Subject-Dependent Emotion Recognition System Based on Multidimensional Electroencephalographic Signals: A Riemannian Geometry Approach

EMAN A. ABDEL-GHAFFAR[1], **YUJIN WU**[2], **AND MOHAMED DAOUDI**[2,3], **(Senior Member, IEEE)**

[1]Electrical Engineering Department, Faculty of Engineering Shoubra, Benha University, Cairo 11629, Egypt
[2]CNRS, Centrale Lille, UMR 9189 CRIStAL, University of Lille, 59000 Lille, France
[3]IMT Lille Douai, Institut Mines-Télécom, Centre for Digital Systems, 59000 Lille, France

Corresponding author: Eman A. Abdel-Ghaffar (eman.mohamed@feng.bu.edu.eg)

ABSTRACT Emotion recognition plays an important role in human computer interaction systems as it helps the computer in understanding human behavior and their decision making process. Using Electroencephalographic (EEG) signals in emotion recognition offers a direct assessment on the inner state of human mind. This study aims to build a subject dependent emotion recognition system that differentiate between high and low levels of valance and arousal, using multidimensional EEG signals. Our system offers a transfer learning- minimum distance to Riemannian mean (TL-MDRM) framework. In this work, we perform two pre-processing stages. In the first stage, we analyze the EEG signals to investigate their non-Gaussianity and determine the most appropriate signal distribution. Using several statistical and goodness of fit tests, T-distribution was found to be the most appropriate distribution. Covariance matrix estimations plays a crucial step in manifold learning technique, based on the most suitable signal distribution the covariance matrix estimation technique is chosen. In the second stage, we perform transfer learning to deal with cross-session variability by generating a unique reference point for each participant and performing affine transformation for the covariance matrices on the symmetric positive definite (SPD) manifold around this point. The results show that, TL process improved the performance even when assuming Gaussian distribution, while assuming T-distribution with TL improved the performance further.

INDEX TERMS Covariance matrix estimation, emotion recognition, multidimensional EEG signals, Riemannian manifold, signal analysis, SPD matrices, transfer learning.

I. INTRODUCTION

In the last few decades there has been a considerably growing attention towards human computer interaction (HCI) systems, but most of those systems are still not efficient in understanding human emotions. The ability to classify human emotional responses to different stimuli opens the door for new innovations in HCI.

The most commonly used methods for extracting human emotional states are facial expressions [1], [2], human voice [3], [4], Electroencephalography (EEG) signals [5]–[7],

The associate editor coordinating the review of this manuscript and approving it for publication was Ludovico Minati.

or by combining multiple modalities for more accurate systems [8]–[10].

Recording EEG signals requires the placement of multiple electrodes at certain locations on the scalp. Due to the current advance in technology, EEG signals capturing devices became wearable, portable, easy to use, and even wireless this makes the use of EEG signals very attractive as it is noninvasive, fast and inexpensive. There exist wide area of applications for the use of EEG-based emotion recognition systems such as, e-learning [11], e-health care [12], entertainment and gaming [13].

EEG signals based emotion recognition systems extracts features from time domain, frequency domain, or joint time

and frequency domains. In time domain techniques, statistical features such as signal power, energy, entropy, . . . , etc. have been excessively used [14]. In frequency domain, features are extracted from different frequency bands and added together to form the feature vectors used for classification [5], [15]. New techniques [16] combining both time and frequency domain features were introduced. Other EEG based emotion recognition systems use raw EEG signals for extracting useful spatial and temporal information's from different channels and different time samples [17].

The use of Riemannian geometry in brain computer interface (BCI) and in studding brain disorders is attracting attention due to its simplicity, robustness, and accuracy. In [18] Fruehwirt *et al.* used Riemannian tangent space mapping in studying Alzheimer's disease. Yuan *et al.* in [19] performed Epileptic seizure detection in the space of the symmetric positive definite (SPD) matrices using Log-Euclidean Gaussian Kernel-Based sparse. In [20] Congedo *et al.* offered a complete review on the use of Riemannian geometry for EEG-based brain computer interface. In our previous work [6] on emotion recognition using EEG signals we used Minimum Distance to Riemannian Mean (MDRM) classifier, for classifying four classes of emotions. Different frequency-bands, channel combinations, and geometric mean generation techniques were examined.

Data in real-world scenarios tends to be corrupted with outliers and/or exhibit heavy tails. In such cases using sample covariance matrices (assuming Gaussian distribution) offers biased results, as it tends to ignore outliers and heavy tail of the data. In [21] J. Charles *et al.* tested EEG signal distribution and found that Laplace distribution proved to be more robust estimator than Gaussian Distribution, and in [22] they also used Laplace distribution for better statistical reduction of multi-channel EEG data. N. Nazmia *et al.* [23] used goodness to fit tests to find the most appropriate EEG signal distribution, they found that the Generalized Extreme Value distribution is the most appropriate distribution for describing the EEG and the Electromyography (EMG) signals.

EEG data recorded on different sessions and/or from different participants tends to have statistical variability. This forms a challenge that faces brain computer interface systems that tries to reuse data from previous sessions/subjects. Transfer learning (TL) is an approach used to overcome this variability. Their are several studies that perform transfer learning on Riemannian geometry framework. In [24] Zanini *et al.* performed cross-session and cross-subject transfer learning, their TL approach is based on modifying MDRM classifier they called their new framework Riemannian alignment (RA)-MDRM. Zanini *et al.* dealt with cross-session/subject variability as a geometric transformation (shift) of covariance matrices on the Riemannian manifold with respect to a reference state, and when the brain is performing a specific task, covariance matrices are shifted in the same direction over the SPD manifold. RA-MDRM technique showed improvement over normal MDRM classifier over

both motor imagery and event-related potentials datasets. He *et al.* [25] aligned the EEG trials in the Euclidean space which makes them more similar and enhances the learning performance for new participants, feature extraction and classification processes are performed on the aligned data. Working in the Euclidean space gave them the advantage of faster computation and also the ability of using various classifiers. In [26] P. Rodrigues *et al.* matched the statistical distribution of two datasets using three steps geometrical transformations (translation, scaling, and rotation) in order to make the shape of the statistical distribution of the data as similar as possible. In [27] Lin *et al.* proposes a machine-learning strategy called robust principal component analysis (RPCA)-embedded transfer learning (TL) frame work aims to generate a personalized cross-day emotion-classification model with less labeled data, while avoiding intra and inter-individual difference. They used the Riemannian distance to measure the between-session similarity and thereby pair most similar auxiliary source sessions to a target session for TL. Yair *et al.* [28] proposed an unsupervised approach for domain adaptation using parallel transportation on the cone manifold of SPD matrices. In [29] Wang *et al.* proposed a domain adaptation SPD matrix network (daSPDnet) to solve subject independent emotion recognition problem. They combined prototype learning with the Riemannian metric and design a new prototype loss, which aims to calculate the geometric mean of the SPD matrix set in the low-dimensional representation layer. Their daSPDnet can extract an intrinsic emotional representation shared between different subject.

The objective of our work is to offer a TL-MDRM framework to classify human emotions and study the effect of using the most appropriate covariance matrix estimation technique (based on determining the closest EEG signal distribution) on the system performance.A two step process was performed. In the first step, we analyse the EEG signals using multiple statistical tests to prove that the signals are heavy tailed, then using two goodness of fit tests we compared between Gaussian distribution, T-distribution and Laplace distribution to find the most appropriate signal distribution. In the second step, we perform cross-session transfer learning, in which we use the pre-trial baseline signals to generate a unique reference point for each participant. Then the points on the SPD manifold are shifted towards this reference point to overcome variability's in different observations. Two experiments were carried out, using two different channel configurations on five different frequency bands.

The rest of this paper is organized as follows; In section (II) we introduce the basic concepts of Riemannian geometry and transfer learning. In section (III) EEG signal analysis is performed using several statistical tests and two goodness of fit tests. In section (IV) our complete methodology is introduced. Results and discussions reported in section (V). In section (VI) we conclude and draw some perspectives of the work.

II. RIEMANNIAN GEOMETRY

A. NOTATIONS

We denote by $M(N) = \left\{ \mathbf{M} \in \mathbb{R}^{N \times N} \right\}$ the space of $N \times N$ square matrices, $S(N) = \left\{ \mathbf{S} \in M(N), \mathbf{S}^T = \mathbf{S} \right\}$ the space in $M(N)$ of symmetric $N \times N$ square matrices. $P(N)$ is an open subset of $S(N)$ defined as $\mathcal{P}(N) = \left\{ \mathbf{P} \in S(N), \mathbf{u}^T \mathbf{P} \mathbf{u} > 0, \forall \mathbf{u} \in \mathbb{R}^N \right\}$. The space of $P(N)$ is the space of symmetric positive definite matrices (SPD), denoted by S_{++}^n. The SPD matrices are the covariance matrices that form a smooth Riemannninan manifold that have a non-positive curvature in the $N(N + 1)/2$ dimensional Euclidean space. Covariance matrix estimation forms a crucial step in manifold learning techniques and is explained in details in section (IV-D).

B. RIEMANNIAN DISTANCE

A non-singular square covariance matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ belongs to the set of symmetric positive-definite (SPD) matrices. Those SPD matrices form a connected Riemannian manifold Sym_N^+ [30]. As the Euclidean distance does not consider the inner curvature of the manifold, it can't be used to measure the distance between two points $\mathbf{A}, \mathbf{B} \in P(N)$. The distance here can be defined as the length of unique shortest path connecting the two points. This path is called the geodesic from $\mathbf{A}$ to $\mathbf{B}$ [31], [32].

