe some of the details of the state of the art methods which have been referred in this study.

FBCSP: FBCSP [12] first uses a group of band-pass filters and CSP algorithm to extract the optimal spatial features from the subject-specific frequency band and then train a classifier to output the final classification results. FBCSP was the best performing method for the BCI competition IV dataset 2a in the competition, and it also won other similar EEG decoding competitions BCI IV dataset 2b [41].

TABLE VI
COMPARISON OF THE KAPPA VALUES OF MULTI-BRANCH 3D CNN AND THE STATE OF THE ART

Subject	Multi-branch 3D CNN	FBCSP [12]	TSSM+SVM [19]	C2CM [21]
1	0.699	0.68	0.77	0.833
2	0.459	0.42	0.33	0.537
3	0.788	0.75	0.77	0.87
4	0.594	0.48	0.51	0.556
5	0.647	0.40	0.35	0.5
6	0.538	0.27	0.36	0.273
7	0.653	0.77	0.71	0.861
8	0.702	0.76	0.72	0.778
9	0.713	0.61	0.83	0.727
mean	**0.644** (+8.4%) (-2.3%) (+12.8%)	0.571	0.594	0.659
Standard deviation	**0.100** (-51.5%) (-51.0%) (-45.7%)	0.184	0.206	0.204

The percentages in different colors mean the corresponding percentage changes between multi-branch 3D CNN and other state of the art methods.

TSSM+SVM: TSSM+SVM [19] first uses a bilinear sub-manifold learning (BSML) algorithm to reduce the SPD matrices space dimensionality in motor imagery BCIs which can be considered as an extension of CSP on covariance matrices in measure of Riemannian distance. Then tangent space of sub-manifold (TSSM) is used to function on the extracted Riemannian sub-manifold. Finally support vector machine (SVM) is applied for classification. This method exhibits strong robustness against a small training dataset, which often occurs in BCI studies.

C2CM: C2CM which is proposed in [21] adopts FBCSP as the data preparation method and use a CNN to extract features. This method outperforms FBCSP, but it needs to change its parameters for different subjects.

The Comparison results are presented in TABLE VI. Overall differences between multi-branch 3D CNN and the state of the art are given in percentages with different colors (i.e., the percentages in red color means the differences between multi-branch 3D CNN and TSSM+SVM). The percentage change calculation equation is defined as (2) where P_d is the difference percentage, I_m is the performance indicator of multi-branch 3D CNN and I_s is the performance indicator of the state of the art.

$$P_d = \frac{I_m - I_s}{I_s} \times 100\% \quad (2)$$

Comparing the multi-branch 3D CNN with TSSM+SVM and FBCSP, it can be found that the multi-branch 3D CNN noticeably outperforms other two methods in most subjects and achieves 8.4% and 12.8% increase in average kappa value and meanwhile 51.5% and 45.7% decrease in standard deviation of different subjects respectively. Comparing the multi-branch 3D CNN with C2CM, it can be found that the multi-branch 3D CNN reaches almost the same average kappa value of C2CM. Whereas the multi-branch 3D CNN achieves a much lower

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2938295, IEEE Transactions on Neural Systems and Rehabilitation Engineering

TABLE VII
THE RESULTS OF EXCLUDING TEST SUBJECT EEG DATA FROM TRAINING DATASET

Subject	Multi-branch 3D CNN*		Multi-branch 3D CNN**
	Kappa	Accuracy (%)	Kappa
1	0.424	49.512	0.699
2	0.341	40.743	0.459
3	0.577	64.509	0.788
4	0.376	44.566	0.594
5	0.477	54.292	0.647
6	0.338	40.469	0.538
7	0.518	58.875	0.653
8	0.527	59.755	0.702
9	0.497	56.836	0.713
mean	0.453	52.173	0.644

Multi-branch 3D CNN* means the multi-branch 3D CNN model trained on the training dataset excluding test subject EEG data, Multi-branch 3D CNN** means the multi-branch 3D CNN model trained on the test subject training dataset.

standard deviation of different subjects, which means the multi-branch 3D CNN performs more robustly on different subjects.

G. The Ability of Overcoming Individual Differences

For the application of BCI system, it is important to build a stable classification model which can overcome individual differences. Once this kind of model is obtained, any subjects can directly use this model to carry out MI-related tasks without any pre-trainings in advance.

To verify if the framework proposed in this study has the ability of overcoming individual differences, in this section, we in turn choose one test subject from all nine subjects and combine the others training EEG data as the training dataset to train the weights of multi-branch 3D CNN network. The classification performance which are tested on each subject testing dataset are shown in TABLE VII. The results were all obtained by averaging fifty results with different model initializations.

