A Combined Rule-Based & Machine Learning Audio-Visual Emotion Recognition Approach

Kah Phooi Seng, *Member, IEEE*, Li-Minn Ang, *Senior Member, IEEE*, and Chien Shing Ooi

Abstract—This paper proposes an audio-visual emotion recognition system that uses a mixture of rule-based and machine learning techniques to improve the recognition efficacy in the audio and video paths. The visual path is designed using the Bi-directional Principal Component Analysis (BDPCA) and Least-Square Linear Discriminant Analysis (LSLDA) for dimensionality reduction and discrimination. The extracted visual features are passed into a newly designed Optimized Kernel-Laplacian Radial Basis Function (OKL-RBF) neural classifier. The audio path is designed using a combination of input prosodic features (pitch, log-energy, zero crossing rates and Teager energy operator) and spectral features (Mel-scale frequency cepstral coefficients). The extracted audio features are passed into an audio feature level fusion module that uses a set of rules to determine the most likely emotion contained in the audio signal. An audio visual fusion module fuses outputs from both paths. The performances of the proposed audio path, visual path, and the final system are evaluated on standard databases. Experiment results and comparisons reveal the good performance of the proposed system.

Index Terms—Emotion recognition, audio-visual processing, rule-based, machine learning, multimodal system

1 Introduction

Emotion recognition is an automated process to identify the affective state of a person and has gained the increasing attention of researchers in the human-computer interaction (HCI) field for various applications like automotive safety [1], gaming experiences [2], mental diagnosis in military service [3], customer services [4], etc. Over the decades, several research efforts have been conducted for audio-visual emotion recognition. In the literature, three main approaches can be broadly distinguished: (i) audio-based approaches, (ii) visual-based approaches, and (iii) audio-visual approaches. Initial works focused on treating the audio data and visual data modalities separately. The audio-based emotion recognition efforts are based on extracting and recognizing the emotional states contained in the human speech signal. An important issue is the selection of the salient features to be used for discriminating the different emotions. Two types of features have been found to be useful for recognizing emotion in speech: prosodic and spectral features. Examples of commonly used prosodic features are pitch and energy and examples of commonly used spectral features are Mel-scale frequency cepstral coefficients (MFCC). Although prosodic features are commonly used in many works [5], [6], [7], [8] some researchers have demonstrated the usefulness of spectral features for speech emotion recognition [9], [10]. The monograph work in [11] further investigated combining different types of features like prosodic and spectral features for audio-based emotion recognition. The visual-based emotion recognition efforts [12], [13], [14] are based on extracting and recognizing the emotional states contained in the human facial expression. An example is a recent work by Tawari & Trivedi [12] which used a representation of image sequences by weighted sums of registered face images where the weights are derived using auditory features. Other examples are the works by [13], [14] which proposed facial expression recognition based on local binary patterns.

Researchers then combined both the audio and visual modalities and found that the combination of these two modalities can improve the recognition performance for emotion [15], [16], [17], [18], [19], [20] and speech [21]. Some earlier works by [15], [16], and [17] proposed their own recognizers which could classify six emotions, and obtained recognition rates of 84, 72 and 97.2 percent respectively. However, these systems were only designed and evaluated based on their own recorded video database that contained only two subjects. Another work by Hoch et al. [18] obtained a high 90.7 percent recognition rate. However, this system was subject-dependent and could only classify three emotions. The work by Wang & Guan [19] proposed an audio-visual emotion recognizer which could achieve a high recognition rate of 82.14 percent. The authors stated in their work that the visual feature representation of their person-independent work is not strong enough, and the visual feature based classification accuracy is low. A recent effort [20] proposed a bimodal emotion recognition system which utilized Kernel Cross-Modal Factor Analysis. Their system achieved 72.47 and 82.22 percent recognition rates when evaluated using the eNTERFACE'05 and RML audio-visual emotion databases respectively. Other recent works on audio-visual emotion recognition include the works by [58], [59], [60], [61].

- *K.P Seng and L.M. Ang are with the School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW 2678, Australia. E-mail: {kseng, lang}@csu.edu.au.*
- *C.S. Ooi is with the Department of Computer Science and Networked Systems, Sunway University, Subang Jaya 47500, Malaysia. E-mail: ocshing@gmail.com.*

Manuscript received 27 Jan. 2015; revised 6 Apr. 2016; accepted 29 Apr. 2016. Date of publication 6 July 2016; date of current version 7 Mar. 2018. Recommended for acceptance by B.W. Schuller. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TAFFC.2016.2588488

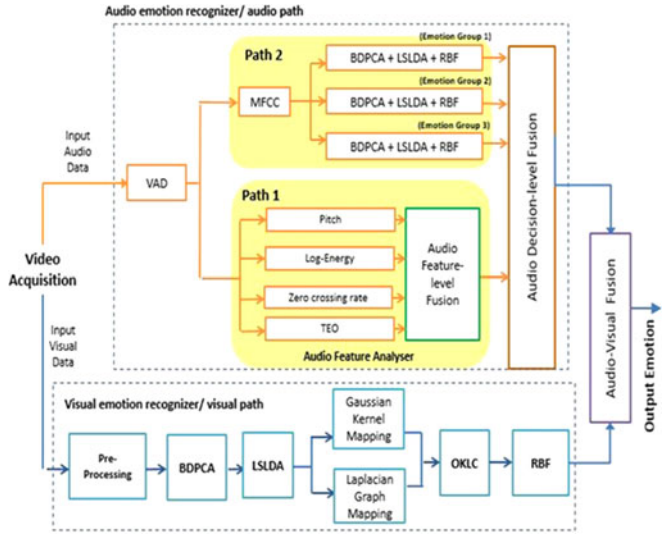


Fig. 1. The proposed audio visual emotion recognition system.

From the above, we observed that the potential to improve the existing performance of the audio-visual emotion recognition is still far from a solved problem due to the limitations in the recognition accuracy. Although there is general agreement amongst researchers on the effectiveness of using a fusion of audio and visual modalities for emotion recognition, the challenge is that there is still uncertainty on how the integration of these modalities can be best accomplished. In particular, the use of multimodal techniques such as audio-visual modalities compared to using audio-only or visual-only modalities lead to a number of additional challenges such as determining the suitable feature extraction techniques and optimal data fusion techniques prior to performing the recognition or classification.

In the literature, two approaches are commonly employed for signal/image-based recognition applications: rule-based and machine learning approaches. Rule-based approaches suppose that the necessary information for decision making is contained in expert knowledge. The rules are obtained from the expert as his or her conditional beliefs, and are of the form IF A THEN B with a weighted uncertainty measure. On the other hand, machine learning approaches aim to improve automatically through experience. A more exact formulation of machine learning can be found in [63] as "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". Thus, an important issue for rule-based approaches is how to design the rules for the extracted feature set. There are different approaches which can be used. The work by Grimm et al. [22] proposed a rule-based fuzzy logic method on the basis of IF–THEN rules to estimate the continuous values of the emotion primitives from acoustic features derived from the speech signal and obtained an overall recognition rate of up to 83.5 percent.

