HAL open science

Fusion of Physiological and Behavioural Signals on SPD Manifolds with Application to Stress and Pain Detection

Yujin Wu, Mohamed Daoudi, Ali Amad, Laurent Sparrow, D Fabien

To cite this version:

Yujin Wu, Mohamed Daoudi, Ali Amad, Laurent Sparrow, D Fabien. Fusion of Physiological and Behavioural Signals on SPD Manifolds with Application to Stress and Pain Detection. International Conference on Systems, Man, and Cybernetics, Oct 2022, Prague, Czech Republic. hal-03883660

HAL Id: hal-03883660

https://hal.science/hal-03883660

Submitted on 3 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Fusion of Physiological and Behavioural Signals on SPD Manifolds with Application to Stress and Pain Detection

Yujin WU
CNRS, UMR 9189 CRIStAL
University of Lille, Centrale Lille
Lille, France 59000
yujin.wu.etu@univ-lille.fr

Mohamed Daoudi
CNRS, UMR 9189 CRIStAL
University of Lille, Centrale Lille
Lille, France 59000
mohamed.daoudi@imt-nord-europe.fr

Ali Amad
Lille Neuroscience & Cognition U1172 INSERM
University of Lille
Lille, France 59000
ali.amad@univ-lille.fr

Laurent Sparrow,
CNRS, UMR 9193 SCALab
University of Lille
Lille, France 59000
laurent.sparrow@univ-lille.fr

Fabien D'Hondt
Lille Neuroscience & Cognition U1172 INSERM
University of Lille
Lille, France 59000
fabien.d-hondt@univ-lille.fr

***Abstract*—Existing multimodal stress/pain recognition approaches generally extract features from different modalities independently and thus ignore cross-modality correlations. This paper proposes a novel geometric framework for multimodal stress/pain detection utilizing Symmetric Positive Definite (SPD) matrices as a representation that incorporates the correlation relationship of physiological and behavioural signals from covariance and cross-covariance. Considering the non-linearity of the Riemannian manifold of SPD matrices, well-known machine learning techniques are not suited to classify these matrices. Therefore, a tangent space mapping method is adopted to map the derived SPD matrix sequences to the vector sequences in the tangent space where the LSTM-based network can be applied for classification. The proposed framework has been evaluated on two public multimodal datasets, achieving both the state-of-the-art results for stress and pain detection tasks.**

***Index Terms*—stress detection, pain detection, multimodal fusion, covariance matrix, symmetric positive definite manifold.**

I. Introduction

Among a wide range of applications for monitoring and controlling human physical and mental health, stress/pain detection has drawn massive attention. Stress is a set of complex physiological, cognitive, and behavioral responses that are triggered by a challenging condition. When exposed to several stressors over a long period of time, a person's mental and physical state can be negatively influenced, which can further result in chronic health problems [1]. In order to detect stress at an early stage to reduce its impact on health conditions, stress detection has been widely used in many fields such as automated driver assistance [2], academic environment [3], social communication [4], etc. On the other hand, pain is another unpleasant sensation that we may encounter in our daily lives. Its accurate assessment plays a key role in diagnosing the condition, monitoring post-operative progress and optimising treatment options [5]. Thus, automated pain detection is another emerging area of healthcare applications designed to precisely assess pain to mitigate subject errors associated with patient self-reporting. A variety of sensors can be applied to collect stress/pain indicators from different perspectives, whereby the corresponding approaches for stress/pain detection can be mainly divided into three categories: 1) physiological based methods via bio-signals (e.g., electroencephalogram (EEG), electrocardiogram (ECG), electrodermal activity (EDA), etc.); 2) behavioural-based methods via physical signals (e.g., facial expressions, speech, body movements, etc.); 3) multimodal-based methods via a combination of physiological and behavioural signals. As complementary information between multimodalities contributes more to the robustness and reliability of the system, therefore, stress/pain detection combining physiological and behavioural indicators has become more attractive. However, how to effectively fuse multimodal data remains an important challenge for such systems. Most existing multimodal-based approaches learn features independently in each modality and eventually integrate them at the feature level or decision level. Thus underlying correlations between multiple modalities are ignored. In this study, we address the above problem by introducing a geometric framework that employs symmetric positive definite (SPD) matrices extracted from physiological and behavioural signals as a joint multimodal feature representation on SPD manifold for stress/pain detection. Continuous multimodal data recording can first be converted into SPD matrix sequences. The tangent space mapping method is then applied to locally flatten the manifold and approximate the SPD matrix sequences by tangent space vector sequences, where an LSTM-based deep neural network can be implemented to learn the context correlations of input sequences for classification. The overview of the proposed method is illustrated in Fig. 1. To the best of our knowledge, this is the first exploration of applying the SPD matrix to fuse behavioural signals and physiological signals for stress/pain

The proposed work was supported by the French State, managed by the National Agency for Research (ANR) under the Investments for the future program with reference ANR-16-IDEX-0004 ULNE.

detection.

II. Related Work

The multimodal framework is promising to improve the performance of stress/pain detection. However, few studies can be found using combined data from different fields (i.e. physiological and behavioural).