Comparing the kappa results of model trained on the training dataset excluding test subject EEG data with model trained on the test subject corresponding training dataset, it can be found that excluding test subject EEG data from training dataset will cause the framework classification performance degrades significantly, which means the individual differences among different subjects influence the framework performance largely. However, when exclude the test subject EEG data from training dataset, multi-branch 3D CNN can still achieve effective classification accuracies (much more than 25% which is the random pick result in the four-class classification case) on all nine subjects. This means the proposed framework does have the ability to overcome individual differences in MI classification tasks.

IV. DISCUSSION

A. Visualization of Intermediate Features

Visualizing intermediate features constructed by multi-branch 3D CNN network can give us a view into how the network extracts spatial and temporal features step by step and why the proposed framework can achieve a better classification performance. Considering this, in this section, we visualized the intermediate features of multi-branch 3D CNN network with C-type input. The network was trained on the training dataset of subject 3 in advance. C-type EEG signals of subject 3 were chosen as the example input because this kind of input has fewer sampling electrodes and the corresponding network also reaches a relatively acceptable performance level, which can assist the experiment in yielding a more concise and easy-to-understand result.

In order to display the three-dimension input and four-dimension features constructed by 3D CNN adequately, we transformed these information into a set of two-dimension oscillograms as shown in Fig. 6. For example, the first graph in Fig. 6 shows the EEG signals represented by input representation method. The distribution of the subgraphs represents the first two dimensions which is the spatial distribution of nine electrodes. Through a similar method, in the following graph of the first layer, nine series of features, which are obtained by sliding a group of 3D convolutional filters along each dimension of the input, are also transformed into nine oscillogram subgraphs. The distribution of the subgraphs represents the spatial distribution of extracted features. Various channels in each subgraph means the features extracted by various filters. All the subsequent graphs follow the same transformation rules.

Several interesting properties can be seen from the visualization results shown in Fig. 6.

Firstly, by observing the input signals, it can be found that the EEG signals sampled from different electrodes are slightly different. These differences which are related to electrodes spatial distribution can be considered as one kind expression form of EEG signals spatial features. The EEG signals sampled from closer electrodes are also more similar in time domain. This similarity reflects the localized characteristic in spatial features of EEG signals.

Secondly, the first layer (i.e. the shared layer) acts as more like a collection of various waveforms under the complementary of different filters. Almost all of the waveform information in EEG signals are retained and some waveform information such as wave troughs present as wave crests in the extracted features. Meanwhile, there still exist similarity between features with close spatial distribution.

Thirdly, as layer going higher, the extracted features become increasingly abstract and less visually interpretable. In the time domain, 3D CNN filters begin to give higher activation values to some higher-level concepts such as certain wave crests and high-frequency features. This process can be recognized as the extraction of EEG temporal features. At the same time, the localized characteristics of features exist in the previous layers begin to disappear and the uniqueness of features from different spatial position becomes more obvious. For example, in the second layer of three branch networks, features from four different positions have more apparent differences than the previous layer. This process can be recognized as the extraction of EEG spatial features. In the third layer, the relationship between represented EEG and features can hardly be seen visually. These presentations carry increasingly less information about the visual contents of EEG signals, but increasingly more information related to classification result.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2938295, IEEE Transactions on Neural Systems and Rehabilitation Engineering

