IJCNN 2019. International Joint Conference on Neural Networks. Budapest, Hungary. 14-19 July 2019

Generalized Alignment for Multimodal Physiological Signal Learning

Yuchi Liu
Australian National University
Canberra, Australia
yuchi.liu@anu.edu.au

Yue Yao
Australian National University
Canberra, Australia
yue.yao@anu.edu.au

Zhengjie Wang
Australian National University
Canberra, Australia
u6259550@anu.edu.au

Josephine Plested
Australian National University
Canberra, Australia
jo.plested@anu.edu.au

Tom Gedeon
Australian National University
Canberra, Australia
tom@cs.anu.edu.au

***Abstract*—Revealing the correspondences and relationships between physiological signals is attractive for bioinformatics and human-computer interaction. Time alignment is a straightforward way to figure out correspondences between time sequential data. However, alignment between multimodal physiological signals is hard to achieve because the similarity metrics are difficult to define if the two physiological signals being investigated are non-linearly correlated, misaligned or quite different in morphology. In this paper, we propose a generalized time alignment method for multimodal physiological signals which (i) learns the feature extractions on physiological signals in a generalized way, and (ii) enables learned features to be in a coordinated space where the similarity between sub-components from two signals can be defined. Furthermore, we applied our alignment based multimodal feature fusion on an evaluation model to perform emotion recognition tasks on the DEAP multimodal physiological signal dataset. The experimental results show that the alignment based feature fusion outperforms the non-aligned feature fusion in most cases.**

***Index Terms*—Deep Learning, Physiological Signals, Canonical Correlation Analysis, Alignment, Recurrent Neural Network,**

I. Introduction

A. Motivation

Human beings comprehend the word through multiple sensory modalities like vision and hearing. Similarly, multimodal physiological signals are primary channels for an intelligent system to understand humans, especially the physiological states of bodies. With the help of wearable or wireless sensors, physiological information can be continually measured and stored as physiological signals. Physiological signals have been widely and successfully used in medical diagnosis [9], activity recognition [14] and entity authentication [6] for many years. Recently, complex activity recognition [34], affective computing [51] and other research areas are actively investigated by utilizing multimodal physiological signals where complementary and supplementary information is sourced. Multiple signals are extracted and fused to improve final system performance.

In order to model multiple physiological signals better, it is worthwhile to find the relationships and correspondences between their sub-components. Previous work has already shown the existence of such correlation by finding oscillatory coupling and the time delay between the central nervous system and peripheral physiological signals through frequency analysis [19]. A power assistant system designed for the disabled person by using Electromyography (EMG) signals estimated from electroencephalography (EEG) signals, shows the significance of investigating relationships across physiological signals [28], [29].

However, as a key topic in multimodal learning, sub-component alignment between multimodal physiological signals could not be realized in a generalized way in the literature. There are two main challenges to achieve such alignment. The first one is that two physiological signals may have very different morphology [43]. In this case, similarity (alignment metrics) between sub-components is hard to define. Secondly, there are not labeled datasets for multimodal physiological signals alignment. Unlike multimodalities in other areas like audio-visual, where alignment of sub-components could be labeled by the human, even experts in the physiological area could not clearly figure out the correlated sub-parts across multimodal physiological signals. Thus, alignment for multimodal biosignals could only be achieved in an unsupervised manner and evaluation metrics for alignment results is also lacking.

On the other hand, most previous works related to fusing information from multimodal physiological signals did not use any hidden alignment information in series sub-elements. Such correlations are successfully utilized in other semantic modalities like audio-visual, visual-text, audio-text by recent works which are based on temporal models [24], [45], [48]. However, as for physiological signals, simple concatenation of features from misaligned pairs could suffer from asynchronization [12]. To avoid this, most feature extraction procedures and feature fusion procedures deal with physiological signals as a whole in the literature. Therefore, misalignment limits model choices in terms of information fusion strategies for physiological signals learning.

978-1-7281-2009-6/$31.00 ©2019 IEEE

B. Contribution

In order to tackle the above challenges, we propose a generalized multimodal physiological signal alignment methodology, which makes the following contributions:

- In data processing and feature extraction phases, we propose a uniform way, called Physio2Video, to generate multimodal physiological feature videos. Thus, the multimodal continuous physiological signals alignment problem can be converted to the video frames alignment problem.
- The proposed frame encoders based on unsupervised learning can non-linearly project the physiological feature frames into a common space where time warping is feasible.
- Our alignment method can be considered as the guidance of feature level information fusion in multimodal learning. The experiment results show that alignment based feature fusion performs better than misaligned feature fusion of sub-elements.

II. Related Works

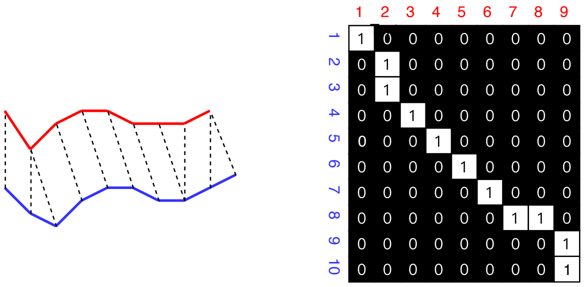
A. Align Multimodal Series

Alignment on multimodal series aims to explicitly find optimal matches between their sub-frames [5].

Dynamic time warping (DTW) like dynamic programming methods are widely used unsupervised ways to archive temporal alignment of time series [31]. Their aim is to find the optimal time warping for instances according to manually predefined similarity metrics between them. The most common similarity measurement way is computing distance between frames from two series based on Frobenius Norm (Euclidean norm). Given two time series, $X = [x_1, x_2, x_3, ..., x_{n_x}] \in \mathbb{R}^{d \times n_x}$ and $Y = [y_1, y_2, y_3, ..., y_{n_y}] \in \mathbb{R}^{d \times n_y}$ with the same feature space d, DTW aims to optimize:

$$
\begin{aligned}
&arg \min_{P_x, P_y} \|XP_x - YP_y\|_F^2 \\
&\text{s.t.} \quad P_x \in \{0,1\}^{n_x \times n} \\
&\qquad\;\; P_y \in \{0,1\}^{n_y \times n}
\end{aligned} \tag{1}
$$

where P_x and P_y denote alignment path matrices (filled with binary values) for time frames in X and Y respectively [citation at here]. There are n alignment pairs after DTW. Considering the joint alignment path matrices $P_{xy} = P_x P_y^T$ and any k-th alignment pair $k(i_x, i_y)$, $P_{xy}(i_x, i_y)$ is 1, in which $0 \leq k < n, 0 \leq i_x < n_x, 0 \leq i_y < n_y$. All other positions in P_{xy} are filled with 0. Although there are exponentially many possible ways to align X and Y with respecting to n_x and n_y, DTW can get the optimal solution of Equation 1 in $O(n_x n_y)$ by using dynamic programming. Fig. 1 gives an example of dynamic time warping between two time series. Some approximate dynamic time warping algorithms like FastDTW have achieved linear time and space complexity [37]. Note that we consider $F2$-norm as the similarity here but it should be set according to real cases. For example, Tapaswi et al. [10] constructed a character based similarity to realize video to text alignment. In another work [2], audio-to-text alignment is done by dynamic programming based on a manually created sound-to-grapheme correspondence matrix.



$$
\begin{array}{c|ccccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\hline
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
7 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
8 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}
$$

Fig. 1. One example of dynamic time warping. In the left part, sub-points in red time series and blue time series are aligned by dotted lines. Alignment matrix P_{xy} filled with 0 and 1 in the right part infers the alignment path. Each 1 in P_{xy} stands for one dotted line (alignment pair) in the left.