Multimodal data fusion: The multimodal fusion techniques can be divided into three categories: early, intermediate and late fusion. In an early fusion, features acquired by different sensors were fused into a unique representation before performing a learning task. Compared with the other two fusions, early fusion is widely explored in literature. Aigrain et al. [6] captured body video, high-resolution facial images and physiological signals from 25 subjects during a mental arithmetic test for stress detection, where 101 features from behavioural and physiological signals were extracted to train an SVM classifier. Rastgoo et al. [7] proposed a multimodal fusion framework for driver stress detection where the ECG signal, vehicle data, and contextual data were each entered into separate CNNs to extract modality-specific features. The LSTM-based network was trained on the concatenated multimodal feature embeddings to detect stress levels. Werner et al. [8] collected facial distances and gradient-based features from video frames which were combined with the statistical features calculated from biological signals. The resulting multimodal vectors were employed to train a random forest model for pain assessment. The multimodal fusion approaches described above show promising results for stress/pain detection. However, when simply splicing features from each modality for classification, only correlations within individual modalities were considered and the potential of applying inter-modal interactions to further boost performance is neglected.

Symmetric Positive Definite (SPD) matrices: Recently, covariance-based representations have gained great popularity in computer vision and machine learning. This success can be explained by three major advantages. Firstly, several characteristics can be fused into a single tensor and deliver higher-order statistical information. Secondly, the covariance matrices are symmetric positive definite (SPD) matrices with well-established mathematical theoretical properties [9], [10]. In addition, the SPD matrices have shown impressive accomplishments in many real-world applications such as pedestrian detection [11], facial expression recognition [12], brain-computer interfaces [13], etc. However, all of these applications concentrate solely on behavioural perspectives or physiological perspectives. Liu et al. [14] proposed a multimodal emotion recognition approach using video and audio modalities. Covariance matrix, linear subspace, and Gaussian distribution were built from facial video frames and regarded as points on Riemannian manifolds. Subsequently, the similarity matrix calculated using different Riemann kernels is fed into multiple classifiers (i.e., SVM, partial least squares, and logistic regression). However, they only constructed the SPD matrix for the video modality, the audio features were extracted using an existing toolkit, and fed into the same type of classifier as the video modality. In the end, the final fusion of the two modalities was established on the decision level. Thus, the exploration of inter-modal correlations using SPD matrix is still missing here. In the work of [10], they presented a more general covariance-based SPD representation, containing additional cross-covariance information from different time steps for action recognition. Inspired by [10], we migrate this new type of tensor representation to the scenario of multimodal stress/pain detection and show its effectiveness in this paper. Different from the previous work, we fuse behavioural and physiological information into one single SPD matrix-based representation, which not only incorporates intra-modal correlations, but also allows for exchanges across two modalities. To the best of our knowledge, this is the first use of the geometry of SPD manifold matrices to merge physiological and behavioural signals.

III. Proposed Method

A. Symmetric positive definite (SPD) matrix for multimodal signal

Let $\mathbf{x_i} = [v_1, ..., v_D]^T \in \mathbb{R}^D, D \geq 2$ represent a multimodal signal vector comprising the behavioural and physiological signals at the i-th timestamp, where the number of signals is denoted by D. A short-segment can be extracted from the continuous signal recordings of a trial, resulting in a centered matrix $\mathbf{X} = [\mathbf{x_1}, ..., \mathbf{x_i} ..., \mathbf{x_N}] \in \mathbb{R}^{D \times N}$, where N is the number of time instants for each segment. The outer product operation (denoted by the symbol $\otimes$) is then performed on all signal column vector pairs $(\mathbf{x_i}, \mathbf{x_j})$ for i and j $= 1, ..., N$ in $\mathbf{X}$ and consequently produces a partition matrix :

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{x_1} \otimes \mathbf{x_1} & \cdots & \mathbf{x_1} \otimes \mathbf{x_N} \\ \mathbf{x_2} \otimes \mathbf{x_1} & \cdots & \mathbf{x_2} \otimes \mathbf{x_N} \\ \vdots & \ddots & \vdots \\ \mathbf{x_N} \otimes \mathbf{x_1} & ... & \mathbf{x_N} \otimes \mathbf{x_N} \end{bmatrix} \in \mathbb{R}^{DN \times DN} \quad (1)$$

where the element of $\mathbf{\Omega}$ at position (i, j) is given by:

$$\mathbf{\Omega}(\mathbf{i}, \mathbf{j}) = \mathbf{x_i} \otimes \mathbf{x_j} = \mathbf{x_i}\mathbf{x_j}^T \in \mathbb{R}^{D \times D} \quad (2)$$

The sample covariance matrix $\mathbf{S}$ which is a SPD matrix can be derived from the diagonal elements of $\mathbf{\Omega}$:

$$\mathbf{S} = \frac{1}{N-1} \sum_{i=1}^{N} \mathbf{\Omega}(\mathbf{i}, \mathbf{i}) = \frac{1}{N-1} \sum_{i=1}^{N} \mathbf{x_i}\mathbf{x_i}^T \in \mathbb{R}^{D \times D} \quad (3)$$

The second SPD matrix defined as cross-covariance can be extracted from the off-diagonal elements of $\mathbf{\Omega}$, which contains the correlation information of the signal vectors at different timestamps and is denoted by $\mathbf{C}$:

$$\begin{aligned} \mathbf{C} &= \frac{1}{N^2 - N} \sum_{i=1, j=1, i \neq j}^{N,N} \mathbf{\Omega}(\mathbf{i}, \mathbf{j}) \\ &= \frac{1}{N^2 - N} \sum_{i=1, j=1, i \neq j}^{N,N} \mathbf{x_i}\mathbf{x_j}^T \in \mathbb{R}^{D \times D} \end{aligned} \quad (4)$$