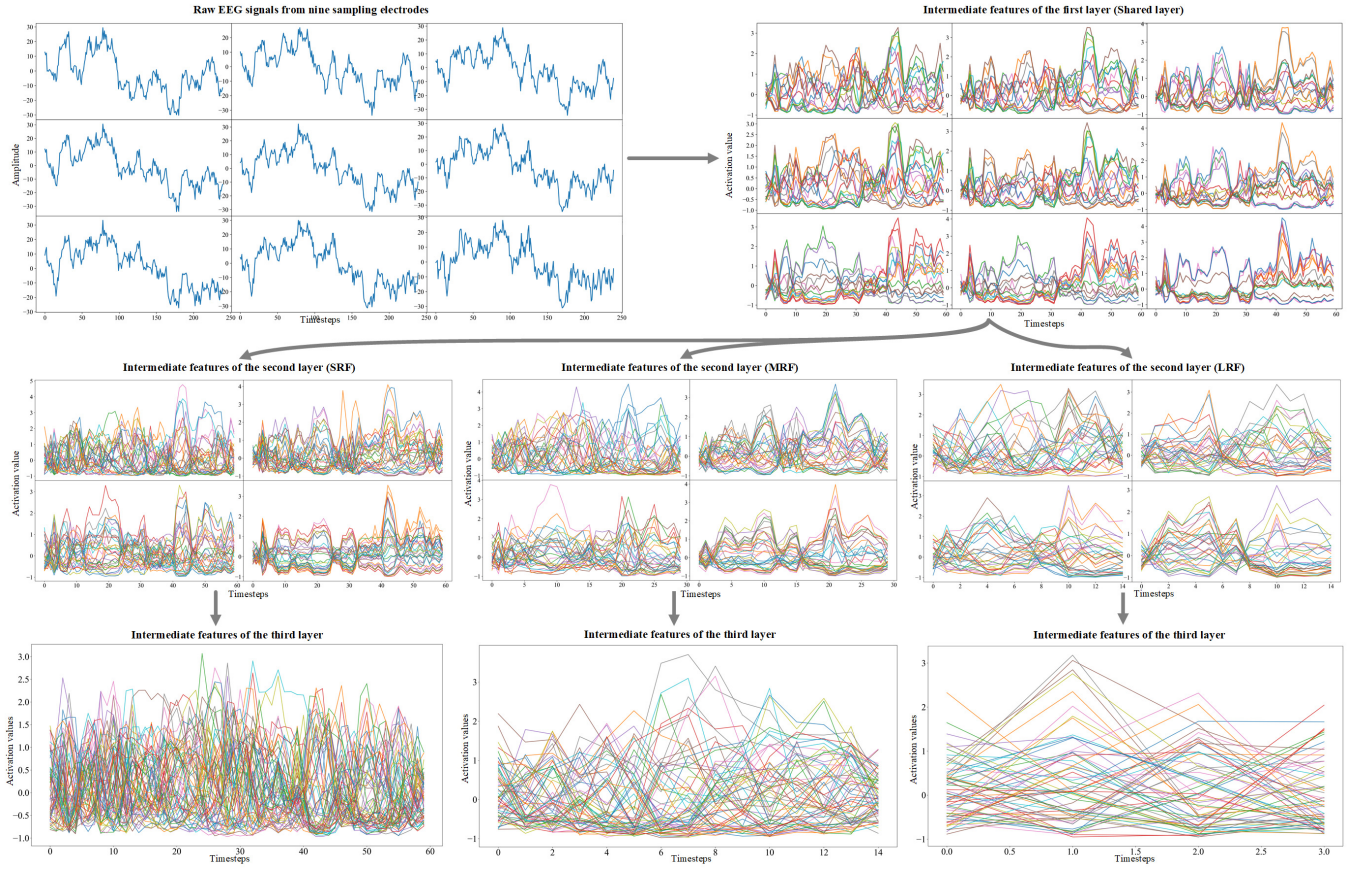

Fig. 6. EEG signals from nine C-type sampling electrodes (the first graph) and the features extracted by multi-branch 3D CNN network (the following graphs)

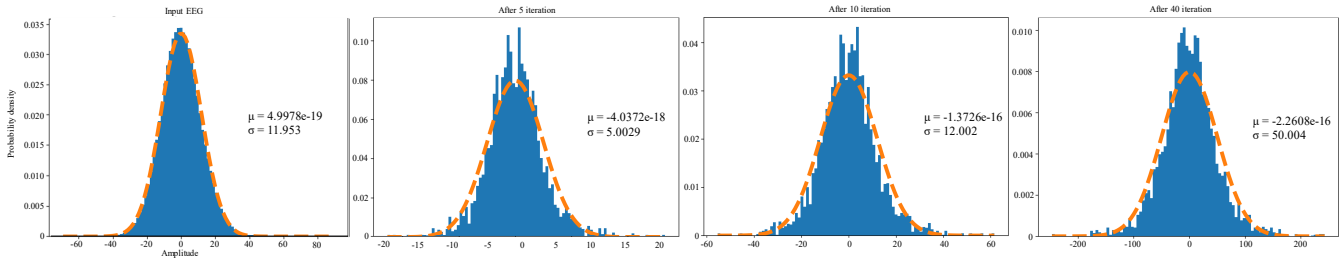

Fig. 7 Distribution histograms of input dataset and estimated signals at different iteration steps.

