24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Learning Conjunctive Information of Signals in Multi-sensor Systems

Seong-Eun Moon[1] and **Jong-Seok Lee**[1]

Abstract. This paper proposes a novel deep learning method for extraction of the conjunctive information that describes the relationship between signals in multi-sensor systems to enhance the performance of the given classification task. The signals obtained from different sensors included in the multi-sensor systems are closely related. Handcrafted metrics have been used to extract the relationship between the signals in some work, which is hardly optimal for the given task. Our proposed method learns the pair-wise relationship from data to maximize the performance of the given task, which is fully data-driven, multi-aspect, and target-oriented. We demonstrate the effectiveness of the proposed method on a toy example and two real-world problems, i.e., activity recognition using accelerometer signals and emotional video classification using brain signals.

1 Introduction

Multi-sensor systems have been actively utilized for monitoring or tracking the states of the environment and users because of the rich information obtained from them. For example, manufacturing process monitoring [17], action recognition [15], and health-care [11] are conducted by using multi-sensor systems. In most existing deep learning approaches, the signals from multi-sensor systems are considered independently. However, multiple sensors in such systems simultaneously measure the same object or the objects that are closely related. Therefore, the relationship between the obtained signals provides the information from different points of view for the objects' state, which is expected to contribute to the enhanced performance for the given task.

There are several cases where the signals obtained by multiple sensors have close relationship and provide the complementary information to each other. For example, when accelerometers are used for monitoring the human activity [7], the signals captured by these sensors provide the information related to the activity from different perspectives (i.e., body parts). Problems related to the climate are also examples involving multi-sensor systems. The local weather status is a partial observation of the complex system where weather conditions at different regions influence on each other. Thus, the weather forecasting is often conducted based on weather conditions acquired from multiple regions. Brain signals such as the electroencephalography (EEG) signal are also examples, where the relationship between the signals from different brain regions is importantly considered. It is well known that the brain is a functional network, i.e., different brain regions are functionally correlated [12]. The functional relationship between the brain signals measured at different locations, called connectivity, is often used for analysis of the functional network of the brain.

Graphs are a way to represent the relationship between the signals. For example, the sensors in a system can be regarded as the nodes of a graph, then the relationship between the sensors can be encoded as the edges between the nodes. Handcrafted metrics such as the physical distance between sensors [2] and correlation between signals [4] are often used for measuring the relationship. However, these manually designed metrics can measure only particular aspects of the relationship. Moreover, a proper metric needs to be chosen for each task, which may not be even optimal for the task. Therefore, it is necessary to develop a method to design a metric that can reflect various aspects of the relationship between signals and is optimized to solve the given problem.

In this paper, we propose an end-to-end deep learning model to learn the relationship between signals in the multi-sensor systems from data, which is expected to be optimal for the given classification task. The relationship extracted by the proposed method is called *conjunctive information* in the sense that the extracted information connects two signals obtained from different sensors and describes diverse types of relationship between them. Our method is distinctive from the existing methods in that the proposed method 1) extracts the relationship between signals to maximize the performance of the given task in an end-to-end manner in contrast to the existing graph-based approaches, 2) enables to obtain the comprehensive information of the relationship unlike the existing metrics considering only a specific aspect, and 3) fully depends on the given data so that it is applicable to various domains and tasks.

The remainder of this paper is organized as follows. After summarizing the related work in Section 2, we explain the proposed method in Section 3. Section 4 shows the effectiveness of the proposed method based on three experiments, i.e., a toy simulation and two real-world problems. Then, we conclude the paper in Section 5.

2 Related work

Multi-sensor systems have been used in various real-world applications, and deep learning approaches to extract meaningful information from the signals in such systems have been proposed. However, many approaches typically regard the signals in multi-sensor systems separately [10, 18, 19]

In a few studies, manually designed metrics are used for extracting the relationship between EEG signals. In [4], absolute values of the Pearson correlation coefficients (PCCs) between the EEG signals or physical distances between the EEG electrodes are obtained to represent the connection strength between the EEG signals measured from different electrodes for a video identification task. In our previ-

[1] Yonsei University, Republic of Korea, email: {se.moon, jongseok.lee}@yonsei.ac.kr