Generally, it is hard to directly define the similarities between sub-components from multi-view sets. Canonical correlation analysis (CCA) can be used to project multi-view moralities into a common feature space with maximized Pearson correlation where the similarity can be easier to define. Therefore CCA based DTW, named as canonical time warping (CTW) is a generalized way to apply DTW like dynamic programming methods on different multidimensional features. CTW was firstly proposed by Zhou and Torre where they accurately aligned two behavioral time series spatially and temporally on CMU-Multimodal Database [53]. They also extended CTW to generalized time warping (GTW) which allow alignment between more than two multimodal series [52]. Given two time series $X \in \mathbb{R}^{d_x \times n_x}, Y \in \mathbb{R}^{d_y \times n_y}$ with varying feature dimensions and time frames, CTW has the follwing objective function after combining CCA and DTW:

$$
\begin{aligned}
&\underset{W_x, W_y, P_x, P_y}{\text{argmax}} \quad corr(W_x^T X P_x, W_y^T Y P_y) \\
&\qquad s.t. \quad W_x = [w_x^1, ..., w_x^d] \in \mathbb{R}^{d_x \times d}, \\
&\qquad\qquad\;\; W_y = [w_y^1, ..., w_y^d] \in \mathbb{R}^{d_y \times d}, \\
&\qquad P_x \in \{0,1\}^{n_x \times n}, P_y \in \{0,1\}^{n_y \times n},
\end{aligned} \tag{2}
$$

where W_x and W_y are linear projections and P_x and P_y are the alignment matrices for projected $W_x^T X$ and $W_y^T Y$. The correlation maximization problem can be transformed as a trace maximization problem [53]:

$$
\begin{aligned}
&\underset{W_x, W_y, P_x, P_y}{\text{argmax}} \quad \text{trace}(W_x^T X P_x P_y^T Y^T W_y) \\
&\qquad s.t. \quad W_x^T X P_x P_x^T X^T W_x = I, \\
&\qquad\qquad\;\; W_y^T Y P_y P_y^T Y^T W_y = I, \\
&\qquad\qquad\;\; W_x = [w_x^1, ..., w_x^d] \in \mathbb{R}^{d_x \times d}, \\
&\qquad\qquad\;\; W_y = [w_y^1, ..., w_y^d] \in \mathbb{R}^{d_y \times d}, \\
&\qquad P_x \in \{0,1\}^{n_x \times n}, P_y \in \{0,1\}^{n_y \times n},
\end{aligned} \tag{3}
$$

The optimum is attained by fixing projections (or fixing alignment path) and optimizing the other one alternately. The solutions to compute the biggest trace in Equation 3 are inherited from the Singular Value Decomposition (SVD) based CCA optimization solution in [20]. Algorithm 1 describes the CTW optimizing procedure which is similar to that in [53]. Σ_{ij} represents the cross-covariance and r_i is the regularization parameter.

Algorithm 1 CTW optimization algorithm

```
  Input: $\bar{X}$(centralized X), $\bar{Y}$(centralized Y)
  Output: $W_x, W_y, P_x, P_y$
begin
  Initialize $W_x \in \mathbb{R}^{d_x \times d}, W_y \in \mathbb{R}^{d_y \times d}$
  repeat
    Get alignment matrix $P_x, P_y$ by applying dynamic time warping on $(W_y^T \bar{X}, W_y^T \bar{Y})$.
    $\hat{\Sigma}_{xx} = \frac{1}{d_x - 1} \bar{X} P_x P_x^T \bar{X}^T + r_x I$
    $\hat{\Sigma}_{yy} = \frac{1}{d_y - 1} \bar{Y} P_y P_y^T \bar{Y}^T + r_y I$
    $\hat{\Sigma}_{xy} = \frac{1}{d - 1} \bar{X} P_x P_y^T \bar{Y}^T + r_{xy} I$
    $M = \hat{\Sigma}_{xx}^{-1/2} \hat{\Sigma}_{xy} \hat{\Sigma}_{yy}^{-1/2}$
    $U, V =$ Singular Value Decomposition$(M)$
    $W_x = \hat{\Sigma}_{xx}^{-1/2} U$
    $W_y = \hat{\Sigma}_{yy}^{-1/2} V$
  until trace(M) maximized
end
```

To address the incapability of CCA in exploring non-linear relationships between modalities, deep canonical correlation analysis (DCCA) replaced the linear transformation in CCA by deep neural networks to extract non-linear correspondence or similarity [1]. Inspired by DCCA, Yin and Chen proposed a deep metric learning autoencoder to enable extracted deep spatiotemporal information of human motion to be compared and aligned [50]. Recently, researchers presented the deep canonical time warping (DCTW) method which used two CNN encoders to automatically learn deep representations of series where the representations are maximally correlated and could also be aligned by DTW like dynamic programming [39]. They successfully applied DCTW on audio and visual streams with less alignment error compared to CTW and GTW. There are no closed-form solutions like CTW, back-propagation is used to optimize DCTW objective.

Similar to DCCA and DCTW, other recent works also use deep learning methods to find similarity or correspondence measurement metrics between multi-views but in a supervised manner. Karpathy and Li align sentence snippets to the visual regions by learning a shared multi-modal embedding where similarity could be computed [25]. The image region embedding and sentence embedding are produced through Convolutional Neural Network (CNN) and Bidirectional Recurrent Neural Network (BRNN). Temporal Regression Localizer (CTRL) align sentences with temporal ordering information to video clips with offsets [16]. They also use a deep visual encoder and a deep sentence encoder to find the coordinated representation of multimodalities.

As for physiological signals, alignment is often used between signals coming from the same source or having similar morphology. For instance, DTW algorithm was used to detect Alzheimer disease by aligning gait (foot movement) signals collected from patients with Alzheimer disease and healthy people [40]. Researchers [15] have used Euclidean distance based similarity to align partially correlated signals (Electrocardiogram, Arterial Blood Pressure and Photo Plethysmogram) with similar repetitive morphology and make temporal segmentation based on the alignment path. However, there is not a generalized way in the literature to align physiological signals which may not share similar morphology.

B. Physiological Feature Extraction by Deep Learning

Deep learning based methods have become popular in recent years as alternative feature extraction methods to obtain physiological features [21]–[23]. One category is end-to-end models which directly consider the sampled data as input. A CNN model encodes every 30second window data in a long sleeping signal (EEG, EMG, EOG) into one feature vector and combines every output of each window together to feed into the classifier [11]. This simple model achieves state-of-the-art performance on sleep stage classification. A recent work recognizes the corresponding visual stimulus of the input EEG signals where a LSTM encoder is trained to extract useful representations for classification [38]. In another category, physiological signals are transformed into spectrograms before being input into deep neural networks where both time domain and frequency domain information are contained. EMG signals collected on the arm are transformed into spectrograms and a convolutional neural network is applied to perform image classification with hand gesture labels in [13]. Another similar work predicting limb moving also uses spectrograms as the model input but splits the spectrogram into continuous time steps [47]. Each time step will be put into a LSTM network to get a representation for the original signal.