B. Visualization of 3D CNN Filters

In addition to learning how multi-branch 3D CNN network extracts spatial and temporal features in the previous section, it is also important to know which kinds of features the framework is likely to give more activation values, which can let us gain a deeper understanding of the network mechanism. For this purpose, we displayed the signals that each filter was meant to respond to, which is equivalent to inspecting the filters learned from 3D CNN. Gradient descent was applied to the value of the input signal of 3D CNN so as to maximize the activation of a specific 3D CNN filter from the last 3D CNN layer. The EEG data of subject 3 was used as experimental data and the corresponding pre-trained multi-branch 3D CNN network with C-type input was used for achieving a more concise and easy-to-understand result. Note that the input signal was first zero initialized and the current estimated signal was subtracted by the mean value of itself after each gradient update.

An interesting phenomenon is shown in Fig. 7. The preprocessed 3D CNN input dataset which is a combination of training dataset and testing dataset obeys a Gaussian distribution, as well as the signals which are estimated by gradient descent. As the iteration steps growing, the standard deviation of estimated signals are continuously growing. A meaningful estimated signal obtained through gradient descent should obey the same Gaussian distribution as the input dataset obeys. Because of this, during the gradient descent process, we kept training the input signals until the standard deviation of the restored signal reached the same level as the input dataset. The estimation results of 3D CNN filters are shown in Fig. 8. To show the results more clearly, all the estimation signals are converted to spectrum form.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2938295, IEEE Transactions on Neural Systems and Rehabilitation Engineering