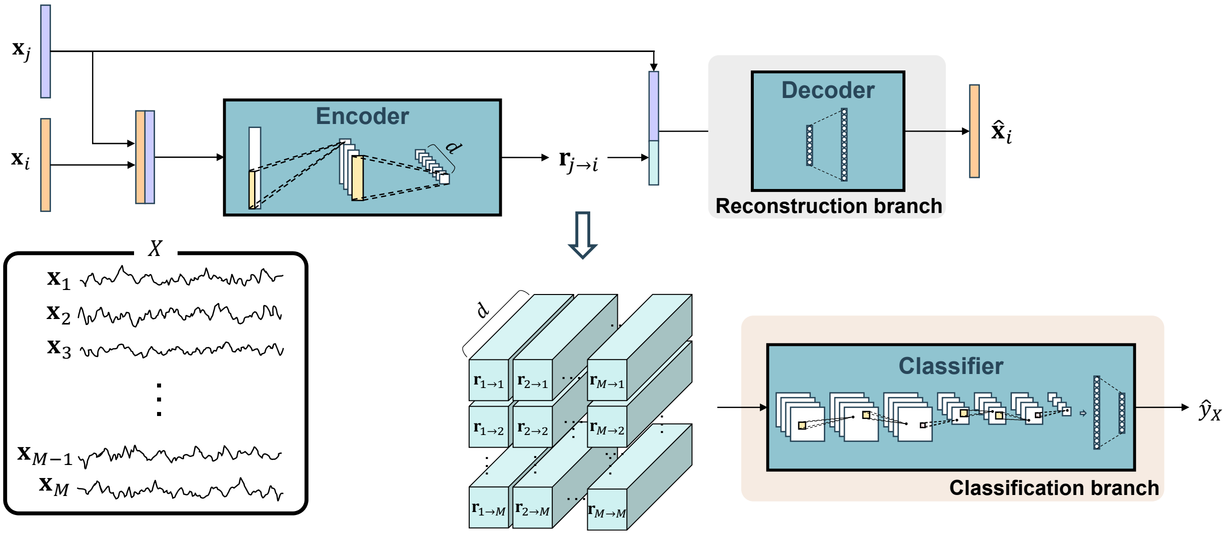


Figure 1: Overview of the proposed method.

ous work [9], PCC, phase locking value (PLV), and phase lag index are used for describing the relationship between EEG signals, and the extracted relational information is applied for emotion classification. Although these approaches are shown to yield better classification accuracy for the target tasks than those without consideration of relationship information, the aforementioned limitations of handcrafted metrics still exist.

A few studies try to infer the relationship between signals based on the graph theory. A variational autoencoder-based graph neural network is proposed to find latent structures of dynamic graphs in [5]. However, only categorical relationships are considered, whereas our proposed method obtains continuous-valued measures of conjunctive information. Moreover, unlike our method, a priori knowledge of the graph structure is required. In [3], a deep learning approach for graph generation is proposed, where the graph generator network is trained by using hypergradients obtained from the classification results of graph convolutional networks. However, this approach targets graph node classification tasks, thus the learned relationship is between data samples from different objects, whereas our method considers the relationship observed within a system. In addition, graph-based approaches usually suffer from high computational and memory complexity [16].

3 Proposed method

Figure 1 illustrates the proposed deep learning method. When a set of time series $X = \{\mathbf{x}_1, ..., \mathbf{x}_M\}$ measured by M sensors is given, the goal of the proposed method is to learn the pair-wise directional relationship between the time series in X and achieve the given task, i.e., prediction of the target label y_X in this work. Particularly, the learned relationship, i.e., the conjunctive information, is expected to reflect various aspects of the relationship between signals rather than only a certain characteristic such as similarity, correlation, and causality.

For two time series $\mathbf{x}_i$ and $\mathbf{x}_j$ obtained from the i-th and j-th sensors, the conjunctive information from $\mathbf{x}_j$ to $\mathbf{x}_i$ is extracted by the encoder E, which can be described as

$$\mathbf{r}_{j\to i} = E(\mathbf{x}_j, \mathbf{x}_i). \tag{1}$$

Here, the conjunctive information is not limited to be a scalar, and its dimension can be made different by adjusting the output size of the encoder. The extracted conjunctive information $\mathbf{r}_{j\to i}$ is concatenated with $\mathbf{x}_j$ in the reconstruction branch, which is used for reconstruction of $\mathbf{x}_i$ by the decoder D, i.e.,

$$\hat{\mathbf{x}}_i = D(\mathbf{x}_j, \mathbf{r}_{j\to i}). \tag{2}$$

Note that the encoder and decoder models are shared across all pairs of signals. The deep learning models that are able to extract the latent relationship from a pair of signals can be used as the encoder, and those that can reconstruct the signals from a given latent vector can be used as the decoder. Most of the supervised learning approaches accord to these conditions, including convolutional neural networks (CNNs) and recurrent neural networks.

The extracted conjunctive information is used for the given classification task in the classification branch. Particularly, we employ a CNN as the classifier, therefore, the conjunctive information values are reshaped to a tensor $\mathbf{R}$. The element of $\mathbf{R}$ at (i, j, k) is defined as

$$\mathbf{R}_{i,j,k} = E_k(\mathbf{x}_j, \mathbf{x}_i), \tag{3}$$

where the subscript k of $E(\cdot)$ indicates the k-th element of the output of the encoder. Therefore, the size of the tensor $\mathbf{R}$ is $M \times M \times d$ when the conjunctive information is a d-dimensional vector. The obtained $\mathbf{R}$ is inputted to the classifier C that predicts the class label of X, which can be written as

$$\hat{y}_X = C(\mathbf{R}). \tag{4}$$

The reconstruction loss between the reconstructed signal ($\hat{\mathbf{x}}_i$) and the original signal ($\mathbf{x}_i$) and the classification loss of the predicted label $\hat{y}_X$ are used as the feedback for training. The weight parameters of the decoder and classifier (w_D and w_C) are learned based on the reconstruction and classification losses, respectively, which can be written as

$$w_D^* = \min_{w_D} \mathcal{L}\left(\mathbf{x}_i, D\left(\mathbf{x}_j, E(\mathbf{x}_j, \mathbf{x}_i)\right)\right) \tag{5}$$

and

$$w_C^* = \min_{w_C} \mathcal{L}\left(y_X, C(\mathbf{R})\right), \tag{6}$$

where $\mathcal{L}(\cdot)$ refers to a loss function between given two inputs.