There are several metrics used to measure the geodesic distances, each of them is more or less suitable based on the application. For any two SPD matrices $\mathbf{A}$ and $\mathbf{B}$, the Affine Invariant Riemannian Metric (AIRM) between them is defined as [30], [33]:

$$d(\mathbf{A}, \mathbf{B}) = \| \log \left(\mathbf{A}^{\frac{-1}{2}} \mathbf{B} \mathbf{A}^{\frac{-1}{2}} \right) \|_F \qquad (1)$$

where $\|\cdot\|_F$ is the *Frobenius norm* of a matrix. One of the important proprieties of this Riemannian metric in Eq. 1 is that it's invariance to affine transformations by any invertible matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$,

$$d(\mathbf{A}, \mathbf{B}) = d(\mathbf{D} \mathbf{A} \mathbf{D}^T, \mathbf{D} \mathbf{B} \mathbf{D}^T) \qquad (2)$$

This property of Riemannian distance is called the congruence invariance, which means that, the distance between any two SPD matrices is invariant with respect to any linear invertible transform in the data space [24], [26]. This property will be used in the transfer learning process (Section IV-E).

C. MEAN OF SPD MATRICES

The geometric mean is a suitable descriptor for the center of mass of the points on the SPD manifold. The Riemannian center of mass of m elements $\mathbf{A}_1, \ldots, \mathbf{A}_m$, called Karcher mean [32], [34] is defined as:

$$\mathbf{G}(\mathbf{A}_1, \ldots, \mathbf{A}_m) = \underset{\mathbf{A} \in \mathcal{P}(n)}{\operatorname{argmin}} \sum_{i=1}^m d^2(X, \mathbf{A}_i) \qquad (3)$$

with $d(\cdot, \cdot)$ is defined in Eq. (1). The notation $\operatorname{argmin} f(X)$ means the point X_0 at which the function f reaches its minimum value. The minimum in Eq. (3) is obtained at a unique

point $\mathbf{G}$ which represents the geometric mean and that forms the solution for the matrix equation:

$$\sum_{i=1}^m \log \left(\mathbf{A_i}^{\frac{-1}{2}} X \mathbf{A_i}^{\frac{-1}{2}} \right) = 0 \qquad (4)$$

Eq. (4) has a closed-form solution only for $m = 2$, for three or more matrices, there is no closed form solution and iterative algorithms should be used [26], [30].

D. MINIMUM DISTANCE TO RIEMANNIAN MEAN CLASSIFIER (MDRM)

In Euclidean spaces large number of standard classifiers could be used, but they are not suitable in our case as the space of S_{++}^n is non linear. A very simple and efficient classifier called Minimum Distance to Riemannian Mean (MDRM), which is based on nearest neighbor classifier could be used.

Given l the set of all labeled classes $l_i \in (l_1, l_2, \ldots, l_k)$, where k is the number of classes. During the training stage the mean for each class is generated $\hat{\mathbf{M}}(l_i)$. In the test stage, the geometric mean of the new observation is generated $\mathbf{M}$. The distance between $\mathbf{M}$ and each class mean $\hat{\mathbf{M}}(l_i)$ is computed. The new observation belongs to the class l according to the classification rule:

$$\hat{l} = \underset{l \in l_1, l_2, \ldots, l_k}{\operatorname{argmin}} \left\{ d \left(\mathbf{M}, \hat{\mathbf{M}}(l) \right) \right\} \qquad (5)$$

where $\hat{\mathbf{M}}(l)$ is the Riemannian mean of class l_k, $\mathbf{M}$ is the covariance matrix representing the mean of test observation, and $\hat{l}$ is the predicted class label of $\mathbf{M}$.

MDRM classifier works in the same way regardless of the data dimension (number of electrodes) and with any number of classes.

E. TRANSFER LEARNING

Zanini *et al.* [24] offered a modification over MDRM classifier, they called their new framework Riemannian alignment RA-MDRM. They dealt with cross session/subject variability as a geometric transformation (shift) of covariance matrices on the Riemannian manifold with respect to a reference state. When the brain is performing a specific task, covariance matrices are shifted in the same direction over the SPD manifold. Their method showed great improvement over normal MDRM in solving classification problem on motor imagery and event-related potentials datasets.

In RA-MDRM they first estimate the covariance matrix of the rest state. For motor imagery data the rest state is the recorded EEG data in the time window in which the participant is not engaged in the experiment, while in event-related potentials they used non-target stimuli as the rest state.

In RA-MDRM the covariance matrices representing the rest state is estimated $(\mathbf{R}_k^n)$, where k is the number of covariance matrices in the rest state, n is the session number. Then compute the mean of those matrices (denoted as $\overline{\mathbf{R}}$, $\overline{\mathbf{R}}^{(1)}$ and $\overline{\mathbf{R}}^{(2)}$ represents the reference points for

session 1 and 2 respectively). $\overline{\mathbf{R}}$ is used as a reference point to reduce the cross-session/subject variability by performing the transformation:

$$\widetilde{\mathbf{C}}_i^n = \left(\overline{\mathbf{R}}^{(n)}\right)^{\frac{-1}{2}} \mathbf{C}_i^n \left(\overline{\mathbf{R}}^{(n)}\right)^{\frac{-1}{2}} \qquad (6)$$

where $\mathbf{C}_i^n$, and $\widetilde{\mathbf{C}}_i^n$ are the ith covariance matrix in the nth session before and after shifting respectively.

Because of the congruence invariance propriety in Eq. (2) the transformation in Eq. (6) does not change the distance between the points that belong to the same session/subject.

III. EEG SIGNAL ANALYSIS

Covariance matrix estimation forms a vital and crucial step in manifold learning techniques, recently several researches tackled the problem of covariance matrix estimation from high dimensional data and from heavy tail distribution data. In [35] Ke *et al.* offered a method for estimating a stronger sample covariance matrix by introducing element-wise and spectrum wise truncation operators, and their M-estimator counterparts. Wei *et al.* [36] proposed an estimator of the covariance matrix under weak assumptions on the underlying distribution. In this work, we focus on understanding the underlying signal distribution. We use simple approaches for determining the most appropriate type of covariance matrix estimation technique by first determining the correct distribution of the EEG data.

In this section we analyse the EEG signals in DEAP dataset to determine the most appropriate distribution to which those signals belong. Based on the closest signal distribution we choose the most accurate technique for covariance matrix generation, we used two methods for covariance matrix generation (see section IV-D).

It is common when dealing with EEG signals to assume that, they exhibit Gaussian distribution. This assumption provides only a modest approximation for EEG data as a random variable [21]–[23]. In [21] J. Charles *et al.* performed Chi-square (χ^2) test, comparing EEG signals probability density function (pdf) against Gaussian distribution and Laplace distribution, they found that Laplace distribution proved to be more robust estimator. In [22] they also used Laplace distribution for better statistical reduction of multi-channel EEG data. Nazmia *et al.* [23] performed two Goodness-of-Fit tests Kolmogorov-Smirnov and Anderson Darling. They tested the EEG and Electromyography (EMG) signal distributions against Exponential distributions, Generalized Pareto distribution and Generalized Extreme Value distribution, they found that the Generalized Extreme Value distribution is the most appropriate distribution for describing the EMG and EEG signals.

In this work, we first perform several statistical tests on the EEG signals to determine whether they follow a Gaussian distribution or not. Then we use two goodness of fit tests Anderson Darling (A^2) and Watson (U^2) to compare the EEG signals distribution against Gaussian distribution, Laplace distribution and T-distribution.

A. OUTLIERS, SKEWNESS AND KURTOSIS TESTS

Outliers are points in data that differs significantly from other values they often indicate that the data has a heavy-tailed distribution and high skewness and that assuming a normal distribution is a modest assumption in this case.

In this part, we analyse the EEG signals (signals from all the 32 participants in DEAP dataset 40 trial each is used in this analysis) using different spatial window sizes to detect the percentage of outliers in each window-size. Through this work we define the outlier as a data value that is greater than three scaled median absolute deviations (MAD) away from the median. Leys *et al.* [37] stated that detecting outliers in data using median absolute deviation forms a more robust measure of data spread than using standard deviation around the mean. Scaled median absolute deviations for a random variable A, having N samples is defined as:

$$MAD = b * (M(|(A_i - M(A))|)) \qquad (7)$$

where M is the median, b is a scaling factor and $i = 1, 2, \ldots, N$. Then, Skewness and Kurtosis tests are used to describe the shape of the distribution and check for the non-Gaussianity of the EEG signals.

Skewness is a measure of the asymmetry of the probability distribution. Skewness close to zero is considered a symmetric distribution. A positive skewness implies a long left tail, which means that, the mass of the distribution is concentrated on the left part. A negative skewness on the other hand indicates a long right tail.

For measuring if the data are light-tailed or heavy-tailed relative to a Gaussian distribution Kurtosis is used. The kurtosis of the Gaussian distribution equals three. Kurtosis greater than three indicates that the data distribution have heavy-tail and Kurtosis less than three means that the data have light-tail.

Given a Random variable X having N samples and s standard deviation, Skewness and kurtosis are calculated as follows [38]:

$$Skew = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{s}\right)^3,$$

$$Kurt = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{s}\right)^4 \qquad (8)$$

Table 1 shows the percentage of outliers in each window (window size varies from one to eight seconds), the percentage of signals with light-tail, heavy-tail,-long left-tail, long right-tail, median kurtosis, and median degree of freedom.