In this paper, we adapted an approach for audio-based rule development based on the works in [23], [24] which was motivated by the psychological study of emotions by Schlosberg [25] in which emotions are represented in three dimensions: activation (arousal), potency (power) and evaluation (pleasure). In these works, the authors remarked that the optimal feature set to be used strongly depends on the emotions to be separated, and that using one global feature set for the discrimination of all emotions is suboptimal. They proposed a three-stage classification technique to recognize six emotions (anxiety, happiness, anger, neutral, boredom and sadness) using a set of rules, and showed that their three-stage rule-based approach could outperform the single-stage approach. In our work, we propose a different set of rules to recognize a different set of six emotions (anger, happy, sad, disgust, surprise and fear). These are the six universal emotional states proposed by the psychologist Paul Ekman and also the emotions found in the eNTERFACE and RML databases which we used.

Learning from the experiences of previous researchers and from our own investigations, this paper proposes an audio-visual emotion recognition system that uses a mixture of rule-based and machine learning approaches to improve the recognition efficacy. Fig. 1 shows an overview of the proposed system which contains one visual path and two audio sub-paths A1 and A2 in parallel. An input video stream is first split into an audio path and a video path. For the video path, the input images are pre-processed to detect and locate the facial region. Then, a combination of feature extraction techniques is designed using the Bi-directional Principal Component Analysis (BDPCA) and Least-Square Linear Discriminant Analysis (LSLDA) to extract and discriminate the visual features amongst the six emotion classes. A new data fusion scheme called Optimized Kernel-Laplacian Radial Basis Function (OKL-RBF) neural classification is proposed for the back-end processing in the visual path. Kernel and Laplacian mappings are first used to compute the similarities of the attribute-based and the relation-based (graph) information from the visual features in parallel. The kernels and Laplacians matrices are then optimised and merged to form an optimal Kernel-Laplacian matrix. Together with the mappings and optimized clustering, the radial basis function (RBF) neural network performs the classification.

For the audio path, the first sub-path A1 is designed based on the audio prosodic features. It contains an Audio Features Analyzer to extract some important prosodic features such as pitch, log-energy, zero-crossing rate (ZCR) and Teager energy operator (TEO). It also contains a module called Audio Feature-Level Fusion to perform fusion at the feature level. The second sub-path A2 is designed based on audio spectral features. This path consists of Mel-scale frequency cepstral coefficients feature extraction followed by three parallel sub-paths for three sets of emotion groups. An Audio Decision-Level Fusion module is also proposed to fuse the information from both sub-paths A1 and A2. A decision making mechanism is included in the fusion module to decide the most likely audio emotion. Finally, an Audio-Visual Fusion module is designed to fuse the output streams from the audio and visual paths. This allows the proposed system to perform human emotion recognition over video conferencing. The remainder of the paper is organized as follows: Section 2 presents the details for the visual path, while Section 3 presents the details for the audio path. Section 2 also gives the details for the OKL-RBF which has been newly designed to extract the essential features from the training data and improve the emotion recognition efficacy. The fusion of the audio and visual paths is

discussed in Section 4. Section 5 gives experiments and comparisons on standard databases. Finally, some concluding remarks are given in Section 6.

2 Visual Path

The visual path can be seen in the bottom block diagram in Fig. 1. This section presents the details for each block in the visual path before the outputs from both the visual and audio paths are fused at the final stage.

2.1 Pre-Processing

During the pre-processing stage, face detection and localization are performed to extract out the facial region and remove the unwanted background information. A popular face detection technique called Viola-Jones (VJ) algorithm [26] was used. The output of the VJ algorithm returns a bounding box containing the facial region which is cropped out from the original image, resized and passed to the feature extractor.

2.2 Feature Extraction

Two popular techniques for feature extraction are principal component analysis (PCA) and linear discriminant analysis (LDA). PCA is an unsupervised technique and is used to reduce the image dimensionality by providing the core features to represent an image data. LDA is a supervised learning technique and is used to enhance the separation of the extracted features amongst the different classes. PCA and LDA methods have been used extensively in the research area of face recognition. In our approach, the original PCA and LDA algorithms have been further enhanced to improve the recognition efficacy. We used the Bi-directional Principal Component Analysis [27] and the Least-Square Linear Discriminant Analysis [28] for dimensionality reduction and class discrimination. These techniques were found to give high recognition efficacy in our simulations. The BDPCA and LSLDA modules are tightly-coupled and have been designed to work in conjunction to extract the essential features from the input facial data for training and classification. In the proposed design, the two modules are cascaded with the output of the BDPCA being used as input into the LSLDA. In PCA, the two-dimensional facial images have to be transformed into one-dimensional vectors prior to performing the subspace analysis, leading to a high-dimensional vector space. The advantage of the BDPCA over the PCA is that it finds the optimal projection subspaces in the row and column directions of images without needing the matrix to vector transformation, resulting in a small covariance matrix.

2.3 Optimized Kernel-Laplacian Radial Basis Function (OKL-RBF) Neural Classification

During the training phase, the extracted training facial features or data from the earlier LSLDA stage are passed to the Gaussian Kernel Mapping and the Laplacian Graph Mapping modules to obtain kernel and Laplacian matrices respectively, which are then passed to the OKLC module. The kernel mapping serves to extract the attribute-based information about the similarities of the data in the high-dimensional feature space whereas the Laplacian mapping serves to extract the information about the relationships within the data in the low-dimensional weighted space. The relation information can be inferred as a graph structure, where the data samples are represented as vertices connected by non-negatively weighted undirected edges. In our approach, the role of the OKLC (which stands for optimized *k*-means Laplacian clustering [29]) is to find the optimal coefficients for the kernel and Laplacian matrices, and form the optimal Kernel-Laplacian matrix, Ω_{OKLC}, which will be used for subsequent training of the RBF neural network. During the testing phase, the extracted testing features or data from the LSLDA stage is mapped again using the Gaussian Kernel Mapping and the Laplacian Graph Mapping techniques, so that the kernel and Laplacian matrices for the testing data can be obtained. The additive model [30] is then used to merge them with the optimal coefficients obtained from OKLC during the training phase. The final stage involves computing a Kernel-Laplacian matrix for the test data, which allows the trained RBF neural network to perform the classification.

At the start, the BDPCA+LSLDA extracted features, $X_{BDPCA\&LSLDA}$are stored in an $a \times j$ matrix, x, where j denotes the number of features and a denotes the number of samples. The Gaussian kernel matrix, G and the Laplacian matrix,$\tilde{L}$ can be constructed using kernel mapping and Laplacian graph mapping as shown in Equation (1),

$$G = \exp\left(-\frac{\|x_a - x_j\|}{2\sigma^2}\right), \quad \tilde{L} = D^{-1/2} W D^{-1/2}, \tag{1}$$

where W denotes the affinity matrix, D denotes an $a \times a$ diagonal degree matrix which is constructed from the sum of rows of W, and $\boldsymbol{\sigma}$ denotes the kernel width as shown in Equation (2).

$$W = w_{aj} = \exp\left(-\frac{\|x_a - x_j\|}{\sigma^2}\right), \quad D = d_a = \sum\nolimits_j w_{aj}. \tag{2}$$