The training of the encoder is based on both losses, i.e.,

$$w_E^* = \min_{w_E} \mathcal{L}\left(\mathbf{x}_i, D\left(\mathbf{x}_j, E(\mathbf{x}_j, \mathbf{x}_i)\right)\right) + \lambda \mathcal{L}\left(y_X, C(\mathbf{R})\right), \tag{7}$$

where w_E represents the weight parameters of the encoder and λ is a value to control the relative importance of the classification loss. Therefore, the encoder learns the relationship from $\mathbf{x}_j$ to $\mathbf{x}_i$, which provides a clue for the reconstruction of $\mathbf{x}_i$ using $\mathbf{x}_j$ and, at the same time, is related to the target label. In the case of $i = j$, the learning of the conjunctive information becomes fully oriented to the classification task, which enables to extract the target-related features from individual signals.

4 Experiments

We first provide a toy example for better understanding of the learned conjunctive information in the proposed method. Then, the effectiveness of our method is demonstrated based on two real-world problems, i.e., the activity recognition and the emotional video classification.

The reconstruction loss is obtained as a root mean squared error (RMSE) between the reconstructed and original signals, and the cross entropy loss is employed as the classification loss. The value of λ is set to 1 for all experiments, which roughly corresponds to the balanced contribution of the reconstruction and classification losses for the training of the encoder.

The structures of the encoder, decoder, and classifier are designed differently for each problem. In all problems, the rectified linear units (ReLU) activation function is employed for the proposed model except for the last layers of the decoder and classifier, for which the linear activation function and softmax activation function are used, respectively. The Adam optimizer is used for training with a learning rate of 0.0001. The batch size is 300, 256, and 256 for the toy example, activity recognition, and emotional video classification, respectively. The dropout scheme with a probability of 0.5 is applied to the fully connected (FC) layers. The training is stopped if the validation loss does not decrease for successive 20 epochs, and the network that shows the best validation classification performance is selected for the test. The proposed model is implemented in PyTorch. The experiments are conducted using a PC equipped with an Intel Xeon E5 CPU and an NVIDIA Tesla K80 GPU.

4.1 Toy simulation

We define a 3-class classification problem for a two-sensor system as the toy example. A sine wave is consistently measured from the first sensor,

$$\mathbf{x}_1 = \sin t, \tag{8}$$

and the signal obtained from the second sensor differs depending on the class as follows:

$$\mathbf{x}_2 = \begin{cases} 2\sin t & \text{if } y_X = 0 \\ \sin\left(t + \frac{\pi}{2}\right) & \text{if } y_X = 1 \\ n \in \mathcal{N}(0,1) & \text{if } y_X = 2, \end{cases} \tag{9}$$

where $X = \{\mathbf{x}_1, \mathbf{x}_2\}$, and y_X corresponds to the class label of X.

The signals obtained from the two sensors are highly correlated in the first class ($y_X = 0$) and in a causal relationship for the second class ($y_X = 1$). The signal of the second sensor for the third class ($y_X = 2$) is Gaussian random noise, which means that the two sensors are independent. The training, validation, and test of the model are conducted using 3000, 300, and 300 data samples (i.e., 1000, 100, and 100 samples per class), respectively. Each data sample is individualized by adding Gaussian random noise of 5% in amplitude. The length of the samples is 384, and the sampling rate is 5 Hz.

Table 1: Results of the toy example.

y_X	**Direction**	$r_{j\to i}$	**RMSE**	**Accuracy**
0	1 → 2	0.017	1.165	1.000
	2 → 1	0	0.497	
1	1 → 2	1.740	0.699	1.000
	2 → 1	0.158	0.566	
2	1 → 2	0	1.037	1.000
	2 → 1	3.228	0.659	

The encoder is implemented as a CNN that consists of two convolutional layers having eight 1×129 kernels and one 2×128 kernel, respectively. As the number of kernels in the last layer of the encoder is one, the conjunctive information is obtained as a scalar value. A FC layer with 128 hidden neurons is employed as the decoder. The classifier consists of a single convolutional layer having thirty-two 3×3 kernels and a FC layer having 128 hidden neurons.

Table 1 shows the results. The reconstruction errors are obtained as average RMSEs for the test data sequences in the corresponding class. As the conjunctive information $r_{j\to i}$ acts as the supplementary information for cross-signal reconstruction in the reconstruction branch, its value can be interpreted as the amount of information that is required to reconstruct $\mathbf{x}_i$ based on $\mathbf{x}_j$. Note that in all cases, the classification accuracy is 100%.

Overall, the extracted conjunctive information values correspond to the preassigned relationships between the two sensors, which proves that the proposed method is able to measure various aspects of inter-signal relationship. When the two signals are perfectly correlated (i.e., $y_X = 0$), the extracted conjunctive information has small values because the signal is reproducible based on the counterpart signal without much information of their relationship. The smaller value of $r_{2\to1}$ is probably because the decoder more frequently faces $\mathbf{x}_1$ than $\mathbf{x}_2$ and the reconstruction of $\mathbf{x}_1$ is relatively easy. Larger values of conjunctive information are obtained for the second class, where the delay of a quarter period exists, compared to the case of the perfect correlation. For the third class, the conjunctive information from the first sensor to the second sensor shows a small value despite of their independence. This reflects that $\mathbf{x}_2$ is reconstructed without the relationship information with $\mathbf{x}_1$ because $\mathbf{x}_2$ is a random signal. In the opposite direction, the decoder tries to reconstruct $\mathbf{x}_1$ based on the random signal, therefore, the conjunctive information has a large value in this case.