The covariance $\mathbf{S}$ and cross-covariance $\mathbf{C}$, are then combined in a symmetric manner to form a new block matrix which remains a SPD matrix and is denoted by $\mathbf{P}$:

$$\mathbf{P} = \begin{bmatrix} \mathbf{S} & \mathbf{C} & \cdots & \mathbf{C} \\ \mathbf{C} & \mathbf{S} & \cdots & \mathbf{C} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C} & \mathbf{C} & \cdots & \mathbf{S} \end{bmatrix} \in \mathbb{R}^{(m \times D) \times (m \times D)} \quad (5)$$

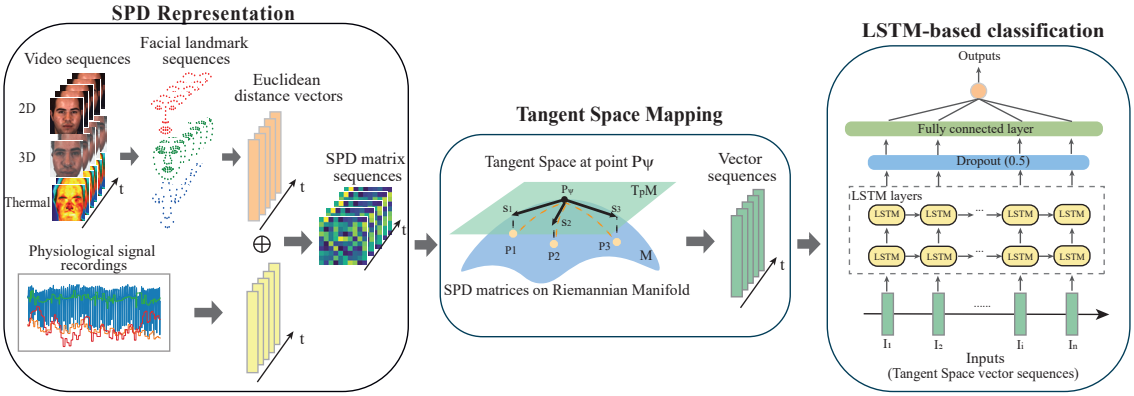

Fig. 1. Overview of the proposed framework. First, the SPD matrix sequences that incorporate the correlation information between multimodal data (i.e. physiological and behavioural signals) can be extracted from the segmented data records. Subsequently, the tangent space mapping projects the SPD matrix sequences to the vector sequences in the tangent space. Finally, these vectors can be used as input to the LSTM-based classification network for stress/pain recognition.

where m is the dimension of the new SPD matrix $\mathbf{P}$, in other words, $\mathbf{P}$ is composed of m blocks of $\mathbf{S}$ and $m(m-1)$ blocks of $\mathbf{C}$. A larger m corresponds to a higher computational cost, while the information ratio of the covariance to the cross-covariance decreases [10].

B. Riemannian Geometry of symmetric positive definite (SPD) Matrices

Tangent Space Mapping: A number of metrics have been proposed for SPD matrices. In this study, we consider one of the most popular metrics, the log-Euclidean metric [15], for its simplicity to compute. Let $\mathbf{P_i}$ and $\mathbf{P_j}$ be any two points on the SPD manifold then the log-Euclidean metric is defined as:

$$\delta_R(\mathbf{P_i}, \mathbf{P_j}) = \|\log(\mathbf{P_i}^{-1}\mathbf{P_j})\|_F \tag{6}$$

where $\|\cdot\|_F$ is the Frobenius norm operator, and $\log(.,.)$ is the matrix log operateur. The geometric mean [16], [17] of a set of SPD matrices $\{\mathbf{P_1}, \mathbf{P_2}, \ldots, \mathbf{P_I}\}, I \geq 1$, $\mathbf{P_i} \in \mathcal{P}(n)$ can be derived using the Riemannian geodesic distance (6):

$$\psi(\mathbf{P_1}, \mathbf{P_2}, \ldots, \mathbf{P_I}) = \arg\min_{\mathbf{P} \in P(n)} \sum_{i=1}^{I} \delta_R(\mathbf{P}, \mathbf{P_i})^2 \tag{7}$$

The tangent space at $\mathbf{P} \in \mathcal{M}$, denoted by $T_{\mathbf{P}}\mathcal{M}$, is a real vector space containing all tangent vectors to $\mathcal{M}$ at $\mathbf{P}$. The exponential map $\mathrm{Exp}_{\mathbf{P}}(\cdot) : T_{\mathbf{P}}\mathcal{M} \rightarrow \mathcal{M}$ and its inverse, the logarithm map, $\mathrm{Log}_{\mathbf{P}}(\cdot) : \mathcal{M} \rightarrow T_{\mathbf{P}}\mathcal{M}$ are defined over Riemannian manifolds for exchanging between the manifold and its tangent space at $\mathbf{P}$. A tangent vector can be mapped to a point on the manifold through the exponential operator. The logarithm operator can project a point on the manifold to the tangent space [18]. In the work of [16], tangent space mapping was introduced which employs the tangent space that lies at the geometric mean (7) of the entire set of SPD matrices: $\mathbf{P}_\psi = \psi(\mathbf{P_i}, i = 1, \ldots, I)$. Each SPD matrix $\mathbf{P_i}$ can be projected into the tangent space as a vector s_i with a dimension of $\frac{n(n+1)}{2}$:

$$s_i = vec({\mathbf{P}_\psi}^{-\frac{1}{2}} \mathrm{Log}_{\mathbf{P}_\psi}(\mathbf{P_i}) {\mathbf{P}_\psi}^{-\frac{1}{2}}) \in \mathbb{R}^{\frac{n(n+1)}{2}} \tag{8}$$

where the vector operator of one SPD matrix $\mathbf{P}$ is defined as

$$vec(\mathbf{P}) = [p_{1,1}, \sqrt{2}p_{1,2}, \sqrt{2}p_{1,3}, \cdots, p_{2,2}, \sqrt{2}p_{2,3}, \cdots, p_{n,n}] \tag{9}$$

$p_{i,j} \in \mathbf{P}$ is the element of $\mathbf{P}$. This operation is designed to achieve vectorization by weighting the upper triangular part of the matrix, where the diagonal elements and off-diagonal elements are multiplied by the unit weights and weights of $\sqrt{2}$, respectively [19].

