Proceedings of the Pakistan Academy of Sciences 52 (1): 27–38 (2015)
Copyright © Pakistan Academy of Sciences
ISSN: 0377 - 2969 (print), 2306 - 1448 (online)

Pakistan Academy of Sciences

Research Article

Bimodal Human Emotion Classification in the Speaker-Dependent Scenario

Sanaul Haq[1], Tariqullah Jan[2*], Asiya Jehangir[2], Muhammad Asif[1], Amjad Ali[3], and Naveed Ahmad[4]

[1]Department of Electronics, University of Peshawar, Peshawar, Pakistan
[2]Department of Electrical Engineering, University of Engineering & Technology, Peshawar, Pakistan
[3]Department of Electrical Engineering, Sarhad University of Science and Information Technology, Peshawar, Pakistan
[4]Department of Computer Science, University of Peshawar, Peshawar, Pakistan

Abstract: This paper investigates the recognition of expressed emotion from speech and facial expressions for the speaker-dependent task. The experiments were performed to develop a baseline system for the audio-visual emotion classification, and to investigate different ways of combining the audio and visual information to achieve better emotion classification. The extracted features were composed of 106 audio and 240 visual features. The audio features consisted of pitch, energy, duration and MFCC features, whereas the visual features were related to positions of the 2D marker coordinates. The Plus *l*-Take Away *r* algorithm was used for feature selection based on the Mahalanobis distance, Bhattacharyya distance, and KL-divergence as selection criteria. The feature selection was followed by feature reduction using the PCA and LDA, and classification using the Gaussian classifier. Both unimodal and bimodal approaches were used for emotion classification. The audio-visual fusion was investigated at two different levels: feature-level and decision-level. The emotion classification results comparable to human performance were achieved on the SAVEE database.

Keywords: Multimodal emotion recognition, feature selection, distance measures, classification, emotional database

1. INTRODUCTION

The interaction between human and machine is becoming more interesting with the development in technology. Human beings communicate with each other through speech, but its verbal content does not carry all the information conveyed. Additional information includes vocalised emotions, facial expressions, hand gestures and body language as well as biometric indicators [1, 2]. From the human perspective, the human-machine interaction will be more natural and attractive if machines are able to recognize human emotions and respond accordingly [3]. On the other hand, recognition of the user's expressed emotion can improve the reliability of communication in dialogue [4].

Automatic emotion recognition has many important applications including affect-sensitive automobile systems, emotional intelligent customer services systems, and game and film industries [5]. The field of emotion recognition has attracted researchers from various disciplines and current research has made significant progress in several areas including acquisition of emotional databases, feature extraction and selection, and classification and fusion of modalities [5, 6].

Previous studies have mainly focused on unimodal approaches (e.g., speech, facial expressions) for emotion recognition. The modalities have largely been treated independently

Received, July 2014; Accepted, February 2015
*Corresponding author: Tariqullah Jan; Email: tariqullahjan@uetpeshawar.edu.pk

and the interrelation between them has not been explored. In actual fact, speech and facial gestures are highly correlated and coordinated, and the relationship between these two modalities is affected by emotions and linguistic content [7]. Humans express their emotion through both speech and gesture, and it has been suggested that an ideal emotion recognizer should be based on multimodal information [1, 8].

The reliability of an emotion recognizer is based on several factors including the quality of emotional data used to build the system. Popular audio databases include the AIBO corpus, Berlin database (EMODB) and Danish database [9–11]. Visual databases include the Cohn-Kanade and MMI databases [12, 13]. Examples of audio-visual databases are GEMEP, Facial Motion Capture database, IEMOCAP, Belfast Naturalistic database and HUMAINE database [7, 14–17]. These databases are either acted or natural.

Audio and visual features of different types have been investigated for the analysis of emotion. Important acoustic features include pitch, formants, duration, spectral energy, and mel frequency cepstral coefficients (MFCCs). Audio features have been used at utterance-level [18, 19], as well as at frame-level [20, 21]. Vision-based emotion recognition is primarily based on facial expressions, since the face plays the most important role in conveying emotion. Facial features can be divided into two categories: geometric and appearance [22]. The techniques of Pantic and Bartlett [22] and Chankg et al [23] are based on geometric features, while [24, [25] used appearance features.

Feature selection and reduction techniques are commonly used to discard uninformative, redundant and noisy information. The processes of feature selection and reduction improve both the classification accuracy and computational efficiency. For emotion recognition, different types of feature search technique have been used including sequential forward selection [19], sequential floating forward selection [26], genetic algorithms [27] and best-first [28]. Feature reduction techniques include PCA and LDA [29, 30].

The choice of classifier plays a crucial role in any pattern recognition problem. Commonly used classifiers are Gaussian mixture model [31], hidden Markov model [20], neural network [32], support vector machine [33], and adaptive boosting [34]. Multimodal approaches have been adopted to improve the emotion classification by fusion of data at feature [2], decision [35], and model [36] levels.

Most research in the area of emotion recognition is based on using a single modality (e.g., audio or visual), and less progress has been made in terms of multimodal approaches. This research aims to achieve better emotion classification by combining the audio and visual modalities. The following sections present the SAVEE database, method, experimental results and conclusion.

2. SURREY AUDIO-VISUAL EXPRESSED EMOTION (SAVEE) DATABASE

The design of an automatic emotion recognizer is based on many factors, and one of the important factors that can affect its performance is the emotional database used to build its models representing human emotions. Emotional behaviour databases of acted and spontaneous emotions have been recorded for emotion analysis. The attributes of an emotional database that affect the performance of an emotion recognizer include emotion categories, number of speakers, modalities and quality of the data [14].

2.1 Corpus Design

We used SAVEE database [37] for our analysis. The database consists of data from four British male speakers, with an average age of 30 years, in Ekman's six basic emotions (anger, fear, disgust, sadness, happiness and surprise) [38] plus neutral.