4.2 Activity recognition

4.2.1 Database

The OPPORTUNITY dataset [13] is employed for the activity recognition experiment. Particularly, the signals obtained from twelve accelerometers attached on the subjects' body are used for the classification between *null*, *stand*, *walk*, *sit*, and *lie*, where *null* indicates the unrecognized activities.

The data sequences are normalized within $[-1.0, 1.0]$ for each sensor channel, then segmented into 0.5-second-long sequences with a 50% overlap according to the protocol of the OPPORTUNITY challenge [1]. As a result, the shape of data sequences becomes 36×15, which corresponds to M = (12 accelerometers) $\times$ (3 axes) = 36 and a time length of 0.5 second for a sampling rate of 30 Hz. We exclude the data sequences where signals for two or more sensors are missing. The data sequences are divided into training, validation, and test data based on the protocol [1], which assigns

Table 2: Network architectures for activity recognition.

		Type	**Output shape**	**Kernel size**
E	1	convolution	$2\times9\times8$	1×7
	2	convolution	$1\times1\times d$	2×9
D	1	dense	128	-
	2	linear	15	-
C	1	convolution	$36\times36\times32$	3×3
	2	convolution	$36\times36\times32$	3×3
	3	max-pooling	$18\times18\times32$	2×2
	4	convolution	$18\times18\times32$	3×3
	5	convolution	$18\times18\times32$	3×3
	6	max-pooling	$9\times9\times32$	2×2
	7	dense	128	-
	8	softmax	5	-

24369, 3324, and 6995 data sequences for each, respectively.

4.2.2 Setup

The structures of encoder E, decoder D, and classifier C are described in Table 2. The classification performance is measured in terms of the classification accuracy and weighted F_1 score to consider the imbalanced distribution of the activity classes.

We implement the random forest and k-nearest neighbors (k-NN) for comparison as traditional machine learning methods. For the random forest, the number of trees and the number of features are set to 100 and squared root of the input dimension, respectively. The maximum depth, minimum leaf size, and minimum node size to be split are fine-tuned based on the validation data. The value of k in the k-NN method is set to 5.

In addition, two existing deep learning methods, i.e., the deep convolutional LSTM (DeepConvLSTM) [10] and residual bidirectional LSTM (ResBidirLSTM) [20], are compared with the proposed method. The former employs two LSTM units after four convolutional layers that conduct convolution operations along the time axis to consider the temporal dynamics. The latter is a bidirectional LSTM network with residual connections, which enables to extract the information reflecting forward and backward temporal directions while preventing the gradient vanishing problem. In these methods, the data from all sensors is inputted to the models after concatenation, which can be seen as an early integration scheme that does not explicitly consider relationship between the signals.

4.2.3 Results

Table 3 summarizes the results. Our proposed method outperforms the conventional machine learning methods and deep learning approaches in terms of the accuracy and weighted F_1 score. This particularly implies that the explicitly modeled conjunctive information is more effective for the activity recognition, for which the relationship between sensors has been considered only implicitly by the sensor integration scheme.

While the proposed approach is capable of extracting directed conjunctive information, undirected information can be also obtained using a symmetric $\mathbf{R}$ that is obtained by copying the extracted upper triangular part and pasting it to the lower triangular part. The directed cases show slight superiority over the undirected cases in Table 3, but the performance difference between them is not significant.

The running time of the best case (directed, $d = 32$) is 1.53 hours and 2.58 seconds for training and test, respectively.

Table 3: Results of activity recognition. The F_1 score of the random forest is not available because it fails to recognize the *lie* class.

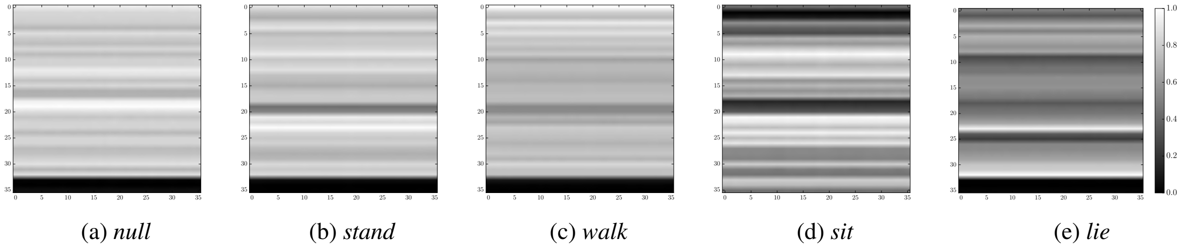
		Model	**Acc**	**F_1**
Baseline		k-NN	0.462	0.469
		random forest	0.608	-
		ResBidirLSTM [20]	0.699	0.685
		DeepConvLSTM [10]	0.734	0.737
Proposed	undirected	$d = 8$	0.803	0.795
		$d = 12$	0.812	0.803
		$d = 16$	**0.816**	**0.809**
		$d = 24$	0.810	0.804
		$d = 32$	0.809	0.803
	directed	$d = 8$	0.806	0.798
		$d = 12$	0.807	0.801
		$d = 16$	0.819	0.812
		$d = 24$	0.807	0.800
		$d = 32$	**0.821**	**0.814**

We conduct further analysis for the best case. First, to examine the importance of the reconstruction branch, the encoder is trained only based on the classification loss. The obtained classification accuracy and F_1 score are 0.782 and 0.755, respectively. This implies that the explicit feedback from the reconstruction branch is crucial for learning of the conjunctive information.

