# Video Analytics for Customer Emotion and Satisfaction at Contact Centers

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

***Abstract*—Due to the high levels of competition in a global market, companies have put more emphasis on building strong customer relationships and increasing customer satisfaction levels. With technological improvements in information and communication technologies, a highly anticipated key contributor to improve the customer experience and satisfaction in service episodes is through the application of video analytics, such as to evaluate the customer's emotions over the full service cycle. Currently, emotion recognition from video is a challenging research area. One of the most effective solutions to address this challenge is to utilize both the audio and visual components as two sources contained in the video data to make an overall assessment of the emotion. The combined use of audio and visual data sources presents additional challenges, such as determining the optimal data fusion technique prior to classification. In this paper, we propose an audio–visual emotion recognition system to detect the universal six emotions (happy, angry, sad, disgust, surprise, and fear) from video data. The detected customer emotions are then mapped and translated to give customer satisfaction scores. The proposed customer satisfaction video analytics system can operate over video conferencing or video chat. The effectiveness of our proposal is verified through numerical results.**

***Index Terms*—Customer experience, customer satisfaction, emotion recognition, radial basis function (RBF), video analytics.**

## I. Introduction

OVER the past few decades, there has been an increasing emphasis on a company's ability to produce high-quality products or services. The development of products or services that meet or exceed the customer expectations is the key to the success of the business in an increasingly competitive and global market. The building of strong customer relationships and the understanding of customer needs or customer satisfaction attributes is essential for product or service success [1], and is heavily dependent on the associated customer satisfaction and experience levels [62], [63]. Although the relation between customer satisfaction and customer retention is not yet fully conclusive, some research works have shown that customer satisfaction does have a direct influence on customer retention [2] and loyalty [3], [4], and the competitiveness of a company. To place emphasis on the importance of customer relationships in business, organizations are increasingly using advanced information and communication technologies (ICT) to improve the customer service experience [64]. To achieve this objective, an organization can develop or employ a Customer Relationship Management system (CRM), which is a system for managing a company's interactions with its current and future customers.

The work by Richards and Jones [65] gives a definition of CRM as "a set of business activities supported by both technology and processes that is directed by strategy and is designed to improve business performance in an area of customer management." The CRM often involves using technology to organize, automate, and synchronize a company's sales, marketing, customer service, technical support operations, etc. A contact center or also known as customer interaction center is a central point of any organization from which all customer contacts are managed and is generally a part of an organization's CRM. The customer's satisfaction evaluated over the full service cycle is one of the most important parameters when judging the overall call center's quality [5], [6]. Contact center management is a new challenge for customer interaction services that should be integrated into a CRM strategy. Modernizing the CRM with new technologies to perform customer experience and satisfaction analytics is also highly desirable for contact centers.

Two current forms of analytics commonly used in contact centers for understanding customer experience are speech-based analytics [7], [8] and text-based analytics [9], [10]. Contact or call centers use audio or speech analytics to analyze thousands of hours of recorded calls to help improve customer experience, gain insight into customer behavior, and identify potential product or service issues [66]. Speech-based analytics handles phone conversations, whereas text-based analytics handles written forms of communications, including text messages, email, chat sessions, blog posts, web forums, review sites, and social media. Emotion analytics is the technology that is used to identify how a customer perceives a product, the presentation of a product, or an interaction with a company representative. It can either be a stand-alone application or built on top of a CRM system, using audio mining techniques and a correlation engine to match a caller's words with emotions. To provide more insight into the caller's emotional state, the caller's tone of voice and how often a particular word is repeated are monitored. Dashboards that indicate the emotions of both the caller and the call center agent allow supervisors to know which calls are going smoothly and which ones may require an intervention.

Some examples of using speech-based analytics for analyzing customer emotion experiences from recorded call center

Manuscript received February 25, 2016; revised December 27, 2016; accepted February 18, 2017. Date of publication May 2, 2017; date of current version May 15, 2018. This paper was recommended by Associate Editor H. Yu. *(Corresponding author: Kah Phooi Seng.)*

The authors are with the School of Computing and Mathematics, Charles Sturt University, Sydney, NSW 2127, Australia (e-mail: jasmine.seng@gmail.com; lang@csu.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2017.2695613

dialogues can be found in the research works presented in [67]–[69]. An early work by Lee *et al*. [67] studied data from users engaged in a spoken dialogue with a machine agent over the telephone for a call center application to distinguish utterances with negative emotions from utterances with nonnegative emotions. The authors employed pitch and energy features from the acoustic signal with principal component analysis (PCA) and a linear discriminant classifier and achieved an accuracy of 80% and 75% for male and female utterances, respectively. Another work by Vidrascu and Devillers [68] used a 10-h dialogue corpus in French recorded in a French Medical emergency call center and obtained a correct detection rate of about 82% between negative and positive emotions using paralinguistic cues and compared several classification methods including support vector machines (SVM) and decision trees. More recently, Pappas *et al.* [69] proposed a method to classify speech windows into either containing the anger emotion or not. Their technique does not require speech recognition, or utterance segmentation, and can thus be applied toward real-world recordings from call centers.

Recently, video-based calls (e.g., Skype Video) have become more popular as a mode of communication among people. As the technology matures and becomes more mainstream, the interactive communication capabilities of contact centers are anticipated to extend from text and audio or speech communications to video-based communications. A highly anticipated key contributor for improving the customer experience for service episodes in contact centers is through the application of video analytics to evaluate the customer's emotions over the full service cycle. The processing of video-based data for analytics requires new technical challenges to be overcome. Video-based data contains two data components or sources: audio-based data and visual-based data. The use of two forms of data sources presents additional challenges such as determining the optimal data fusion technique prior to classification.