C. Multimodal Fusion

Previous surveys [5], [18] have shown that the main categories of multimodal fusion are feature level fusion, decision level fusion, and hybrid fusion. In early multimodal learning work [4], [35], [44] based on traditional machine learning methods like Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Naive Bayes, and K-Nearest Neighbor (KNN), decision fusion is applied because these classification methods could not tackle multimodal data with different distribution and morphology. Classifiers are trained separately on different modalities. Nevertheless, if feature selection methods like Kernel Discriminant Analysis (KDA) and Correlation based Feature Selection are introduced, the selected feature can be fused to perform feature level information fusion [17], [30], [41]. Hybrid fusion is successfully performed on the CMU speaker identification task in this work [46] by training a multi-stream hidden Markov model (HMM) on audio, video, and fused audio-video features.

With the rise of deep learning technology in recent years, feature level fusion has been extensively studied as joint representation learning. Audio-Video speech recognition performance has been improved significantly in the former by learning a joint feature representation of audio and video through autoencoders. [32]. Time sequential features can be fused by RNN model if they are time aligned. Audio-video information is fused by LSTMs in these works [33], [45] to perform audio-video emotion recognition tasks.

However, the performance of multimodal fusion between temporal features will suffer if the different modalities are misaligned [46]. Addressing multimodal fusion on misaligned time series has not been solved in the literature. This paper will give a general solution for misaligned multimodal physiological series fusion.

III. Data Processing Model

Our technique is to convert multimodal physiological signals to multiple physiological feature videos, an approach we term Physio2Video. It allows us to convert the signals alignment problem into a video frames alignment problem. The Physio2Video approach uses separate techniques for creating videos from EEG signals and other peripheral physiological signals. In this case, EEG2Video, EMG2Video and GSR2Video are described below.

A. EEG2Video

The EEG2Video idea is from Bashivan's work [8]. As shown in Figure 2, a small period of EEG signals is transformed to a colored image with the help of electrodes location information. For each piece of EEG signal, three frequency bands are extracted using Fourier Transform. They are theta (4-7Hz), alpha (8-13Hz) and beta (13-30Hz) bands which are commonly used in EEG analysis [7]. Then the average of each frequency will be calculated to assign three scalar value for each channel (electrode). Next, a polar projection is used to project 3-D electrode position to 2D position, thereby creating a 2D map. Finally, the Clough-Tocher scheme is used to interpolate blank areas. In this way, not only time-frequency information but also location information for different channels of EEG signals are extracted.

In our case, a long EEG trial will be cut into small pieces by time windows with 2-second window size and 1.5-second overlaps. For each time window of EEG signal, it will be transformed into an EEG image, which is considered as one frame of EEG video. Overlaps enable the generated feature EEG image video to show the continuous change. We remove the 3 seconds pre-trial time for each trial in DEAP (60 seconds left) so that the tensor shape for each EEG feature video is

$$X = (seq\_len \times channel_x \times H \times W)$$

where $seq\_len = 117$, $channel = 3$, $H = 32$ and $W = 32$. seq_len is the fame numbers which is 117 for each EEG video. $channel_x$ means three scalar value for theta, alpha and beta bands. H and W are height and width respectively for one EEG feature video frame. Figure 3 shows an example of produced EEG feature video which consists of EEG feature time frames with continuous changes.

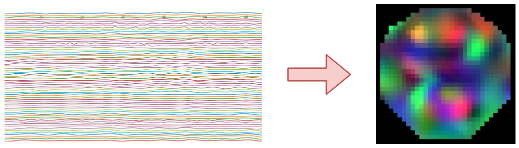


Fig. 2. EEG to image [49].

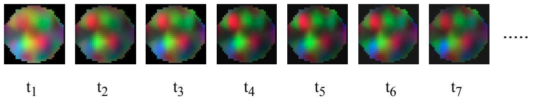


Fig. 3. Time Frames in EEG feature video with continues change.

B. EMG2Video and GSR2Video

The idea of converting peripheral physiological signals into feature videos is inspired by recent success of spectrogram representation for the wave-like data [47]. EMG2Video and GSR2Video are derived from all of the parameters of the spectrogram variables. Similar to the EEG2video, a long peripheral physiological signal trial is cut into small pieces with overlaps, using the same settings as EEG2Video. For each piece of the peripheral physiological signal which is generated by this sliding window, the waveform signal is transformed into spectral coefficients by using Fourier Transform on each window. As a result, a time-frequency spectrogram is created for the whole physiological signal by using the Fast Fourier Transform (FFT). For a certain type of physiological signal, the number of spectrograms with time-frequency domain information should be the same as channel numbers of the physiological signal. The length of the time axis and the size of the frequency axis are considered as the time frame numbers and the feature size for each time frame. For example, we have two channels for EMG signals so that there are two spectrograms showing time-frequency information for each channel (Figure 4). The video for EMG can be shown as the following tensor:

$$Y : (seq\_len, channel_y, F)$$

where $seq\_len = 117, channel_y = 2, F = 128$. seq_len and $channel_y$ refer to the video frame numbers and video channel numbers. F is the number of coefficients which are computed in a time-frequency frame. We use all coefficients (128) produced by FFT here.

GSR feature video could be generated in the same way but with different channel numbers for each time frame, that is:

$$Z : (seq\_len, channel_z, F)$$

where $channel_z = 1$ because we only have one channel for GSR signal in our case.

IV. Generalized Alignment Model

The goal of our generalized alignment model is to find the alignment between sub-components of two physiological sig-

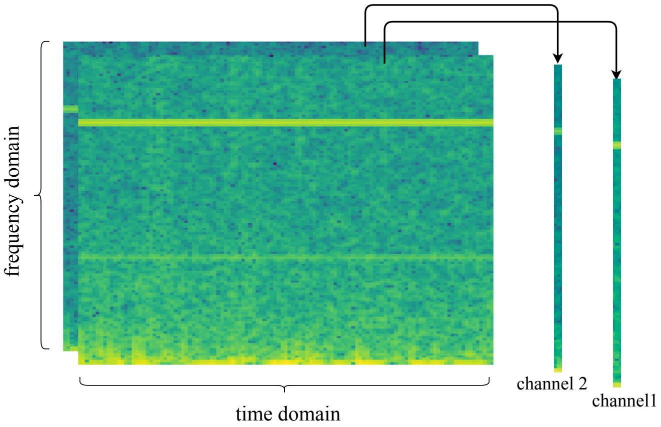


Fig. 4. EMG time-frequency spectrograms for two EMG signal channels. Each time step in the time domain contains frequency domain information from two different EMG signal channels, which are stored in two vectors with size F. As for GSR signals, we can generate a similar spectrogram to present GSR feature video.

nals with different modalities. Before proceeding with alignment, we have already obtained three time series signal feature video X, Y, and Z for EEG, EMG, and GSR respectively by following the data processing phase. The alignment results between sub-frames of the three feature videos correspond to the alignment results between sub-components of multimodal physiological signals.

A. Physiological Signal Feature Frame Encoders

We use our proposed feature video frame deep encoders to non-linearly project two signal feature videos into a coordinated feature space, where alignment between sub-components could be performed. Our encoders for the above videos are all based on a CNN approach although they vary between different signal feature videos.