Next, the reconstruction performance is evaluated in terms of the reconstruction RMSE. Figure 2 illustrates the RMSE values after normalization within [0,1] for all pairs of sensors (36×36) depending on the activity class, where darker colors indicate smaller RMSEs. The averages and standard deviations of RMSEs are 0.227 ± 0.033, 0.601 ± 0.062, 0.666 ± 0.071, 0.546 ± 0.055, and 0.483 ± 0.121 for Figures 2a-2e, respectively. Therefore, although the difficulty of cross-signal reconstruction changes depending on the activity class, the variation of reconstruction errors across the sensor pairs within the same class is relatively small. This has two implications. First, the extracted conjunctive information provides useful information for the cross-sensor reconstruction. If the reconstruction of one signal is only based on the other signal in the pair, the reconstruction errors would vary significantly among the sensor pairs because different sensors measure accelerometer signals from different body parts. Moreover, the reconstruction errors are largely different across the activity classes for the same sensor pairs. This indicates that the relationship between the signals significantly varies depending on the class and thus provides useful information for classification.

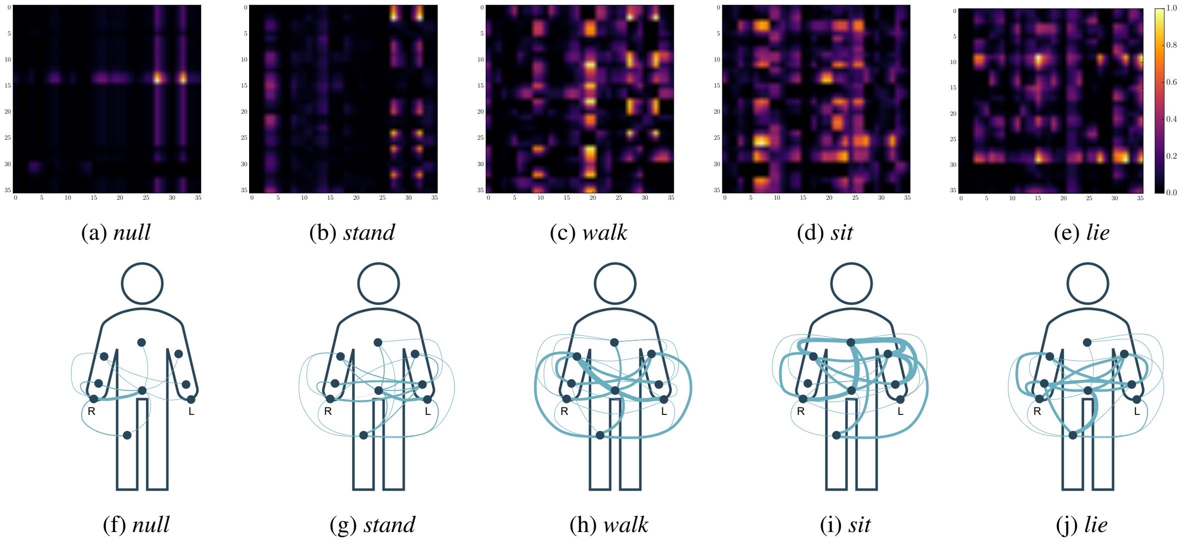
In addition, the Grad-CAM approach [14] is applied to verify the significance of particular sensor pairs for the activity recognition. The results are shown in Figure 3, which are obtained by averaging the Grad-CAM heatmaps of randomly selected 40 test data for each activity class. While the arrangement of sensors in $\mathbf{R}$ follows the description in [13] during the training and test processes, the Grad-CAM heatmaps are reshaped depending on the sensor positions in order to ease the analysis. A brighter region of the figures indicates higher contribution of the corresponding conjunctive information. In addition, the Grad-CAM results are also visualized as connections of sensor pairs on the body, where brighter regions in the Grad-CAM results are represented as thicker connection lines.

It is noticeable that the highlighted locations in the Grad-CAM result vary depending on the activity type. For example, the conjunctive information values related to *hip* are contributive to the recognition of *walk* in Figures 3c and 3h, whereas the recognition of *sit* is mostly influenced by the conjunctive information values within the upper body (Figures 3d and 3i).



(a) *null* (b) *stand* (c) *walk* (d) *sit* (e) *lie*

Figure 2: Reconstruction errors in RMSE depending on the activity class. A pixel value in each panel represents the error of reconstruction of the signal from the sensor in the row using the signal from the sensor in the column.



(a) *null* (b) *stand* (c) *walk* (d) *sit* (e) *lie*

(f) *null* (g) *stand* (h) *walk* (i) *sit* (j) *lie*

Figure 3: Results of Grad-CAM analysis for activity recognition. (a)-(e) are Grad-CAM heatmaps (0-2: *left hand*, 3-5: *left wrist*, 6-11: *left arm*, 12-14: *right hand*, 15-17: *right wrist*, 18-23: *right arm*, 24-26: *back*, 27-29: *hip*, 30-35: *right leg*), and (f)-(j) visualize the Grad-CAM results on body illustrations, where the strength of connections between sensors is obtained as an average of the corresponding Grad-CAM heatmap values. Note that closely located sensors are merged in (f)-(j) for better visualization. Thicker lines indicate more significant contribution for activity recognition (i.e., brighter colors in the heatmaps).