From Table 1 we can see that the existence of outliers increase by increasing the window size and their values are significant, they can not be overlooked and should be taken into consideration. The kurtosis values are higher than three, which means that the percentage of heavy-tailed signals are much more than light-tailed signals. The skewness value is mostly grater than zero around 52% of the signals have a

**TABLE 1.** Percentage of outliers in each window (window size varies from 1 to 8 seconds), signals with long right-tail, long left-tail, light-tail, heavy-tail, median Kurtosis, and median degree of freedom.

Window size in seconds	1	2	3	**4**	5	6	7	8
Outliers	2.13	5.76	9.62	**13.64**	17.69	21.86	26.1	30.27
Long-Right tail	51.97	52.37	52.55	**52.57**	52.66	52.87	52.85	52.95
Long-Left tail	48.03	47.63	47.45	**47.43**	47.34	47.13	47.15	47.05
Heavy-Tailed	65.9	81.88	88.32	**91.82**	93.86	95.38	96.36	97.08
Light-Tailed	34.1	18.12	11.68	**8.18**	6.14	4.62	3.64	2.92
Median-Kurtosis	3.288	3.829	4.27	**4.63**	4.875	5.103	5.27	5.428
Median-degree of freedom	17.91	8.39	6.488	**5.7**	5.33	5.049	4.88	4.75
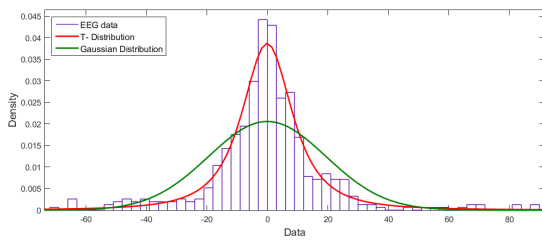
(a) Probability density function (PDF).
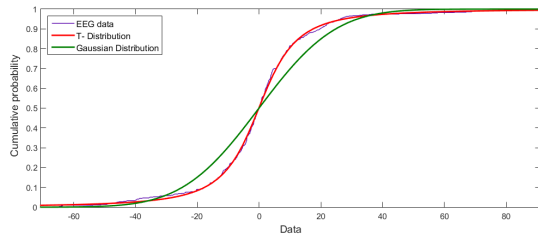
(b) Cumulative Distribution Function (CDF).
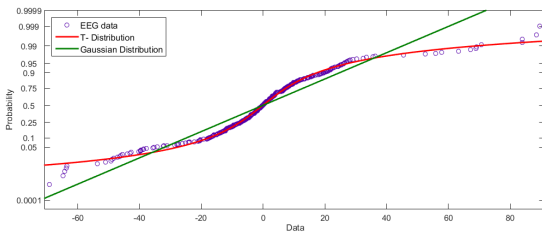
(c) Quantile-Quantile plot (QQP).

FIGURE 1. A comparison between EEG signal distribution against both T-distribution and Gaussian distribution. As an illustrative example we used the EEG signal from user number 19, trial 40, electrode *F*3, using 4s window size.

long right tail and the rest of the signals (around 47%) have a long left tail. Testing for outliers, skewness and kurtosis show that the EEG signals have heavy-tails and does not follow Gaussian distribution.

B. GOODNESS OF FIT TESTS

In this section we perform two goodness to fit tests to compare the EEG signal distribution against two distributions that are close to Gaussian but has heavier tails.

TABLE 2. A^2 (GD), A^2 (TD), A^2 (GD & TD) and A^2 (LD) are the results from Anderson Darling test that fails to reject the null hypothesis of Gaussian distribution, T-distribution, both (Gaussian distribution & T-distribution), and Laplace distribution respectively. U^2 (LD) is the result from Watson statistic test that fails to reject the null hypothesis of Laplace distribution.

A^2(LD)	A^2 (TD)	A^2(GD & TD)	A^2(LD)	U^2(LD)
27.047%	97.435%	25.52%	0.0378%	2.852%

The T-distribution is a probability distribution that is similar to the Gaussian Distribution (GD) with a bell shape but has heavier tails, i.e. it tends to have values that exist far from its mean. The existence of heavier-tails could be detected by a parameter of T-distribution called the degree of freedom, the higher the degree of freedom the more the distribution becomes close to Gaussian distribution (median degree of freedom is shown in the last row in Table 1. T-distribution tends to have bigger values of Kurtosis and smaller values of degree of freedom than Gaussian distribution. Laplace Distribution [39] also known as a double exponential distribution also have a bell shape but it is very sharp in the center and is used to model symmetric data with long tails.

1) Anderson Darling goodness of fit test (A^2): Is based on the cumulative probability distribution of data it is a modification of the Kolmogorov-Smirnov test and is known to gives more weight to the observations in the tails of the distribution. It is more sensitive to the existence of outliers and is better in detecting departure form normality specially in the tails of the distribution [40].
2) Watson statistic test (U^2): Is suggested to be the most power full when testing for Laplace distribution against other symmetric distributions. This test is quite powerful and provides equal sensitivity to the tails as to the median of the empirical distribution function [41].

The Anderson Darling test and Watson test for Laplace distribution is performed using *R* package "lawstat" [42], while the Anderson Darling test for Gaussian distribution and T-distribution is performed using MATLAB statistical toolbox.

Table 2 shows The percentage of EEG signals that fails to reject the null hypothesis at 0.05 significance level for the Gaussian distribution, T-distribution, or Laplace distribution using Anderson Darling test and Watson test. In Fig. 1 we show a comparison between EEG signal distribution against both Gaussian distribution and T-distribution using the Probability Density Function (PDF), Cumulative Distribution Function (CDF), and Quantile-Quantile plot (QQP). Results in Table 2 and Fig. 1 is generated from using 4s window size (the temporal window size that will be used in the emotion classification task). From Table 2 and Fig. 1 we can see that the EEG signal distribution is closer to T-distribution than Gaussian distribution and Laplace distribution.

IV. METHODOLOGY

A. DATASET

DEAP dataset [43] is a multimodal dataset that is used in analyzing human affective state. The electroencephalogram (EEG) and other physiological signals of 32 individuals (16 males and 16 females, with average age 26.9) were recorded while each of them was watching 40 one-minute music videos. EEG signals were recorded at a sampling rate of 512 Hz using 32 active AgCl electrodes placed around the scalp according to the 10-20 international positioning system [44]. The dataset was recorded in two sessions separated by a break (20 observations per session) and in two different labs participants 1-22 were recorded in Twente and participant 23-32 in Geneva [43].

The DEAP dataset has two versions, the first one is the original without pre-processing, while in the second one EEG signals were down sampled to 128 Hz, the Electrooculography (EOG) artifacts were removed, the signals were filtered from 4 to 45 Hz. In the pre-processed version each observation (trial) is 63s, in which the first 3s are baseline signals before the participant starts to be engaged in the experiment. In this work, we are studying emotion classification based on EEG signals using the pre-processed version and the 3s baseline signals are used to generate the reference point for each participant.

We performed our experiment once on the 18 electrodes placed on the upper half of the scalp and another time on the 32 electrodes placed around the complete scalp. The positions of our 18 electrodes are: FP1-FP2, AF3-AF4, F3-F4, F7-F8, FC1-FC2, FC5-FC6, C3-C4, T7-T8, FZ and CZ. The 32 electrodes are composed of the previous 18 electrodes plus CP1-CP2, CP5-CP6, P3-P4, P7-P8, PO3-PO4, O1-O2, PZ and OZ.

B. EMOTION LABELING

In DEAP dataset, participants were asked to label each trial by giving a score between 1 and 9 to rate the levels of arousal, valence, liking and dominance for each of the 40 one-minute videos. We divided the two dimensional emotion plane (see Fig. 2) into four classes according to the scores given by participants to valence (V) and arousal (A).
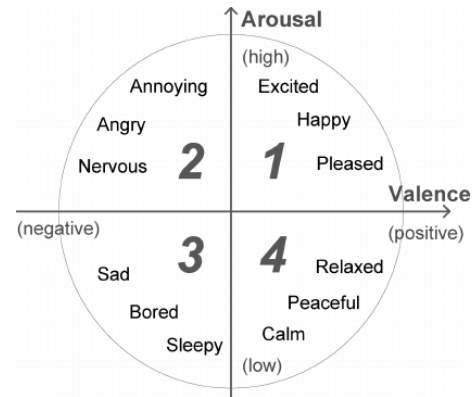

FIGURE 2. Valence-Arousal model for human emotions.

TABLE 3. Emotion classification labels. The valence and arousal scores used for labeling, the number of trials in the dataset that belongs to each label, and the average number of trials per participant belongs to each label.

Emotion Label	Scores	# Trials	#Trials/Participant
High Valance	$V > 5$	707	22
Low Valance	$V \leq 5$	575	18
High Arousal	$A > 5$	737	23
Low Arousal	$A \leq 5$	543	17

In this work, we use score 5 as a threshold with is commonly used when working with DEAP dataset [7], [15]. Two different binary classification problems were introduced for subject-dependent emotion recognition: The discrimination of low/high arousal (LA/HA), and low/high valance (LV/HV). Table 3 shows emotion classification labels, the number of trials in the dataset that belongs to each label, and the average number of trials per participant belongs to each label. Since there are a balanced number of observations in each class accuracy can be used for performance measure.