The text material for the database was selected from the TIMIT database [39], which consists of phonetically-diverse sentences. The text material consisted of 15 sentences for each of the six emotions and 30 sentences for the neutral. The distribution of sentences in this way resulted in 120 utterances per actor and 480 utterances in total.

2.2 Data Recording

The data were recorded using 3dMD's 4D capture system [40] at the University of Surrey, UK. The 3dMD's system covers 180 degrees of the face. The sampling rate for audio data was 44.1 kHz, while that for video was 60 fps. To extract facial

features, each actor's frontal face was painted with 60 markers. Markers were painted on the forehead, eyebrows, cheeks, lips and jaw.

2.3 Data Processing and Annotation

Both the audio and visual data were annotated. The audio data were labelled at the phone level, and facial markers were tracked for each frame of the visual data.

2.4 Subjective Quality Evaluation

Quality of the recorded data was checked in terms of expressed emotions by performing the subjective evaluation. The SAVEE database was evaluated by 20 subjects (10 male, 10 female) with an average age of 25 years. The audio, visual and audio-visual data were evaluated at utterance level.

Each actor's data were evaluated by 10 subjects. The classification accuracy for seven emotion classes averaged over four actor's data and 10 evaluators was 66.5 % for audio, 88.0% for visual, and 91.8 % for audio-visual data. Overall, at least 8 out of 10 evaluators were able to recognize the expressed emotions for 441 out of 480 utterances under bimodal scenario, indicating that the database contain a high quality recorded data.

3. METHOD

The speaker-dependent emotion classification was performed by adopting a method comprising three main steps, please see Fig.1. The first step was feature extraction, in which audio features consisting of pitch, duration, energy and spectral envelope, and visual features consisting of 2D coordinates of facial markers were extracted. The next step consisted of feature selection and reduction. The Plus *l*-Take Away *r* algorithm (sequential forward backward selection) was used for feature selection based on three criteria: Bhattacharyya distance, Mahalanobis distance and KL-divergence. Feature selection was followed by feature reduction using the PCA and LDA transformation techniques. Finally, different emotion categories were classified using the Gaussian classifiers.

3.1 Feature Extraction

For the speaker-dependent emotion classification features were extracted at utterance-level consisting of 106 audio and 240 visual features. The details of the audio and visual features and their extraction are given below.

3.1.1 Audio Features

The audio features were related to pitch (f_0), duration, energy and spectral envelop. These features were extracted using the Speech Filing System [41] and HTK [42], as shown in Fig. 2a.

Pitch Features: The fundamental frequency (f_0) was extracted using the Speech Filing System based on RAPT algorithm. Features related to f_0 contour were minimum and maximum of mel frequency; mean and standard deviation of first and second Gaussian of mel frequency; minimum, maximum, mean and standard deviation of mel frequency first order difference.

Duration Features: Semi-automated phone labels were used to extract duration features. The phone labelling was performed in two steps: first automatic labelling of the audio was performed using the HTK, and second the Speech Filing System was used to correct the automatic phone labels based on listening assisted by the waveform and spectrogram. The following duration features were extracted: voiced speech duration, unvoiced speech duration, sentence duration, average voiced phone duration, average unvoiced phone duration, voiced-to-unvoiced speech duration ratio, average voiced-to-unvoiced phone duration ratio, speech rate, voiced-speech-to-sentence duration ratio, and unvoiced-speech-to-sentence duration ratio.

Energy Features: The energy features were extracted by first filtering the signal in different bands using a Butterworth filter and then calculating the energy at frame level using a Hamming window having a duration of 25 ms. The step size was 10 ms. The following energy features were extracted: mean and standard deviation of total log energy; minimum, maximum, range, mean and standard deviation of normalised energies in the original speech signal and speech signal in the frequency bands 0-0.5 kHz, 0.5-1 kHz, 1-2 kHz, 2-4 kHz and 4-8 kHz; minimum, maximum, range, mean and standard deviation of first order difference of normalised energies in the original speech signal and speech signal in the same frequency bands.

Spectral Features: The spectral envelope features were extracted at utterance level using the HTK: mean and standard deviation of 12 MFCCs.

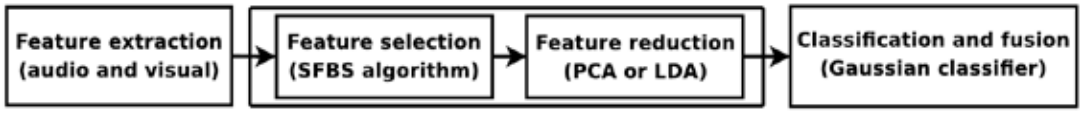

Fig. 1. Block diagram of emotion classification method for the speaker-dependent scenario.

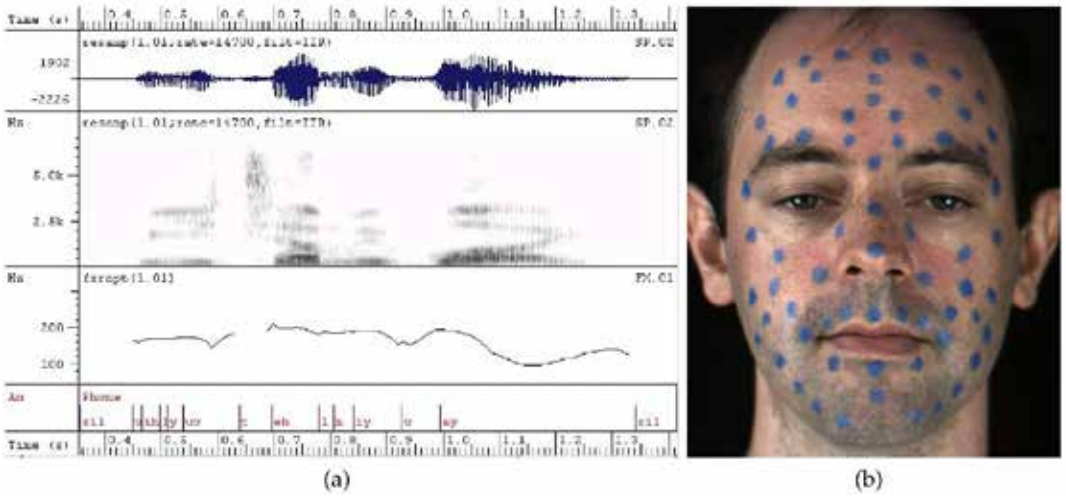

Fig. 2. (a) Audio feature extraction with Speech Filing System software; (b) video data with marker locations.