Given an EEG feature video $X \in \mathbf{R}^{seq\_len \times channel_x \times H \times W}$, we perform 4 layers of 2 dimensional (2D) convolution based neural network with batch normalization, pooling and dropout techniques over each time frame $x \in \mathbf{R}^{channel_x \times H \times W}$ to perform EEG feature frame encoding. For each time step (frame) of EMG feature video $Y \in \mathbf{R}^{seq\_len \times channel_y \times F}$ and GSR feature video $Z \in \mathbf{R}^{seq\_len \times channel_z \times F}$, there is only one frequency domain feature vector with size F for each channel. Thus, the input of EMG frame encoder and GSR frame encoder should be $y \in \mathbf{R}^{channel_y \times F}$ and $z \in \mathbf{R}^{channel_z \times F}$. We apply 1D convolution over the input frames with 5 convolution layers in total for both EMG and GSR cases. The embedding results from the three frame encoders are all one dimension feature vectors of size of 64. Detailed parameters of building blocks for the encoders are shown in Table I. We use F_x, F_y and F_z to donate encoders for X, Y and Z respectively.

TABLE I
THE DEEP CNN STRUCTURES FOR EEG FRAME ENCODER AND EMG FRAME ENCODER. PARAMETERS $cin, cout, k, p$ AND s IN BRACKETS REFER TO INPUT CHANNEL NUMBERS, OUTPUT CHANNEL NUMBERS, KERNEL SIZE, PADDING AND STRIDE. THE GSR FRAME ENCODER HAS THE SAME STRUCTURE AS THE EMG FRAME ENCODER BUT WITH $cin = 1$ IN THE FIRST BLOCK.

EEG frame encoder	**EMG frame encoder**
Input ($3 \times 32 \times 32$)	**Input** (2×128)
Conv2d ($cin = 3, cout = 16,$ $k = 3, p = 1$) BatchNorm2d, ReLU Maxpooling2d($k = 2 \times 2, s = 2$), Doupout(0.1)	Conv1d ($cin = 2, cout = 16,$ $k = 3, p = 1$) BatchNorm1d, ReLU Maxpooling1d($k = 2, s = 2$), Doupout(0.1)
Conv2d ($cin = 3, cout = 16,$ $k = 3, p = 1$) BatchNorm2d, ReLU Maxpooling2d($k = 2, s = 2$), Doupout(0.1)	Conv1d ($cin = 2, cout = 16,$ $k = 3, p = 1$) BatchNorm1d, ReLU Maxpooling1d($k = 2, s = 2$), Doupout(0.1)
same as last block	same as last block
same as last block	same as last block
	same as last block
Flatten into a one dimension vector	
Output 64	**Output** 64

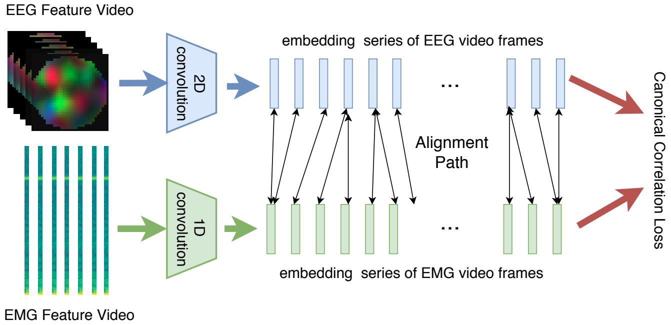

EEG Feature Video
2D convolution
embedding series of EEG video frames
...
Alignment Path
Canonical Correlation Loss
1D convolution
...
embedding series of EMG video frames
EMG Feature Video

Fig. 5. The structure of the generalized alignment model. Video frame embedding is generated for each frame by the feature video encoders F_x and F_y. The unsupervised training process achieves alignment of video frames and learning frame feature encoders jointly.

B. Alignment and Canonical Correlation Loss

We have three series of vectors $F_x(X) \in \mathbf{R}^{d \times n_x}$, $F_y(Y) \in \mathbf{R}^{d \times n_y}$ and $F_z(Z) \in \mathbf{R}^{d \times n_z}$ with the same dimension size d after performing feature video encoding. In order to learn an appropriate non-linear deep encoding and achieve reasonable alignment, we try to optimize the canonical correlation of two aligned embedding series. Given EEG frames embedding $F_x(X)$ and EMG frames embedding $F_y(Y)$ as an example:

$$\begin{aligned} \underset{F_x, F_y, P_x, P_y}{\operatorname{argmax}} \quad & corr(F_x(X)P_x, F_y(Y)P_y) \\ P_x \in \{0,1\}^{n_x \times n}&, P_y \in \{0,1\}^{n_y \times n}, \end{aligned} \tag{4}$$

where P_x and P_y is the alignment path through dynamic time warping. Notice that n_x and n_y are equal in our case because they are processed from the same trial with the same sliding window and step size. However, n is not less than n_x and n_y since alignment will copy frames if they are aligned more than once. As mentioned before, there are no closed form solutions for deep CNN based encoders F_x, F_y. We use back-propagation here and the negative canonical correlation between aligned frame embedding series from X and Y is considered as the loss function:

$$L_{corr} = -corr(H_x, H_y). \tag{5}$$

where $H_x = F_x(X)P_x \in \mathbf{R}^{d\times n}$ and $H_y = F_y(Y)P_y \in \mathbf{R}^{d\times n}$. similar to the CTW optimization methods introduced, we consider $\bar{H}_x$ and $\bar{H}_y$ as centralized matrices and construct a matrix $M = \hat{\Sigma}_{xx}^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_{yy}^{-1/2}$, where

$$\begin{aligned} \hat{\Sigma}_{xx} &= \frac{1}{d-1}\bar{H}_x\bar{H}_x^{\,T} + r_x I, \\ \hat{\Sigma}_{yy} &= \frac{1}{d-1}\bar{H}_y\bar{H}_y^{\,T} + r_y I, \\ \hat{\Sigma}_{xy} &= \frac{1}{d-1}\bar{H}_x\bar{H}_y^{\,T}. \end{aligned} \tag{6}$$

The canonical correlation between two aligned video frame embedding series is the trace normalization of M. Thus our final canonical correlation loss can be written as:

$$L_{corr} = -corr(H_x, H_y) = \|M\|_{tr} = tr(M^T M)^{-1/2}. \tag{7}$$

During the training process, two encoders will be optimized gradually to enable two embedded series sourced from the same stimulation to be more correlated and make their alignment more intuitive. Figure 5 shows the training pipeline to learn EEG feature video frame encoder and EMG feature video frame encoder cooperatively by using canonical correlation loss on two aligned embedded series.

V. Alignment Evaluation Model

Since there is no dataset labeling sub-components' alignment for multimodal physiological signals, we could not evaluate our generalized alignment methodology directly. In this case, we propose a feature fusion model targeting multimodal time series classification and compare the classification accuracy of our alignment based fusion and other feature fusion strategies.