In this paper, a new intelligent customer satisfaction video analytics system is proposed for use in contact centers. The system consists of an intelligent audio–visual emotion recognition (I-AVER) module followed by a customer satisfaction analytics and monitoring module. The audio–visual recognition module is designed to recognize the universal six emotions that are happy, angry, sad, disgust, surprise, and fear. The remainder of this paper is organized as follows. Section II gives an overview of analytics in call centers and reviews some related works on audio–visual emotion recognition technologies. Section III presents the proposed audio–visual emotion recognition system and gives details for the audio and visual paths, followed by the fusion of these paths. Experiments on the audio–visual emotion recognition system are presented in Section IV. This section also includes some analysis of the proposed system toward noisy and poor quality visual data, which may occur in contact centers. The transformation from the audio and visual-based emotions to the customer satisfaction levels is presented in Section V. Finally, some conclusions are given in Section VI.

## II. Background, Related Work, and Motivation

Although online communications have been gaining in popularity, the telephone continues to be a useful point of customer contact for companies. A call (telephony-based) center, which is a traditional contact center, is a centralized office used to handle a large volume of calls. A call center utilizes several very mature technologies such as automatic call distributors and interactive voice response (IVR) systems. The interaction may be managed by the IVR system or may require the intervention of a human agent. Other technologies integrated into the call center may include speech recognition, voice recognition, speech analytics, and many other technologies to improve agent productivity and customer satisfaction. The customer's satisfaction evaluated over the full service cycle is one of the most important parameters when judging the overall call center's quality [11], [12]. Today, because customers contact companies by various means such as telephone, e-mail, online chat, fax, social media, websites, instant messaging, etc., many companies refer to their call center as a contact center (or a multichannel contact center). The contact center management is a new challenge for customer interaction services. Modernizing with new technologies such as emotion detection and recognition is highly desirable for contact centers.

Fig. 1 shows the hype cycle for contact center infrastructure from Gartner [13] and some existing and emerging technologies. The figure shows that technologies related to the routing and prioritization of customer interactions, such as IP-based and speech recognition for contact center applications, and technologies to optimize certain aspects of contact center operations (e.g., contact center quality management and contact center workforce management) are mature. Similarly, some operational models, including presence-based contact routing, contact center as a service and hosted contact centers, are still emerging and expected to mature in two to five years' time. Technologies targeting newer customer contact channels and activities, such as video chat for customer service, proactive communications applications and services, CSS suites, and wireless devices in contact centers, are still gaining in maturity.

The video chat for customer service involves a one-way or two-way live video streaming between a customer and an agent and will be broadly available over mobile devices, on websites, and at kiosks to assist with a wide range of capabilities. The video chat session can either stream one way from the agent to the customer or could stream both ways, should the customer have a camera enabled. Some advanced analytic capabilities (such as audio mining and speech analytics, contact center interaction analytics, and emotion detection and recognition), as well as advanced omnichannel integration capabilities found in customer engagement hubs, hold promise for improved levels of customer service but are expected to take up to ten years to reach mainstream adoption.

The remaining review of work in this section will discuss some emerging technology and recent research works of emotion detection and recognition using audio-based, visual-based and audio–visual data. Initial efforts of researchers focused on recognizing human emotions using the audio modality from the emotional states enclosed in the speech signals. Some examples of these are the audio-based emotion recognition efforts by Sobol-Shikler and Robinson [14] and Wu and Liang [15]. The next efforts emerged with the growing popularity of image processing and researchers began to explore the visual