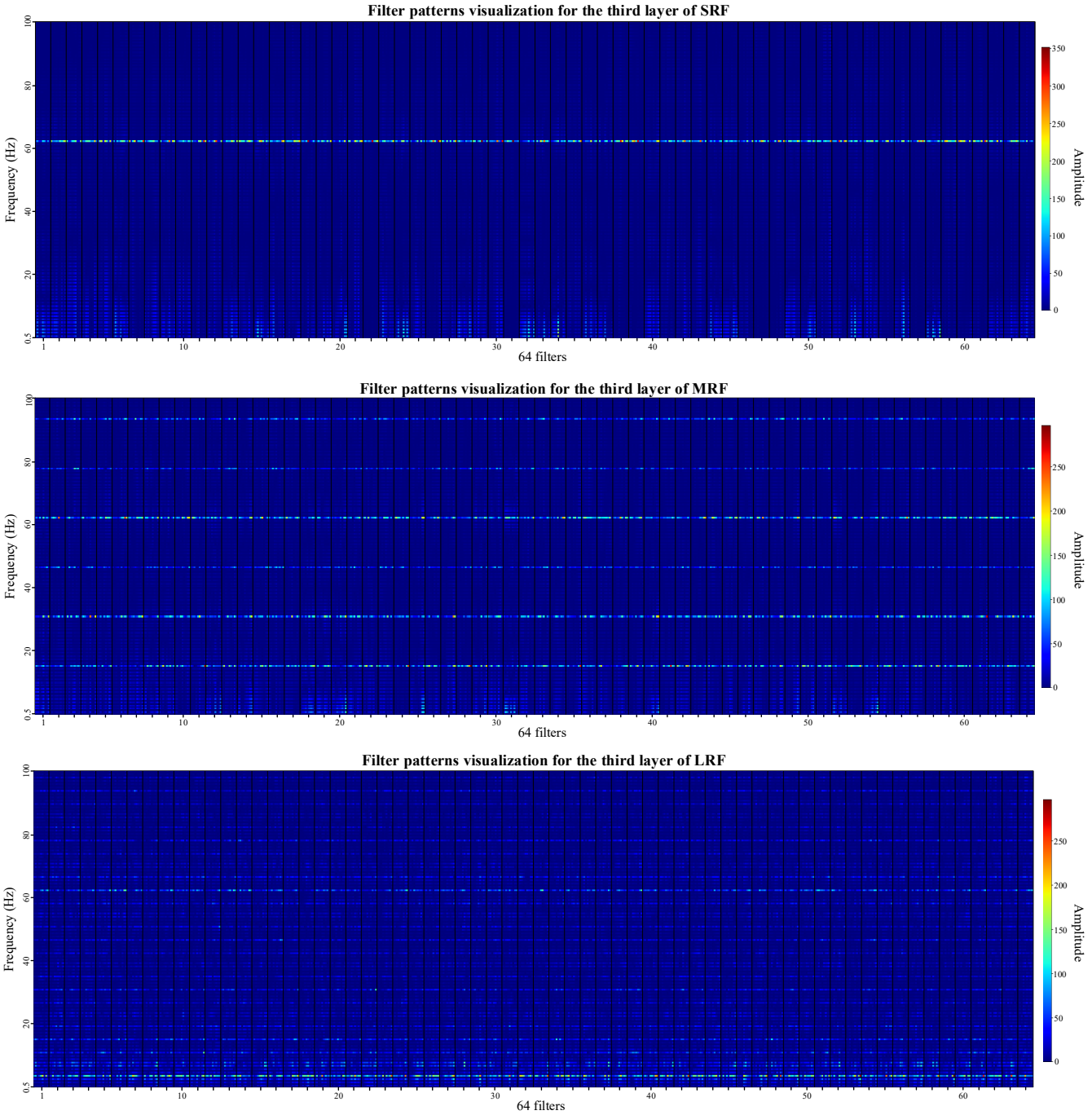

Fig. 8. Visualization of all 3D CNN filters frequency spectrum patterns from the last layers of SRF branch (shown in the graph in the 1st row), MRF branch (shown in the graph in the 2nd row) and LRF branch (shown in the graph in the 3rd row) based on the weights trained on subject 3. Each graph has sixty-four columns (separated by black lines) which refer to sixty-four filters. Each filter has nine columns of pixels which means the frequency spectrums of nine sampling electrodes. These nine electrodes are FCz, C2, CP4, C1, CPz, P2, CP3, P1 and POz from left to right in each column. The color of each pixel represents the amplitude of corresponding frequency.

As can be seen, the filters spectrum patterns of each branch network are all different after training, which means each filter is seeking to find a unique pattern in EEG signal which is crucial for MI classification. Besides, influenced by parameters (size and stride) and architecture of 3D CNN, the filters can only focus on signals in certain frequency bands when extract MI related features. As the network becomes more sophisticated, the receptive frequency bands become denser. For examples, SRF mainly focuses on the frequency around 62.5 Hz, but MRF can simultaneously focus on about five different specific frequencies. LRF focuses on even more kinds of receptive frequencies, especially the frequencies lower than 20 Hz. With the help of denser receptive frequency bands, 3D CNN can learn features from more kinds of frequencies. However, this property cannot significantly improve the network performance, sometimes even degrades the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2019.2938295, IEEE Transactions on Neural Systems and Rehabilitation Engineering

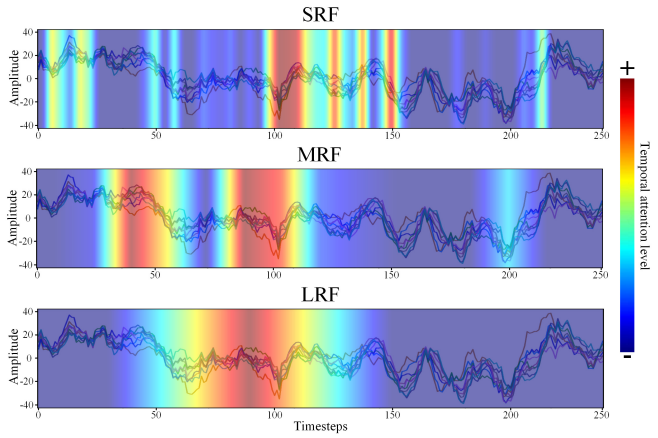

Fig. 9. Grad-CAM visualizations for three branch networks (SRF, MRF and LRF). The red regions correspond to more important EEG signals for Hand (left) MI class of subject 3. The superimposed EEG has nine C-type channels of signals which were plotted in different colors.

performance, which can be seen from the experimental results in Sec.III.B.1) that LRF perform worse than SRF and MRF. This seems to be caused by the noise signals in certain frequency bands which affect the networks abilities of features extraction and generalization. Even so, combining different single branch networks can substantially improve the framework performance, which can also be seen from Sec.III.B.1). Besides, considering in conjunction with the experimental results in Sec.III.B.2), the complementary of different receptive frequency bands of three branch networks is also one of the reasons for alleviating overfitting issues. It is worth mentioning that through a limited research we find the number of filters will largely affect the network final performance. This phenomenon can be explained by the differences among filters, which means more filters can cover more extracted feature types, but will inevitably slow down the whole MI classification framework and bring about overfitting problems.

Overall, based on the phenomenon described above, further optimization methods of 3D CNN employed in EEG-based MI classification area can be carried out in future study, such as adjusting the parameters or architectures of 3D CNN to meet the frequency characteristics of specific MI classification tasks.