A. Align Video Frames Based on Trained Frame Encoders

We also take EEG and EMG case as an example. After the unsupervised training procedure, we get two video frame encoders which enable frame embedding series to be aligned in a coordinated space. Assuming $F_x(X) \in \mathbf{R}^{d\times n_x}$ and $F_y(Y) \in \mathbf{R}^{d\times n_y}$ are projected embedding series, alignment path $P_x \in \{0,1\}^{n_x\times n}$ and $P_y \in \{0,1\}^{n_y\times n}$ can be obtained by using DTW. The alignment path for frame embedding series correspond to the alignment for the original signal feature videos. Therefore, signal feature videos $X \in \mathbf{R}^{n_x\times channel_x\times H\times W}$ and $Y \in \mathbf{R}^{n_y\times channel_y\times F}$ can be aligned into $X' \in \mathbf{R}^{n\times channel_x\times H\times W}$ and $Y' \in \mathbf{R}^{n\times channel_y\times F}$.

B. Task-Related Frame Encoders

The trained encoders mainly extract coordinated representations across multimodal physiological signals since the training objective is to maximize the canonical correlation of encoding results. With respect of the evaluation tasks, it is important to extract task related representations for classification performance. Thus, we use extra frame encoders $(F_x{}', F_y{}')$ for aligned videos (X', Y') and optimize them based on the gradient produced by the final classification objective. They share the same structures with previous encoders (F_x, F_y) so that the vector size of frames embedding are d for both EEG and EMG. Since the input videos (X', Y') are already aligned, we could intuitively concatenate the aligned frame embedding pairs into $a_i = \left[F_x{}'(x_i'), F_y{}'(y_i')\right]$, $i \in \mathbf{R}^n$ to perform sub-component feature fusion.

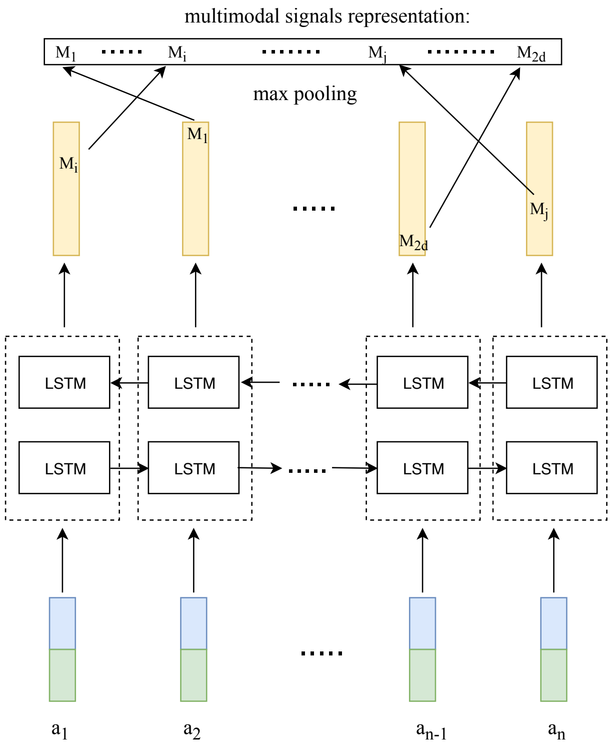


Fig. 6. Illustration of the frame embedding series encoder. Concatenated blue and green blocks refer the aligned frame embeddings sourced from EEG and EMG. Max pooling operation extract the maximum value over each channel from concatenated hidden states to form the joint representation of EEG & EMG.

C. Aligned Embedding Series Classification

For a sequence of n concatenated time frame embeddings $\{a_i\}_{i=1,...,n}$, a Bidirectional Long Short Term Memory (LSTM) network with max pooling is applied to encode the aligned frame embedding series into fixed-size representations which are related to specific tasks. The encoder structure is shown is Figure 6. The hidden states in this encoder during forward propagation could be written as:

$$\begin{aligned} \overrightarrow{h}_i &= \overrightarrow{LSTM}_i(a_1, ..., a_i), \\ \overleftarrow{h}_i &= \overleftarrow{LSTM}_i(a_d, ..., a_i), \\ h_i &= \left[\overrightarrow{h}_i, \overleftarrow{h}_i\right], \end{aligned} \tag{8}$$

where $\overrightarrow{h}_i$ and $\overleftarrow{h}_i$ refer the hidden state for the forward LSTM and backward LSTM at the time step i. h_i is the is the concatenation of $\overrightarrow{h}_i$ and $\overleftarrow{h}_i$. We choose the maximum value (max pooling) over each dimension among n hidden states ($\{h_i\}_{i=1,\ldots,n}$) to form a fixed-size representation for the input embedding series. In our case, the dimension size for each aligned and concatenated frame embedding a_i is $2d$ and we keep the same size for the hidden state in each LSTM unit. Therefore, the multimodal representations after max pooling for aligned frame embedding series can be written as $r \in \mathbf{R}^{4d}$ because of the bidirectional structure. Finally, two fully connected layers are applied to compute the cross-entropy loss.

VI. Experiments

A. Datasets

We used the DEAP multimodal dataset [26] to perform and validate our proposed alignment method. This emotion analysis dataset records physiological signals from 1280 trials on 32 healthy participants where 40 trials were conducted by each participant. Every participant gives ratings between 1 to 9 on four criteria(Arousal, Valence, Dominance, and Likeness) to 40 different music videos separately. Arousal and Valence are two criteria widely investigated in affective computing. The emotion classification tasks on the DEAP dataset are commonly considered as two-class classification problems for each criterion where subjective-ratings are assigned into high and low with the threshold of 5. Their preprocessed data contains 40 signal channels for each trial with 128 Hz sampling rate and 63 seconds length including 3 seconds pre-trial time in the beginning. There are 32 channels of EEG signals (based on 10-20 system) and other 8 peripheral physiological signals including 2 EOG channels, 2 EMG channels, 1 GSR channel, etc.

B. Experiment Setup

We perform experiments on three physiological signals from the DEAP dataset: EEG, EMG and, GSR. Following the Physio2Video methods, we prepare three kinds of physiological signal feature videos for the above three signals respectively. Since there are 1280 trials for 32 participants (40 trials for each one) in the DEAP dataset in total, we get 3×2018 feature videos.

Two video frame encoders are trained for each video pair: (EEG & EMG), (EEG & GSR), and (EMG & GSR) according to the canonical correlation loss defined at Equation 7, so in total 6 encoders are trained. We use Stochastic Gradient Descent (SGD) as the optimizer with a learning rate of 0.001, a weight decay of 5×10^{-4} and a mini-batches size of 16. We use video frame encoder pairs in this training to enable video frames to be aligned after non-linear embedding.

The evaluation model trained on DEAP emotion classification task is determined based on the 5-fold cross-validation on all 1280 trials. The limited sample size is the main reason to choose cross-validation. Except for the classification accuracy, F_1 score is also measured to evaluate the robustness of models because of the binary classification target. We also use SGD optimizer here with the same learning rate, weight decay and mini-batch size as the training of above encoders used for alignment. If the training accuracy for each epoch decreases, we multiply the learning rate by 0.9. The training will be stopped if the learning rate is less than 5×10^{-6}.

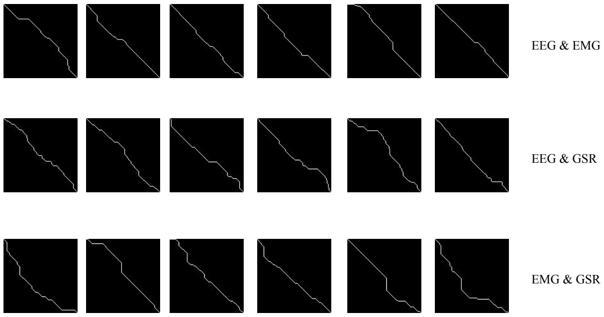


Fig. 7. Alignment path for physiological signals in three different alignment cases.