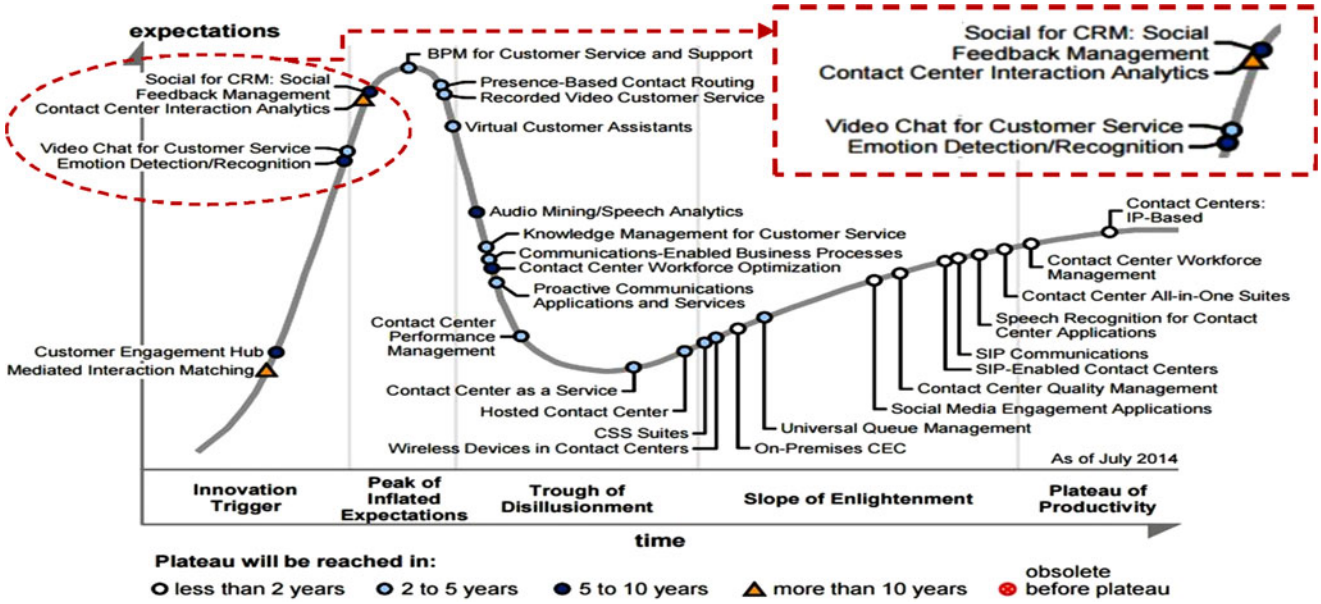


Fig. 1. Contact center infrastructure cycle from Gartner [13].

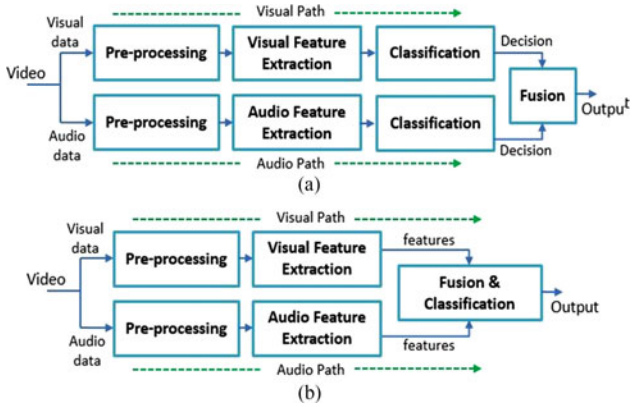


Fig. 2. Two generic structures of audio–visual recognition systems.

modality for emotion recognition (see [16], [17]) based on extracting and recognizing the emotional states enclosed in human facial expressions. The audio and visual modalities were then combined and researchers found that emotion (Song *et al.* [18]) and speech (Chin *et al.* [19]) recognition systems using the integration of these two modalities can give better performance. This is because emotion is generally expressed through several modalities in human to human interaction. Although many researchers agree that the fusion of these two modalities can significantly improve the recognition accuracy, the challenge is that there is still uncertainty on how the integration of these modalities toward emotion recognition can be best accomplished [20]. Fig. 2 illustrates two generic structures of audio–visual recognition systems that are used in emotion recognition.

Several researchers have proposed techniques for the audio–visual emotion recognition problem. Cheng-Yao *et al.* [21], De Silva and Chi [22], and Chen *et al.* [23] proposed their own recognizers, which could classify six emotions and achieved rates of 84%, 72%, and 97.2% of recognition accuracy, respectively. In another work, the researchers in [24] proposed a subject-dependent system and could classify three emotions, although a high 90.7% recognition accuracy could be obtained. Zeng *et al.* [25] worked on person-dependent audio–visual emotion recognition to classify positive and negative emotions from 20 subjects and obtained an average accuracy of 89%. Another effort by Zeng *et al.* [26] showed that their audio–visual emotion recognition system gave a rate of 72.42% of accuracy in person-independent experiments and a rate of 96.3% in person-dependent experiments, on a video database that consists of 20 subjects with more emotions. The common weaknesses of these works is that they are either person-dependent, utilizing a testing database with an insufficient number of subjects, or with an insufficient number of emotions.

A research effort by Wang and Guan [27], proposed a better audio–visual emotion recognizer that could achieve a high accuracy rate of 82.14% on their own six classes database. However, they commented 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 later effort by the same researchers [28] proposed a better version of their bimodal emotion recognition system that utilized kernel cross-modal factor analysis. Their system achieved around 72.47% to 82.22% rates of accuracy when they evaluated their work using two standard audio–visual emotion databases. It is also important to note that all the single modal or multimodal emotion recognition research works reviewed up to this stage emphasize only using the attribute-based information about the data objects. In real-world applications, the relation (graph-based) information should also be taken into consideration to improve the recognition accuracy.

## III. Intelligent Audio–Visual Emotion Recognition System Module

Fig. 3 shows an overview of the proposed I-AVER system module, which is slightly different from the generic structures

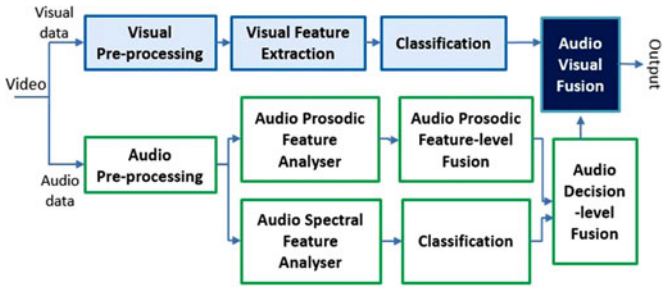


Fig. 3. Overview of the intelligent AVER system module.