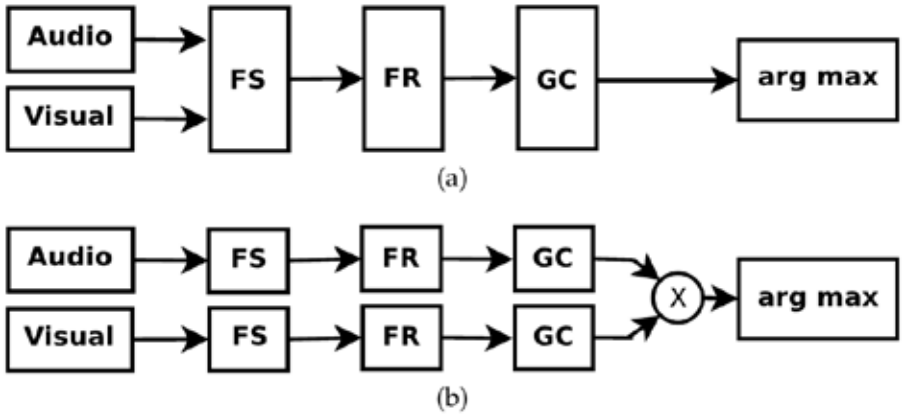

Fig. 3. Block diagrams of the audio-visual fusion at two levels: (a) feature-level; (b) decision-level. FS: Feature Selection; FR: Feature Reduction; GC: Gaussian Classifier.

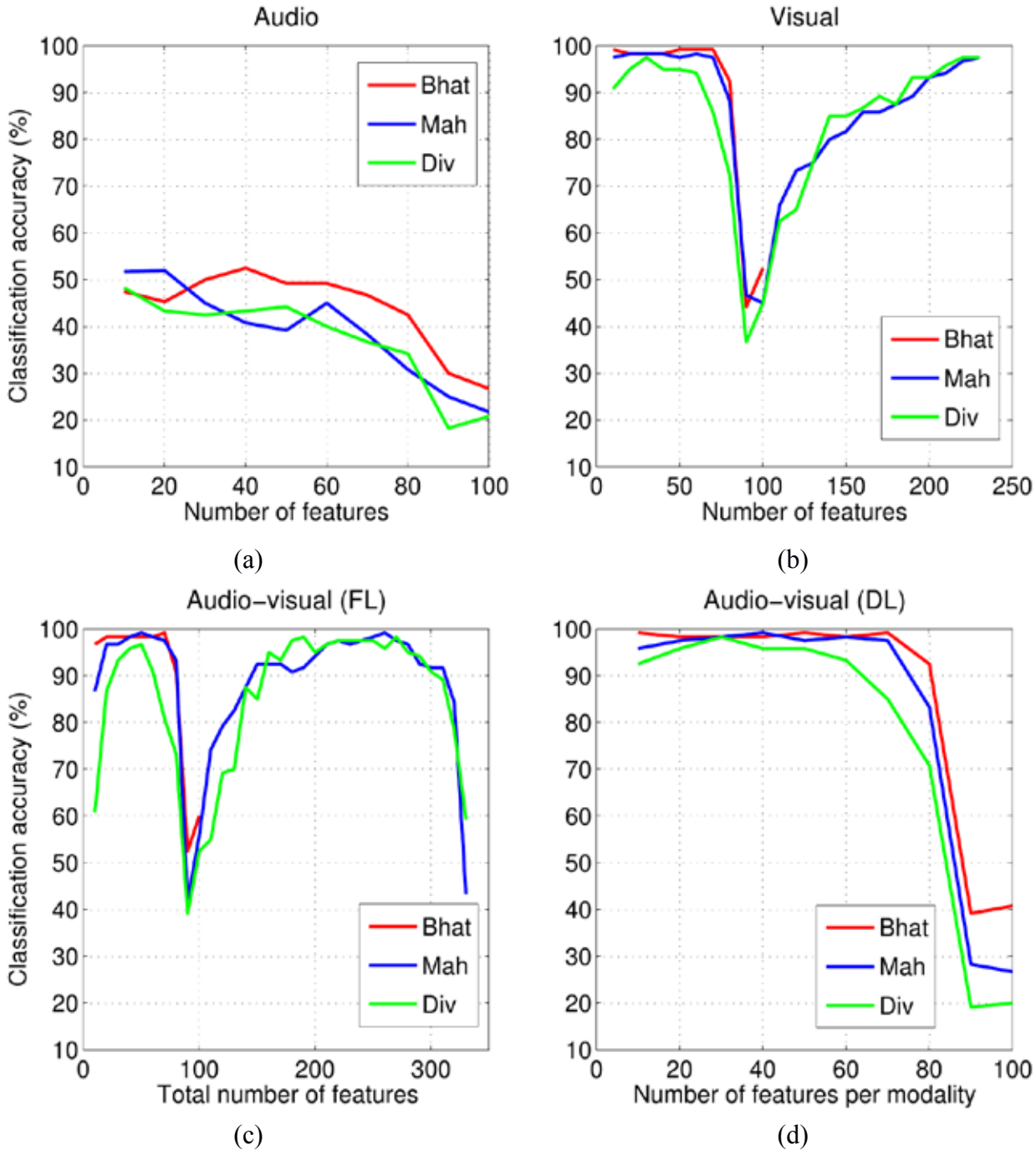

Fig. 4. Average classification accuracy (%) achieved for seven emotions on the subject KL data using the LDA-transformed features selected with the Bhattacharyya (Bhat), Mahalanobis (Mah) and KL-divergence (Div) measures. The results were obtained for the (a) audio, (b) visual, and audio-visual modalities fused at (c) feature-level (FL), and (d) decision-level (DL).