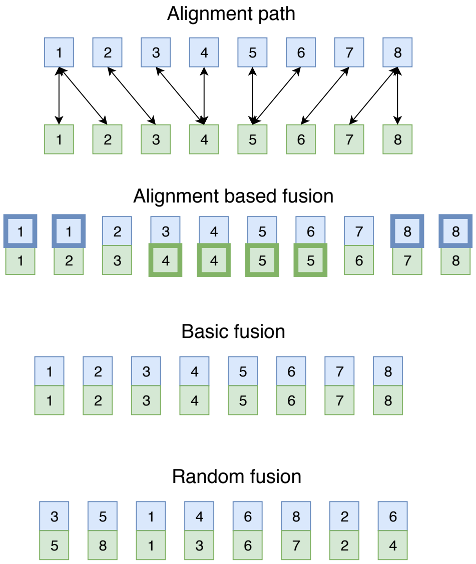


Fig. 8. Given the alignment path, alignment based feature fusion will concatenate corresponding pairs in alignment path. If one feature vector is aligned more than once, it will be copied. Basic fusion will simply concatenate two feature vectors with the same time step. Random fusion will shuffle two series and concatenate feature vectors with the same new time step.

C. Results

Figure 7 shows the alignment path examples for three types of multimodal physiological signals alignment cases based on our pre-trained video frame encoders defined in Section IV. These real alignment paths correspond to the alignment demo in Figure 1.

Before evaluating the above alignment results, we perform emotion recognition on each single physiological feature video by using a RCNN model which has the same modeling pipeline as our evaluation model but without concatenating other features in Figure 6. The unimodal performance of the RCNN model on each signal are shown in Table II.

TABLE II
5-FOLD CLASSIFICATION ACCURACY AND F_1 SCORE FOR EEG, EMG, AND GSR

	Arousal		Valence	
	accuracy	F_1	*accuracy*	F_1
EEG	**0.6149**	0.7040	0.5657	0.6972
EMG	0.5758	**0.7301**	**0.5807**	**0.7338**
GSR	0.5761	0.7279	0.5539	0.7121

We perform our RCNN based evaluation model with three different feature fusion strategies (including our alignment based fusion) on DEAP emotion recognition tasks. There are two baseline fusion strategies: Base Fusion and Random Fusion. Base Fusion will naively align video frames which are in the same time step and concatenate their corresponding embeddings which are extracted from task related CNN encoders. Random fusion will shuffle time steps of frames randomly for each input feature video and randomly concatenated two frame embeddings. Figure 8 visualizes the three different fusion mechanisms.

In the (EEG & EMG) feature fusion case in Table III, our alignment based fusion model performs best for Arousal classification on both accuracy (0.6583) and F_1 (0.7423) criteria. It did not achieve the best accuracy on Valence but is the most robust one in terms of F_1 score. Table IV and Table V show the results of the fusion cases of (EEG & GSR) and (EMG GSR) respectively. The results in these two tables are similar with Table III where alignment based fusion performs best on three measurement (accuracy for Arousal, F_1 for Arousal, F_1 for Valence) among four measurements in total. Also, we could observe that our methodology always performs best on accuracy and F_1 score for Arousal.

TABLE III
5-FOLD CLASSIFICATION ACCURACY AND F_1 SCORE ON BASELINE MODELS AND ALIGNMENT BASED MODEL BY USING EEG AND EMG

	Arousal		Valence	
	accuracy	F_1	*accuracy*	F_1
Basic	0.6361	0.7403	**0.6211**	0.6255
Random	0.6294	0.7024	0.5609	0.7031
Alignment	0.6583	**0.7423**	0.6133	**0.7265**

TABLE IV
5-FOLD CLASSIFICATION ACCURACY AND F_1 SCORE ON BASELINE MODELS AND ALIGNMENT BASED MODEL BY USING EEG AND GSR

	Arousal		Valence	
	accuracy	F_1	*accuracy*	F_1
Basic	0.6273	0.6751	0.5781	0.6901
Random	0.5986	0.71846	**0.6016**	0.6789
Alignment	**0.64378**	**0.7204**	0.5977	**0.7209**

TABLE V
5-FOLD CLASSIFICATION ACCURACY AND F_1 SCORE ON BASELINE MODELS AND ALIGNMENT BASED MODEL BY USING EMG AND GSR

	Arousal		Valence	
	accuracy	F_1	*accuracy*	F_1
Basic	0.5906	0.7094	0.**5641**	0.7046
Random	0.5922	0.6973	0.5498	0.7094
Alignment	**0.5945**	**0.7101**	0.5586	**0.7127**

Handcrafted features based	**Results**
[26]	Arousal (acc: 0.5700, F_1: 0.5330) , Valence (accuracy: 0.6270, F_1: 0.6080)
[3] (participant-specific)	Arousal (accuracy: 0.7306, F_1: -), Valence (accuracy: 0.7314, F_1: -)
[51] (participant-specific)	Arousal (accuracy: 0.7719, F_1: 0.6901), Valence (accuracy: 0.7617, F_1: 7243)
Deep features based	**Results**
[42]	Arousal (accuracy: 0.5120, F_1: -), Valence (accuracy: 0.6090, F_1: -)
[27]	Arousal (accuracy: 0.6420, F_1: -), Valence (accuracy: 0.5840, F_1: -)
[36]	Arousal (accuracy: 0.6590, F_1: -), Valence (accuracy: -, F_1: -)
our EEG & EMG fusion	Arousal (accuracy: 0.6583, F_1: 0.7432), Valence (accuracy: 0.6133, F_1: 7209)

TABLE VI
PERFORMANCE COMPARISON WITH OTHER WORKS.

Based on three multimodal fusion tables and the unimodal performance table before, we could see that adding one more modality will improve the original unimodal classification accuracy on Arousal and Valence regardless the added modality and feature fusion strategy in most cases. This demonstrates the efficiency of multimodal fusion.

Many other works in Table VI investigated multimodal fusion based emotion recognition on DEAP dataset. However, their experiments setup and model evaluation metrics are every different because there are no evaluation standards for the DEAP dataset.

VII. DISCUSSION

Although some work [3], [51] in Table VI achieved over 0.7 classification accuracy on the DEAP dataset, it can not be concluded that their multimodal fusion frameworks are better than ours. Their experiment sets are based on participant-specific data which means that the final results are the averaged performance on each participant. This is much easier since the variation of data distribution in one subject is smaller than across subjects. This type of experimental setting is common when the model relies on manually defined features like (Power Spectral Density) PSD and Relative Power Energy (RPE) because the manually defined feature usually varies across different participants. It is hard for a model to extract deep information if the model is not deep neural network based. Therefore we list another three deep neural network based models [27], [36], [42] to compare with our results. All of them trained their models by using cross-subjects settings. The results show that our alignment based fusion model is competitive compared with others.

Considering the unbalanced emotion labels of Arousal and Valence in the DEAP dataset, the classification accuracy is less representative than the F_1 score to evaluate the classification performance [51]. Therefore our model still shows its effectiveness even though the alignment based fusion does not get the best classification accuracy on Valence because we achieved best on each F_1 criteria in all three fusion occasions.