C. TEMPORAL WINDOWING

The EEG signals during each trial were recorded for 63s (60s trials and a 3s pre-trial baseline), which is much longer than the time needed for recognizing emotion states. For identifying the emotion state, EEG signals are windowed into short segments. Thammasan *et al.* [45] tested emotion recognition accuracy with window duration that varies from 1 to 8 seconds their result showed that increasing the window size reduces performance. Mohammadi *et al.* [15] stated that the emotion hold time is between 2 to 4 seconds and found that their best performance was achieved using 4s window. Through this work we use a window size of 4s with 50% overlap during the emotion recognition stage and 1s window with 50% overlap during reference point generation (used for cross-session transfer learning process (Section IV-E)).
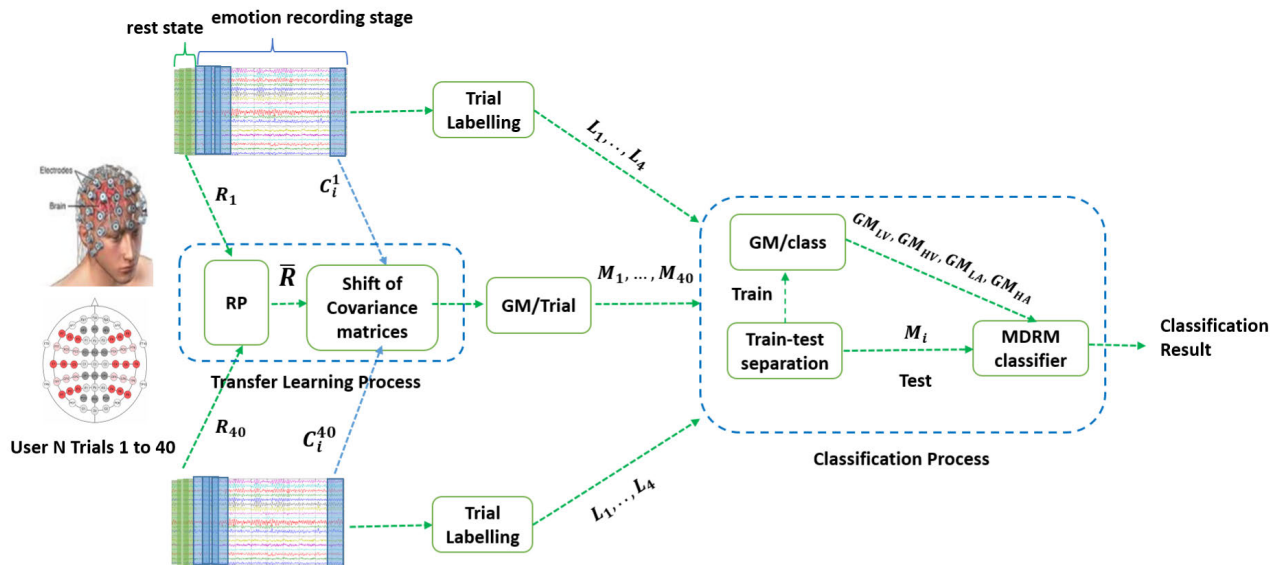
FIGURE 3. Proposed Subject-dependent EEG based emotion recognition system first experiment. For each user N ($N = 1, 2, \ldots, 32$), a two steps transfer learning process is performed. In the first step we generate a reference point ($R_{i, \ i=1,\ldots,40}$) from the rest state of each user observation, then a common reference point ($\bar{R}$) for each participant is generated from his 40 trials recorded during the two sessions (20 trials per session). Then, all the covariance matrices C_i^j (the ith covariance matrix in the jth trial) are shifted on the manifold towards this reference point (see section (IV-E)). Each trial is labeled (low valance/high valance (*LV/HV*), and low arousal/high arousal (*LA/HA*)). The geometric mean for each trial (*GM/Trial*) is generated. In the classification process, observations are divided into training set and testing set. From the training set, the geometric mean for each class (*GM/Class*) is generated. In the testing process the geometric mean of each test observation is generated. The classification process is performed using minimum distance to Riemannian mean (*MDRM*) classifier (see section (IV-F)).

D. COVARIANCE MATRIX ESTIMATION TECHNIQUES

The EEG signals recorded from N electrodes, each electrode data forms a time series $\mathbf{x}_k(t)$ where $k = 1, \ldots, N$. Each time domain signal is divided into small overlapping windows (in this work we use, 4s window with 50% overlap, this gives us 29 windows with 512 sample in each window). Let $\mathbf{W}_{ik}$ refers to window i where $i = 1, \ldots, 29$ coming from electrode k where $k = 1, \ldots, N$. Each window is a vector containing n samples. Convolution is performed between each window and the corresponding windows coming from the N electrodes to generate 29 covariance matrices $\mathbf{C}_i$, $i = 1, \ldots, 29$. We denote by $\mathbf{X} \in \mathbb{R}^{N \times n}$ a given EEG recording epoch recorded from N electrodes and having n samples per window. The covariance matrix $\mathbf{C}$ between N random variables is a square matrix that can be calculated from $\mathbf{X}$, $\mathbf{C} \in \mathbb{R}^{N \times N}$ [30].

In section (III) we performed EEG signal analysis and clarified that EEG signals are corrupted with outliers and exhibit heavy tails. In this case using sample covariance matrix will offer a biased estimation. In section (III-B) we performed Anderson Darling goodness of fit test and Watson statistic test to clarified that the EEG signals are closer to T-distribution than Gaussian distribution or Laplace distribution. In this work we estimate the covariance matrix using two different method, sample covariance, and T-distribution covariance.

1) Sample Covariance: Assuming that the EEG signal distribution is Gaussian distribution ignoring the effect of outliers and heavy-tails, the sample covariance matrix

it is given by:

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i,$$

$$\mathbf{C} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^T \quad (9)$$

2) T-distribution Covariance: By assuming the observed data follows multivariate Student's t distribution, the parameters (mean vector, covariance matrix, degree of freedom, $\ldots$, etc) can be directly learned from the raw data via maximum likehood estimation (MLE). In [46] Rui Zhou *et al.* proposed an algorithm based on the generalized expectation maximization GEM) method to obtain the estimator. In this work, we generate T-distribution covariance using *fit* : *mvt*() function in *R* package *fitHeavyTail* [47]. In [48] Rui Zhou *et al.* offers a detailed explanation for covariance matrix estimation under heavy tail using *fit* : *mvt*() function.

E. CROSS-SESSION TRANSFER LEARNING

In this work we deal with the variability in EEG signals recorded by the same subject on two different sessions by finding a unique reference point for each subject and performing affine transformation for covariance matrices around this point. The variability between different sessions results from changes in electrodes positioning, environmental conditions, and subject physiological state.
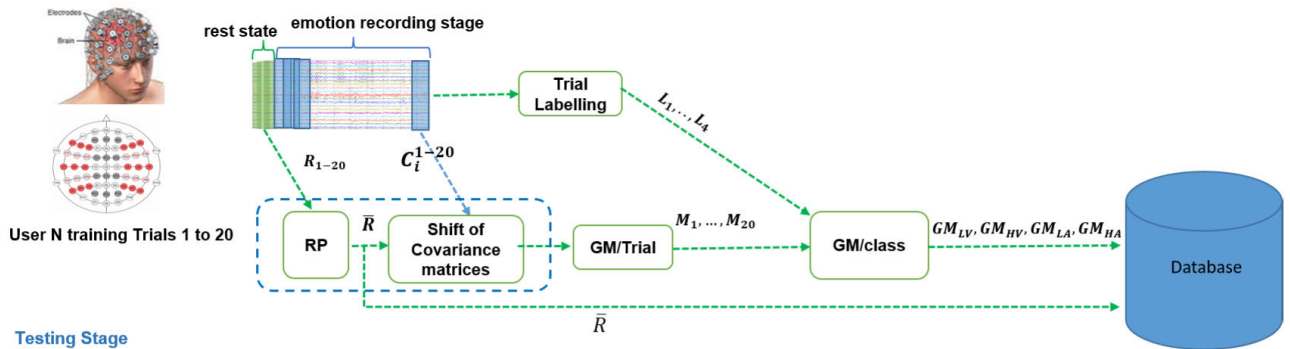
FIGURE 4. Proposed Subject-dependent EEG based emotion recognition system second experiment. During the training stage (20 trials per participant); for each participant N ($N = 1, 2, \ldots, 32$) we generate a reference point (R_i, $_{i=1,\ldots,20}$) from the rest state of each user observation, then a common reference point ($\bar{R}$) for each participant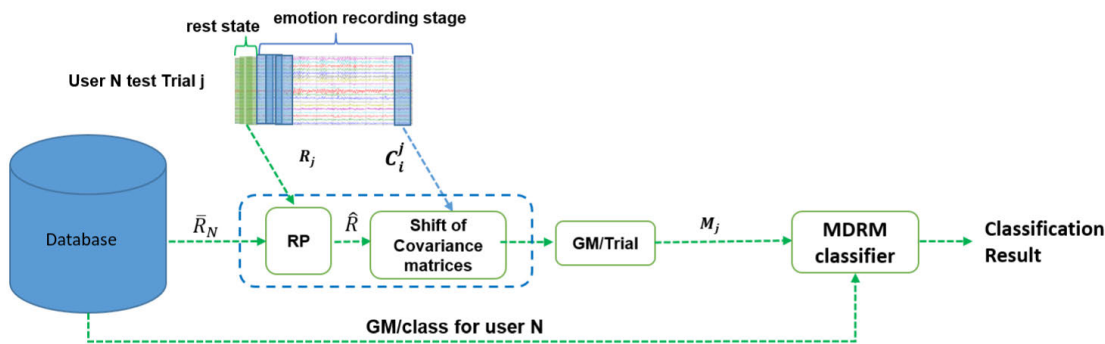 is generated from his 20 training trials recorded during the training sessions (the $\bar{R}$ per participant is stored in the database. Then, all the covariance matrices C_i^j (the ith covariance matrix in the jth trial) are shifted on the manifold towards this reference point (see section (IV-E)). Each trial is labeled (low valance/high valance (*LV/HV*), and low arousal/high arousal (*LA/HA*)). The geometric mean for each trial (*GM/Trial*) is generated. From the training set, the geometric mean for each class (*GM/Class*) is generated and stored in the database. In the testing stage we use the observations recorded in the second session. For user N test trial J, the reference point from the test observation is generated R_j and added to his reference point stored in the database ($\bar{R}_N$) to generate his new reference point $\hat{R}$, then the covariance matrices in this test observation is shifted around this new reference point. The geometric mean for this test trial is generates M_j. Using the participant geometric mean for each class stored in the database the classification process is performed using minimum distance to Riemannian mean (*MDRM*) classifier (see section (IV-F)).