C. Classification of SPD matrix sequences

Let us consider a set $\mathcal{L} = \bigcup_{q=1}^{Q} \mathcal{L}^q$ consisting of data from Q subjects. The qth subset $\mathcal{L}^q$ is represented by $\mathcal{L}^q = \{([\mathbf{P_i}, ..., \mathbf{P_{i+T}}], y_i), \mathbf{P_i} \in \mathcal{P}(n), i \in [1, |\mathcal{L}^q|]\}$, which is the set of SPD matrix sequences, where T is the length of the sequence and y_i is the corresponding stress/pain labels, such that $y_i = f([\mathbf{P_i}, ..., \mathbf{P_{i+T}}])$. In current literature, few approaches have been suggested to tackle the non-linearity of the SPD manifold. A common method of dealing with the non-linearity is to estimate the manifold-value data by projecting them into the tangent space of a specific point on the manifold (e.g., the geometric mean of the data) [20]. In this work, we will adopt this approach to solve our problem. For each subset, its corresponding geometric mean ${\mathbf{P}_\psi}^q$ can be obtained with the equation (7). Then each SPD matrix in the qth subset $\mathcal{L}^q$ is mapped into the tangent space and the derived corresponding subset of vector sequences is denoted by $s^q = \left\{([s_i, ..., s_{i+T}], y_i), s_i \in \mathbb{R}^{\frac{n(n+1)}{2}}, i \in [1, |s^q|]\right\}$ using the equation (8). The above process will be repeated for each subject's data. In the end, $\mathcal{L}^* = \bigcup_{q=1}^{Q} s^q$ is considered as the input of the LSTM-based deep neural network in Fig. 1. During training, the temporal contextual relationships in the tangent space vector sequence are explored by the 2-layer LSTM network, and the output features are then fed into the fully connected layer, followed by the sigmoid function to obtain the predicted probabilities.

IV. Experiments and Results

To evaluate the validity of the proposed method, we conducted multimodal stress detection experiments on WESAD dataset while multimodal pain detection experiments were carried out on BP4D+ dataset.

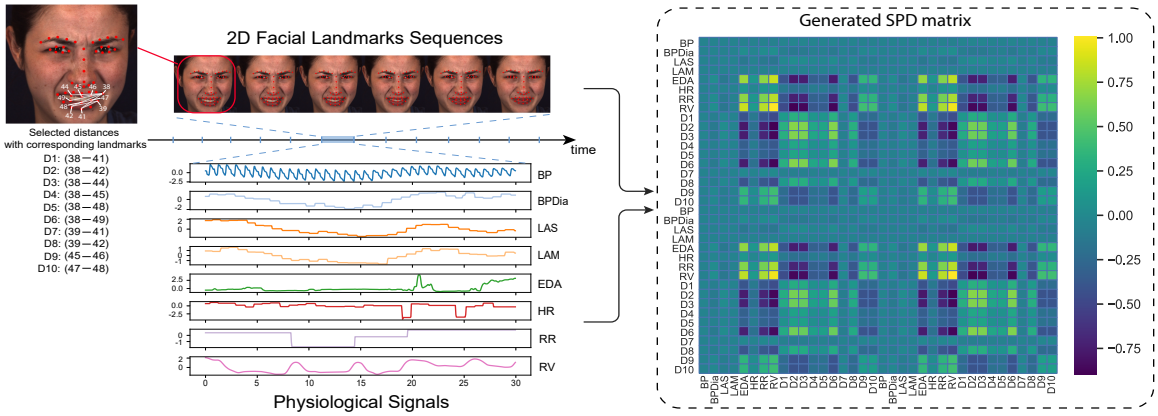

Fig. 2. The multimodal SPD representation generated by a pain sample in the BP4D+ dataset, where correlations within and across two modalities (i.e., vision and physiology) can be observed. (D1-D10: 10 distances automatically selected based on Anova F-value, BP: raw blood pressure, BPDia: diastolic blood pressure, LAS: systolic blood pressure, LAM: mean blood pressure, EDA: electrodermal activity, HR: heart rate, RR: respiration rate, and RV: respiration volts.)

A. Datasets

The **WESAD dataset** [21] is one of the most widely used public datasets for stress and affect recognition. In a restricted laboratory setting, multimodal data consisting of motion and physiological signals from 15 subjects were captured by two wearable devices, a wrist sensor and a chest sensor, respectively, and the experimental protocol was designed to stimulate three different emotional states (baseline, stress, amusement) in the participants. Based on previous work [21]–[23], a binary stress detection problem (*stress* vs. *non-stress*) can be formulated on the WESAD dataset by fusing baseline class and amusement class to form the *non-stress* class.