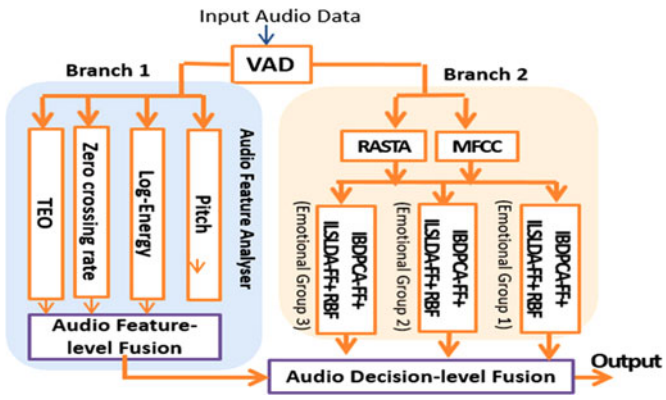


Fig. 4. Audio path module of I-AVER.

shown in Fig. 2. The module also consists of two paths (audio path and visual path) to extract the emotion information from the audio-based data and visual-based data contained in video. The visual path consists of the blocks of preprocessing, feature extraction, and classification. On the other hand, the audio path has two parallel sub-branches after the audio preprocessing. The first sub-branch consists of an audio prosodic feature analyzer followed by the feature-level fusion. The second sub-branch processes the audio spectral features followed by the classification. Both audio sub-branches are merged in the audio decision-level fusion block. The final block called the audio–visual fusion module is used to dynamically fuse the outputs from both the audio and visual paths. This allows the I-AVER system to operate in a dynamic mode to perform emotion recognition over video conferencing. Details of each block in Fig. 3 will be discussed in next sections.

### *A. Audio Path of Intelligent Audio–Visual Emotion Recognition*

Fig. 4 shows the architecture for the audio path for the I-AVER. The input signal is first preprocessed by the Voice Activity Detector (VAD) [35]. Then, the preprocessed signal will be passed to Branch 1 and Branch 2 in parallel. Branch 1 consists of the audio prosodic feature analyzer and audio feature-level fusion. In Branch 1, we use the method proposed in [34] where the audio prosodic features such as zero-crossing rate (ZCR), teager energy operator (TEO), pitch, and log-energy are further utilized. The prosodic features are used to form some empirical rules to select the specific classifier in Branch 2, and the output of the classifier is taken to be the recognized audio emotion. Branch 2 consists of the spectral feature extraction to extract the MFCC [30] and RASTA [31] spectral features. This is followed by three IBDPCA-FF+ILSLDA-FF+RBF sub-branches with two-class radial basis function (RBF) neural classifiers for three sets of emotion groups: {angry, happy}, {sad, disgust}, and {surprise, fear}. Each RBF classifier is designed to also output a reject class [32], [33]. During operation, the three sets of classifiers are input with the same features from the input speech signal. Two scenarios are possible: 1) if only one classifier responds with a valid class and the remaining two classifiers respond with the reject class; and 2) any other combination. For scenario 1), the output of the valid classifier is taken to be the recognized audio emotion. For scenario 2), an additional stage would be required to select from the three classifiers. In the design, the outcome from Branch 1 is only used if the outcome from Branch 2 is indeterminate [i.e., scenario 2)].

*1) Preprocessing:* During preprocessing, the VAD technique using the short-time zero crossing rate (STZCR) and short-time energy (STE) features [35] is used to detect the speech segments in the audio signal. The speech signal $x(m)$ is divided into $n$ number of frames. The STE and STZCR are calculated and compared to determine the presence or absence of speech, and the silence or unvoiced segments are removed. The STE calculation gives the energy in the frame, whereas the calculated STZCR gives a measure of the weighted average of the number of times the speech signal changes signs within a specific time window.

*2) Emotion Groups Feature Extraction and Classification:* The feature extraction flow begins with extracting two types of spectral features from the speech: MFCC and RASTA. The MFCCs are the cepstral coefficients derived from a mel-scale frequency filter bank. The MFCC feature has been identified to be one of the most influential audio features. The details for the MFCC computations can be found in [30]. The relative spectral (RASTA) [31] features suppress the spectral components that change more slowly or quickly than the typical range of speech changes. The work in [36] showed the effectiveness of using RASTA features based on perceptual linear prediction for audio-based emotion recognition. The extracted MFCC and RASTA spectral features are concatenated and passed to three parallel sub-branches for feature extraction and classification. The details of the feature extraction techniques (IBDPCA-FF and ILSLDA-FF) and RBF classification will be discussed later in Section III-B during the description of the visual path.

In our proposed approach, each audio sub-branch deals with two emotion classes and each RBF neural network only has to perform a two-class classification. The output of each RBF network includes the reject class [32], [33]. RBF networks have the ability to reject input patterns from novel classes not in the training data. Outputs from the RBF network that exceed a threshold value are classed as rejections. The six emotions are divided into three pairwise groups. As a result, three two-class emotion groupings (EGs) are formed, which are the EG1 {angry, happy}, EG2 {sad, disgust}, and EG3 {surprise, fear}. For scenario 1), if only one classifier responds with a valid class

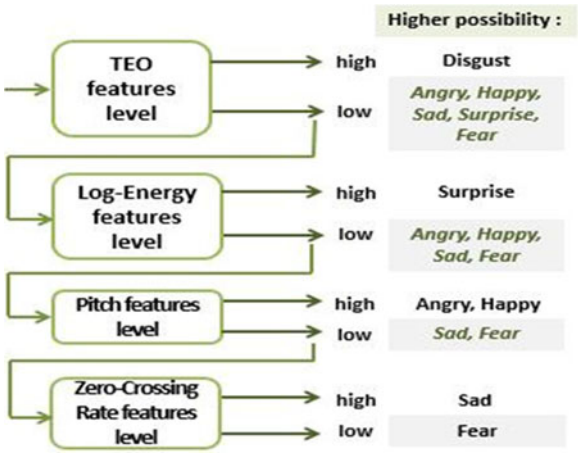


Fig. 5. Flowchart showing mechanism to assign weightings for EGs.