Although temporal information has been lost for these randomly fused features, the neural network could still extract time irrelevant features which are useful for emotion recognition. Also, randomly shuffled feature series will be considered as different input samples. In this case, random fusion will augment the training dataset dramatically. Therefore, we find random fusion performs better on accuracy than our alignment based method on Valence criteria in Table IV, but this small difference is likely due to noise being just 0.004 difference.

In Table III and Table V, basic fusion works better on the classification accuracy of Valence, again only on accuracy, and the difference is 0.008. It may again be noise or caused by an inappropriate step size. In Figure 9, the sub-intervals from the two stream correspond with each other in the real case if they have the same letter (we use A as an example). As we can see in this figure, The real interval of misalignment is 0.25 second, but the smallest step we have to perform is 0.5 which is the step size in our Physio2Video technique. It will keep the misalignment in this case after we perform time warping. However, using a too small step size will increase the total time step numbers in the series. This will dramatically increase the computation cost of the training process of generalized alignment model in Figure 5.

VIII. Conclusion

We present a generalized alignment methodology for multimodal physiological signals. As prepossessing, wave-liked digital signals are transformed into physiological feature videos by our Pysio2Video technique. We trained two deep CNN based frame encoders to convert two physiological feature videos into two vector series where their canonical correlations are maximized. Our alignment results of vectors in the coordinated space represent the alignment for their corresponding video frames. In order to evaluate the alignment results, we introduced the RCNN based multimodal fusion model to perform emotion recognition on the DEAP dataset. Experiments results show that our alignment based multimodal feature fusion performs better than the other two baseline feature fusion models with potential or actual misalignment. Thus, we can see that our alignment result could relieve the misalignment problem in the multimodal fusion domain.

References

[1] Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K., 2013. Deep canonical correlation analysis. In International Conference on Machine Learning, 12471255.

[2] Anguera, X.; Luque, J.; and Gracia, C., 2014. Audio-to-text alignment for speech recognition with very limited resources. In Fifteenth Annual Conference of the Inter-national Speech Communication Association.

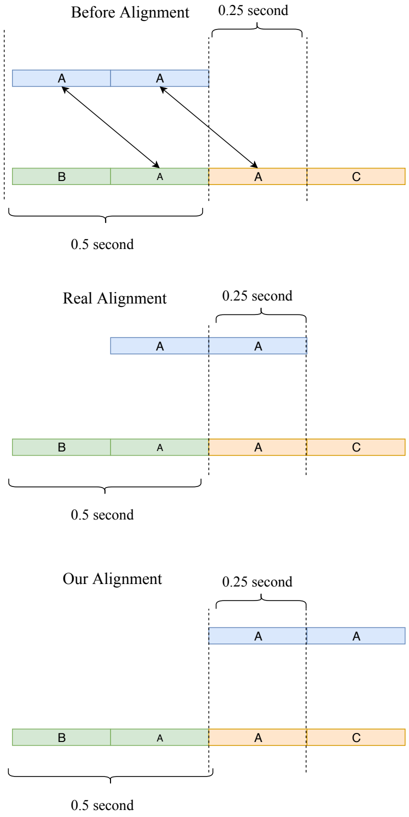


Fig. 9. Given two time series which have the upper and lower distribution in this figure. There are two sub-intervals in the upper one and four sub-intervals in the lower one. Sub-intervals with the same letter correspond with each other and they are in the same time step if they have the same color. The two sub-intervals which are denoted as A in the upper one correspond with the two sub-interval which are also donated as A in the lower one. The misalignment of two series is only 0.25 seconds and they should be aligned as such in the 'Real Alignment' condition. However, the step size is 0.5 seconds so it will keep the misalignment in our case as shown in 'Our Alignment' condition.

[3] Atkinson, J. and Campos, D., 2016. Improving bci-based emotion recognition by combining eeg feature selection and kernel classifiers.Expert Systems with Applications, 47 (2016), 3541.

[4] Bahrepour, M.; Meratnia, N.; Taghikhaki, Z.; and Havinga, P. J., 2011. Sen-sor fusion-based activity recognition for parkinson patients. In Sensor Fusion-Foundation and Applications. InTech.

[5] Baltruaitis, T.; Ahuja, C.; and Morency, L.-P., 2018. Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, (2018).

[6] Bao, S.-D.; Zhang, Y.-T.; andShen, L.-F., 2006. Physiological signal based entity authentication for body area sensor networks and mobile health care systems. In2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, 24552458.IEEE.

[7] Bashivan, P.; Bidelman, G. M.; and Yeasin, M., 2014. Spectro temporal dynamics of the eeg during working memory encoding and maintenance predicts individ-ual behavioral capacity.European Journal of Neuroscience, 40, 12 (2014),

[8] Bashivan, P.; Rish, I.; Yeasin, M.; and Codella, N., 2015. Learning representations from eeg with deep recurrent-convolutional neural networks. arXiv preprintarXiv:1511.06448, (2015).

[9] Besson, M.; Von Czettriz, G.; and Bax, R., 1999. Wireless medical diagnosis and monitoring equipment. US Patent 5,957,854.

[10] Bojanowski, P.; Lajugie, R.; Grave, E.; Bach, F.; Laptev, I.; Ponce, J.; and Schmid,C., 2015. Weakly-supervised alignment of video with text. In Proceedings of the IEEE international conference on computer vision, 44624470.

[11] Chambon, S.; Galtier, M. N.; Arnal, P. J.; Wainrib, G.; and Gramfort, A., 2018. Adeep learning architecture for temporal sleep stage classification using multivariateand multimodal time series. IEEE Transactions on Neural Systems and Rehabilitation Engineering, (2018).

[12] Chen, S. And Jin, Q., 2015. Multi-modal dimensional emotion recognition using recurrent neural networks. In Proceedings of the 5th International Workshop on Au-dio/Visual Emotion Challenge, 4956. ACM.

[13] Côté-Allard, U.; Fall, C. L.; Drouin, A.; Campeau-Lecours, A.; Gosselin, C.;Glette, K.; Laviolette, F.; and Gosselin, B., 2018. Deep learning for electromyo-graphic hand gesture signal classification by leveraging transfer learning.arXivpreprint arXiv:1801.07756, (2018).

[14] Gallese, V.; Fadiga, L.; Fogassi, L.; and Rizzolatti, G., 1996. Action recognition inthe premotor cortex.Brain, 119, 2 (1996), 593609.

[15] Ganeshapillai, G. And Guttag, J. V., 2011. Weighted time warping for temporalsegmentation of multi-parameter physiological signals. (2011).

[16] Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R., 2017. Tall: Temporal activity localization via language query. arXiv preprint arXiv:1705.02101, (2017).

[17] Gao, L.; Bourke, A. K.;and Nelson, J., 2011. A system for activity recognition using multi-sensor fusion. In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, 78697872. IEEE.

[18] Gravina, R.; Alinia, P.; Ghasemzadeh, H.; and Fortino, G., 2017. Multi-sensor fu-sion in body sensor networks: State-of-the-art and research challenges.Information Fusion, 35 (2017), 6880.

[19] Grosse, P.; Cassidy, M.; andBrown, P., 2002. Eegemg, megemg and emgemgfrequency analysis: physiological principles and clinical applications.Clinical Neurophysiology, 113, 10 (2002), 15231531.