The conjunctive information values related to the arms or hands are frequently attended for the classification. For instance, the conjunctive information within the upper body and that between *hip* and the upper body parts significantly contributes to the classification of *walk*, *sit*, or *lie* in Figure 3. Since subjects were instructed to perform the given gestures such as cleanup, eating, and opening/closing doors, the arms and hands are expected to be in dynamic states while the movements of the other parts are relatively static or regular. Therefore, while the target activities are mostly driven by the states of the lower body, the incoordinated relationship with the upper body provides additional information for recognizing the locomotion, which is effectively captured by the proposed method.

4.3 Emotional video classification

4.3.1 Database

We employ the DEAP dataset [6] for emotional video classification using EEG signals. The classification task is to identify which video was watched by a subject among 40 videos based on the 32-channel EEG signals measured during the watching period, where the video was supposed to invoke the subject's emotional response. The preprocessed EEG signals provided in the dataset is used, which underwent noise removal, downsampling to 128 Hz, filtering, etc. The EEG sequences are divided into 3-second-long segments with an overlap of two seconds. Then, randomly selected 10% of the EEG sequences are used as test data, and another 10% are employed for validation. The rest is used for training.

4.3.2 Setup

Table 4 describes the network architectures of our method in this experiment. The structures are similar to those used for the activity recognition. However, the kernel sizes of convolutional layers are slightly changed to match the data size, and more numbers of hidden nodes in the FC layers and kernels in the convolutional layers are employed to reflect the higher complexity of the cross-signal reconstruction and classification task. As the test data is well-balanced (368 ± 18 sequences for each class), the classification accuracy is used as the performance measure. The cases of undirected conjunctive information are examined for this experiment due to the excessive computational complexity of the directed cases.

The random forest and k-NN are implemented for comparison, where the power spectral density of EEG sequences is used as features because of the high dimension of the raw EEG signals. The parameters of the random forest and the k-NN are determined in the same way to that for the activity recognition.

Table 4: Network architectures for video classification.

		Type	Output shape	Kernel size
E	1	convolution	2×128×8	1×129
	2	convolution	$1\times1\times d$	2×128
D	1	dense	256	-
	2	linear	384	-
C	1	convolution	32×32×32	3×3
	2	convolution	32×32×64	3×3
	3	max-pooling	16×16×64	2×2
	4	convolution	16×18×128	3×3
	5	convolution	16×16×256	3×3
	6	max-pooling	8×8×256	2×2
	7	dense	256	-
	8	softmax	40	-

Table 5: Results of video classification.

	Model	Accuracy
Baseline	k-NN	0.462
	random forest	0.465
	GCNN [4]	0.653
	ConnCNN-TE [9]	0.554
	ConnCNN-PCC [9]	0.677
	ConnCNN-PLV [9]	0.731
	ConnCNN-all [9]	0.721
Proposed	$d = 8$	0.929
	$d = 12$	**0.979**
	$d = 16$	0.952
	$d = 24$	0.935
	$d = 32$	0.930

In adddition, a graph convolutional neural network (GCNN) approach [4] is compared, which considers the relationship between sensors to conduct the convolution on graphs. Particularly, the physical distance between EEG electrodes is employed as the connection strength between the sensors.

A CNN-based deep learning approach [9] (denoted as ConnCNN) that utilizes brain connectivity features is also implemented. In this approach, the handcrafted connectivity metrics are employed to obtain the connectivity matrix that corresponds to $\mathbf{R}$ in our proposed model, then the connectivity matrix is inputted to the CNN classifier. Bandpass filters are applied to the raw EEG signals to separate signals into delta, theta, low alpha, high alpha, alpha, low beta, mid beta, high beta, beta, and gamma frequency bands, and the transfer entropy (TE), PCC, or PLV is calculated from each pair of the band-limited EEG signals as connectivity metric values. The obtained connectivity values are used for constructing the connectivity matrix, which has a size of 32×32×10 ($M \times M \times$(number of frequency bands)). Furthermore, the case of multiple connectivity features (ConnCNN-all) is examined by employing the TE, PCC, and PLV at the same time, for which the TE, PCC, and PLV connectivity matrices are concatenated along the depth dimension and thus the size of the connectivity matrix becomes 32×32×30.

4.3.3 Results

Table 5 shows the classification results. The proposed method results in the best classification accuracy of 0.979 (for $d = 12$), outperforming both the conventional machine learning methods and the connectivity-based deep learning methods. This demonstrates that the learned conjunctive information obtained by the proposed method is more effective than the manually designed connectivity features. For this case, the running time is measured as 28.69 hours and 25.09 seconds for training and test, respectively.

The ConnCNN-all scheme is less effective even compared to the case with a single connectivity metric (i.e., ConnCNN-PLV). This indicates that the simple combination of multiple handcrafted metrics is not effective enough to model various aspects of the relationship between EEG signals. In contrast, the proposed method extracts extensive relationship between EEG signals from different brain regions effectively based on the data, which results in the outstanding classification performance.

When the reconstruction branch is omitted in our method, the classification accuracy drops to 0.938, which shows the importance of the reconstruction branch as in the case of activity recognition.