The OKLC is a strategy of learning the optimal weighted convex linear combinations of multiple kernels and Laplacians. The OKLC objective function, J_{OKLC} which is needed to be solved is shown in Equation (3),

$$\begin{aligned} &\underset{A,\vec{\theta}}{\text{maximum}} \; J_{OKLC} = trace\left(A^T\left(\tilde{\mathbf{L}} + \mathbf{G}\right)A\right) \\ &\text{where} \quad \tilde{\mathbf{L}} = \sum_{i=1}^{r} \theta_i \tilde{L}_i, \quad \mathbf{G} = \sum_{j=1}^{s} \theta_{j+r} G_{cj}, \quad \sum_{i=1}^{r} \theta_i^{\delta} = 1, \quad \sum_{j=1}^{s} \theta_{j+r}^{\delta} = 1, \\ &\qquad \theta_l \geq 0, \quad l = 1, \ldots, (r+s), \\ &\qquad A^T A = I_K, \end{aligned} \tag{3}$$

where $\theta_1, \ldots, \theta_r$ and $\theta_{r+1}, \ldots, \theta_{r+s}$ are the optimal coefficients assigned to the Laplacians and kernels respectively and $\delta \in \{0,2\}$ is a sparseness control parameter. The symbol $\mathbf{G}$ denotes the combined kernel matrices of multiple G_{cj}, and the symbol $\tilde{\mathbf{L}}$ denotes the combined Laplacian matrices of multiple $\tilde{L}_i$. In this paper, we used an equivalent number of kernel matrices, s and Laplacian matrices, r. The training phase of the algorithm can be described as follows:

1. The initial calculations to combine the kernel matrices, $\mathbf{G}^{(0)}$, and the Laplacian matrices, $\tilde{\mathbf{L}}^{(0)}$ are started

with a random initial guess of the coefficients for kernels, $\theta_1^{(0)},\ldots,\theta_r^{(0)}$ and Laplacians,$\theta_{r+1}^{(0)},\ldots,\theta_{r+s}^{(0)}$, respectively. These combinations provide the initial Kernel-Laplacian matrix, $\Omega_{OKLC}^{(0)}$ as shown in Equation (4).

$$\begin{aligned}\tilde{\mathbf{Ł}}^{(0)} &= \sum_{i=1}^{r}\theta_i^{(0)}\tilde{L}_i^{(0)}, \mathbf{G}^{(0)} = \sum_{j=1}^{s}\theta_{j+r}^{(0)}G_{cj}^{(0)}.\\ \Omega_{OKLC}^{(0)} &= \tilde{\mathbf{Ł}}^{(0)} + \mathbf{G}^{(0)}\end{aligned} \tag{4}$$

2. Next, based on the computed $\Omega_{OKLC}^{(0)}$, eigenvectors are computed and sorted in descending order based on its eigenvalues. The resulting eigenvector is then applied to *k*-means clustering to obtain the weighted scalar cluster membership matrix which is denoted as $A^{(0)}$. The initial iteration index, $\boldsymbol{\gamma}$ is set as 0.
3. To discriminate the cluster assignment $A^{(\boldsymbol{\gamma})}$, the affinity matrix, $F^{(\boldsymbol{\gamma})}$ is computed as shown in Equation (5).

$$F_{ab} = \begin{cases} +1 & \text{if } A_{ab} > 0, \quad a = 1,\ldots,N, \quad b = 1,\ldots,k \\ -1 & \text{if } A_{ab} = 0, \quad a = 1,\ldots,N, \quad b = 1,\ldots,k \end{cases}, \tag{5}$$

where a denotes the index of the samples, and b denotes the index of the cluster.

4. The $\boldsymbol{\gamma}$th iteration's coefficients of Laplacians, $\theta_1^{(\gamma)},\ldots,\theta_r^{(\gamma)}$ and coefficients of kernels, $\theta_{r+1}^{(\gamma)},\ldots,\theta_{r+s}^{(\gamma)}$ are then obtained by adopting the algorithm of SIP-LSSVM-MKL [31] as shown in Equations (6) and (7),

$$\theta_1^{(\gamma)},\ldots,\theta_r^{(\gamma)} \longleftarrow \text{SIP-LSSVM-MKL}(\tilde{L}_1,\ldots,\tilde{L}_r,F^{(\gamma)}), \tag{6}$$

$$\theta_{r+1}^{(\gamma)},\ldots,\theta_{r+s}^{(\gamma)} \longleftarrow \text{SIP-LSSVM-MKL}(G_1,\ldots,G_{cs},F^{(\gamma)}) \tag{7}$$

where $\tilde{L}_i$ denotes the ith number of the Laplacian matrix in $\tilde{\mathbf{Ł}}$ and G_{cj} denotes the jth number of the kernel matrix in $\mathbf{G}$.

5. Using the additive model, the new combination of $\mathbf{G}$ and $\tilde{\mathbf{Ł}}$, named $\Omega_{OKLC}^{(\gamma+1)}$ is obtained as shown in Equation (8).

$$\begin{aligned}\tilde{\mathbf{Ł}}^{(\gamma)} &= \sum_{i=1}^{r}\theta_i^{(\gamma)}\tilde{L}_i^{(\gamma)}, \mathbf{G}^{(\gamma)} = \sum_{j=1}^{s}\theta_{j+r}^{(\gamma)}G_{cj}^{(\gamma)}.\\ \Omega_{OKLC}^{(\gamma+1)} &= \tilde{\mathbf{Ł}}^{(\gamma)} + \mathbf{G}^{(\gamma)}\end{aligned} \tag{8}$$

6. A new weighted scalar cluster membership matrix, $A^{(\boldsymbol{\gamma})}$, can then be computed by using the eigenvalue decomposition on $\Omega_{OKLC}^{(\gamma+1)}$.
7. After $A^{(\boldsymbol{\gamma})}$ is obtained, the error of the clustering assignment matrix, ΔA is calculated using Equation (9).

$$\Delta A = \frac{\left\|A^{(\gamma+1)} - A^{(\gamma)}\right\|^2}{\left\|A^{(\gamma+1)}\right\|^2}. \tag{9}$$

8. If the value of the error, ΔA is larger than the initially set threshold error ε, the iteration ($\boldsymbol{\gamma}$+1) is repeated from Steps 4 to 8. The optimized coefficients for kernels, $\theta_1^{(\gamma)},\ldots,\theta_r^{(\gamma)}$ and Laplacians, $\theta_{r+1}^{(\gamma)},\ldots,\theta_{r+s}^{(\gamma)}$ are obtained once the condition ($\Delta A < \varepsilon$) is satisfied.
9. At the output of the OKLC stage, the Kernel-Laplacian matrix, $\Omega_{OKLC}^{(\gamma+1)}$, optimized coefficients for kernels, $\theta_1^{(\gamma)},\ldots,\theta_r^{(\gamma)}$ and optimized coefficients for Laplacians,$\theta_{r+1}^{(\gamma)},\ldots,\theta_{r+s}^{(\gamma)}$ are obtained. The parameter $\Omega_{OKLC}^{(\gamma+1)}$ is subsequently passed to train the RBF neural network.
10. At the input of each neuron in the input layer, the activation of the hidden units is calculated based on the distance, d_i between the input vector, $\Omega_{OKLC(i)}$ and the center of the neuron, $\boldsymbol{\mu}_j$ as shown in Equation (10),

$$d_i = \left\|\Omega_{OKLC_{(i)}} - \mu_j\right\|, \tag{10}$$

where i denotes the number of input samples and j denotes the number of hidden units.

11. When the distance obtained is smaller than the spread width or width of the receptive field $\boldsymbol{\sigma}_j$ in the input space, the appreciable value can be obtained by the radial basis function $\boldsymbol{\phi}_j$ as shown in Equation (11).

$$\phi_j(\mathrm{x}) = \exp\left(-\frac{\left\|\Omega_{OKLC_{(i)}} - \mu_j\right\|}{2\sigma_i^2}\right). \tag{11}$$

12. By applying the basis function to the distance, the output of the neuron can be formed in the output layer. Linear regression is then performed to predict the targeted outputs, $\boldsymbol{y}$ as shown in Equation (12).

$$output = \sum_{i=1}^{j} w_j\phi_j(\mathrm{x}), \tag{12}$$

where w_j are the weights between the hidden and output layers, and the bias term has been set to 1.