3.1.2 Visual Features

The visual features were extracted by painting markers on actors' foreheads, eyebrows, cheeks, lips and jaws, as shown in Fig. 2b. After data capture, the markers were manually labelled for the first frame of a sequence and automatically tracked for the remaining frames using a marker tracker. The tracked marker x and y coordinates were normalised relative to a reference point at the bridge of the nose by subtracting the mean displacement and rotating for the correction of the head pose [7]. Finally, 240 visual features were obtained from the 2D marker coordinates as the mean and standard deviation of the adjusted marker coordinates. The facial markers were divided into upper, middle and lower sections. The upper region included region above the eyes, the lower region contained area below the upper lip, and the middle region covered the cheek region.

3.2 Feature Selection and Reduction

The feature selection improves the classification accuracy and makes the algorithm faster by removing the uninformative, redundant or noisy information. For the speaker-dependent emotion classification, we adopted a two-step process: feature selection with Plus l-Take Away r algorithm, followed by feature reduction with PCA and LDA.

3.2.1 Feature Selection

Feature selection was performed using a standard algorithm based on the discriminative criterion function. The Plus l-Take Away r algorithm [43] is a feature search method based on some criterion function. It combines the sequential forward selection (SFS) and sequential backward selection (SBS) algorithms to achieve better results.

At each step, l numbers of features are included to the current feature set and r numbers of features are discarded. The process continues until the required feature set size is achieved. The feature search was performed with $l = 2$ and $r = 1$, i.e., one feature was added at each step. We used this algorithm for feature selection based on three different criteria: Mahalanobis distance, Bhattacharyya distance, and KL-divergence [44]. These distance measures have been used as dissimilarity measures in different applications including speaker recognition [44], emotion recognition [45] and texture retrieval [46].

Mahalanobis distance: It is used to define the similarity between two classes [47]. The Mahalanobis distance between two normally distributed classes ω_i and ω_j is defined as

$$d_{Mah} = \sqrt{(\mu_i - \mu_j)^T (P_i\Sigma_i + P_j\Sigma_j)^{-1}(\mu_i - \mu_j)} \qquad (1)$$

where μ_i and μ_j are the means, $\mathbf{\Sigma}_i$ and $\mathbf{\Sigma}_j$ are the covariances, and P_i and P_j are the prior probabilities of classes ω_i and ω_j, respectively. The prior probabilities are calculated as $P_i = (n_i - 1)/(n_i + n_j - 2)$ and $P_j = (n_j - 1)/(n_i + n_j - 2)$, where n_i and n_j denote the numbers of samples from classes ω_i and ω_j, respectively. The Mahalanobis distance is scale invariant and it takes into account the correlation between variables. The Mahalanobis and Euclidean distances are equivalent when the covariance term $P_i\,\mathbf{\Sigma}_i + P_j\,\mathbf{\Sigma}_j$ is equal to the identity matrix.

Bhattacharyya distance: It is another way of defining the separability between two classes [48]. For normally distributed classes it is given by

$$d_{Bhat} = \frac{1}{8}(\mu_i - \mu_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\mu_i - \mu_j) + \frac{1}{2}\ln \frac{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|}{\sqrt{|\Sigma_i||\Sigma_j|}} \qquad (2)$$

The Bhattacharyya distance consists of two components: the first term defines the class separability based on class means, whereas the second term provides the class separability based on class covariance matrices. The first term represents the Mahalanobis distance using an average covariance matrix.

Kullback-Leibler (KL) divergence measure: The divergence measure provides the dissimilarity between two classes based upon information theory [49]. For two normally distributed classes ω_i and ω_j, the KL-divergence is defined as

$$J_{Div} = \frac{1}{2} tr\left[(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j)(\mu_i - \mu_j)^T\right] + \frac{1}{2} tr\left[(\Sigma_i - \Sigma_j)(\Sigma_j^{-1} - \Sigma_i^{-1})\right] \qquad (3)$$

where tr denotes the matrix trace operation. The above-mentioned relation consists of two components: the first term provides the difference between two classes using the class means, while the second term provides the difference based on covariance matrices. In this way, the divergence

measure defines the separation between two classes based on both the class means and covariance matrices.

Features were normalised prior to applying the feature selection by using the Z-norm (i.e., mean subtraction and division by standard deviation). The data was assumed to be normally distributed and full covariance matrices were used for computation, similar to other studies [44]–[46]. The above-mentioned distance measures provide the separability between two classes. For *m* number of classes the separability measure is obtained by averaging it over all binary combinations of the classes, and is given by

$$J = \sum_{i=1}^{m-1} \sum_{j>i} P_i P_j J_{ij} \quad (4)$$

where J_{ij} is the separability measure between two classes ω_i and ω_j, whereas P_i and P_j are the prior probabilities of classes ω_i and ω_j, respectively.

3.2.2 Feature Reduction

Statistical methods can be used to reduce the dimensionality of a feature set. This is achieved by applying the linear transformation, $\mathbf{y} = \mathbf{W}\mathbf{x}$, where $\mathbf{y}$ denotes the feature vector in reduced feature space, $\mathbf{W}$ is the transformation matrix, and $\mathbf{x}$ is the original feature vector. PCA technique [50] is used to extract the important characteristics of high-dimensional data and to remove the uninformative and noisy data. The LDA method [51] provides the separation between classes based on the ratio of between-class variance to within-class variance. We applied LDA by using the covariance of all training data rather than between-class variance in order to compare the LDA and PCA for different numbers of features. PCA and LDA methods involve feature centring, whitening, covariance computation and eigen decomposition. For feature reduction, we applied both PCA and LDA as linear transformation techniques to the selected features.