In this work, we performed two experiments. In the first experiment (Fig. 3) we used the 3s baseline signals from the EEG data recorded during the two sessions together to generate a unique reference point for each participant. In the second experiment (shown in Fig. 4) we used data from one session for training and the second for testing. Each participant reference point is generated during training stage from the baseline signals in the training set, then during testing the 3s base line signal in test observation is used to adjust the participant reference point. A complete illustration for the proposed two experiments is given in section (IV-F). The effect of using a reference point and shifting the covariance matrices towards it by performing affine transformation helps in reducing the bias in the EEG data recorded in two different sessions. Fig. 5 shows each class data (in both sessions) before and after shifting and Fig. 6 shows the effect of all classes covariance matrices in both sessions before and after shifting.

F. EMOTION CLASSIFICATION

In this work, we are examining subject dependent emotion recognition, which means that the train and test observation of each subject is independent of other observations from other subjects. Two different binary classification problems were examined for subject-dependent emotion recognition: The discrimination of low/high valence (LV/HV), and low/high arousal (LA/HA). Let $\left\{ \left(\mathbf{M}^{(1)}, l^{(1)} \right), \ldots, \left(\mathbf{M}^{(n)}, l^{(n)} \right) \right\}$ be a training set of labeled observations. Where $\mathbf{M}^{(i)}$ is the center of mass for observation i and $l^{(i)}$ is the corresponding emotion label for that observation, $l^{(i)} \in \{HV, LV, HA, LA\}$ in a certain frequency band.

Emotion classification processes was performed using the EEG data recorded from the 31 participants in DEAP dataset, participant number 20 was excluded as his class 3 (Low valance) observations exist only in session 2, there are no LV observations in the first session.
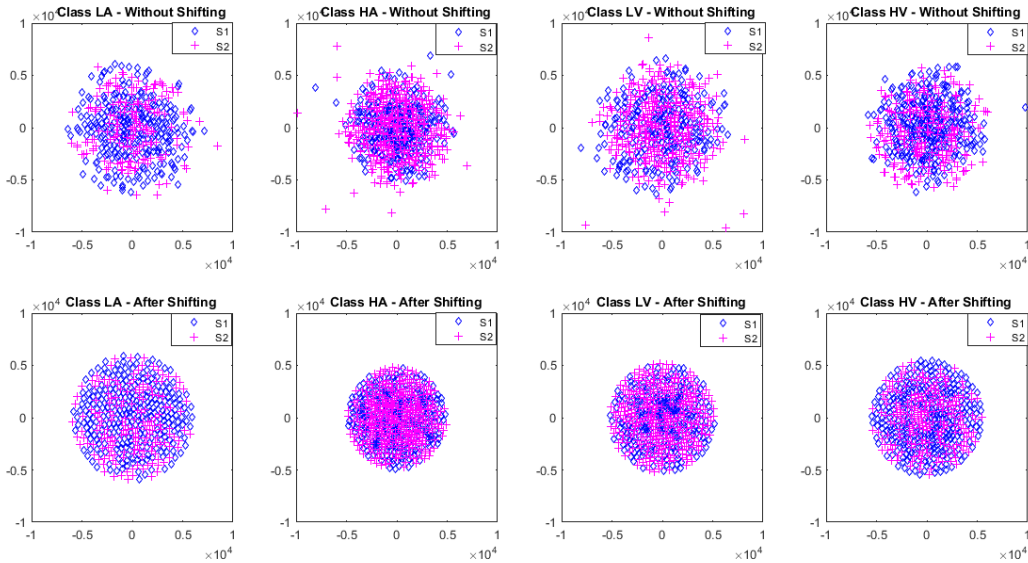
FIGURE 5. Original covariance matrices of subject 1 (as an example) in both session 1 and session 2 in each of the four classes (LA, HA, LV and HV) before and after shifting (Eq. (6)). Visualization obtained through t-SNE method using the Riemannian distance (Eq. (1)).
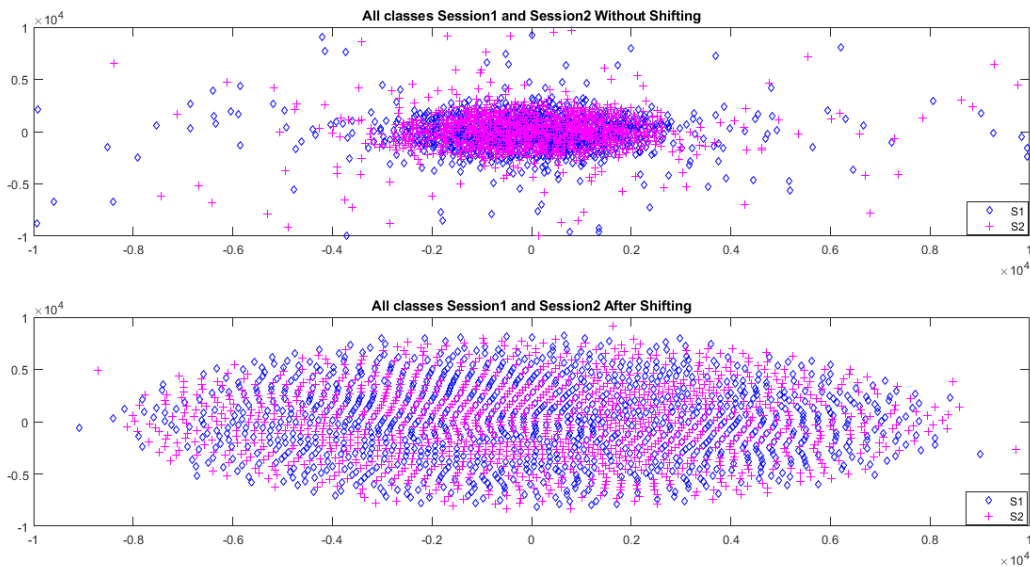
FIGURE 6. Original covariance matrices of subject 1 (as an example) in both session 1 and session 2 before and after shifting (Eq.(6)). Visualization obtained through t-SNE method using the Riemannian distance (Eq. (1)).

Five different frequency bands were examined, the signal without frequency separation, theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz) and gamma ($>$30 Hz). Two different experiments were performed:

1) EXPERIMENT 1
In the first experiment (Illustrated in Fig. 3). The 40 observations (trials) of each participant recorded during the two sessions are used to generate the reference point.

During the training stage:

1) The first 3s (rest state) in each observation is divided into one second windows with 50% overlapping (5 windows, 128 sample in each window),

2) Covariance matrices are generated from each corresponding window 5 covariance matrices from each observation $\mathbf{R}_k^j$, k is the number of covariance matrices in the rest state in each observation, j is the number of observations.

TABLE 4. Average classification results for experiment 1 (Reference point generated from both sessions). Sample covariance without transfer learning, Sample covariance with transfer learning and T-distribution covariance with transfer learning.

Freq. Band	Sample Cov. without TL		Sample Cov. with TL		T-dist Cov.with TL	
	A	V	A	V	A	V
All	53.95 ± 10.9	54.52 ± 9.5	56.26 ± 10	59.31 ± 9	58.33 ± 8.3	61.1 ± 8.6
Theta	$\mathbf{55.61 \pm 9.3}$	$\mathbf{55.81 \pm 10}$	$\mathbf{75.63 \pm 7.7}$	$\mathbf{73.97 \pm 7.6}$	$\mathbf{86.3 \pm 5.28}$	$\mathbf{87.1 \pm 5.31}$
Alpha	57.91 ± 9.8	55.06 ± 9.2	64.45 ± 8.1	65.63 ± 7.4	67.11 ± 8.6	67.18 ± 8.28
Beta	53.25 ± 10.2	54.68 ± 9.6	63.77 ± 8.6	66.15 ± 8.7	67.67 ± 8.6	71.03 ± 8.3
Gamma	46.09 ± 10.8	46.13 ± 9.5	58.98 ± 10.4	61.66 ± 7.9	60 ± 11.4	62 ± 7.2

(a) 18 Channels

Freq. Band	Sample Cov. without TL		Sample Cov. with TL		T-dist Cov.with TL	
	A	V	A	V	A	V
All	62.66 ± 3.47	56.76 ± 2.15	67.28 ± 0.49	60.87 ± 1.9	68.79 ± 1.02	65.76 ± 2.1
Theta	$\mathbf{62.3 \pm 2.32}$	$\mathbf{56.04 \pm 4.7}$	$\mathbf{84.93 \pm 2.14}$	$\mathbf{83.37 \pm 1.86}$	$\mathbf{86.37 \pm 2.1}$	$\mathbf{88.78 \pm 2.16}$
Alpha	63.46 ± 2.96	56.67 ± 2.22	75.25 ± 2.85	71.41 ± 3.04	75.94 ± 2.82	71.84 ± 2.4
Beta	60.73 ± 2.66	58.34 ± 2.82	72.87 ± 1.41	66.53 ± 2.92	73.1 ± 2.68	73.52 ± 2.94
Gamma	65.37 ± 3.98	58.05 ± 3.84	67.08 ± 5.23	60.84 ± 3.312	74.63 ± 3.89	73.15 ± 2.93

(b) 32 Channels

TABLE 5. Average classification results experiment 2 (Reference point generated from training session only). Sample covariance without transfer learning, Sample covariance with transfer learning and T-distribution covariance with transfer learning.