During the testing phase, the extracted features on the testing samples from the LSLDA stage are mapped to obtain the matrices of the test kernels, $\boldsymbol{G}_{test}$ and the test Laplacians, $\tilde{L}_{test}$. The optimized kernels and Laplacians are obtained from Step 9. The additive model is then applied to combine $\boldsymbol{G}_{test}$ and $\tilde{L}_{test}$ using Equation (4). The theory of the additive model or Rayleigh Quotient provides the optimization of multi-source learning. This strategy combines the outputs of different feature-mappings after each of them have been optimized separately. As the result, the corresponding testing data is in a new form, Ω_{OKLC_test}, and the previously trained RBF parameters are utilized to perform classification on Ω_{OKLC_test}.

3 Audio Path

The audio path can be seen in the top block diagram in Fig. 1. This section presents the details for each block in the audio path. The Path A2 contains three sub-paths in parallel for three emotion groups. Each emotion group is designed to contain two emotion classes. The three emotion groups are: Group 1 (angry and happy), Group 2 (sad and disgust)

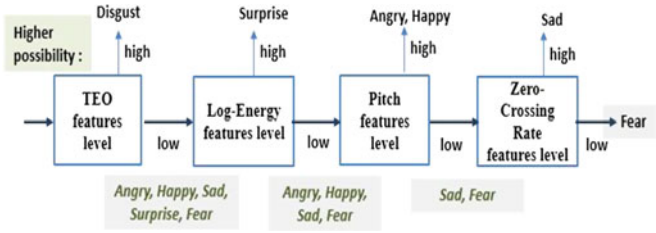


Fig. 2. The feature-level fusion module.

and Group 3 (surprise and fear). The design of the emotion groups is discussed later in this section.

3.1 Pre-Processing

The Voice Activity Detector (VAD) [31] is used for pre-processing the speech signal to eliminate the background noise and segment out the non-speech portions of the audio signal. The VAD technique uses the short-time energy (STE) and short-time zero-crossing rate (STZCR) features. First, the speech signal $x(m)$ in the time domain is divided into n number of frames. The STE detection is performed to determine the energy within each frame or segmented voice signal. Next, the STZCR is calculated from the weighted average from the number of times the speech signal changes sign within a time window. The STE and STZCR are compared to determine the existence of speech in the signal. The unwanted signal frames (e.g., silence or unvoiced) are then segmented out.

3.2 Audio Analyzer & Feature-Level Fusion (Path A1)

The Audio Feature Analyzer has been designed based on our earlier investigations on audio prosodic features [32] and extracts important features (such as pitch, log-energy, ZCR, and TEO) from the input signal. The pitch extraction method [33] calculates the distance between the zero crossing points of the signal. The speech signal $x(m)$ in the time domain is first divided into n number of frames by windowing, $w(n)$ and is denoted as $s(m)$. Then the pitch is obtained from its periodicity R as shown in Equation (13),

$$R(k) = \sum_{m=0}^{L-k-1} s(m)s(m+k), \tag{13}$$

where L denotes the window length, and k refers to the representation of the pitch period of a peak. The log-energy [34] indicates the total squared amplitude in a segment of speech and can be formulated as shown in Equation (14),

$$E = \log_{10}\left(\sum_{i=1}^{N} x^2\right), \tag{14}$$

where N denotes the number of frames and x denotes the sample of the speech. The zero crossing rate (ZCR) [35] calculates the weighted average of the number of times the speech signal changes sign within a particular time window, and can be calculated as shown in Equation (15),

$$\mathrm{sgn}\{x\} = \begin{cases} 1 & if \quad x(n) \geq 0 \\ -1 & if \quad x(n) < 0 \end{cases}, \tag{15}$$

where n refers to the current sample. The TEO [36] detects the nonlinear component which changes appreciably between different emotional speech signals. An energy tracking operator for the speech signal is shown in Equation (16),

$$\Psi[s(n)] = s^2(n) - s(n+1)s(n-1), \tag{16}$$

where $\Psi[s(n)]$ are the coefficients and $s(n)$ is the sampled speech signal. Our investigations have shown that these four prosodic features used in conjunction give good discrimination features amongst the different emotion classes.

The following rules/heuristics are formulated: (i) The TEO feature can be used to discriminate the disgust emotion class from the other five emotion classes; (ii) The ZCR feature can be used to distinguish the sad, disgust and surprise classes from the remaining three emotion classes; (iii) The angry and surprise emotion features tend to have higher pitch values whereas the sad and fear emotions have lower values; and (iv) The log-energy feature can be used to distinguish the surprise emotion class. Fig. 2 shows an algorithmic description of the above. Note that the angry and happy emotions cannot be discriminated using the four prosodic features, and we put these two emotions into one emotion group. The MFCC spectral features from the Path A2 will be used to recognize the individual emotion within the group. Fig. 2 also shows that the disgust and surprise emotions should be put into different groups since they can be discriminated using the TEO and log-energy features. We performed analysis and experimented with the remaining two emotions (sad and fear) in different permutations and finally ended up with the following three emotion groups: Group 1 (angry and happy), Group 2 (sad and disgust) and Group 3 (surprise and fear). We note that the authors in [23] reached a similar grouping with the angry and happy emotions placed in one group, and the sad and fear (anxiety) emotions placed in different groups. For the three emotion groups, we associate weightings W_{Group1}, W_{Group2}, and W_{Group3} to determine the most probable emotion group. These weightings will then be used to influence the RBF neural classification in the Path A2.

The flow of the fusion starts by first comparing the computed TEO value with a pre-set or threshold level. If the computed TEO value is higher, the emotion "disgust" has a high probability and the weights assignment mechanism in the Audio Decision Level Fusion module will assign a heavier weight to W_{Group2} (disgust/sad) from the Path A2. If the TEO value is lower, then the computed log-energy value is compared to a log-energy threshold level. The mechanism assigns a heavier weight to W_{Group3} (surprise/fear) if the computed log-energy value is above the threshold. Next, if the computed log-energy value is lower, then the computed value of the pitch feature is compared with a pitch threshold level. If the computed pitch value is above the threshold, a heavier weight will be assigned to W_{Group1} (angry/happy). Next, the ZCR is examined. A heavier weight will be assigned to W_{Group2} from Path A2 if the computed ZCR is above its threshold, while a heavier weight will be assigned to W_{Group3} if a lower ZCR is obtained. The set of rules in Fig. 2 influence which of the outputs from the three parallel

Fig. 3. The audio decision-level fusion module.

sub-paths in the Path A2 should be emphasized. A point to note is that the groupings of the three emotion pairs in the Path A2 were not randomly selected as described earlier. The rule-based mechanism in the Path A1 work together in conjunction with the RBF neural classifier and the three emotion groups in the Path A2, and form a strong strategy for identifying the correct audio emotion. The Paths A1 and A2 are not independent in the design, and assist and support each other in the decision making process.