Fig. 6. Audio decision-level fusion module.

and the remaining two classifiers respond with the reject class, the output of the valid classifier is taken to be the recognized audio emotion and the audio path is finished.

*3) Audio Decision Level Fusion:* For scenario 2), Branch 1 is further taken to select from the three classifiers. Four audio prosodic features (ZCR, TEO, pitch, and log-energy) are calculated and used in the audio feature-level fusion module. The pitch analyzer divides the speech signal $x(m)$ into $n$ frames using a window $w(n)$ and is denoted as $s(m)$. The following equation shows the calculation of the pitch feature from its periodicity $R$ as:

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

where $k$ denotes the representation of the pitch period of a peak and $L$ is the window length. The log-energy [38] feature indicates the total squared amplitude in a segment of speech and can be formulated as

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

where $x$ denotes the sample of the speech and $N$ refers to the number of frames. The ZCR [39] feature calculates the weighted average of the number of times the speech signal $x(n)$ changes signs within a time window and can be calculated as

$$\operatorname{sgn}\{x\} = \begin{cases} 1, & \text{if} \quad x(n) \geq 0 \\ -1, & \text{if} \quad x(n) > 0 \end{cases}. \tag{3}$$

Equation (4) calculates the TEO [40] feature that detects the significantly changing nonlinear components between speech segments as

$$\Psi\left[s\left(n\right)\right] = s^2(n) - s(n+1)s(n-1). \tag{4}$$

The extracted values for the four prosodic features are then passed into the audio feature-level fusion module.

This module uses a mechanism to assign weightings for the three EGs as shown by the chart in Fig. 5. These weightings will then be used to influence the RBF neural classification in Branch 2. The TEO feature is first compared. A high TEO value assigns a heavier weight to the EG containing the disgust emotion (i.e., EG2). If the TEO value is lower, a further comparison is made with a log-energy threshold. A high log-energy value assigns a heavier weight to the EG containing the surprise emotion (i.e., EG3). If the log-energy value is lower, another comparison is made with a pitch level threshold. A high pitch value assigns a heavier weight to the EG containing the angry and happy emotions (i.e., EG1). If the pitch value is lower, a final comparison is made with a ZCR threshold. A high ZCR value assigns a higher weight to the EG2 (sad emotion) and a lower ZCR value assigns a higher weight to the EG3 (fear emotion).

The sets of conditions in Fig. 5 influence which of the outputs from the three parallel sub-branches in Branch 2 should be emphasized. The mechanism in Branch 1 works together in conjunction with the RBF neural classifier and the three emotion groups in Branch 2, and form an effective strategy for identifying the correct audio emotion.

The audio decision-level fusion module shown in Fig. 6 makes the final decision based on the combined outputs from Branch 1 and Branch 2. The audio feature-level fusion module in Branch 1 assists the decision making via the weights assignment mechanism to assign three sets of weights $(W_{\text{Group 1}}, W_{\text{Group 2}}, W_{\text{Group 3}})$ to the outputs from the three parallel subpaths in Branch 2. The decision is made by considering the output with the heaviest weight. For example, if $W_{\text{Group1}}$ has the highest weight, then the emotion output of $(\text{IBDPCA} - \text{FF} + \text{ILSLDA} - \text{FF} + \text{RBF})_{\text{Group1}}$ from Branch 1 is selected. During this stage, the output values from the classifiers are used without considering the reject class. The output of the selected two-class classifier is taken to be the recognized audio emotion.

### *B. Visual Path of Intelligent Audio–Visual Emotion Recognition*

Fig. 7 shows the architecture for the visual path for the I-AVER. The preprocessing includes a face detection/segmentation step to locate and segment out the face region. The Viola–Jones algorithm [41] is used because of its good detection rate and high calculation speed. This detector is constructed using an integral image and rectangular features similar to Haar wavelets.

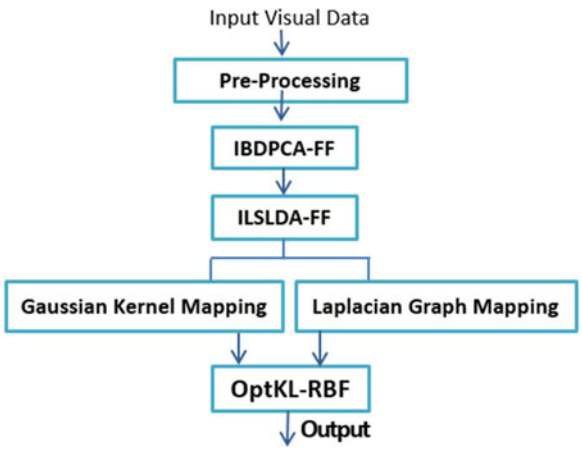


Fig. 7. Visual Path module of I-AVER.

Then, feature extraction techniques called incremental bidirectional principal component analysis with forgetting factor (IBDPCA-FF) and incremental least square linear discriminant analysis with forgetting factor (ILSLDA-FF) are used to extract and discriminate the visual features amongst the different emotion classes [29]. The visual path contains an optimized scheme for data fusion termed optimized kernel-Laplacian radial basis function (OptKL-RBF). OptKL-RBF uses the kernel and Laplacian mappings to calculate the similarities of attribute-based information and relation-based (graph) information from the visual features. The kernels and Laplacians matrices are optimized and merged to form an optimal kernel-Laplacian matrix. The kernel-Laplacian matrix is then used by the RBF neural network to perform the classification.