Principal Component Analysis (PCA): PCA method is widely used for the statistical analysis of data [50]. It has the ability to extract useful information from noisy data by reducing its dimensionality.

Let $\mathbf{X}$ be an $m \times n$ matrix, where m denotes the number of features and n denotes the number of samples. First, the mean value of each feature is subtracted and each feature is divided by its standard deviation to have the same range of variation for different features. Second, we define a matrix $\mathbf{Y}$ of $n \times m$ dimensions.

$$Y = \frac{1}{\sqrt{n-1}} X^T \quad (5)$$

It can be shown that

$$Y^T Y = \Sigma_X \quad (6)$$

where $\mathbf{\Sigma_X}$ denotes the covariance of $\mathbf{X}$. The eigenvectors of $\mathbf{\Sigma_X}$ provides the principal components of $\mathbf{X}$. The Singular Value Decomposition (SVD) of matrix $\mathbf{Y}$ provides the eigenvector matrix $\mathbf{V}$. The columns of matrix $\mathbf{V}$ are the eigenvectors of $\mathbf{Y^T Y} = \mathbf{\Sigma_X}$, and therefore the principal components of $\mathbf{X}$.

The SVD decomposition of a matrix $\mathbf{M}$ is given by

$$M = U \Sigma V^T \quad (7)$$

Here $\mathbf{U}$ and $\mathbf{V}$ are the orthogonal matrices, where the elements of $\mathbf{V}$ are eigenvectors, and $\mathbf{U}$ is the set of vectors defined by $\mathbf{u}_i \equiv (1/\sigma_i)\mathbf{X}\mathbf{v}_i$. $\mathbf{\Sigma}$ is a diagonal matrix with singular values

$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$. Singular values are positive and real, which are obtained by taking the square roots of eigenvalues of a matrix.

Linear Discriminant Analysis (LDA): LDA is another example of the feature reduction techniques, which provides the separation between classes based on the ratio of between-class variance to within-class variance [51]. The criterion function for LDA is given by

$$J(W) = \frac{|W^T \Sigma_B W|}{|W^T \Sigma_W W|} \quad (8a)$$

$$\Sigma_B = \sum_{i=1}^{m} n_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (8b)$$

$$\Sigma_T = \Sigma_W + \Sigma_B \quad (8c)$$

$$\Sigma_T = \sum_{x} (x - \mu)(x - \mu)^T \quad (8d)$$

where $\mathbf{\Sigma_B}$ is the between-class variance, $\mathbf{\Sigma_W}$ denotes the within-class variance, and $\mathbf{\Sigma_T}$ is the total variance matrix. μ is the total mean vector, μ_i denotes the mean vector for class i, and m is the total number of classes. The transformation matrix $\mathbf{W}$ maximises the ratio of between-class variance to within-class variance. The columns of matrix $\mathbf{W}$ contain the eigenvectors corresponding to the largest eigenvalues in

$$\Sigma_B w_i = \lambda_i \Sigma_W w_i \quad (9)$$

In the above relation, if $\mathbf{\Sigma_W}$ is non-singular then the equation can be solved by eigenvalue decomposition of $\mathbf{\Sigma_W^{-1}\Sigma_B}$. Alternately, the eigenvalues can be calculated as the roots of characteristic polynomial

$$|\Sigma_B - \lambda_i \Sigma_W| = 0 \quad (10)$$

and then solving

$$(\Sigma_B - \lambda_i \Sigma_W) w_i = 0 \quad (11)$$

for eigenvectors $\mathbf{w}_i$. The matrix $\mathbf{\Sigma_B}$ is of the rank $m - 1$ or less, since it is the sum of m matrices of rank one or less, and only $m - 1$ of these are independent. There are $m - 1$ non-zero eigenvalues and weight vectors corresponding to these eigenvalues. In the case of isotropic within-class scatter, the eigenvectors are those of $\mathbf{\Sigma_B}$. In general, the solution for $\mathbf{W}$ is not unique and the transformations rotate and scale the axes in different ways. These linear transformations do not make any significant changes to the criterion function $J(\mathbf{W})$ or classifier.

PCA is non-parametric and the solution is unique and independent of any hypothesis about the data probability distribution. These two properties are the weakness as well as the strength of PCA. LDA is closely related to PCA in the sense that both are linear feature reduction techniques. The difference is that PCA does not take into account any information about the classes, while LDA explicitly attempts to model the difference between the classes of data.

3.3 Classification and Fusion of Modalities

The Gaussian classifier utilises the Bayes decision theory for classification. It is assumed that the class-conditional probability $p(x|\omega_i)$ have Gaussian distribution for each class ω_i. The Bayes decision rule is described as

$$i_{Bayes} = arg \max_i P(\omega_i|x) =$$

$$arg \max_i p(x|\omega_i) P(\omega_i) \quad (12)$$

where $P(\omega_i|x)$ denotes the posterior probability, and $P(\omega_i)$ defines the prior class probability. A single component Gaussian was used to model each emotion class ω_i using a diagonal covariance matrix.

The audio-visual emotion classification was performed by the fusion of modalities at feature-level, and at decision-level, as shown in Fig.3.

4. EXPERIMENTAL RESULTS

A detailed analysis was performed to compare the class separability performance of the Mahalanobis distance, Bhattacharyya distance, and KL-divergence measures using the single subject (KL) data from the SAVEE database.