3.3 Emotion Groups Feature Extraction and Classification (Path A2)

The Path A2 flow begins with processing the MFCC from the speech. MFCCs are the cepstral coefficients derived from a mel-scale frequency filter-bank [37]. The MFCC feature has been identified to be one of the most influential audio features. After the MFCC feature extraction, these features are passed to three parallel sub-paths which use similar feature extraction and classification techniques as for the visual path (i.e., BDPCA+LSLDA+RBF). In our approach, each audio sub-path deals with two emotions and each RBF neural network only has to perform two-class classification. The rules in the Feature-Level Fusion module assist the design in the Path A2 by reducing the problem to the two-class classification. The six emotions are divided into three groups of pairs and the emotion pairs are discriminated using the audio prosodic features in the Path A1. The work in [32] gives some further empirical reasons for selecting these two-class groupings by evaluating their MFCC projection features in the PCA+LDA subspace.

3.4 Audio Decision-Level Fusion

The Audio Decision-Level Fusion module makes the final decision based on the outputs from the Path A1 and the Path A2 as shown in Fig. 3. The Audio Feature-Level Fusion module in the Path A1 assists the decision making via the Weights Assignment Mechanism to assign weights for $W_{Group1,}$ $W_{Group2,}$and W_{Group3} to the three outputs from the three parallel sub-paths in the Path A2. Based on $W_{Group1,}$ $W_{Group2,}$and W_{Group3}, the decision is made by considering the output with the heaviest weight. For example, if W_{Group1} has the largest number, then the output of $(\text{BDPCA+LDA+RBF})_{Group1}$ from the Path A2 is considered. If the output from the two-class classification sub-path $(\text{BDPCA+LDA+RBF})_{Group1}$ gives 'Happy', then the final decision for the audio emotion is 'Happy'. From our simulations, we have observed that the rule-based weights assignment mechanism coupled with the two-class classifiers give good performance. A different approach would be to treat the six emotions individually and design a machine learning classifier to select from the six emotions directly without using the rule-base. We have investigated this and the experiments section show comparison results for the audio path when the rule-base has been substituted using a statistical machine learning approach which is similar to the video path.

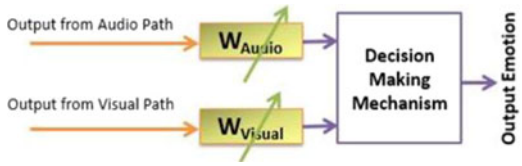


Fig. 4. Audio-visual fusion module.

4 Audio Visual Fusion

After obtaining the outputs from both the audio and visual paths, the final Audio Visual Fusion Module performs the fusion of the outputs. Two common approaches can be used: feature-level fusion and score-level fusion [38]. Our system used a score-level fusion architecture as shown in Fig. 4. To recognize human emotion via video conferencing, it is important that the proposed audio visual emotion recognition be designed to operate on video streams. For the audio path, a sliding window is applied to the speech signal. The speech signal within this window is segmented out and the window continuously slides through the speech signal. The percentage of overlapping can be pre-defined. This mechanism allows the audio path to adaptively and continuously process the speech signal and recognize the emotion at that point. The speech signal will be segmented into multiple portions in a certain length and this leads to a set of multiple outputs from the audio path. Similarly, for the visual path, the image frames are also continuously captured, processed and continuously gives a set of outputs. Thus, there are two sets of outputs from the audio and visual paths for a certain length of video sequence. This mechanism allows the proposed system to detect the emotions over a continuous video sequence. These two sets of outputs contain their respective outcomes or recognized emotions from the audio and the visual paths. Each outcome could be any one of the six emotions. For each path, the same outcomes or recognized emotions are summed to form a score set for the six emotions. There are two score sets or 12 scores in total. The weights, W_{Visual} and W_{Audio} are assigned to these scores. A proportional scoring mechanism was used (e.g., if the output from the visual path contains two outcomes for the angry emotion and one outcome for the sad emotion, then the emotion scores for V_{Angry} and V_{Sad} are 0.67 and 0.33 respectively). Each emotion score from the visual path is multiplied with W_{Visual}, while each emotion score from the audio path is multiplied with W_{Audio}. The weights W_{Visual} and W_{Audio} are assigned to the scores after considering the channel noise level and can be pre-set or adaptively adjusted (e.g., based on the mechanism described in [39]). The module performs the calculation and the emotion with the maximum score is the final decision over a certain period.

TABLE 1
Recognition Accuracy for ORL Database

Method	Recognition rate (%)
LRC [44]	93.00
2DNPP [45]	94.05
2DLDA [46]	95.85
DCV [47]	97.75
$(2D)^2$LDA [48]	98.50
Gaidhane et al. [49]	99.40
Proposed visual path (BDPCA+LSLDA+OKL-RBF)	98.50

TABLE 2
Recognition Accuracy for YALE Database

Method	Recognition rate (%)
DLPP [50]	87.88
Laplacianfaces [51]	88.70
PCA+LDA [52]	93.20
BDPCA+LDA [27]	94.10
2DLDA [46]	95.70
$(2D)^2$LDA [48]	96.50
Gaidhane et al. [49]	97.50
Cheng et al. [53]	99.10
Proposed visual path (BDPCA+LSLDA+OKL-RBF)	99.50

5 Experimental Results

This section presents experiments to evaluate the performance of the proposed visual/audio paths and the proposed final system. Although the focus of the paper is on emotion recognition, we think it is useful to first proof the effectiveness of the visual path using a face recognition task. Thus, the visual path which consists of BDPCA+LSLDA+OKL-RBF techniques was first evaluated using standard face databases such as ORL [40] and Yale [41]. Next, we performed evaluation on the Extended Cohn-Kanade dataset (CK+) [42] which is a visual emotion recognition database. Finally, the performance of the full audio-visual emotion recognition system was evaluated on the standard audio-visual emotion databases eNTERFACE'05 [43] and RML [20].

5.1 Performance Evaluation of Visual Path on ORL and Yale

In this experiment, the visual path operates as a face recognizer and its performance is evaluated using the ORL and Yale databases. The ORL database contains ten different images for each of 40 distinct subjects. The facial images are taken at different times varying the lighting, facial expressions (open or closed eyes, smiling or not smiling), and facial details (glasses or no glasses). The experiments used five images of each subject chosen randomly for training, resulting in a training set of 200 images. The remaining five images were used for testing, resulting in a testing set of 200 images. The recognition rate obtained for the ORL database is shown in Table 1 together with comparison results for some other methods. The proposed visual path as a face recognizer gave a recognition rate of 98.50 percent on the ORL database. It performed better than some recent methods such as LRC [44], 2DNPP [45], 2DLDA [46], DCV [47] and is comparable with $(2D)^2$LDA [48]. It performed only slightly lower than the recent method proposed by Gaidhane et al. [49] (98.5% versus 99.4 percent). We emphasize that our focus is on emotion recognition and not face recognition. As will be shown later, our method performed better than other methods for emotion recognition. For further investigation of the system components, experiments were performed where the OKL-RBF classifier was substituted with two other machine learning (ML) classifiers (RBF and SVM) leaving the feature extraction module unchanged (i.e., BDPCA+LSLDA+RBF and BDPCA+ LSLDA+SVM), and gave recognition rates of 92 and 90 percent respectively. This showed the validity for using the OKL-RBF classifier compared to other ML techniques. To investigate the feature extraction, all the system modules were substituted (i.e., PCA+LDA+RBF) and gave a reduced recognition rate of 86.5 percent compared to BDPCA+LSLDA +RBF with a 92 percent recognition rate.