C. *Temporal Attention Ranges of Different Branch Networks*

To study the temporal relationship between different branch networks, in this part, we adopted Grad-CAM [39] to visualize the temporal attention of each branch network in the multi-branch 3D CNN. Concretely, for each branch network, the gradients of the score (before the soft-max) for MI class were first computed with respect to the feature map of the last convolutional layer. Then the gradients of each feature map were global averaged to obtain the importance weight of the corresponding feature map and a weight combination followed by a ReLU was implemented to obtain a coarse heat map for showing the branch network's temporal attention. Finally, the coarse heat map was up-sampled to input EEG's temporal resolution by using bi-linear interpolation and superimposed with the input EEG.

One Grad-CAM result of three branch networks with C-type input, presented in Fig. 9, shows that, SRF, MRF and LRF branch focus on different ranges of EEG for MI classification. This phenomenon indicates that the multi-branch 3D CNN 's three branch networks complement each other in time domain. This kind of complementary may improve the whole network's performance.

This phenomenon can also be explained by the finding in the previous section (Sec.IV.B). As LRF focuses more on the frequencies lower than 20 Hz, its temporal attention usually has a wider range than other branches, such as SRF which mostly focuses on the frequency around 62.5 Hz. The signals in attention ranges can be considered as the combinations of signals which are more similar to the 3D CNN filters patterns.

D. *3D CNN framework*

In the field of computer vision, many successful methods such as VGG [33], ResNet [36] all adopt deep architecture. These successful experiences can be transferred to the 3D CNN framework which may lead to an improved performance in MI classification tasks. The depth of the 3D CNN framework is an important configuration which needs to be further studied. It is worth mentioning that the 3D CNN block can be replaced by several 2D CNN blocks or even 1D CNN blocks. This kind of replacement can reduce the number of parameters in the network. Inevitably, it cannot fully utilize the relationship between spatial and temporal information which may influence the final classification performance. In terms of weights initialization, we find that appropriate initial weights in the training process are critical for network's final classification performance in the course of the experiments, which is not detailed in this paper. A better weights initialization method still needs to be further studied.

E. *MI-related frequency band*

At present, brain rhythms related to the imagination of movement are known to be broadly concentrated in the μ (12–16 Hz) and β (18–24 Hz) frequency bands [42]. However, in a recent study [43], results show that there are MI-related features in low frequencies such as 6-12 Hz and these features can be extracted effectively by CSP-based methods. This finding indicates that EEG signals in lower frequency bands may be potentially valuable for MI classification but need a different features extraction method instead of CSP-based method. Several studies which focus on the relationship between low-frequency cortical modulations and MI classification tasks including states of knee extension [44], types of grasping [45] and sitting or standing [46] also achieve effective results. The EEG signals used in these studies are mainly low-pass filtered around the frequency lower than 4 Hz, which indicates the features corresponding to frequencies less than 4 Hz are useful for MI tasks.

In this study, the experimental results in Sec.III.C show that there exists MI-related features not only in the frequencies higher than 6 Hz but also in even lower frequency bands such as 0.5-4 Hz. These features can be extracted out through 3D CNN effectively. For instance，from the classification results of 0.5-4 Hz , it can be found that all subjects achieve relatively good results which mean value is 68.808%. Meanwhile, comparing the classification results of 4-38 Hz with the results of 0.5-38 Hz, it can be found that the framework performance decreases substantially without the information in frequency

band of 0.5-4 Hz. Besides, comparing with the classification results of 0.5-4 Hz with the results of 4-38 Hz, several subjects such as subject 8 and subject 9 achieve better results in 4-38 Hz, but for some subjects such as subject 2, the results of 0.5-4 Hz are better. This means MI-related frequency bands are variable among different subjects. A similar phenomenon about the variability of MI-related frequency bands, which happens in higher frequencies has also been mentioned in [47].

Moreover, for most subjects, the results of 0.5-4 Hz and the results of 4-38 Hz are both worse than the results of 0.5-38 Hz. This means there are correlations which are useful for MI classification between signals of different bands. If EEG signals are divided into different frequency bands and considered separately, the classification results may be affected.

F. Multi-branch architecture

In the multi-branch architecture, we can use not only 3D CNN, but also many other methods as the branch networks, including deep learning methods such as RNN and other combined method such as C2CM which is based on CSP. Specifically, because the temporal features in EEG are not fully considered in the CNN, using RNN as a complementary branch network for the CNN-based multi-branch network may improve the classification performance. However, as the number of the branches increases, the computational cost will correspondingly grow, which will lead to more training time and worse real-time performance.