The feature selection with the Mahalanobis distance and KL-divergence criteria was performed using the full set of audio, visual, and audio-visual (feature-level fusion) features. In the case of Bhattacharyya distance, the feature selection process encountered a numerical problem after selecting a certain numbers of features, which is discussed in more detail later in this section. For this reason, the audio features were selected from the pitch, energy, duration, and MFCC features subgroups, whereas the visual features were selected from the upper, middle, and lower region of the face. In the case of audio modality, the proportions of selected features were 13 % (pitch), 49 % (energy), 12 % (duration) and 26 % (MFCC), whereas for the visual modality, the proportions of selected features were 37 % (upper face), 37 % (middle face) and 26 % (lower face). The feature-level fusion was performed by combining 30 % audio and 70 % visual features, as the visual modality performed better than the audio modality. This combination was chosen as it performed better than other combinations, such as 40 % audio, 60 % visual; and 50 % audio, 50 % visual features. For the decision-level fusion, the posterior probabilities obtained for the two modalities were multiplied for equal numbers of selected features from the audio and visual modalities. The weighting of the two modalities were equal. The data were divided into six sets, where in each experiment the training data consist of five sets and the testing data consist of one set. The average results were obtained by combining the results of all six experiments.

The results achieved for the seven emotion categories using the audio, visual, and audio-visual modalities are plotted in Fig.4. These results were obtained for the LDA 6 features. The best results achieved for the audio modality were 53 % (standard error (SE): 7.2, 40 features (ft.)), 52 % (SE: 7.5, 20 ft.) and 48 % (SE: 13.5, 10 ft.),

whereas for the visual modality it were 99 % (SE: 1.6, 50 ft.), 98 % (SE: 3.3, 20 ft.) and 98 % (SE: 3.3, 30 ft.) using the Bhattacharyya, Mahalanobis and KL-divergence measure-based features, respectively. In the bimodal scenario, the best results for the feature-level fusion were 99 % (SE: 1.6, 70 ft.),

99 % (SE: 1.6, 50 ft.) and 98 % (SE: 2.1, 190 ft.), whereas for the decision-level fusion it were 99 % (SE: 1.6, 50 ft.), 99 % (SE: 1.6, 40 ft.) and 98 % (SE: 3.3, 30 ft.) for the Bhattacharyya, Mahalanobis and KL-divergence measure-based features, respectively.

The LDA-transformed features performed better than the PCA-transformed features. The results obtained for the PCA 20 components are discussed here. The best results for the audio modality were 42 % (SE: 6.5, 50 ft.), 42 % (SE: 5.5, 10 ft.) and 38 % (SE: 7.9, 30 ft.), whereas for the visual modality it were 98 % (SE: 3.3, 50 ft.), 98 % (SE: 3.3, 80 ft.) and 98% (SE: 3.3, 140 ft.) for the Bhattacharyya, Mahalanobis and KL-divergence measure-based features, respectively. In the case of bimodal scenario, the best results for the feature-level fusion were 90 % (SE: 3.0, 20 ft.), 89 % (SE: 3.9, 130 ft.) and 80 % (SE: 10.1, 170 ft.), whereas for the decision-level fusion it were 98 % (SE: 3.3, 70 ft.), 98 % (SE: 3.3, 40 ft.) and 95 % (SE: 6.2, 30 ft.) for the Bhattacharyya, Mahalanobis and KL-divergence measure-based features, respectively.

A comparison of the human and machine performance for the seven emotion categories is shown in Table 1. We achieved classification accuracy comparable to human performance. In general, the visual modality performed better than the audio modality, and the performance of fusion at decision-level was better than the fusion at feature-level, especially in the case of PCA. A much higher accuracy was obtained for the visual modality alone. For this reason, no significant improvement was observed in the classification accuracy when the two modalities were combined. The overall performance of Bhattacharyya distance was better than the KL-divergence. In comparison to the Mahalanobis distance, the Bhattacharyya distance performed slightly better for the audio and visual modalities in the case of LDA, whereas for the bimodal scenario the results were comparable. In the case of PCA, a comparable performance was achieved for both the Bhattacharyya and Mahalanobis distances.

Table 1. Comparison of human and machine average classification accuracies (%) for seven emotions on the SAVEE. FL: Feature-Level; DL: Decision-Level; Bhat.: Bhattacharyya distance; Mah.: Mahalanobis distance; Div.: KL-Divergence measure.

Modality	Human	Machine (PCA)			Machine (LDA)		
		Bhat.	Mah.	Div.	Bhat.	Mah.	Div.
Audio	67 ± 2.5	42 ± 6.5	42 ± 5.5	38 ± 7.9	53 ± 7.2	52 ± 7.5	48 ± 13.5
Visual	88 ± 0.6	98 ± 3.3	98 ± 3.3	98 ± 3.3	99 ±1.6	98 ± 3.3	98 ± 3.3
Audio-visual (FL fusion)	92 ± 0.1	90 ± 3.0	89 ± 3.9	80 ± 10.1	99 ± 1.6	99 ± 1.6	98 ± 2.1
Audio-visual (DL fusion)	92 ± 0.1	98 ± 3.3	98 ± 3.3	95 ± 6.2	99 ± 1.6	99 ± 1.6	98 ± 3.3

For the Bhattacharyya distance, the feature selection starts with the full set of features but it encounters a problem after selecting a certain number of features. The Bhattacharyya distance for two normally distributed classes is defined as

$$d_{Bhat} = \frac{1}{8}(\mu_i - \mu_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1}$$

$$(\mu_i - \mu_j) + \frac{1}{2}\ln \frac{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|}{\sqrt{|\Sigma_i||\Sigma_j|}}$$

where μ_i and μ_j are the means, while $\mathbf{\Sigma}_i$ and $\mathbf{\Sigma}_j$ are the covariance matrices of classes ω_i and ω_j, respectively. The Bhattacharyya distance consists of two components: the first term defines the class separability based on class means, whereas the second term provides the class separability based on class covariance matrices. The first term does not cause any problems, but the second term becomes infinite after selecting a certain number of features. The distance measure is averaged over all binary combinations of the emotion classes for different numbers of selected features, and when one or more of these combinations fails, it leads to the failure of the feature selection process. This problem is caused by the denominator of second term, which consists of the product of determinants of the two covariance matrices. If the values of the two determinants are very small, their product results in a zero value, and thus returns an infinite value for the second term. It was

observed that the feature selection process was affected by the small amount of training data. The problem of singularity can be avoided by using a large amount of training data and reducing the number of features [46]. To overcome this limitation of the Bhattacharyya distance, we selected features from the subgroups of features for each of the audio and visual modalities.