Freq. Band	Sample Cov. without TL		Sample Cov. with TL		T-dist Cov.with TL	
	A	V	A	V	A	V
All	53.85 ± 0.33	51.11 ± 0.23	56.03 ± 2.07	56.22 ± 0.96	58.21 ± 2.61	60.48 ± 2.89
Theta	$\mathbf{50.6 \pm 0.69}$	$\mathbf{51.63 \pm 2.28}$	$\mathbf{75.25 \pm 0.86}$	$\mathbf{73.21 \pm 0.87}$	$\mathbf{75.35 \pm 1.8}$	$\mathbf{76.71 \pm 2.59}$
Alpha	53.14 ± 0.28	51.35 ± 0.56	63.85 ± 1.63	65.02 ± 2.01	67.03 ± 3.1	66.62 ± 2.73
Beta	52.81 ± 2.45	51.8 ± 1.37	60.38 ± 2.28	60.12 ± 2.04	62.21 ± 0.26	60.81 ± 0.8
Gamma	45.7 ± 1.5	46.03 ± 0.81	58.67 ± 2.28	57.61 ± 0.73	59.85 ± 1.75	57.87 ± 1.55

(a) 18 Channels

Freq. Band	Sample Cov. without TL		Sample Cov. with TL		T-dist Cov.with TL	
	A	V	A	V	A	V
All	61.24 ± 0.08	51.75 ± 1.59	63.17 ± 1.95	59.5 ± 3.34	64.18 ± 8.03	60.7 ± 6.59
Theta	$\mathbf{53.63 \pm 0.71}$	$\mathbf{51.78 \pm 0.95}$	$\mathbf{76.78 \pm 3.47}$	$\mathbf{74.57 \pm 4.5}$	$\mathbf{79.74 \pm 3.9}$	$\mathbf{80.11 \pm 2.63}$
Alpha	53.46 ± 3.13	51.64 ± 1.59	65.94 ± 8.6	65.57 ± 10.02	69.31 ± 1.21	68.9 ± 0.24
Beta	54.23 ± 1.62	52.34 ± 1.18	62.32 ± 2.12	62.4 ± 2.84	72.48 ± 2.5	66.7 ± 0.33
Gamma	53.69 ± 1.53	50.86 ± 1.18	65.63 ± 0.93	59.84 ± 1.58	67.25 ± 0.37	60.4 ± 2.1

(b) 32 Channels

3) The geometric mean (denoted as $\bar{\mathbf{R}}$) from $\mathbf{R}_k^j$ is generated. $\bar{\mathbf{R}}$ represents the unique reference point for that participant.

4) Each participant 40 observations are divided to training set and testing set. 70% of the trials as used for training and 30% are used for testing.

5) The 60s left in each training observation is divided into 4s window with 50% overlapping. Consider $\mathbf{C}_i^j$, the ith covariance matrix in the jth observation.

6) Using $\bar{\mathbf{R}}$ as the reference point and substituting in Eq. (6), affine transformation is performed on each

TABLE 6. Results from 18 channels (Experiment 1) each participant in the theta frequency-band.

User Number	Sample Cov. without TL		Sample Cov. with TL		T Dist. Cov. with TL	
	A	V	A	V	A	V
1	60.5 ± 8.5	61.4 ± 9.4	76 ± 8.7	71.7 ± 16	$\mathbf{83.5 \pm 10.9}$	$\mathbf{95.7 \pm 2.7}$
2	58.8 ± 12.8	75.7 ± 8.5	90.3 ± 0.8	87.7 ± 0.4	$\mathbf{95.5 \pm 3.4}$	$\mathbf{95.1 \pm 3.7}$
3	69.7 ± 8.9	51.4 ± 15.6	88.7 ± 0.6	90.5 ± 1	$\mathbf{98.0 \pm 1.4}$	$\mathbf{93.8 \pm 5.8}$
4	43.5 ± 12.3	51 ± 15.5	57.3 ± 14.8	79 ± 3.7	$\mathbf{96.3 \pm 3.1}$	$\mathbf{93.5 \pm 5.4}$
5	52.4 ± 7.8	36.3 ± 6.5	$\mathbf{67.1 \pm 13.7}$	47.3 ± 9.4	50.3 ± 9.0	$\mathbf{57.0 \pm 12.3}$
6	41.4 ± 6.3	42.2 ± 13	72.4 ± 8.7	61.1 ± 9.6	$\mathbf{82.6 \pm 9.5}$	$\mathbf{90 \pm 6.2}$
7	57.3 ± 9.1	47.8 ± 10	$\mathbf{86.3 \pm 8}$	$\mathbf{81.4 \pm 2.7}$	59.5 ± 12.3	73.3 ± 8.4
8	63.4 ± 6	70.2 ± 6.4	64.6 ± 6.5	53.1 ± 11	$\mathbf{96.0 \pm 3}$	$\mathbf{98.6 \pm 0.9}$
9	61.5 ± 13.1	75 ± 12.2	77.5 ± 6.2	76.7 ± 5.3	$\mathbf{93.8 \pm 4.0}$	$\mathbf{88.3 \pm 8.5}$
10	59.5 ± 14.0	76.7 ± 6.2	57.1 ± 5.2	75 ± 15.5	$\mathbf{93.3 \pm 6.2}$	$\mathbf{95 \pm 3.7}$
11	56.3 ± 4.9	35.8 ± 5.2	85.8 ± 6.8	88 ± 4	$\mathbf{96.2 \pm 2.1}$	$\mathbf{97 \pm 2}$
12	58.0 ± 14.7	41.2 ± 5.9	52.3 ± 14	47.1 ± 10.7	$\mathbf{87.7 \pm 6.2}$	$\mathbf{95.7 \pm 2.7}$
13	44.5 ± 3.4	46.9 ± 6.9	$\mathbf{90 \pm 2.2}$	89 ± 3.3	86.4 ± 6.7	$\mathbf{92.6 \pm 4.5}$
14	66.7 ± 12.7	56.7 ± 9.1	77.2 ± 8.8	80 ± 4.1	$\mathbf{81.9 \pm 6.6}$	$\mathbf{93.3 \pm 6.2}$
15	43.3 ± 3.3	70 ± 10	88.3 ± 4.1	85 ± 7.5	$\mathbf{99.0 \pm 0.8}$	$\mathbf{98.7 \pm 0.3}$
16	56.7 ± 11.1	55.0 ± 9.3	85 ± 3.3	82 ± 1	$\mathbf{98.3 \pm .3}$	$\mathbf{98.4 \pm 1.2}$
17	50 ± 16.3	50 ± 12.9	63 ± 8.3	84.3 ± 17.4	$\mathbf{97.5 \pm 2.1}$	$\mathbf{96.7 \pm 2.8}$
18	44.5 ± 13.0	55.5 ± 8.5	71 ± 9.6	64.8 ± 8.3	$\mathbf{76.8 \pm 7.5}$	$\mathbf{77.5 \pm 7.8}$
19	71.9 ± 10.7	55 ± 8.7	82.8 ± 9.4	70.5 ± 4.9	$\mathbf{92.5 \pm 6.1}$	$\mathbf{96.7 \pm 2.1}$
21	48 ± 9.9	54.3 ± 16	74.7 ± 19	81.4 ± 5.3	$\mathbf{96.8 \pm 1.8}$	$\mathbf{95.2 \pm 3.1}$
22	45 ± 7.9	38.3 ± 8.5	87.5 ± 4	73.3 ± 11.1	$\mathbf{95.5 \pm 2.5}$	$\mathbf{94.0 \pm 4.0}$
23	62.5 ± 10.3	55.5 ± 5.6	80.8 ± 6.5	69 ± 9.7	$\mathbf{82.9 \pm 7.1}$	$\mathbf{91.5 \pm 7.6}$
24	53 ± 8.8	56.2 ± 6.3	82.7 ± 6.3	78.1 ± 7.8	$\mathbf{94 \pm 3.7}$	$\mathbf{90.2 \pm 7}$
25	44.4 ± 7.5	36.7 ± 8.5	94.4 ± 2.2	93.3 ± 3.3	$\mathbf{96.7 \pm 2.7}$	$\mathbf{98.3 \pm 1.3}$
26	63.1 ± 10.2	62 ± 7.4	62.6 ± 16.8	62.3 ± 14.7	$\mathbf{77.9 \pm 9.5}$	$\mathbf{65.8 \pm 9.2}$
27	65.6 ± 13.3	65.6 ± 12	68.3 ± 10.5	$\mathbf{57.8 \pm 17.7}$	$\mathbf{70.8 \pm 8.7}$	56.1 ± 5.7
28	36.7 ± 8.2	51.5 ± 10.8	53.3 ± 6.7	39 ± 11.2	$\mathbf{54.4 \pm 4.7}$	$\mathbf{52.0 \pm 11.8}$
29	70 ± 8.5	74.5 ± 12.4	77.5 ± 7.5	$\mathbf{86.4 \pm 2.9}$	$\mathbf{80.0 \pm 6.7}$	75.0 ± 10.3
30	65 ± 6.2	45.3 ± 12.4	$\mathbf{83.3 \pm 3.3}$	$\mathbf{87.5 \pm 8.9}$	71.1 ± 6.7	60.8 ± 8.8
31	63.3 ± 4.1	64.9 ± 12.7	75 ± 11.3	66 ± 5.7	$\mathbf{80.0 \pm 8}$	$\mathbf{88.6 \pm 3.5}$
32	48.6 ± 7.2	71.7 ± 11.3	59.2 ± 11.7	68.3 ± 12.2	$\mathbf{96.4 \pm 2.9}$	$\mathbf{93.3 \pm 6.2}$
Mean $\pm$ Std.	55.6 ± 9.3	55.8 ± 9.8	75.6 ± 7.7	74 ± 7.6	$\mathbf{86.3 \pm 5.28}$	$\mathbf{87.1 \pm 5.31}$

covariance matrix in each training observation to overcome cross-session variability.