[20] Hoerl, A. E. And Kennard, R. W., 1970. Ridge regression: Biased estimation for nonorthogonal problems.Technometrics, 12, 1 (1970), 5567.

[21] Yao, Y., Plested, J. and Gedeon, T., 2018, December. Deep Feature Learning and Visualization for EEG Recording Using Autoencoders. In International Conference on Neural Information Processing (pp. 554-566). Springer, Cham.

[22] Yao, Y., Plested, J. and Gedeon, T., 2018, December. A Feature Filter for EEG Using Cycle-GAN Structure. In International Conference on Neural Information Processing (pp. 567-576). Springer, Cham.

[23] Yao, Y., Plested, J., Gedeon, T., Liu, Y. and Wang, Z., 2019, January. Improved Techniques for Building EEG FeatureFilters. Submitted to International Joint Conference on Neural Networks.

[24] Hu, D.; Li, X.; et al., 2016. Temporal multimodal learning in audiovisual speech recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 35743582.

[25] Karpathy, A. And Fei-Fei, L., 2015. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, 31283137.

[26] Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun,T.; Nijholt, A.; and Patras, I., 2012. Deap: A database for emotion analysis; using physiological signals. IEEE Transactions on Affective Computing, 3, 1 (2012), 1831.

[27] Li, X.; Zhang, P.; Song, D.; Yu, G.; Hou, Y.; and Hu, B., 2015. Eeg based emotion identification using unsupervised deep feature learning. (2015).

[28] Liang, H.; Zhu, C.; Yoshioka, M.; Ueda, N.; Tian, Y.; Iwata, Y.; Yu, H.; Duan, F.; and Yan, Y., 2017. Estimation of emg signal for shoulder joint based on eeg signals for the control of upper-limb power assistance devices. In Robotics and Automation(ICRA), 2017 IEEE International Conference on, 60206025. IEEE.

[29] Liang, H.; Zhu, C.; Yoshikawa, Y.; Yoshioka, M.; Uemoto, K.; Yu, H.; Yan, Y.; and Duan, F., 2014. Emg estimation from eegs for constructing a power assist system.In Robotics and Biomimetics (ROBIO), 2014 IEEE International Conference on, 419424.IEEE.

[30] Maurer, U.; Smailagic, A.; Siewiorek, D. P.; and Deisher, M., 2006. Activity recognition and monitoring using multiple sensors on different body positions. In Wearable and Implantable Body Sensor Networks, 2006. BSN 2006. International Work-shop on, 4pp. IEEE.

[31] Mller, M., 2007. Dynamic time warping.Information retrieval for music and motion,(2007), 6984.

[32] Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y., 2011. Multimodal deep learning. In Proceedings of the 28th international conference on machine learning(ICML-11), 689696.

[33] Nicolaou, M. A.; Gunes, H.;andPantic, M., 2011. Continuous prediction of spon-taneous affect from multiple cues and modalities in valence-arousal space.IEEETransactions on Affective Computing, 2, 2 (2011), 92105.

[34] Peng, L.; Chen, L.; Wu, X.; Guo, H.; and Chen, G., 2017. Hierarchical complex activity representation and recognition using topic model and classifier level fusion.IEEE Trans. Biomed. Engineering, 64, 6 (2017), 13691379

[35] Ravi, N.; Dandekar, N.; Mysore, P.;and Littman, M. L., 2005. Activity recognition from accelerometer data. InAaai, vol. 5, 15411546.

[36] Said, A. B.; Mohamed, A.; Elfouly, T.; Harras, K.; and Wang, Z. J., 2017. Multimodal deep learning approach for joint eeg-emg data compression and classification. In Wireless Communications and Networking Conference (WCNC), 2017 IEEE,16. IEEE.

[37] Salvador, S. And Chan, P., 2007. Toward accurate dynamic time warping in lineartime and space.Intelligent Data Analysis, 11, 5 (2007), 561580

[38] Spampinato, C.; Palazzo, S.; Kavasidis, I.; Giordano, D.; Shah, M.; and Souly,N., 2016. Deep learning human mind for automated visual classification. arXivpreprint arXiv:1609.00344, (2016).

[39] Trigeorgis, G.; Nicolaou, M. A.; Schuller, B. W.; and Zafeiriou, S., 2018. Deep canonical time warping for simultaneous alignment and representation learning of sequences.IEEE Transactions on Pattern Analysis Machine Intelligence, , 5 (2018),11281138.

[40] Varatharajan, R.; Manogaran, G.; Priyan, M.; and Sundarasekar, R., 2017.Wearable sensor devices for early detection of alzheimer disease using dynamic time warping algorithm.Cluster Computing, (2017), 110.

[41] Verma, G. K. and Tiwary, U. S., 2014. Multimodal fusion framework: A multireso-lution approach for emotion classification and recognition from physiological sig-nals. NeuroImage, 102 (2014), 162172

[42] Wang, D.andShang, Y., 2013. Modeling physiological data with deep belief net-works. International journal of information and education technology (IJIET), 3, 5 (2013),505.

[43] Wang, X.; Gu, Y.; Xiong, Z.; Cui, Z.; andZhang, T., 2014. Silk-molded flexible, ultra sensitive, and highly stable electronic skin for monitoring human physiological signals.Advanced materials, 26, 9 (2014), 13361342.

[44] Ward, J. A.; Lukowicz, P.; Troster, G.; and Starner, T. E., 2006. Activity recognition of assembly tasks using body-worn microphones and accelerometers. IEEE trans-actions on pattern analysis and machine intelligence, 28, 10 (2006), 15531567.

[45] Wöllmer, M.; Kaiser, M.; Eyben, F.; Schuller, B.; and Rigoll, G., 2013. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework.Image and Vision Computing, 31, 2 (2013), 153163

[46] Wu, Z.; Cai, L.; and Meng, H., 2006. Multi-level fusion of audio and visual fea-tures for speaker identification. In International Conference on Biometrics, 493499. Springer.

[47] Xia, P.; Hu, J.; and Peng, Y., 2018. Emg-based estimation of limb movement using deep learning with recurrent convolutional neural networks.Artificial organs, 42, 5(2018), E67E77.

[48] Yang, X.; Ramesh, P.; Chitta, R.; Madhvanath, S.; Bernal, E. A.; and Luo,J., 2017. Deep multimodal representation learning from temporal data.CoRR,abs/1704.03152, (2017).

[49] Yao, Y., Plested, J. and Gedeon, T., 2018, December. Deep Feature Learning and Visualization for EEG Recording Using Autoencoders. In International Conference on Neural Information Processing (pp. 554-566). Springer, Cham.

[50] Yin, X.andChen, Q., 2016. Deep metric learning autoencoder for nonlinear tem-poral alignment of human motion. In Robotics and Automation (ICRA), 2016 IEEE International Conference on, 21602166. IEEE.

[51] Yin, Z.; Zhao, M.; Wang, Y.; Yang, J.; and Zhang, J., 2017. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model.Computer methods and programs in biomedicine, 140 (2017), 93110.

[52] Zhou, F. and De laTorre, F., 2016. Generalized canonical time warping.IEEE transactions on pattern analysis and machine intelligence, 38, 2 (2016), 279294.

[53] Zhou, F.and Torre, F., 2009. Canonical time warping for alignment of human behavior. In Advances in neural information processing systems, 22862294.