The Yale face database contains 11 different images for each of 15 distinct subjects. The facial images demonstrate variations in lighting conditions (left-light, center-light, right-light), facial expressions (normal, happy, sad, sleepy, surprised and wink), and facial details (glasses or no glasses). The experiments used 90 samples from the 15 subjects for training and the remaining 75 samples were used for testing. The recognition rate obtained for the Yale database is shown in Table 2 together with comparison results for some other methods. The proposed visual path gave a recognition rate of 99.5 percent and outperformed all other previous methods including DLPP [50], Laplacianfaces [51], PCA+LDA [52], BDPCA +LDA [27], 2DLDA [46], $(2D)^2$LDA [48], and the recent methods by Gaidhane et al. [49] and Cheng et al. [53]. The results for PCA+LDA and BDPCA+LDA were obtained from [49]. Tables 1 and 2 have shown the effectiveness of the visual path using a face recognition task. The next sections will show its effectiveness for emotion recognition.

5.2 Performance Evaluation of Visual Path on CK+

In this experiment, the visual path operates as the visual emotion recognizer and its performance is evaluated using the CK+ database. The CK+ is a visual emotion dataset which contains 593 FACS-coded image sequences (digitized into 640×480 pixel arrays) of 123 subjects incorporating various facial expressions. Out of these sequences, 327 sequences contain emotion labels belonging to seven expressions (anger, contempt, disgust, fear, happiness, sadness and surprise) that can be used as the ground truth. The experimental settings were the same as that used in [54]. A subset of six emotions (excluding the contempt emotion) was selected from the dataset. The experiments used 1260 samples from 30 subjects for training and the remaining 180 samples were used for testing. The results for the CK+ database is shown in Table 3. The proposed visual path gave a recognition rate of 96.11 percentand performed better than other methods such as SNMF [54], MFA [55], GSNMF [54], NGE [56], and DSNGE [57]. The results in Table 3 has shown the effectiveness of the visual path using a face emotion recognition task.

5.3 Performance Evaluation on eNTERFACE'05 and RML

In this experiment, the performance of the proposed audio-visual emotion recognition system is evaluated using the

TABLE 3
Recognition Accuracy for CK+ Database

Method	Recognition rate (%)
SNMF [54]	91.60
MFA [55]	92.56
GSNMF [54]	93.50
NGE [56]	94.18
DSNGE [57]	94.82
Proposed visual path	96.11

TABLE 4
Recognition Accuracy for ENTERFACE'05 Database

Emotion recognition system	Recognition rate (%)
Wang et al. [20]	72.47
Dobrisek et al. [58]	77.50
Zhalehpour et al. [59]	78.26
Poria et al. [60]	85.23
Proposed system	86.67

eNTERFACE'05 and RML databases. The eNTERFACE'05 is an audio-visual emotion database which contains 1170 utterances. The utterances were produced by 43 subjects from 14 different nations. The emotions included in this database are happy, angry, disgust, sad, surprise and fear. The database contains video samples from the 43 subjects, expressing the six basic emotions, with a sampling rate of 48,000 Hz and a frame rate of 25 fps. The image frames have a size of 720×576 pixels, with the average size of the face region around 260×300 pixels. We used the same experimental settings as that used in [20]. The experiments used 1279 samples from the 43 subjects. For cross-validation evaluation, the samples were randomly selected from ten subjects, and then 75 percent of the samples were used for training and the remaining 25 percent were used for testing. The entire experiment was repeated ten times and the mean of the ten recognition rates were recorded. The experiments used some parameters set as follows: k_{row} (for visual) $= 11$, k_{col} (for visual) $= 4$, k_{row} (for audio) $= 11$, k_{col} (for audio) $= 5$, and $r = s = 5$.

The recognition accuracy for the eNTERFACE'05 database is shown in Table 4 together with the comparison results for other works. As shown in Table 4, the proposed system achieved an average recognition rate of 86.67 percent and performed better than other methods such as Wang et al. [20], Dobrisek et al. [58], Zhalehpour et al. [59] and Poria et al. [60]. In [60], the authors reported another result where they combined the text modality as well (i.e., audio+video+text) which gave a rate of 87.95 percent. We note that our work performed only slightly lower than the method proposed by Poria et al. (audio+video+text) (86.67 versus 87.95 percent), even though their method used the text modality as well. We also compared our results with the work by Xie and Guan [62]. In this work, the authors presented their results in the form of graphs/charts and numerical results were not given. Estimating the values from their graphical results (Fig. 5 in their paper), we note that our work gave a higher rate than theirs.

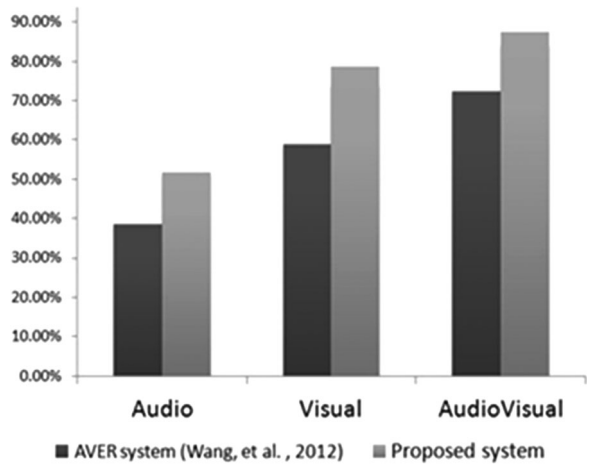


Fig. 5. Comparisons of different modalities for ENTERFACE'05 database.

The recognition rates for the visual and audio paths in our proposed system were 79 and 51.7 percent respectively. For further investigation of the audio path, experiments were performed where the audio path was treated using a six-class statistical machine learning approach (RBF classifier) similar to the visual path (i.e., BDPCA+LSLDA+RBF) without using the rule-base in Path A1. The six audio emotions were considered as six individual classes to be recognized by the RBF classifier. Using this approach, the audio path gave a reduced recognition rate of 46.7 percent. This showed the validity in using the rule-based approach for discriminating amongst the three emotion groups and then using the machine learning approach to recognize the individual emotion within the discriminated two-class group, compared to using the machine learning approach for discriminating the six emotions directly. The optimal feature set to be used depended on the emotions to be separated, and that using one global feature set for the discrimination of all six emotions was suboptimal. The rule-base provided for the discrimination of these emotion classes.

The RML is an audio visual emotion database which contains 720 videos from eight subjects for the six emotions. The audio sampling rate is 22.050 kHz, and the visual data image is in the dimensions of 720×480 pixels. We have chosen 400 video samples to be used in this experiment. Each video sample is truncated to a length of two seconds. Similar settings for the cross-validation used in the eNTERFACE'05 experiment were used. The recognition accuracy for the RML database is shown in Table 5 together with the comparison results for other works. There are fewer results reported for this database compared with eNTERFACE. As shown in Table 5, the proposed system achieved an average recognition rate of 90.83 percent and outperformed the methods by Wang et al. [20] and the deep networks approach by [61]. The recognition rates for the visual and audio paths were 83 and 65.8 percent respectively. When the audio path was treated using the six-class statistical

TABLE 5
Recognition Accuracy for RML Database

Emotion recognition system	Recognition rate (%)
Deep networks [61]	79.72
Wang et al. [20]	82.22
Proposed system	90.83

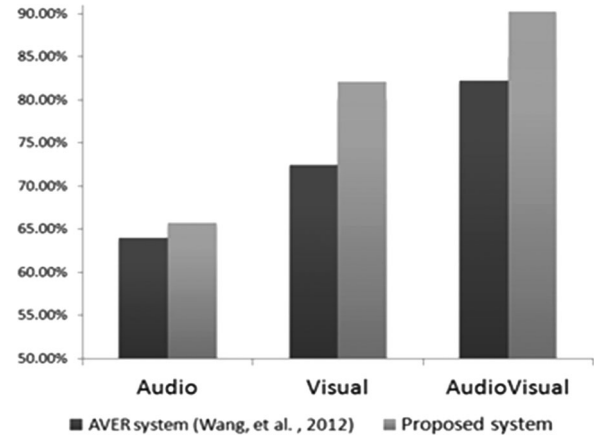


Fig. 6. Comparisons of different modalities for RML database.

machine learning RBF classifier without using the rule-base, the audio recognition rate reduced to 54.2 percent.