7) Each trial from the training set is labeled $l^{(i)} \in$ {HV, LV, HA, LA};

8) The geometric mean for each training trial is generated $\mathbf{M}_i$,

9) The geometric mean (**GM**) for each emotion class is generated $\mathbf{GM}_{HV}, \mathbf{GM}_{LV}, \mathbf{GM}_{HA}, \mathbf{GM}_{LA}$ from the training set.

During the testing stage (for each test observation T):

1) The last 60s is divided into 4s window with 50% overlapping. Consider $\mathbf{C}_i^T$, the ith covariance matrix in the Tth test observation.

2) Using $\bar{\mathbf{R}}$ generated in the training stage as the reference point and substituting in Eq. (6), affine transformation is performed on each covariance matrix. The shifted covariance matrices ($\widetilde{\mathbf{C}}_i^T$) are the ones used in the emotion classification.

3) The geometric mean for the test observation is generated $\mathbf{M}^T$,

4) Riemannian distance between $\mathbf{M}^T$ and $\mathbf{GM}_{HV}, \mathbf{GM}_{LV}$ is calculated and the test observation belongs to the class (LV/HV) to which it has the minimum distance. Then for the LA/HA classification, distance between $\mathbf{M}^T$ and $\mathbf{GM}_{HA}, \mathbf{GM}_{LA}$ is calculated and the test observation belongs to the class (LA/HA) to which it has the minimum distance.

TABLE 7. The accuracy improvement caused by transfer learning and by T-distribution covariance with transfer learning.

			All	Theta	Alpha	Beta	Gamma	Mean ± Std.
Sample Cov. with TL VS Sample Cov without TL	18 Channels	Experiment 1	3.55	19.1	8.68.56	11	14.21	11.28 ± 5.9
		Experiment 2	3.65	23.12	12.19	7.94	12.27	11.83 ± 6.5
	32 Channels	Experiment 1	4.26	24.98	13.26	10.17	2.25	10.98 ± 8.1
		Experiment 2	4.84	22.97	13.21	9.07	10.46	12 .04 ± 6.2
T-dist Cov. with TL VS Sample Cov. with TL	18 Channels	Experiment 1	1.93	11.9	2.07	4.39	0.78	4.21 ± 3.4
		Experiment 2	3.2	3.41	0.56	3.61	9.93	4.14 ± 3.1
	32 Channels	Experiment 1	3.22	1.8	2.39	1.26	0.72	1.88 ± 0.87
		Experiment 2	1.11	4.25	3.35	7.23	1.09	3.4 ± 2.28

2) EXPERIMENT 2

In the second experiment (Illustrated in Fig. 4). We used observations in one session for training and the second for testing. Each participant reference point is generated during training stage from the baseline signals in the training set, then during testing the 3s base line signal in test observation is used to adjust the participant reference point.

The training stage is the same as in Experiment 1 except that, only 20 observations recorded in the training session is used to generate the reference point and the geometric mean of each class.

During the testing stage (for each user N test observation T):

1) The 3s baseline signal is used to generate a reference point from that test observation $\mathbf{R}_T$. This new reference point is used to adjust user N reference point stored in the database ($\mathbf{R}_N$) to generate his new reference point $\hat{\mathbf{R}}$.

2) The last 60s is divided into 4s window with 50% overlapping. Consider $\mathbf{C}_i^T$, the ith covariance matrix in the Tth test observation.

3) Using the new generated reference point $\hat{\mathbf{R}}$ and substituting in Eq. (6), affine transformation is performed on each covariance matrix.

4) The geometric mean for the test observation $\mathbf{M}^T$ is generated, from the shifted covariance matrices ($\widetilde{\mathbf{C}}_i^T$).

5) Using the participant geometric mean for each class generated during the training stage classification process is performed using minimum distance to Riemannian mean (MDRM) classifier.

V. RESULTS AND DISCUSSION

In this work, we used MDRM classifier with transfer learning for subject-dependent emotion recognition based on EEG signals.

Two pre-processing steps were performed. In the first step, the EEG signals were analysed by performing several statistical and goodness of fit tests. We found that T-distribution is most appropriate for describing EEG signals distribution. The signal analysis step helped us in determining the

most appropriate covariance matrix estimation technique. Two covariance matrix estimation techniques were examined, Sample covariance and T-distribution covariance. In the second step, cross-session transfer learning process was performed to overcome the variability in the EEG data recorded in different sessions. Classification was performed using MDRM classifier. Accuracy is used as an index of classification performance. It is considered as a suitable index as, the average number of observations/participant belongs to each label is almost balanced (see Table 3). Three methods were tested: using sample covariance without performing transfer learning process, using sample covariance with performing transfer learning, and finally using T-distribution covariance with performing transfer learning.

Two different experiments were performed for reference point generation. The result for the first experiment (see Table 4) achieved by performing five fold cross validation and averaging the results. In each fold we fuse the observations of each participant recorded during the two sessions, label the data to four classes (section IV-B), shuffle the observations belongs to each class, divide each class observations to 70% train and 30% testing and we made sure that the training data and the test data are entirely disjointed. The result for the second experiment (Table 5) achieved by using the first session for training and the second for testing and then using the second for training and the first for testing and averaging the results.

From Table 4 and 5 we can see that, using transfer learning process improved the results even with sample covariance estimation technique and combining T-distribution covariance with transfer learning process showed better performance in all frequency bands. Theta frequency band gave the best performance 87.1% for valence, 86.3% for arousal in experiment 1 using 18 channels, 88.78% for valence, 86.37% for arousal in experiment 1 using 32 channels, 76.71% for valence, 75.35% for arousal in experiment 2 using 18 channels, and 80.11% for valence, 79.74% for arousal in experiment 2 using 32 channels. It is clear that generating the reference point from the two sessions (Experiment 1) gave better results than generating the reference point from

training session and using minimum information from the test observation to adjust this reference point. Also, using 32 channels placed around the entire scalp gave better results than using only 18 channels placed over the upper half of the scalp. In Table 6 we show the results for each of the 31 participants generated in experiment 1 using 18 channels.

In Table 7 the percentage of improvement offered by using transfer learning and by using T-distribution covariance is shown. Using transfer learning as a pre-processing step improved the results even if the covariance estimation is performed using sample covariance.

VI. CONCLUSION

In this work, we present a scheme to improve the accuracy of MDRM classifier. We build subject-dependent EEG emotion recognition system based on MDRM classifier, with adding a two steps pre-processing stage.

In the first step, we analyze the EEG signals to investigate their non-Gaussianity, we found that the signals are corrupted with outliers, exhibit heavy tails, and that T-distribution is the closest to the actual EEG signals distribution. Based on the previous finding, the covariance matrix estimation was performed. In the second step, we performed cross-session transfer learning by generating a common reference point for each participant from his rest-state.

Performing cross-session transfer learning improved the system performance even when using sample covariance estimation, while combining T-distribution covariance with transfer learning gave the best results. The proposed pre-processing steps could be used in any brain computer interface system based on SPD manifold learning techniques.

REFERENCES

[1] V. V. Salunke and C. G. Patil, "A new approach for automatic face emotion recognition and classification based on deep networks," in *Proc. Int. Conf. Comput., Commun., Control Autom. (ICCUBEA)*, Aug. 2017, pp. 1–5.

[2] I. A. Adeyanju, E. O. Omidiora, and O. F. Oyedokun, "Performance evaluation of different support vector machine kernels for face emotion recognition," in *Proc. SAI Intell. Syst. Conf. (IntelliSys)*, Nov. 2015, pp. 804–806.

[3] R. Lotfidereshgi and P. Gournay, "Biologically inspired speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5135–5139.

[4] J. Umamaheswari and A. Akila, "An enhanced human speech emotion recognition using hybrid of PRNN and KNN," in *Proc. Int. Conf. Mach. Learn., Big Data, Cloud Parallel Comput. (COMITCon)*, Feb. 2019, pp. 177–183.

[5] M. Murugappan, N. Ramachandran, and Y. Sazali, "Classification of human emotion from EEG using discrete wavelet transform," *J. Biomed. Sci. Eng.*, vol. 3, no. 4, pp. 390–396, 2010.

[6] E. A. Abdel-Ghaffar and M. Daoudi, "Emotion recognition from multidimensional electroencephalographic signals on the manifold of symmetric positive definite matrices," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Aug. 2020, pp. 354–359.

[7] Y. Yang, Q. Wu, M. Qiu, Y. Wang, and X. Chen, "Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–7.

[8] D. Datcu and L. J. M. Rothkrantz, "Automatic bi-modal emotion recognition system based on fusion of facial expressions and emotion extraction from speech," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–2.

[9] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, Jan./Mar. 2016.

[10] S. Siddharth, T.-P. Jung, and T. J. Sejnowski, "Impact of affective multimedia content on the electroencephalogram and facial expressions," *Sci. Rep.*, vol. 9, no. 1, p. 16295, Dec. 2019.