The **BP4D+ dataset** [24] is a large-scale multimodal spontaneous emotion database. 140 subjects were required to complete 10 tasks to elicit 10 different emotions, during which 2D RGB images, 3D model sequences, thermal videos and 8 physiological signal sequences with 1.5 million frames were captured by different sensors. In addition, the metadata are also provided, including 2D/3D/thermal facial landmarks, hand-labelled FACS codes and auto-tracked head poses. In this work, we focus on the identification of pain. As the dataset only provides the most facially-expressive segments for four emotions: happy, embarrassment, fear and pain, we therefore performed a pain detection task (*pain* vs. *non-pain*) by combining happy, embarrassment, fear as the *non-pain* class as proposed in [25]. An example of 2D texture images/3D model sequences/thermal images from the *Pain* class and their corresponding facial landmarks is shown in Fig. 3.

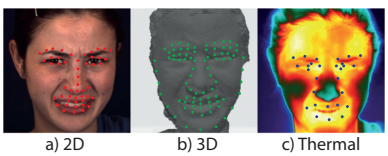

Fig. 3. An example of 2D texture images/3D model sequences/thermal images from the *Pain* class and their corresponding facial landmarks provided in BP4D+ dataset.

B. Data Preprocessing and SPD matrix construction

For WESAD dataset, physiological and motion signals captured by the two sensors are filtered and downsampled to the same frequency, followed by a 10-second non-overlapping segmentation as proposed by [26]. Finally, the SPD matrix series can be calculated from multimodal data segments. For BP4D+ dataset, we first calculated the Euclidean distance between each of the two facial landmarks for each video frame using the provided 2D/3D/thermal facial landmarks, and automatically selected the 10 most discriminative distances by feature selection based on Anova F-value. Then the distance vector and the synchronized physiological signal vector were concatenated together to form a new augmented vector. Subsequently, all the augmented vectors are sliced into 1-second non-overlapping data segments. In the end, the obtained SPD matrix sequences can be extracted from prepared data segments. Fig. 2 shows an SPD matrix of a *pain* class sample in BP4D+ dataset that generated from the 10 facial distances between 2D landmarks and 8 physiological signals. During the pain task, the subject was asked to immerse hands in ice water, and her mouth was involuntarily opened, hence the distances automatically selected were all based on the landmarks in the lip region, which is consistent with findings in the literature that lip movements such as oblique lip raising [27], horizontal lip stretching [28], etc. are related to pain. Both intra-modal and inter-modal correlations can be observed from the SPD matrix. Among all physiological signals, there was a strong correlation between electrodermal activity (EDA) and respiration signal (respiration rate (RR) and respiration volt (RV)) when the subject was suffering from pain. Among the selected distances, the 2nd, 3rd and 6th distances were more correlated with each other. Furthermore, the association between physiological and facial indicators can be explored, i.e. EDA, RR and RV were also positively correlated with the 9th and 10th distances.

C. Implementation and evaluation

All the classification models were implemented using Pytorch. To avoid overfitting, dropout operation was employed after the LSTM layers with a hidden state dimension of

TABLE I
STRESS DETECTION PERFORMANCE ON WESAD DATASET USING THE SAMPLE COVARIANCE MATRIX S, THE CROSS-COVARIANCE MATRIX C AND THE PROPOSED SPD REPRESENTATION P.

Modalities	Wrist Sensor				Chest Sensor				Wrist+Chest Sensors			
	Physio + Motion		Only Physio		Physio + Motion		Only Physio		Physio + Motion		Only Physio	
Features	Accuracy	F1score	Accuracy	F1score	Accuracy	F1score	Accuracy	F1score	Accuracy	F1score	Accuracy	F1score
with S (baseline)	89.23	88.07	90.13	89.73	90.64	90.49	90.66	89.02	92.88	92.78	93.06	93.00
with C (baseline)	90.91	90.04	88.07	87.28	91.30	90.94	89.42	88.44	93.12	92.73	93.08	91.84
with P (m = 2)	94.08	93.81	92.15	92.20	94.55	94.53	92.76	92.61	94.19	94.23	95.29	95.33
with P (m = 3)	93.00	92.38	93.10	92.78	**95.54**	**94.76**	**92.96**	**92.64**	94.71	93.81	95.09	94.15
with P (m = 4)	**94.65**	**93.99**	**93.47**	**93.25**	94.25	93.97	92.12	91.75	**96.35**	**95.76**	**96.88**	**96.44**

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON WESAD DATASET (*Stress* VS. *Non-stress*)

Modality	Methods	Accuracy	F1score
Wrist	Schmidt et al. [21]	87.12	84.11
	Samyoun et al. [29]	92.1	89.7
	Gil-Martin et al. [30]	92.7	92.55
	Huynh et al. [22]	93.14	-
	Lai et al. [23]	94.16	93.62
	Proposed Method	**94.65**	**93.99**
Chest	Samyoun et al. [29]	91.1	90.2
	Schmidt et al. [21]	92.83	91.07
	Gil-Martin et al. [30]	93.10	93.01
	Lai et al. [23]	**96.69**	**96.61**
	Proposed Method	95.54	94.76
Wrist + Chest	Schmidt et al. [21]	92.28	90.74
	Samyoun et al. [29]	94.7	93.4
	Gil-Martin et al. [30]	96.62	96.63
	Lai et al. [23]	**97.75**	**97.74**
	Proposed Method	96.88	96.44

128. The Adam optimizer with a learning rate lr=0.001 was selected to minimize the binary cross-entropy loss function during model training of 50 epoch. Decay coefficients of the first and second moment estimation β_1 and β_2 were set to 0.9 and 0.999, respectively. In the end, the proposed framework is evaluated using Leave-One-Subject-Out cross validation (LOSO-CV) on WESAD dataset followed by [21]–[23] and Subject independent 10-fold cross validation on BP4D+ dataset followed by [25] with two selected metrics: Accuracy and F1score. Both experimental protocols assess the model's capacity to generalize across previously unseen subject data.