*1) Visual Feature Extraction Techniques:* Some feature extraction techniques that have been shown to be effective for visual features include PCA and linear discriminant analysis (LDA). The original PCA and LDA methods involve batch learning and have the disadvantage that the complete training set is required during the training process. If additional training data are incorporated in the future, then retraining on the full set of the training data (old and additional data samples) are needed and this leads to large computational or space requirements. Therefore, incremental learning techniques for feature extraction are needed to address this problem.

A second desirable for emotion recognition is to have the capability of putting more emphasis on recently acquired data or images and less emphasis on earlier observations. We have observed that down-weighting the contribution of earlier observations is important to improve the recognition accuracy. Without this down-weighting, as time progresses, the observation history can become very large to the point of overwhelming the relative contribution of each block of new data and cause the RBF learner to become blind to changes in the observation stream. Thus, a better solution that allows incremental learning and can moderate the balance between old and new observations is highly desirable. In this paper, we propose enhanced versions for PCA and LDA, which can function in the incremental mode and allow for new observations termed IBDPCA-FF and ILSLDA-FF.

The combination of IBDPCA-FF and ILSLDA-FF is proposed to extract and discriminate the visual emotion features amongst the different classes. The IBDPCA-FF is an enhanced version of bidirectional principal component analysis (BDPCA) [44] and it maintains the strengths of the original version. The IBDPCA-FF is based on our earlier work in [29]. The ILSLDA-FF is then used to further enhance the separation of the extracted features amongst the different classes. The ILSLDA-FF is designed based on the LSLDA model [45] and modified to function in the incremental mode. The IBDPCA-FF+ILSLDA-FF can operate in the incremental learning mode and provide weight moderations for both the newly added and the existing training data.

The IBDPCA-FF algorithm consists of two main parts: initialization and incremental learning. The steps of the IBDPCA-FF algorithm are briefly presented as follows.

*a) Initialization:* The initial set of $K$ training samples $X^{(1)} = [X_1, \ldots, X_K]$ and its transposed version $X^{(2)} = [X_1^T, \ldots, X_K^T]$ are first arranged for the initialization stage. These two sets are used for computing the initial subspaces $U_k\mathrm{col}$ and $V_{k\mathrm{row}}$, the singular values for the row and column directions $D^{(1)}$ and $D^{(2)}$, and the mean. The singular value decomposition (SVD) technique is used as

$$[U_{k_{\mathrm{row}}}\ D^{(1)}\ V^{(1)}] = \mathrm{SVD}(X^{(1)}), \quad [V_{k_{\mathrm{col}}}\ D^{(2)}\ V^{(2)}] = \mathrm{SVD}(X^{(2)}). \tag{5}$$

*b) Incremental learning:*

1) The mean updating process updates the mean by taking account of the training data $X_{K+1}$ and the previous mean value $M_A$. In our proposed method, the number of previously trained images $K$ is multiplied by the forgetting factor (where $0 < f < 1.0$ ($f = 1.0$ indicates does not carry any effect to the updates and $f = 0$ indicates complete forgetfulness) to moderate the old data with the new one as

$$M_{A\mathrm{New}} = \frac{fK}{fK+1} M_A + \frac{1}{1+fK} X_{K+1}. \tag{6}$$

2) The centered matrix of the new sample $B$ is then obtained by subtracting the new sample from the old mean as

$$B = X_{K+1} - M_A. \tag{7}$$

3) The orthogonalization method is applied using QR decomposition as follows:

$$[Q_u, R_u] = qr\left(B - U_{k\mathrm{col}} U_{k\mathrm{col}}^{\mathrm{T}} B\right) \tag{8}$$

$$[Q_v, Q_v] = qr\left(B^T - V_{k\mathrm{row}} V_{k\mathrm{row}} B^T\right). \tag{9}$$

4) The newly formed small SVD left singular value can be obtained from

$$\left[\ddot{U}, \ddot{D}_1, \ddot{V}_1\right] = \mathrm{SVD}\begin{pmatrix} fD^{(1)} & {U_{k\mathrm{col}}}^T B \\ 0 & R_u \end{pmatrix} \tag{10}$$

$$\left[\ddot{V}, \ddot{D}_2, \ddot{V}_2\right] = \mathrm{SVD}\begin{pmatrix} fD^{(2)} & {V_{k\mathrm{row}}}^T B^T \\ 0 & R_v \end{pmatrix}. \tag{11}$$

5) The forgetting factor is introduced to integrate into the input value of the compacted singular value to downsize the contributions of the previous data. The Steps 2) to 4) are repeated for each sample in the added set. The final trained subspaces $U_{k\mathrm{col}}$ and $V_{k\mathrm{row}}$ can be formed by using

$$U_{k\mathrm{col}} = [U_{k\mathrm{col}}, Q_u]\ddot{U} \tag{12}$$

$$V_{k\mathrm{row}} = [V_{k\mathrm{row}}, Q_v]\ddot{V}. \tag{13}$$

6) Finally, the newly extracted feature matrix $X_{\mathrm{IBDPCA-FF}}$ can be constructed as

$$X_{\mathrm{IBDPCA-FF}} = U_{k\mathrm{col}}^T X V_{k\mathrm{row}}. \tag{14}$$

The incremental approach for the ILSLDA-FF rearranges the formulations of calculating the LSLDA solution *W* and simplifies the updating computations on the indicator matrix *Y*. The steps of the ILSLDA-FF are briefly presented as follows.