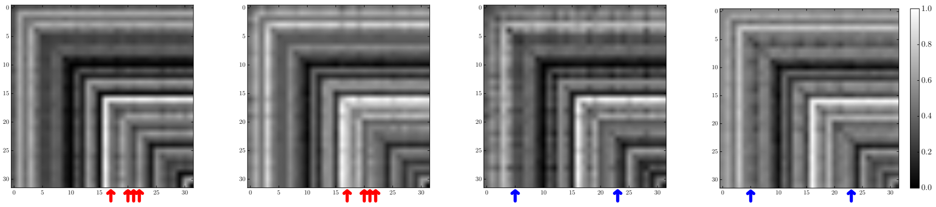
Results of the reconstruction are described in Figure 4. The reconstruction errors are shown for the videos that have distinguished emotional characteristics in terms of the valence and arousal. The valence and arousal scores obtained from the subjects are provided in the dataset. We select the videos having the highest or lowest valence or arousal scores. Distinguished variations depending on class (i.e., video) are observed, such as the 17th, 20th, 21st, and 22nd sensors in Figures 4a and 4b (marked by red arrows) and the 5th and 23rd sensors in Figures 4c and 4d (marked by blue arrows). Those sensors perhaps reflect the emotional responses appearing in the brain activity.

We employ the t-SNE technique [8] to compare the learned conjunctive information values and handcrafted connectivity values. The results are shown in Figure 5, where different colors correspond to different videos. Figure 5a shows the entire t-SNE results of the conjunctive information values, where the data points are clustered according to subjects on a global scale due to the high dependency of the EEG signals on subjects. Therefore, the region corresponding to the data points of subject #17 (marked by a red box in Figure 5a) is enlarged for better visualization in Figure 5b. Figures 5c-5e are also obtained in the same way for TE, PCC, and PLV, respectively. It is apparent that the conjunctive information values are well clustered according to the target class in Figure 5b. In contrast, in Figures 5c-5e, discriminability between classes is much lowered, which explains the worse classification accuracies than that of our method. This demonstrates that our proposed method successfully extracts the data-driven and target-optimized relationship between signals.

5 Conclusion

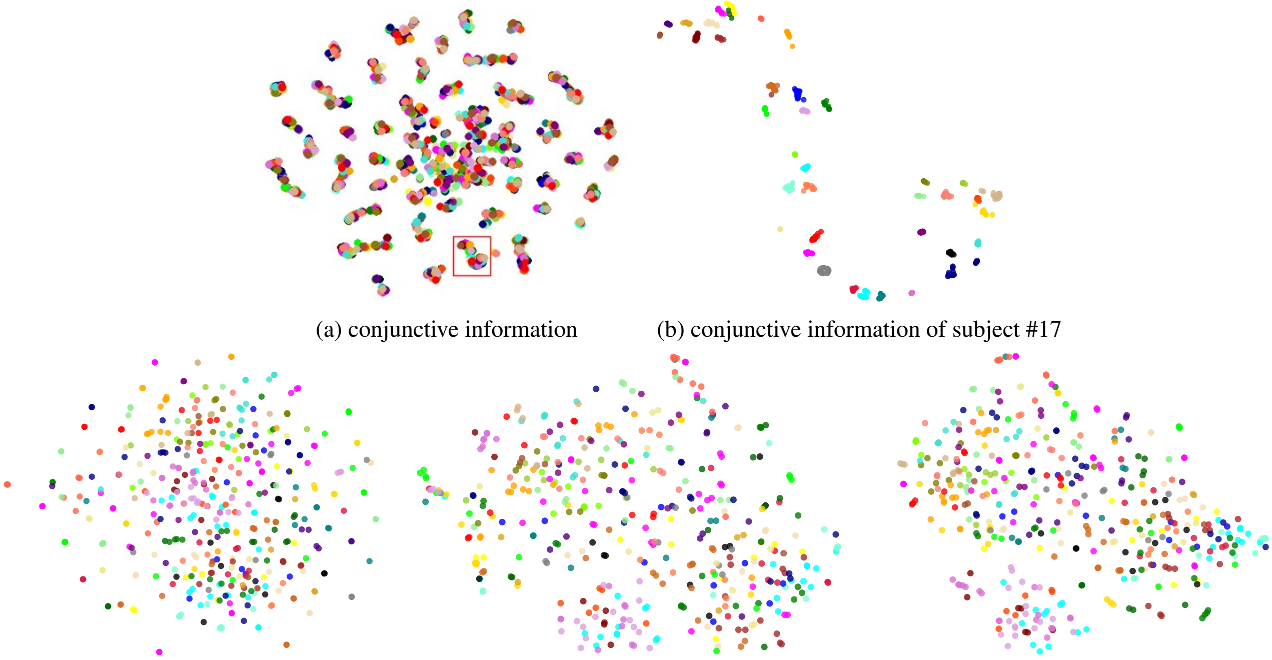
We have proposed a deep learning approach to learn the conjunctive information between signals in multi-sensor systems. The proposed method has two advantages compared to the conventional handcrafted metrics to measure the relationship between signals. First, our method is applicable to datasets from various domains because it is a totally data-driven, task-optimized general framework. Second, our method extracts the comprehensive relationship between signals, yielding improved performance for the given task, while the handcrafted metrics are designed for only specific aspects of the relationship. These advantages were proven by the toy example and two real-world problems of different domains.

In our future work, the proposed method will be extended to heterogeneous multi-sensor systems. The multi-sensor signals considered in this paper were acquired from the same type of sensors (accelerometers or EEG electrodes). However, different types of sensors are often utilized in a single multi-sensor system. In such a case, a mechanism to handle signals of heterogeneous characteristics would be additionally required.