D. Stress detection results on WESAD

Binary stress detection experiments were performed using the wrist/chest-based data of all subjects in WESAD dataset. The evaluation results of different modality combinations are reported in Table I. We first verify the necessity of fusing the sample covariance matrix S and cross-covariance matrix C to form the proposed representation P. From the experimental results, it can be observed that the proposed SPD matrix optimizes the detection performance for all modality combinations, compared to those using only matrix S and matrix C. Secondly, to investigate the impact of increasing the dimensionality of the SPD matrix defined in the equation (5) on the detection results, we performed experiments using representations with different numbers of blocks of S and C (e.g., m=2 meaning that P consists of 2 blocks of S and 2 blocks of C). We found that using more blocks of S and C to compose the proposed representation P slightly enhanced the classification performance. This can be attributed to the increased proportion of cross-covariance information in the high-dimensional SPD matrix, further demonstrating the benefit of correlations between multiple modalities at different instants for the classification task. Besides, we only test up to m=4, since we noticed that sometimes the m=3 case performs best, following the trade-off between classification performance and computational cost. Finally, we also explored whether fusing data from different modalities, i.e., motion signals and physiological signals, could lead to a performance gain. For experiments based on wrist/chest sensor data, combining these two types of data for detection yielded better performance, compared to results based on physiological signals only. When fusing data from both devices (wrist + chest), results using all modalities were not improved, which can be attributed to the redundant information generated by the same motion signals from both devices. When combining all physiological-based modalities, the highest performance (96.88% accuracy and 96.44% F1 score) was obtained with $P(m = 4)$.

Comparison with State-of-the-art: To validate the effectiveness of the proposed method for fusing motion and physiological information, Table II shows the comparison results with 5 state-of-the-art methods using multimodal features. For a fair comparison, only methods that use the same experimental protocol were considered. In the work of Schmidt et al. [21], features from the time and frequency domains are used to train a variety of traditional machine learning classifiers, among which the LDA (Linear discriminant analysis) model achieved the best performance. Samyoun et al. [29] presented GAN/RNN/MLP-based deep model to generate gold standard chest sensor features from wrist data, and classified emulated features with various machine learning algorithms (e.g., Extra Trees, Random Forest, etc.). Gil-Martin et al. [30] proposed a CNN-MLP architecture to extract meaningful features from the Fourier transform (FFT), cube root (CR) and constant q spectral transform (CQT) of signal sub-window. Huynh et al. [22] used filter bank as model input and automatically selected the optimal model for each modality from 10,000 deep neural networks for training. Finally, features of all modalities were concatenated for classification. Lai et al. [23] employed residual-temporal convolution network to process the filtered multimodal signals and proposed various fusion strategies. The above work for comparison simply spliced features

TABLE III
UNIMODAL AND MULTIMODAL PAIN DETECTION PERFORMANCE ON BP4D+ DATASET USING THE SAMPLE COVARIANCE MATRIX S, THE CROSS-COVARIANCE MATRIX C AND THE PROPOSED SPD REPRESENTATION P. (2D/3D/THERMAL: 2D/3D/THERMAL FACIAL LANDMARKS; PHYSIO: ALL PHYSIOLOGICAL SIGNALS).

	Uni-modality								Multi-modality					
Modalities	2D		3D		Thermal		Physio		2D + Physio		3D + Physio		Thermal + Physio	
Features	Accuracy	F1score	Accuracy	F1score	Accuracy	F1score	Accuracy	F1score	Accuracy	F1score	Accuracy	F1score	Accuracy	F1score
with S (baseline)	88.76	85.58	86.47	85.83	81.46	80.10	77.53	72.84	91.29	90.16	87.07	86.18	82.28	80.38
with C (baseline)	86.96	86.43	88.84	85.85	82.60	82.11	76.70	72.41	90.49	87.72	90.26	87.36	84.24	78.56
with P(m = 2)	91.44	89.24	91.01	88.84	82.82	82.70	81.82	81.72	92.86	90.03	92.53	90.20	89.04	88.73
with P(m = 3)	**91.59**	**89.46**	91.13	89.04	83.18	82.93	**83.24**	**82.42**	93.04	90.85	**92.66**	**90.47**	88.82	88.69
with P(m = 4)	90.94	88.56	**91.27**	**89.30**	**83.53**	**83.37**	82.81	81.42	**93.45**	**91.37**	92.54	90.25	**89.07**	**88.96**

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS ON BP4D+ DATASET (*pain* VS. *non-pain*).