G. Overfitting issues

In MI classification task, overfitting issue does arise from time to time because of the small training dataset. Early stopping method and cropped training strategy are two common methods to mitigate the overfitting issues [48], [49]. For early stopping method, it needs to separate a validation set from the limited training dataset. This operation will make the training dataset smaller, which may lead to a poorer training effect. Besides, results in Sec. III.B.2) show that, the cropped training strategy cannot avoid overfitting thoroughly. While the multi-branch structure shows great resistibility to overfitting. This structure can also be applied in other MI classification methods which may help to mitigate overfitting to some extent.

V. Conclusion

In this work, 3D representation method and multi-branch 3D CNN were proposed to tackle MI classification tasks. Experimental results show that this framework can achieve good performance in MI classification tasks and significantly improve the robustness on different subjects with appropriate filtering and initial weights. It also demonstrates that the multi-branch structure can mitigate overfitting when the dataset is small and the branch networks in the multi-branch structure can complement each other in time domain. In addition, the proposed MI classification framework is general enough that it can be migrated into other EEG related areas such as seizure detection, cognitive load classification, etc.

References

[1] Pfurtscheller G, Da Silva F H L. Event-related EEG/MEG synchronization and desynchronization: basic principles[J]. Clinical neurophysiology, 1999, 110(11): 1842-1857.

[2] Tang Z, Sun S, Zhang S, et al. A brain-machine interface based on ERD/ERS for an upper-limb exoskeleton control[J]. Sensors, 2016, 16(12): 2050.

[3] V. Bostanov, "BCI competition 2003—Data sets Ib and IIb: Feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram," IEEE Trans. Biomed. Eng., vol. 51, no. 6, pp. 1057–1061, Jun. 2004.

[4] W. Y. Hsu and Y. N. Sun, "EEG-based motor imagery analysis using weighted wavelet transform features," J. Neurosci. Methods, vol. 176, no. 2, pp. 310–318, 2009.

[5] B. Obermaier, C. Neuper, C. Guger, and G. Pfurtscheller, "Information transfer rate in a five-classes brain-computer interface," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 9, no. 3, pp. 283–288, Sep. 2001.

[6] D. P. Burke, S. P. Kelly, P. de Chazal, R. B. Reilly, and C. Finucane, "A parametric feature extraction and classification strategy for braincomputer interfacing," IEEE Trans. Neural. Syst. Rehabil. Eng., vol. 13, no. 1, pp. 12–17, Mar. 2005.

[7] Ramoser H, Muller-Gerking J, Pfurtscheller G. Optimal spatial filtering of single trial EEG during imagined hand movement[J]. IEEE transactions on rehabilitation engineering, 2000, 8(4): 441-446.

[8] R. Zhang et al., "Control of a wheelchair in an indoor environment based on a brain-computer interface and automated navigation," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 24, no. 1, pp. 128–139, Jan. 2016.

[9] P. Wei, W. He, Y. Zhou, and L. Wang, "Performance of motor imagery brain-computer interface based on anodal transcranial direct current stimulation modulation," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 21, no. 3, pp. 404–415, May 2013.

[10] Novi Q, Guan C, Dat T H, et al. Sub-band common spatial pattern (SBCSP) for brain-computer interface[C]//2007 3rd International IEEE/EMBS Conference on Neural Engineering. IEEE, 2007: 204-207.

[11] R. Zhang, P. Xu, L. Guo, Y. Zhang, P. Li, and D. Yao, "Z-score linear discriminant analysis for EEG based brain-computer interfaces," PLoSOne, vol. 8, pp. 1–7, 2013.

[12] Ang K K, Chin Z Y, Zhang H, et al. Filter bank common spatial pattern (FBCSP) in brain-computer interface[C]//2008 ieee international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE, 2008: 2390-2397.

[13] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification, 2nd ed. New York:John Wiley, 2001.

[14] Cortes C. V. Vapnik Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

[15] Yang H, Sakhavi S, Ang K K, et al. On the use of convolutional neural networks and augmented CSP features for multi-class motor imagery of EEG signals classification[C]//2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2015: 2620-2623.

[16] Wu S L , Liu Y T , Hsieh T Y , et al. Fuzzy Integral With Particle Swarm Optimization for a Motor-Imagery-Based Brain–Computer Interface[J]. IEEE Transactions on Fuzzy Systems, 2016, 25(1):1-1.

[17] Ko L W , Lu Y C , Bustince H , et al. Multimodal Fuzzy Fusion for Enhancing the Motor-Imagery-Based Brain Computer Interface[J]. IEEE Computational Intelligence Magazine, 2019, 14(1):96-106.