5. CONCLUSIONS

Classification of seven emotion classes was performed on the SAVEE database using the Mahalanobis distance, Bhattacharyya distance, and KL-divergence measures as feature selection criteria. The LDA-transformed features performed better than the PCA-transformed features. The overall best results were achieved with LDA 6 features and PCA 20 features (components).

In general, better results were achieved for the visual modality in comparison to the audio modality, and the fusion at decision-level performed better than the fusion at feature-level, especially in the case of PCA. The overall performance of Bhattacharyya distance was better than the KL-divergence. In comparison to the Mahalanobis distance, the Bhattacharyya distance performed slightly better for the audio and visual modalities in the case of LDA, whereas for the bimodal scenario it was comparable. In the case of PCA, a comparable performance was achieved for both the Bhattacharyya and Mahalanobis distances.

Classification accuracy comparable to human was achieved on the SAVEE database. Differences existed between the classification accuracies of the machine and humans. The possible reasons are differences in the training data, i.e., the machine was trained/tested in a speaker-dependent scenario but humans were adapted to a small amount of data and evaluation was performed in a speaker-independent fashion, the task was discrete emotion classification, and the expressed emotions may have been lacking in naturalness.

In future the current method will be applied to the data of all speakers of SAVEE database and the method will be extended to other databases. It will be interesting to investigate the emotion classification both in the speaker-dependent and speaker-independent scenarios.

6. ACKNOWLEDGEMENTS

We are thankful to Dr P.J.B. Jackson, CVSSP, University of Surrey, UK for providing us the SAVEE database for our analysis.

7. REFERENCES

1. Sebe, N., I. Cohen & T.S. Huang. Multimodal Emotion Recognition. In: *Handbook of Pattern Recognition and Computer Vision*. World Scientific (2005).
2. Kessous, L., G. Castellano & G. Caridakis. Multimodal Emotion Recognition in Speech-Based Interaction using Facial Expression, Body Gesture and Acoustic Analysis. *Journal on Multimodal User Interfaces* 3(1): 33–48 (2010).
3. Picard, R.W. *Affective Computing*. MIT Press, Cambridge (1997).
4. Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz & J.G. Taylor. Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine* 18(1): 32–80 (2001).
5. Zeng, Z., M. Pantic, G.I. Roisman & T.S. Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(1): 39–58 (2009).
6. Haq, S. & P.J.B. Jackson. Multimodal Emotion Recognition. In: *Machine Audition: Principles, Algorithms and Systems*. IGI Global Publisher, USA (2010).
7. Busso, C. & S.S. Narayanan. Interrelation Between Speech and Facial Gestures in Emotional Utterances: A Single Subject Study. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8): 2331–2347 (2007).
8. Pantic, M., N. Sebe, J.F. Cohn & T. Huang. Affective Multimodal Human Computer Interaction. In *Proc. ACM Int'l Conf. on Multimedia*, Singapore, Hilton, p. 669–676, (2005).
9. Batliner, A., C. Hacker, S. Steidl, E. Noth, S. D'Arcy, M. Russel & M. Wong. "You Stupid Tin Box" - Children Interacting with the AIBO Robot: A Cross-Linguistic Emotional Speech Corpus. In: *Proc. Int'l Conf. on Language Resources and Evaluation*, Lisbon, Portugal, p. 171–174 (2004).
10. Burkhardt, F., A. Paeschke, M. Rolfes, W. Sendlmeier & B. Weiss. A Database of German Emotional Speech. In *Proc. Interspeech*, Lisbon, Portugal, p. 1517–1520 (2005).
11. Engberg I.S. & A.V. Hansen. *Documentation of Danish Emotional Speech Database (DES)*. Center for Person Kommunikation, Dept. of Comm. Tech., Inst. of Elect. Sys., Aalborg Univ., Denmark (1996).