Modality	Methods	Accuracy	F1score
Physiology	Hinduja et al. [25]	77.7	30.0
	Proposed Method	**81.82**	**81.72**
Vision	Szczapa et al. [31]	81.86	77.34
	Choo et al. [32]	84.03	83.98
	Huang et al. [33]	86.43	85.71
	Proposed Method	**91.59**	**89.46**
Vision + Physiology	$^{(D)}$Szczapa et al. [31]	82.77	76.32
	$^{(F)}$Szczapa et al. [31]	84.32	78.83
	$^{(F)}$Huang et al. [33]	87.94	87.16
	$^{(D)}$Choo et al. [32]	89.08	88.68
	$^{(D)}$Huang et al. [33]	89.36	89.13
	$^{(F)}$Choo et al. [32]	89.80	89.46
	Proposed Method	**93.45**	**91.37**

$^{(D)}$: Decision level fusion, $^{(F)}$: Feature level fusion.

from different modalities for prediction and thus ignored the cross-modality correlations. Overall, our proposed method using the joint SPD representation achieves the state-of-the-art results on wrist sensor data and competitive results on chest/wrist+chest sensor data, respectively, demonstrating its efficiency for integrating multimodal data.

E. Pain detection results on BP4D+

To further assess the validity of the proposed method, we conducted unimodal and multimodal pain detection experiments on the BP4D+ dataset. The evaluation results are summarised in Table III. Similar to the process performed on WESAD, we first consider the results obtained using S or C alone as baseline to explore the importance of combining them in the proposed representation. Based on the results in Table III, we reach the same conclusion that the joint SPD representation can improve classification performance. In addition, we can infer that increasing the dimension of P can further boost performance for all modalities. In the end, we also noticed that all the multimodal settings exhibit performance gains compared to unimodal detection results. Among four unimodality (i.e. physiological signal, 2D/3D/thermal facial landmarks), the trained model has the best performance using 2D facial landmarks where recognition accuracy and F1score achieved 91.59%, 89.46% respectively. In the multimodal experiments, the best results with accuracy and F1score of 93.45% and 91.37% can be observed with 2D + Physio setting.

Comparison with State-of-the-art: Table IV shows the comparison results with 4 state-of-the-art methods using 2D facial landmarks and physiological signals. Due to the diversity of problem formulations and experimental settings, only a few pain detection efforts can be directly compared to our framework. In the work of [25], a random forest classifier was trained on features consisting of physiological signals and face action units for pain detection. Here we only presented the comparison results based on physiological signals with them, as we did not use AUs for the detection. Our proposed method improves the accuracy by about 4%. Moreover, our framework achieved a more balanced pain detection with an F1score of 81.72%. Since most pain detection datasets contain only vision-related information, very little pain recognition work has been carried out based on data from two different domains, i.e. vision and physiology. Therefore, to validate the effectiveness of our proposed approach on fused multimodal data, state-of-the-art pain recognition methods that accept only visual data were implemented and combined with our physiological signal-based model for comparison. For a fair comparison, only the facial landmark-based methods were considered. We used the code provided by the authors, and if the code was not available, we followed the parameters provided in their article. Szczapa et al. [31] represented the facial landmark sequences as trajectories on the Riemannian manifold. Each point of the trajectory is a Gram matrix computed from the 2D facial landmarks. Then the Global Alignment Kernel (GAK) was used to calculate the similarity matrix between the trajectories, which was used as feature for SVR-based (Support Vector Regression) pain estimation. To compare the classification performance, we replaced the SVR with an SVM (Support Vector Machine). Huang et al. [33] used a 1D CNN based architecture to extract discriminate features from the normalized distance between 2D facial landmarks for pain recognition. Choo et al. [32] employed a dual-layer 3D CNN for capturing the spatial-temporal features of the 2D facial landmark sequences. When comparing the performance of pain recognition based on solely visual information, our model performs better as shown in Table IV. To compare the performance based on multimodal information, we used two fusion techniques that are commonly used in the literature, feature level fusion and decision level fusion, respectively. We first note that the performance of all the vision-based models used for comparison is improved when

combined with our physiology-based model, providing side evidence that our model learns discriminative physiological features. Secondly, our model outperforms other multimodal approaches, both in terms of feature level fusion and decision level fusion, which confirms that the correlation between two modalities is well captured by the proposed method and that inter-modal communication can further contribute to the classification performance. Overall, our method achieves the state-of-the-art results on both unimodal data and multimodal data, validating again its effectiveness.

V. Conclusion

In this work, we explore for the first time the feasibility of SPD matrix-based representations for efficiently fusing physiological and behavioural signals, which can capture simultaneously correlation information within and across modalities. Tangent space mapping converts the generated SPD matrix time series into linear vectors for its application to the LSTM-based classification. The effectiveness of the proposed method was evaluated on public stress and pain detection datasets. In the end, the proposed framework shows the state-of-the-art results on both stress and pain detection tasks, respectively.

References

[1] S. Gedam and S. Paul, "A review on mental stress detection using wearable sensors and machine learning techniques," *IEEE Access*, vol. 9, pp. 84 045–84 066, 2021.

[2] L. Mou, C. Zhou, P. Zhao, B. Nakisa, M. N. Rastgoo, R. Jain, and W. Gao, "Driver stress detection via multimodal fusion using attention-based cnn-lstm," *Expert Systems with Applications*, vol. 173, p. 114693, 2021.

[3] A. Sano, A. Phillips, A. Yu, A. Mchill, S. Taylor, N. Jaques, C. Czeisler, E. Klerman, and R. Picard, "Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones," 2015, pp. 1–6.

[4] S. Sahu, E. Kithani, M. Motwani, S. Motwani, and A. Ahuja, "Stress Detection of Office Employees Using Sentiment Analysis," in *Emerging Technologies in Data Mining and Information Security*, A. E. Hassanien, S. Bhattacharyya, S. Chakrabati, A. Bhattacharya, and S. Dutta, Eds. Singapore: Springer Singapore, 2021, pp. 143–153.