(a) low valence (b) high valence (c) low arousal (d) high arousal

Figure 4: Reconstruction errors for emotional video classification depending on emotional characteristics of videos.



(a) conjunctive information (b) conjunctive information of subject #17

(c) TE of subject #17 (d) PCC of subject #17 (e) PLV of subject #17

Figure 5: Results of t-SNE analysis on the conjunctive information obtained by the proposed method and the handcrafted brain connectivites. Different colors indicate different target classes (i.e., videos).

In this work, we designed our method for systems involving rather small numbers of sensors, i.e., sensors placed on users' body for activity recognition and those on the scalp for EEG monitoring, which need to consider wearability and comfortability. In some other cases, a large number of sensors may be involved, e.g., climate or financial data, where the complexity to consider all pair-wise combinations may matter. Considering such applications, future work on enhancing scalability of the proposed method would be worth investigating.

ACKNOWLEDGEMENTS

This work was supported by the Institute for Information & communications Technology Promotion (IITP) grant funded by the Ministry of Science and ICT (MSIT) (R7124-16-0004, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding).

REFERENCES

[1] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Troster, Jose del R. Millan, and Daniel Roggen, 'The opportunity challenge: a benchmark database for on-body sensor-based activity recognition', *Pattern Recognition Letters*, **34**(15), 2033–2042, (2013).

[2] Cen Chen, Kenli Li, Sin G. Teo, Xiaofeng Zou, Kang Wang, Jie Wang, and Zeng Zeng, 'Gated residual recurrent graph neural networks for traffic prediction', in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 485–492, (2019).

[3] Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He, 'Learning discrete structures for graph neural networks', in *Proceedings of the International Conference on Machine Learning*, pp. 1–13, (2019).

[4] Soobeom Jang, Seong-Eun Moon, and Jong-Seok Lee, 'EEG-based video identification using graph signal modeling and graph convolutional neural network', in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 3066–3070, (2018).

[5] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel, 'Neural relational inference for interacting systems', in *Proceedings of the International Conference on Machine Learning*, pp. 2688–2697, (2018).

[6] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras, 'DEAP: a database for emotion analysis using physiological signals', *IEEE Transactions on Affective Computing*, **3**(1), 18–31, (2012).

[7] Yonggang Lu, Ye Wei, Li Liu, Jun Zhong, Letian Sun, and Ye Liu, 'Towards unsupervised physical activity recognition using smartphone accelerometers', *Multimedia Tools and Applications*, **76**(8), 10701–

10719, (2017).

[8] Laurens van der Maaten and Geoffrey Hinton, 'Visualizing data using t-SNE', *Journal of Machine Learning Research*, **9**, 2579–2605, (2008).

[9] Seong-Eun Moon, Soobeom Jang, and Jong-Seok Lee, 'Convolutional neural network approach for EEG-based emotion recognition using brain connectivity and its spatial information', in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 2556–2560, (2018).

[10] Francisco Javier Ordonez and Daniel Roggen, 'Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition', *Sensors*, **16**(1), 115:1–25, (2016).

[11] Alexandros Pantelopoulos and Nikolaos G. Bourbakis, 'A survey on wearable sensor-based systems for health monitoring and prognosis', *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, **40**(1), 1–12, (2010).

[12] Jonathan D. Power, Alexander L. Cohen, Steven M. Nelson, Gagan S. Wig, Kelly Anne Bames, Jessica A. Church, Alecia C. Vogel, Timothy O. Laumann, Fran M. Miezin, Bradley L. Schlaggar, and Steven E. Petersen, 'Functional network organization of the human brain', *Neuron*, **72**(4), 665–678, (2011).

[13] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Forster, Gerhard Troster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, Jakob Doppler, Clemens Holzmann, Marc Kurz, Gerald Holl, Ricardo Chavarriaga, Hesam Sagha, Hamidreza Bayati, Marco Creatura, and Jose del R. Millan, 'Collecting complex activity datasets in highly rich networked sensor environments', in *Proceedings of the International Conference on Networked Sensing Systems*, pp. 233–240, (2010).

[14] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Bata, 'Grad-CAM: visual explanations from deep networks via gradient-based localization', in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, (2017).

[15] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu, 'Deep learning for sensor-based activity recognition: a survey', *Pattern Recognition Letters*, **118**, 3–11, (2019).

[16] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu, 'A comprehensive survey on graph neural networks', in *arXiv:1901.00596*, pp. 1–22, (2019).

[17] Xiaoya Xu and Qingsong Hua, 'Industrial big data analysis in smart factory: current status and research strategies', *IEEE Access*, **5**, 17543–17551, (2017).

[18] Jian Bo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy, 'Deep convolutional neural networks on multichannel time series for human activity recognition', in *Proceedings of the International Joint Conference on Artifical Intelligence*, pp. 3995–4001, (2015).

[19] Li Zhang, Hongli Gao, Juan Wen, Shichao Li, and Qi Liu, 'A deep learning-based recognition method for degradation monitoring of ball screw with multi-sensor data fusion', *Microelectronics Reliability*, **75**, 215–222, (2017).

[20] Yu Zhao, Rennong Yang, Guillaume Chevalier, Ximeng Xu, and Zhenxing Zhang, 'Deep residual bidir-LSTM for human activity recognition using wearable sensors', *Mathematical Problems in Engineering*, **2018**, 7316954:1–13, (2018).