1) The first step is to perform the initial stage based on the required inputs: current training set $X_{\mathrm{IBDPCA-FF}}$ with their corresponding indicator $Y_{\mathrm{IBDPCA-FF}}$, and current mean. The transformation of input image matrix into row vectors is required prior to the computation. Hence, after transforming $X_{\mathrm{IBDPCA-FF}}$ and $Y_{\mathrm{IBDPCA-FF}}$, they are both assumed as the initial training set $X_i = [x_1^T; ...; x_{K-1}^T; x_K^T]$ and its indicators in vectors $Y_i = [y_1^T; ...; y_{K-1}^T; y_K^T]$, respectively.
2) Using these parameters, $T^{-1}$ can be obtained using $(XX^T)^{-1}$, while $T_0$ is initialized as the identity matrix $I_{d\times d}$, and $W_0$ is zero. Another value of the forgetting factor for this algorithm $f_{\mathrm{LDA}}$ is used.
3) The incremental approach is to solve $W_{i+1}$. The new vector for indicator $\mathrm{y_{i+1}}$ can be defined as

$$y_{i+1}(c) = \begin{cases} 1, & \text{if } X_{i+1} \text{ belongs to class } c \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

4) The new indicator matrix, $Y_{i+1}$ can be updated by concatenating $Y_i$ with the obtained $y_{i+1}$. The $T_i^{-1}x_{i+1}$ can be computed as follows:

$$T_i^{-1}x_{i+1} = x_{i+1} + \sum_{k=0}^{i-1} \frac{\left(T_k^{-1}x_{k+1}\right)\left(T_k^{-1}x_{k+1}\right)^T}{1 + x_{i+1}^T T_i^{-1} x_{i+1}}(x_{i+1}). \tag{16}$$

5) This updated $T_i^{-1}x_{i+1}$ can then be used as the parameter to solve the updated LDA solution $W_{i+1}$ as

$$W_{i+1} = W_i + \frac{\left(T_i^{-1}x_{i+1}\right)}{f_{\mathrm{LDA}}^{i+1} + x_{i+1}^T\left(T_i^{-1}x_{i+1}\right)}(y_{i+1}^T - x_{i+1}^T W_i). \tag{17}$$

6) Finally, the obtained solution of $W_{\mathrm{LDA}} = W_{i+1}$ is utilized to project the IBDPCA-FF extracted feature matrix in vector columns to obtain the IBDPCA-FF+ILSLDA-FF extracted feature matrix as

$$X_{\mathrm{IBDPCA-FF\,\&\,ILSLDA-FF}} = W_{\mathrm{LDA}}^T \tilde{X}. \tag{18}$$

The output from the proposed IBDPCA-FF + ILSLDA-FF technique is then passed to a newly designed neural classifier termed the OptKL-RBF scheme, which will be discussed next.

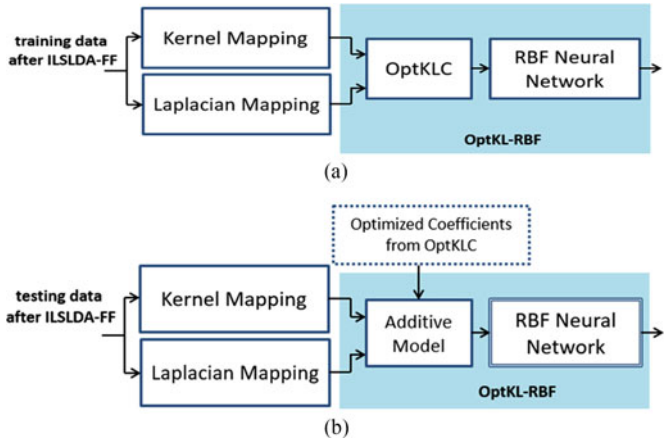


Fig. 8. (a) Training and (b) testing phases of OptKL-RBF classification.

*2) Optimized Kernel Laplacian Radial Basis Function Neural Classification:* Fig. 8(a) and (b) shows the training and testing phases for the proposed OptKL-RBF neural classification. During the training phase, the extracted training facial features or data from the earlier ILSLDA-FF stage are passed to the Gaussian kernel mapping and the Laplacian graph mapping submodules to obtain kernel and Laplacian matrices, respectively, which are then passed to the optimal *k*-means Laplacian clustering (OptKLC) submodule. This submodule is designed based on the OptKLC technique described by Yu *et al.* [46]. The objective of the OptKLC is to find the optimal coefficients for the kernel and Laplacian matrices to form the optimal kernel-Laplacian matrix ($\Omega_{\mathrm{OptKLC}}$). The $\Omega_{\mathrm{OptKLC}}$ will then be used for training the RBF neural network. For the testing phase, the features from the ILSLDA-FF stage are 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 matrices are then merged with the optimal coefficients obtained from OptKLC during the training phase using the additive model [47]. Finally, a kernel-Laplacian matrix is calculated for the test data, and input into the trained RBF network for classification.

The steps of the OptKL-RBF algorithm are briefly presented as follows.