[5] P. Werner, D. Lopez-Martinez, S. Walter, A. Al-Hamadi, S. Gruss, and R. W. Picard, "Automatic recognition methods supporting pain assessment: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 530–552, 2022.

[6] J. Aigrain, M. Spodenkiewicz, S. Dubuisson, M. Detyniecki, D. Cohen, and M. Chetouani, "Multimodal stress detection from multiple assessments," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 491–506, 2018.

[7] M. N. Rastgoo, B. Nakisa, F. Maire, A. Rakotonirainy, and V. Chandran, "Automatic driver stress level classification using multimodal deep learning," *Expert Systems with Applications*, vol. 138, p. 112793, 2019.

[8] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Automatic pain recognition from video and biomedical signals," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 4582–4587.

[9] R. Bhatia, *Positive Definite Matrices*. Prinecton University press, 2009.

[10] J. Zhang, Z. Feng, Y. Su, and M. Xing, "Cross-covariance matrix: Time-shifted correlations for 3D action recognition," *Signal Processing*, vol. 171, p. 107499, Jun. 2020.

[11] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on riemannian manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713–1727, 2008.

[12] N. Otberdout, A. Kacem, M. Daoudi, L. Ballihi, and S. Berretti, "Automatic analysis of facial expressions based on deep covariance trajectories," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 10, pp. 3892–3905, 2020.

[13] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Classification of covariance matrices using a Riemannian-based kernel for BCI applications," *Neurocomputing*, vol. 112, pp. 172–178, Jul. 2013.

[14] M. Liu, R. Wang, S. Li, Z. Huang, S. Shan, and X. Chen, "Video modeling and learning on Riemannian manifold for emotion recognition in the wild," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 113–124, Jun. 2016.

[15] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Log-Euclidean metrics for fast and simple calculus on diffusion tensors." *Magnetic Resonance in Medicine*, vol. 56, no. 2, pp. 411–421, Aug. 2006.

[16] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass Brain–Computer Interface Classification by Riemannian Geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, Apr. 2012.

[17] M. Moakher, "A differential geometric approach to the geometric mean of symmetric positive-definite matrices," *SIAM J. Matrix Analysis Applications*, pp. 735–747, 2005.

[18] M. Faraki, M. Palhang, and C. Sanderson, "Log-euclidean bag of words for human action recognition," *IET Computer Vision*, vol. 9, no. 3, pp. 331–339, jun 2015.

[19] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian Detection via Classification on Riemannian Manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713–1727, Oct. 2008.

[20] S. Jayasumana, R. I. Hartley, M. Salzmann, H. Li, and M. T. Harandi, "Kernel methods on riemannian manifolds with gaussian RBF kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2464–2477, 2015.

[21] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection," in *Proceedings of the 2018 on International Conference on Multimodal Interaction - ICMI '18*, Boulder, CO, USA, 2018, pp. 400–408.

[22] L. Huynh, T. Nguyen, T. Nguyen, S. Pirttikangas, and P. Siirtola, *StressNAS: Affect State and Stress Detection Using Neural Architecture Search*. New York, NY, USA: Association for Computing Machinery, 2021, p. 121–125.

[23] K. Lai, S. N. Yanushkevich, and V. P. Shmerko, "Intelligent stress monitoring assistant for first responders," *IEEE Access*, vol. 9, pp. 25 314–25 329, 2021.

[24] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, "Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 3438–3446.

[25] S. Hinduja, S. J. Canavan, and G. Kaur, "Multimodal fusion of physiological signals and facial action units for pain recognition," in *15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020, Buenos Aires, Argentina, November 16-20, 2020*. IEEE, 2020, pp. 577–581.

[26] P. Sarkar and A. Etemad, "Self-supervised ecg representation learning for emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[27] C. J. Patrick, K. D. Craig, and K. M. Prkachin, "Observer judgments of acute pain: facial action determinants," *Journal of Personality and Social Psychology*, vol. 50, no. 6, pp. 1291–1298, Jun. 1986.

[28] L. LeResche, "Facial expression in pain: A study of candid photographs," *Journal of Nonverbal Behavior*, vol. 7, no. 1, pp. 46–56, Sep. 1982.

[29] S. Samyoun, A. Sayeed Mondol, and J. A. Stankovic, "Stress Detection via Sensor Translation," in *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. Marina del Rey, CA, USA: IEEE, May 2020, pp. 19–26.

[30] M. Gil-Martin, R. San-Segundo, A. Mateos, and J. Ferreiros-Lopez, "Human stress detection with wearable sensors using convolutional neural networks," *IEEE Aerospace and Electronic Systems Magazine*, vol. 37, no. 1, pp. 60–70, 2022.

[31] B. Szczapa, M. Daoudi, S. Berretti, P. Pala, A. D. Bimbo, and Z. Hammal, "Automatic estimation of self-reported pain by interpretable representations of motion dynamics," in *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*. IEEE, 2020, pp. 2544–2550.

[32] K. W. Choo and T. Du, "Pain detection from facial landmarks using spatial-temporal deep neural network," in *Thirteenth International Conference on Digital Image Processing (ICDIP 2021)*, X. Jiang and H. Fujita, Eds., vol. 11878, International Society for Optics and Photonics. SPIE, 2021, pp. 593 – 597.

[33] Y. Huang, L. Qing, S. Xu, L. Wang, and Y. Peng, "Hybnet: a hybrid network structure for pain intensity estimation," *Vis. Comput.*, vol. 38, no. 3, pp. 871–882, 2022.