For further analysis, experiments to evaluate the effectiveness of the separate audio modality, visual modality, and combined audio-visual modalities were performed and compared to the work by Wang et al. [20] for the eNTERFACE and RML databases as shown in Fig. 5 and 6 respectively. This is one of the few reported works where the authors have reported the individual recognition rates for the audio and visual modalities as well as the fused audiovisual modalities for both the eNTERFACE and RML databases. The results presented in [20] for the individual audio and visual modalities are in graphical form and we have estimated their results from their diagrams. These results further verified the good performance of the proposed system.

6 Conclusion

We have proposed a combined rule-based and machine learning approach to solve the audio-visual emotion recognition problem. The main challenges were to determine the suitable feature extraction techniques and optimal data fusion techniques prior to classification. To address these challenges, we proposed the BDPCA+LSLDA for feature extraction (dimensionality reduction and class discrimination). For classification, we proposed an optimal data fusion technique for training an RBF neural classifier to fuse the kernel and Laplacian matrices for the visual path, and a combined rule-based mechanism and two-class RBF classifiers for the audio path. Numerical results were presented to verify the effectiveness of our proposal. Moreover, the proposed system is also suitable to recognize human emotion from a video stream such as for video conferencing applications as briefly described in Section 4. Our future work will look at practical problems associated with this application such as the length of the window segment and percentage of overlap for the audio path, and the number of frames within a block for the visual path. The work can also be applied towards assessing customer satisfaction scores for a customer relationship management system over video conferencing or video chat where the emotion components contained within the audio and visual data can be used as the representations for the customer satisfaction attributes.

Acknowledgments

The authors would like to thank the anonymous reviewers and editors for their constructive feedback and helpful suggestions.

References

[1] A. Tawari and M. M. Trivedi, "Audio visual cues in driver affect characterization: Issues and challenges in developing robust approaches," in *Proc. Int. Joint Conf. Neural Netw.*, 2011, pp. 2997–3002.

[2] E. Hudlicka and J. Broekens, "Foundations for modelling emotions in game characters: Modelling emotion effects on cognition," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interaction Workshops*, 2009, pp. 1–6.

[3] S. Tokuno, et al., "Usage of emotion recognition in military health care," in *Proc. Defense Sci. Res. Conf. Expo*, 2011, pp. 1–5.

[4] D. Morrison, R. Wang, L. C. De Silva, and W. L. Xu, "Real-time spoken affect classification and its application in call-centres," in *Proc. 3rd Int. Conf. Inform. Tech. Appl.*, 2005, pp. 483–487.

[5] I. Luengo, E. Navas, I. Hernaez, and J. Sanchez, "Automatic emotion recognition using prosodic parameters," in *Proc. INTERSPEECH*, 2005, pp. 493–496.

[6] S. G. Koolagudi, N. Kumar, and K. S. Rao, "Speech emotion recognition using segmental level prosodic analysis," in *Proc. Int. Conf. Devices Commun.*, 2011, pp. 1–5.

[7] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[8] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies," *APSIPA Trans. Signal Inf. Process.*, vol. 3, pp. 1–18, 2014.

[9] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Commun.*, vol. 52, pp. 613–625, 2010.

[10] J. Pribil and A. Pribilova, "Comparison of complementary spectral features of emotional speech for German, Czech, and Slovak," in *Proc. 2011 Int. Conf. Cogn. Behavioural Syst.*, 2012, pp. 236–250.

[11] K. S. Rao and S. G. Koolagudi, *Robust Emotion Recognition Using Spectral and Prosodic Features, SpringerBriefs in Electrical and Computer Engineering*. Berlin, Germany: Springer Science & Business Media, 2013.

[12] A. Tawari and M. M. Trivedi, "Face expression recognition by cross modal data association," *IEEE. Trans. Multimedia*, vol. 15, no. 7, pp. 1543–1552, Nov. 2013.

[13] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, pp. 803–816, 2009.

[14] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," *Comput. Vis. Image Understanding*, vol. 115, pp. 541–558, 2011.

[15] C.-Y. Chen, Y.-K. Huang, and P. Cook, "Visual/acoustic emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2005, pp. 1468–1471.

[16] L. C. De Silva and P. C. Ng, "Bimodal emotion recognition," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recogn.*, 2000, pp. 332–335.

[17] L. S. Chen, H. Tao, T. S. Huang, T. Miyasato, and R. Nakatsu, "Emotion recognition from audiovisual information," in *Proc. IEEE 2nd Workshop Multimedia Signal Process.*, 1998, pp. 83–88.

[18] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, vol. 2, 2005, pp. 1085–1088.

[19] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 936–946, Jun. 2008.

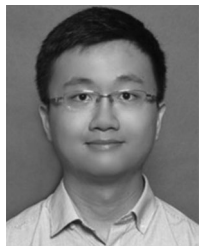
[20] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 597–607, Jun. 2012.

[21] S. W. Chin, K. P. Seng, and L. M. Ang, "Lips contour detection and tracking using watershed region-based active contour model and modified H∞", *IEEE Trans. Circuits Syst. Video Tech.*, vol. 22, no. 6, pp. 869–874, Jun. 2012.