[11] F.-R. Lin and C.-M. Kao, "Mental effort detection using EEG data in E-learning contexts," *Comput. Educ.*, vol. 122, pp. 63–79, Jul. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360131518300794

[12] M. Ali, A. H. Mosa, F. Al Machot, and K. Kyamakya, "EEG-based emotion recognition approach for e-healthcare applications," in *Proc. 8th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Jul. 2016, pp. 946–950.

[13] M. N. Giannakos, K. Sharma, and E. Niforatos, "Exploring EEG signals during the different phases of game-player interaction," in *Proc. 11th Int. Conf. Virtual Worlds Games Serious Appl. (VS-Games)*, Sep. 2019, pp. 1–8.

[14] D. Nie, X.-W. Wang, L.-C. Shi, and B.-L. Lu, "EEG-based emotion recognition during watching movies," in *Proc. 5th Int. IEEE/EMBS Conf. Neural Eng.*, Apr. 2011, pp. 667–670.

[15] Z. Mohammadi, J. Frounchi, and M. Amiri, "Wavelet-based emotion recognition system using EEG signal," *Neural Comput. Appl.*, vol. 28, no. 8, pp. 1985–1990, Aug. 2017, doi: 10.1007/s00521-015-2149-8.

[16] S. K. Hadjidimitriou and L. J. Hadjileontiadis, "Toward an EEG-based recognition of music liking using time-frequency analysis," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 12, pp. 3498–3510, Dec. 2012.

[17] W. Tao, C. Li, R. Song, J. Cheng, Y. Liu, F. Wan, and X. Chen, "EEG-based emotion recognition via channel-wise attention and self attention," *IEEE Trans. Affect. Comput.*, early access, Sep. 22, 2020, doi: 10.1109/TAFFC.2020.3025777.

[18] W. Fruehwirt, M. Gerstgrasser, P. Zhang, L. Weydemann, M. Waser, R. Schmidt, T. Benke, P. Dal-Bianco, G. Ransmayr, D. Grossegger, H. Garn, G. W. Peters, S. Roberts, and G. Dorffner, "Riemannian tangent space mapping and elastic net regularization for cost-effective EEG markers of brain atrophy in Alzheimer's disease," 2017, *arXiv:1711.08359*.

[19] S. Yuan, W. Zhou, Q. Wu, and Y. Zhang, "Epileptic seizure detection with log-Euclidean Gaussian kernel-based sparse representation," *Int. J. Neural Syst.*, vol. 26, no. 3, May 2016, Art. no. 1650011.

[20] M. Congedo, A. Barachant, and R. Bhatia, "Riemannian geometry for EEG-based brain-computer interfaces; A primer and a review," *Brain-Comput. Interfaces*, vol. 4, no. 3, pp. 1–20, 2017.

[21] J. Chalrles, J. Sclabassi, and M. Sun, "Non-Gaussian modeling of EEG data," in *Proc. 1st Joint BMES/EMBS Conf.*, Atlanta, GA, USA, 1999, p. 1023.

[22] P. J. Charles, M. Sun, and R. J. Sclabassi, "Statistical reduction of EEG data," in *Proc. 22nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Chicago, IL, USA, Jul. 2000, pp. 1394–1395.

[23] N. Nazmi, S. A. Mazlan, H. Zamzuri, and M. A. A. Rahman, "Fitting distribution for electromyography and electroencephalography signals based on goodness-of-fit tests," *Proc. Comput. Sci.*, vol. 76, pp. 468–473, Jan. 2015.

[24] P. Zanini, M. Congedo, C. Jutten, S. Said, and Y. Berthoumieu, "Transfer learning: A Riemannian geometry framework with applications to brain–computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 5, pp. 1107–1116, May 2018.

[25] H. He and D. Wu, "Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 2, pp. 399–410, Feb. 2020.

[26] P. Rodrigues and C. Jutten, "Riemannian Procrustes analysis: Transfer learning for brain–computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 8, pp. 2390–2401, Aug. 2019.

[27] Y.-P. Lin, "Constructing a personalized cross-day EEG-based emotion-classification model using transfer learning," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 5, pp. 1255–1264, May 2020.

[28] O. Yair, M. Ben-Chen, and R. Talmon, "Parallel transport on the cone manifold of SPD matrices for domain adaptation," *IEEE Trans. Signal Process.*, vol. 67, no. 7, pp. 1797–1811, Apr. 2019.

[29] Y. Wang, S. Qiu, X. Ma, and H. He, "A prototype-based SPD matrix network for domain adaptation EEG emotion recognition," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107626. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320320304295

[30] M. Congedo, A. Barachant, and E. K. Koopaei, "Fixed point algorithms for estimating power means of positive definite matrices," *IEEE Trans. Signal Process.*, vol. 65, no. 9, pp. 2211–2220, May 2017.

[31] R. Bhatia, *Positive Definite Matrices*. Princeton, NJ, USA: Princeton Univ. Press, 2009.

[32] F. Nielsen and R. Bhatia, *Matrix Information Geometry*. Berlin, Germany: Springer-Verlag, 2013, ch. 2.

[33] M. Moakher, "A differential geometric approach to the geometric mean of symmetric positive-definite matrices," *SIAM J. Matrix Anal. Appl.*, vol. 26, no. 3, pp. 735–747, Jul. 2005.

[34] D. A. Bini and B. Iannazzo, "A note on computing matrix geometric means," *Adv. Comput. Math.*, vol. 35, pp. 175–192, Jun. 2011.

[35] Y. Ke, S. Minsker, Z. Ren, Q. Sun, and W.-X. Zhou, "User-friendly covariance estimation for heavy-tailed distributions," 2018, *arXiv:1811.01520*.

[36] X. Wei and S. Minsker, "Estimation of the covariance structure of heavy-tailed distributions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017, pp. 2859–2868.

[37] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *J. Exp. Social Psychol.*, vol. 49, no. 4, pp. 764–766, 2013.

[38] G. Huang, *EEG Signal Processing and Feature Extraction*. Singapore: Springer, 2019, ch. 17.

[39] C. Forbes, M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions*. Hoboken, NJ, USA: Wiley, 2011, ch. 26.

[40] B. W. Yap and C. H. Sim, "Comparisons of various types of normality tests," *J. Statist. Comput. Simul.*, vol. 81, no. 12, pp. 2141–2155, Dec. 2011.

[41] P. Puig and M. A. Stephens, "Tests of fit for the Laplace distribution, with applications," *Technometrics*, vol. 42, no. 4, pp. 417–424, 2000.

[42] J. L. Gastwirth, Y. R. Gel, W. L. W. Hui, and V. Lyubchich. (2020). *Tools for Biostatistics, Public Policy, Law*. R Package Version 3.4. [Online]. Available: https://CRAN.R-project.org/package=lawstat

[43] S. Koelstra, C. Muhl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis;using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan./Mar. 2012.

[44] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems," *NeuroImage*, vol. 34, no. 4, pp. 1600–1611, Feb. 2007.

[45] N. Thammasan, K.-I. Fukui, and M. Numao, "Application of deep belief networks in eeg-based dynamic music-emotion recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 881–888.

[46] R. Zhou, J. Liu, S. Kumar, and D. P. Palomar, "Robust factor analysis parameter estimation," in *Proc. Int. Conf. Comput. Aided Syst. Theory (EUROCAST)*, 2019, pp. 3–11, doi: 10.1007/978-3-030-45096-0_1.

[47] D. P. Palomar and R. Zhou. (2020). *fitHeavyTail: Mean Covariance Matrix Estimation under Heavy Tails*. R Package Version 0.1.2. [Online]. Available: https://CRAN.R-project.org/package=fitHeavyTail

[48] D. P. Palomar and R. Zhou. *Mean Vector and Covariance Matrix Estimation Under Heavy Tails*. Accessed: Dec. 31, 2019. [Online]. Available: https://cran.r-project.org/web/packages/fitHeavyTail/vignettes/CovarianceEstimationHeavyTail.html
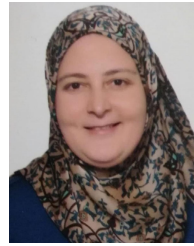
EMAN A. ABDEL-GHAFFAR was born in Cairo, Egypt, in 1976. She received the B.Sc., M.Sc., and PhD. degrees in electrical engineering from the Faculty of Engineering Shoubra, Benha University, Cairo, Egypt, in 1999, 2004, and 2010, respectively. She had her M.Sc. in studying image compression techniques, and studied multi-modal biometric fusion in her PhD. She is currently working as a Lecturer at the Faculty of Engineering Shoubra, Benha University. Her current research interests include biometrics, brain–computer interaction, data compression, multimedia security, speech processing, and image processing.

YUJIN WU received the B.S. degree in electronic information from Xidian University, Xi'an, China, in 2018, and the E.E. degree in computer and electronics from the Polytechnique School, Grenoble Alpes University, France, in 2019. She is currently pursuing the Ph.D. degree with the University of Lille, France. Her research interests include stress detection, multi-modal emotion recognition, and computer vision.

MOHAMED DAOUDI (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the University of Lille, France, in 1993. He is currently a Full Professor of computer science at IMT Lille Douai and the Head of the CRIStAL Laboratory, Image Group, (UMR CNRS 9189). He has published over 150 papers in some of the most distinguished scientific journals and international conferences. His research interests include pattern recognition, shape analysis, and computer vision. He is fellow of IAPR. He has served as the General Chair of the IEEE International Conference on Automatic Face and Gesture Recognition, 2019. He is an Associate Editor of *Image and Vision Computing* journal, the IEEE TRANSACTIONS ON MULTIMEDIA, *Computer Vision and Image Understanding*, the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, and the *Journal of Imaging*.

• • •