[18] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Information Theory, vol. 13, no. 1, pp. 21-27, 1967.

[19] Xie X , Yu Z L , Lu H , et al. Motor Imagery Classification Based on Bilinear Sub-Manifold Learning of Symmetric Positive-Definite Matrices[J]. IEEE Trans Neural Syst Rehabil Eng, 2017, 25(6):504-516.

[20] Hema C R , Paulraj M P , Yaacob S , et al. Recognition of motor imagery of hand movements for a BMI using PCA features[C]// International Conference on Electronic Design. IEEE, 2008.

[21] Sakhavi S, Guan C, Yan S. Learning temporal information for brain-computer interface using convolutional neural networks[J]. IEEE transactions on neural networks and learning systems, 2018 (99): 1-11.

[22] Schirrmeister R T, Springenberg J T, Fiederer L D J, et al. Deep learning with convolutional neural networks for EEG decoding and visualization[J]. Human brain mapping, 2017, 38(11): 5391-5420.

[23] Jouppi N P, Young C, Patil N, et al. In-datacenter performance analysis of a tensor processing unit[C]//2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA). IEEE, 2017: 1-12.

[24] Dai M, Zheng D, Na R, et al. EEG Classification of Motor Imagery Using a Novel Deep Learning Framework[J]. Sensors, 2019, 19(3): 551.
[25] Bashivan P, Rish I, Yeasin M, et al. Learning representations from EEG with deep recurrent-convolutional neural networks[J]. arXiv preprint arXiv:1511.06448, 2015.
[26] Luo T, Chao F. Exploring spatial-frequency-sequential relationships for motor imagery classification with recurrent neural network[J]. BMC bioinformatics, 2018, 19(1): 344.
[27] Tabar Y R, Halici U. A novel deep learning approach for classification of EEG motor imagery signals[J]. Journal of neural engineering, 2016, 14(1): 016003.
[28] Naeem M, Brunner C, Leeb R, et al. Seperability of four-class motor imagery data using independent components analysis[J]. Journal of neural engineering, 2006, 3(3): 208.
[29] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1): 221-231.
[30] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
[31] Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. CoRR, abs/1311.2901,2013. Published in Proc. ECCV, 2014.
[32] Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks[J]. arXiv preprint arXiv:1312.6229, 2013.
[33] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[34] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.
[35] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.
[36] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[37] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the thirteenth international conference on artificial intelligence and statistics. 2010: 249-256.
[38] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
[39] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 618-626.
[40] Fleiss J L, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability[J]. Educational and psychological measurement, 1973, 33(3): 613-619.
[41] Tangermann M, Müller K R, Aertsen A, et al. Review of the BCI competition IV[J]. Frontiers in neuroscience, 2012, 6: 55.
[42] Fazli S, Grozea C, Danóczy M, et al. Subject independent EEG-based BCI decoding[C]//Advances in Neural Information Processing Systems. 2009: 513-521
[43] Suk H I, Lee S W. Subject and class specific frequency bands selection for multiclass motor imagery classification[J]. International Journal of Imaging Systems and Technology, 2011, 21(2): 123-130.
[44] úbeda Andrés, Azorín José M, Dario F , et al. Estimation of Neuromuscular Primitives from EEG Slow Cortical Potentials in Incomplete Spinal Cord Injury Individuals for a New Class of Brain-Machine Interfaces[J]. Frontiers in Computational Neuroscience, 2018, 12:3-.
[45] Agashe H A, Paek A Y, Zhang Y, et al. Global cortical activity predicts shape of hand during grasping[J]. Frontiers in Neuroscience, 2015, 9:121.
[46] Bulea T C , Saurabh P , Atilla K , et al. Sitting and standing intention can be decoded from scalp EEG recorded prior to movement execution[J]. Frontiers in Neuroscience, 2014, 8.
[47] Ramoser H, Muller-Gerking J, Pfurtscheller G. Optimal spatial filtering of single trial EEG during imagined hand movement[J]. IEEE transactions on rehabilitation engineering, 2000, 8(4): 441-446.
[48] Tayeb Z, Fedjaev J, Ghaboosi N, et al. Validating deep neural networks for online decoding of motor imagery movements from EEG signals[J]. Sensors, 2019, 19(1): 210.
[49] Bentlemsan M, Zemouri E T T, Bouchaffra D, et al. Random forest and filter bank common spatial patterns for EEG-based motor imagery classification[C]//2014 5th International Conference on Intelligent Systems, Modelling and Simulation. IEEE, 2014: 235-238.