[22] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, pp. 787–800, 2007.
[23] M. Lugger and B. Yang, "Psychological motivated multi-stage emotion classification exploiting voice quality features," *Speech Recognition, Technologies and Applications*, F. Mihelic and J. Zibert, Eds., I-Tech, Vienna, Austria, Nov. 2008, pp. 395–410.
[24] M. Lugger and B. Yang, "An incremental analysis of different feature groups in speaker independent emotion recognition," in *Proc. 16th Int. Congr. Phonetic Sci.*, 2007, pp. 2149–2152.
[25] H. Schlosberg, "Three dimensions of emotion," *Psychological Rev.*, vol. PR-61, no. 2, pp. 81–88, 1954.
[26] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comp. Vis.*, vol. 57, pp. 137–154, 2004.
[27] W. Zuo, D. Zhang, J. Yang, and K. Wang, "BDPCA plus LDA: A novel fast feature extraction technique for face recognition," *IEEE Trans. Syst. Man Cybern. Part B: Cybern.*, vol. 36, no. 4, pp. 946–953, Aug. 2006.
[28] Y.-R. Yeh and Y.-C. F. Wang, "A rank-one update method for least squares linear discriminant analysis with concept drift," *Pattern Recogn.*, vol. 46, pp. 1267–1276, 2013.
[29] S. Yu, L. C. Tranchevent, W. Glanzel, J.A. Suykens, B. De Moor, and Y. Moreau, "Optimized data fusion for K-means Laplacian clustering," *Bioinformatics*, vol. 27, pp. 118–126, 2011.
[30] S. Yu, L. C. Tranchevent, B. Moor, and Y. Moreau, "Rayleigh quotient-type problems in machine learning," in *Proc. Kernel-Based Data Fusion Mach. Learn.*, vol. 345, 2011, pp. 27–37.
[31] X. Yang, B. Tan, J. Ding, J. Zhang, and J. Gong, "Comparative study on voice activity detection algorithm," in *Proc. Int. Conf. Electric. Control Eng.*, 2010, pp. 599–602.
[32] C. S. Ooi, K.P. Seng, L. M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," *Expert Syst. Appl.*, vol. 41, pp. 5858–5869, 2014.
[33] Z. Xufang, D. O'Shaughnessy, and M.-Q. Nguyen, "A processing method for pitch smoothing based on autocorrelation and cepstral F0 detection approaches," in *Proc. Int. Symp. Signals Syst. Electron.*, 2007, pp. 59–62.
[34] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Detection of stress and emotion in speech using traditional and FFT based log energy features," in *Proc. Joint Conf. 4th Int. Conf. Inf. Commun. Signal Process.*, vol. 3, 2003, pp. 1619–1623.
[35] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *Proc. Amer. Soc. Eng. Edu.*, 2008, pp. 1–7.
[36] D. A. Cairns, J. H. L. Hansen, and J. F. Kaiser, "Recent advances in hypernasal speech detection using the nonlinear teager energy operator," in *Proc. 4th Int. Conf. Spoken Lang*, 1996, pp. 780–783.
[37] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for MFCC feature extraction," in *Proc. 4th Int. Conf. Signal Process. Commun. Syst.*, 2010, pp. 1–5.
[38] G. Chetty, M. Wagner, and R. Goecke, "A multilevel fusion approach for audiovisual emotion recognition," *Emotion Recognition: A Pattern Analysis Approach*, A. Konar and A. Chakraborty, Eds., John Wiley & Sons, 2015, pp. 437–460.
[39] Y. W. Wong, K. P. Seng, and L. M. Ang, "Audio-visual recognition system in compression domain," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 21, no. 5, pp. 637–646, May 2011.
[40] *Our Database of Faces*, AT&T Laboratories, Cambridge, U.K., 1992.
[41] Yale University Face Database, (1997). [Online]. Available: http://cvc.yale.edu/projects/yalefaces/yalefaces.html
[42] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recogn. Workshops*, 2010, pp. 94–101.
[43] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The ENTERFACE'05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops*, 2006, Art. no. 8.
[44] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106–2112, Nov. 2010.
[45] H. Zhang, Q. M. J. Wu, T. W. S. Chow, and M. Zhao, "A two-dimensional neighborhood preserving projection for appearance-based face recognition," *Pattern Recogn.*, vol. 45, no. 5, pp. 1866–1876, 2012.
[46] L. Ming and B. Yuan, "2D-LDA: A statistical linear discriminant analysis for image matrix," *Pattern Recogn. Lett.*, vol. 26, no. 5, pp. 527–532, 2005.
[47] Y. Wen, "An improved discriminative common vectors and support vector machine based face recognition approach," *Expert Syst. Appl.*, vol. 39, no. 4, pp. 4628–4632, 2012.
[48] S. Noushath, G. H. Kumar, and P. Shivakumara, "$(2D)^2$LDA: An efficient approach for face recognition," *Pattern Recogn.*, vol. 39, no. 7, pp. 1396–1400, 2006.
[49] V. H. Gaidhane, Y. V. Hote, and V. Singh, "An efficient approach for face recognition based on common eigenvalues," *Pattern Recogn.*, vol. 47, pp. 1869–1879, 2014.
[50] W. Yu, X. Teng, and C. Liu, "Face recognition using discriminant locality preserving projections," *Image Vis. Comput.*, vol. 24, pp. 239–248, 2006.
[51] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.
[52] J. Yang and J. Y. Yang, "Why can LDA be performed in PCA transformed space," *Pattern Recog.*, vol. 36, no. 2, pp. 563–566, 2003.
[53] Q. Cheng, H. Zhou, J. Cheng, and H. Li, "A minimax framework for classification with applications to images and high dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2117–2130, Nov. 2014.
[54] Z. Ruicong, M. Flierl, R. Qiuqi, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Trans. Syst. Man Cybern. Part B: Cybern.*, vol. 41, no. 1, pp. 38–52, Feb. 2011.
[55] S. Yan, D. Xu, B. Zhang, Q. Yang, H.-J. Zhang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[56] J. Yang, S. Yan, Y. Fu, X. Li, and T. Huang, "Non-negative graph embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2008, pp. 1–8.
[57] H.-W. Kung, Y.-H. Tu, and C.-T. Hsu, "Dual subspace nonnegative graph embedding for identity-independent expression recognition," *IEEE Trans. Inform. Forensics Sec.*, vol. 10, no. 3, pp. 626–639, Mar. 2015.
[58] S. Dobrisek, R. Gajsek, F. Mihelic, N. Pavesic, and V. Struc, "Towards efficient multi-modal emotion recognition," *Int. J. Adv. Robot. Syst.*, vol. 10, no. 53, pp. 1–10, 2013.
[59] S. Zhalehpour, Z. Akhtar, and C. E. Erdem, "Multimodal emotion recognition based on peak frame selection from video," *Signal Image Video Process.*, vol. 10, pp. 1–8, 2015.
[60] S. Poria, E. Cambria, A. Hussain, and G.-B. Huang, "Towards an intelligent framework for multimodal affective data analysis," *Neural Netw.*, vol. 63, pp. 104–116, 2015.
[61] C. Fadil, R. Alvarez, C. Martinez, J. Goddard, and H. Rufiner, "Multimodal emotion recognition using deep networks," in *Proc. Latin Amer. Cong. Biomed. Eng*, 2014, pp. 813–816.
[62] Z. Xie and L. Guan, "Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2013, pp. 1–6.
[63] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

Kah Phooi Seng received the BEng and PhD degrees from the University of Tasmania in Australia. She is currently the adjunct professor with the School of Computing and Mathematics, Charles Sturt University. Before returning to Australia, she was a professor and department head of computer science & networked system at Sunway University. Before joining Sunway University, she was an associate professor with the School of Electrical and Electronic Engineering, Nottingham University. Her research interests include intelligent visual processing, multimodal signal processing, artificial intelligence, multimedia wireless sensor network, affective computing, the development of intelligent system, and multimodal Big Data analytics. She has published more than 230 papers in journals and international refereed conferences. She is a member of the IEEE.

Li-Minn Ang received the BEng and PhD degrees from Edith Cowan University in Australia. He is currently a senior lecturer in computing with Charles Sturt University, Australia and was previously an associate professor with Nottingham Malaysia. His research interests include visual information processing, embedded systems and wireless sensor networks, reconfigurable computing, the development of real-world computer systems, large-scale data gathering in wireless multimedia sensor systems, big data analytics for sensor networks, and multimedia Internet-of-Things. He has published more than 100 papers in journals and international refereed conferences, and is the author of the book *Wireless Multimedia Sensor Networks on Reconfigurable Hardware*. He is a senior member of the IEEE and a fellow of the Higher Education Academy (United Kingdom).



Chien Shing Ooi received the BEng. degree (1st class) from the University of Nottingham and the MSc degree in computer science (research) from Lancaster University. He is currently the software engineer in Espressoft Systems. His research interests include image processing, audio signal processing and emotion recognition.