12. Kanade, T., J. Cohn & Y. Tian. Comprehensive Database for Facial Expression Analysis. In *Proc. IEEE Int'l Conf. Face and Gesture Recognition*, Grenoble, France, m p. 46–53 (2000).
13. Pantic, M, M. Valstar, R. Rademaker & L. Maat. Web-Based Database for Facial Expression Analysis. In *Proc. ACM Int'l Conf. Multimedia and Expo*, Amsterdam, The Netherlands, p. 317–321 (2005).
14. Douglas-Cowie, E., N. Campbell, R. Cowie & P. Roach. Emotional Speech: Towards a New Generation of Databases. *Speech Communication* 40(1-2): 33–60 (2003).
15. Douglas-Cowie, E., R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir & K. Karpouzis. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In *Proc. Int'l Conf. on Affective Computing and Intelligent Interaction*, Lisbon, Portugal, p. 488–500 (2007).
16. Bänziger, T., H. Pirker & K.R. Scherer. GEMEP-GEneva Multimodal Emotion Portryals: A Corpus for the Study of a Multimodal Emotional Expressions. In *Proc. LREC Workshop on Corpora for Research on Emotion and Affect*, Genova, Italy, p. 15–19 (2006).
17. Busso, C., M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee & S.S. Narayanan. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation* 42(4): 335–359 (2008).
18. Haq, S. & P.J.B. Jackson. Speaker-Dependent Audio-Visual Emotion Recognition. In *Proc. Int'l Conf. on Auditory-Visual Speech Processing*, Norwich, UK, p. 53–58 (2009).
19. Hassan, A. & R.I. Damper. Emotion Recognition from Speech using Extended Feature Selection and a Simple Classifier. In *Proc. Interspeech*, Brighton, UK, p. 2043–2046 (2009).
20. Lin, Y. & G. Wei. Speech Emotion Recognition Based on HMM and SVM. In *Proc. Int'l Conf. Machine Learning and Cybernetics*, Guangzhou, China, p. 4898–4901 (2005).
21. Neiberg, D., K. Elenius & K. Laskowski. Emotion Recognition in Spontaneous Speech Using GMMs. In *Proc. Interspeech*, Pittsburgh, Pennsylvania p. 809–812 (2006).
22. Pantic, M. & M.S. Bartlett. Machine Analysis of Facial Expressions. In: *Face Recognition.* I-Tech Education and Publishing, Vienna, Austria, p. 377–416 (2007).
23. Chang, Y., C. Hu, R. Feris & M. Turk. Manifold Based Analysis of Facial Expression. *Image and Vision Computing* 24(6): 605–614 (2006).
24. Bartlett, M.S., G. Littlewort, M. Frank, C. Lainscsek, I. Fasel & J. Movellan. Fully Automatic Facial Action Recognition in Spontaneous Behavior. In *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, Southampton, UK, p. 223–230 (2006).
25. Whitehill, J. & C.W. Omlin. Haar Features for FACS AU Recognition. In *Proc. Int'l Conf. on Automatic Face and Gesture Recognition*, Southampton, UK, p. 217–222 (2006).
26. Haq, S., P.J.B. Jackson & J.D. Edge. Audio-Visual Feature Selection and Reduction for Emotion Classification. In *Proc. Int'l Conf. on Auditory-Visual Speech Processing*, Tangalooma, Australia, p. 185–190 (2008).
27. Casale, S., A. Russo & S. Serranoa. Multistyle Classification of Speech Under Stress using Feature Subset Selection Based on Genetic Algorithms. *Speech Communication* 49(10-11): 801–810 (2007).
28. Gunes, H. & M. Piccardi. Affect Recognition from Face and Body: Early Fusion vs. Late Fusion. In *Proc. IEEE Int'l Conf. on Systems, Man, and Cybernetics*, Hawaii, USA, p. 3437–3443 (2005).
29. Kim, E.H., K.H. Hyun & Y.K. Kwak. Improvement of Emotion Recognition from Voice by Separating of Obstruents. In *Proc. IEEE Int'l Symposium on Robot and Human Interactive Communication*, Hatfield, UK, p. 564–568 (2006).
30. Busso, C., Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann & S. Narayanan. Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. In *Proc. Int'l Conf. on Multimodal Interfaces*, State College, PA, USA, p. 205–211 (2004).
31. Neiberg, D. & K. Elenius. Automatic Recognition of Anger in Spontaneous Speech. In *Proc. Interspeech*, Brisbane, Australia, p. 2755–2758 (2008).
32. Petrushin, V.A. Emotion in Speech: Recognition and Application to Call Centers. In *Proc. Artificial Neural Networks in Engineering*, St. Louis, Missouri, USA, p. 7–10 (1999).
33. Ashraf, A.B., S. Lucey, J.F. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. Solomon & B.J. Theobald. The Painful Face – Pain Expression Recognition Using Active Appearance Models. In *Proc. ACM Int'l Conf. on Multimodal Interfaces*, Nagoya, Japan, p. 9–14 (2007).
34. Bartlett, M.S., G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, & J. Movellan. Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, San Diego, CA, USA, p. 568–573 (2005).
35. Zeng, Z., J. Tu, M. Liu, T.S. Huang, B. Pianfetti, D. Roth, & S. Levinson. Audio-Visual Affect Recognition. *IEEE Transactions on Multimedia*, 9(2): 424–428 (2007).

36. Petridis, S. & M. Pantic. Audiovisual Discrimination Between Laughter and Speech. In *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, USA, p. 5117–5120 (2008).
37. Jackson, P. & S. Haq. *Surrey Audio-Visual Expressed Emotion (SAVEE) Database*. http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/References.html (2014).
38. Ekman, P., W.V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W.A. LeCompte, T. Pitcairn, P.E. Ricci-Bitti, K. Scherer & M. Tomita. Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion. *Journal of Personality and Social Psychology* 53(4): 712–717 (1987).
39. Fisher, W.M., G.R. Doddington & K.M. Goudie-Marshall. The DARPA Speech Recognition Research Database: Specifications and Status. In *Proc. DARPA Workshop on Speech Recognition*, p. 93–99, Palo Alto, California (1986).
40. 3dMD. *3dMD 4D Capture System*. http://www.3dmd.com/ (2014).
41. Huckvale. M. *Speech Filing System*. http://www.phon. ucl.ac.uk/resource/sfs/ (2014).
42. Young, S., G. Evermann, D. Kershaw, D. Moore, J. Odell, D. Ollason, V. Valtchev & P. Woodland. *Hidden Markov Model Toolkit*. http://htk.eng.cam.ac.uk/ (2014).
43. Kittler, J. Feature Set Search Algorithms. In: *Pattern Recognition and Signal Processing.* p. 41–60. Sijthoff & Noordoff International Publishers, The Netherlands (1978).
44. Campbell, J.P. Speaker Recognition: A Tutorial. *Proceedings of the IEEE* 85(9): 1437–1462 (1997).
45. Wang, Y. & L. Guan. Recognizing Human Emotional State From Audiovisual Signals. *IEEE Transactions on Multimedia*, 10(5): 936–946 (2008).
46. Comaniciu, D., P. Meer & D. Tyler. Dissimilarity Computation Through Low Rank Corrections. *Pattern Recognition Letters* 24(1-3): 227–236 (2003).
47. Mahalanobis, P.C. On the Generalised Distance in Statistics. In *Proc. National Institute of Sciences of India*, vol. 2, p. 49–55, (1936).
48. Bhattacharyya, A. On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distributions. *Bulletin of the Calcutta Mathematical Society* 35:99–109 (1943).
49. Kullback, S. *Information Theory and Statistics*. Dover Publications, New York (1968).
50. Shlens, J. *A Tutorial on Principal Component Analysis*. Systems Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla (2005).
51. Duda, R.O., P.E. Hart & D.G. Stork. *Pattern Classification*. John Wiley & Sons, USA, (2001).