*Initial:* After feature extraction, the IBDPCA-FF and ILSLDA-FF features ($X_{\mathrm{IBDPCA-FF\&ILSLDA-FF}}$) are in a matrix *x* of dimensions $a \times j$, where *a* is the number of samples, and *j* is the number of features. The following equation is then used to calculate the kernel and Laplacian matrices $H_1$ and $H_2$ using kernel mapping and Laplacian mapping as

$$H_1 = \exp\left(-\frac{\|x_a - x_j\|}{2\sigma^2}\right), \quad H_2 = D^{-1/2}\, W D^{-1/2} \tag{19}$$

where *W* is the affinity matrix, *D* is an $a \times a$ diagonal matrix, and $\sigma$ is the width of the kernel as follows:

$$W = w_{aj} = \exp\left(-\frac{\|x_a - x_j\|}{\sigma^2}\right), \quad D = d_a = \sum\nolimits_j w_{aj}. \tag{20}$$

The OptKLC determines the optimal weighted convex linear combinations of kernels and Laplacians using the objective

function $J_{\mathrm{OptKLC}}$ as

$$\underset{A,\beta}{\text{maximum}}\ J_{\mathrm{OptKLC}} = \text{trace}\left(A^T\ \left(\tilde{\boldsymbol{H}}_2 + \boldsymbol{H}_1\right) A\right)$$

where

$$\tilde{\boldsymbol{H}}_2 = \sum_{i=1}^{r} \beta_i \tilde{H}_{2_i},\ \ \boldsymbol{H}_1 = \sum_{j=1}^{s} \beta_{j+r} H_{1_{cj}},$$

$$\sum_{i\,=\,1}^{r} \beta_i^{\delta} = 1 \sum_{j\,=\,1}^{s} \beta_{j+r}^{\delta} = 1$$

$$\beta_l \geq 0,\ l = 1,\ \ldots,\ (r+s),\ A^T A = I_k. \tag{21}$$

$\beta_1, \ldots \beta_r$ give the optimal Laplacian coefficients and $\beta_{r+1}, \ldots \beta_{r+s}$ give the optimal kernels coefficients. The combined kernel matrices of multiple $H_{1_{cj}}$ are denoted by $\boldsymbol{H}_1$, and the combined Laplacian matrices of multiple $\tilde{H}_{2_i}$ are denoted by $\tilde{\boldsymbol{H}}_2$.

The training phase of the algorithm can be described using the following steps.

*Step 1:* An initial guess of the coefficients for kernels ($\beta_1^{(0)},\ \ldots,\ \beta_r^{(0)}$) and Laplacians ($\beta_{r+1}^{(0)},\ \ldots,\ \beta_{r+s}^{(0)}$) are randomly made. These gives the initial kernel matrix $\boldsymbol{H}_1{}^{(0)}$ and initial Laplacian matrix $\tilde{\boldsymbol{H}}_2^{(0)}$. Equation (22) is used for combination to give the initial kernel-Laplacian matrix $\boldsymbol{\Omega}_{\mathrm{OptKLC}}^{(0)}$.

$$\tilde{\boldsymbol{H}}_{\boldsymbol{2}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{r}} \beta_i^{(0)} \tilde{H}_{2_i}^{(0)},\quad \boldsymbol{H}_{\boldsymbol{1}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{s}} \beta_{j+r}^{(0)} H_{1_{cj}}^{(0)}$$

$$\boldsymbol{\Omega}_{\mathrm{OptKLC}}^{(0)} = \tilde{\boldsymbol{H}}_{\boldsymbol{2}}^{(0)} + \boldsymbol{H}_{\boldsymbol{1}}{}^{(0)}. \tag{22}$$

*Step 2:* Eigenvector decomposition is performed on $\boldsymbol{\Omega}_{\mathrm{OptKLC}}^{(0)}$, and the calculated eigenvectors are sorted in descending order based on its eigenvalues. *k*-means clustering is applied to the resultant eigenvectors to give the weighted cluster membership matrix $A^{(0)}$. The initial iterative index $\gamma$ is 0.

*Step 3:* Equation (23) is used to calculate the cluster assignment $A^{(\gamma)}$ and the affinity matrix $F^{(\gamma)}$.

$$F_{ab} = \begin{cases} +1 & \text{if } A_{ab} > 0, \quad a = 1, \ldots, N, \quad b = 1, \ldots, k \\ -1 & \text{if} A_{ab} = 0, \quad a = 1, \ldots, N, \quad b = 1, \ldots, k \end{cases} \tag{23}$$

where *a* is the sample index, and *b* is the cluster index.

*Step 4:* Equations (24) and (25) are used to calculate the $\gamma$th iteration's Laplacian coefficients ($\beta_1^{(\gamma)},\ \ldots,\ \beta_r^{(\gamma)}$) and kernels coefficients ($\beta_{r+1}^{(\gamma)},\ \ldots,\ \beta_{r+s}^{(\gamma)}$) using the SIP-LSSVM-MKL [17]

$$\beta_1^{(\gamma)},\ \ldots,\ \beta_r^{(\gamma)} \leftarrow \text{SIP} - \text{LSSVM} - \text{MKL} \times \left(\tilde{H}_{2_1},\ \ldots,\ \tilde{H}_{2_r},\ F^{(\gamma)}\right) \tag{24}$$

$$\beta_{r+1}^{(\gamma)},\ \ldots,\ \beta_{r+s}^{(\gamma)} \leftarrow \text{SIP} - \text{LSSVM} - \text{MKL} \left(H_{1_1},\ \ldots, H_{1_{cs}},\ F^{(\gamma)}\right). \tag{25}$$

$\tilde{H}_{2_i}$ is the *i*th number of the Laplacian matrix in $\tilde{\boldsymbol{H}}_2$, and $H_{1_{cj}}$ is the *j*th number of the kernel matrix in $\boldsymbol{H}_1$.

*Step 5:* Equation (26) is used to combine $\boldsymbol{H}_1$ and $\tilde{\boldsymbol{H}}_2$ to give the next kernel-Laplacian matrix $\boldsymbol{\Omega}_{\mathrm{OptKLC}}^{(\gamma+1)}$

$$\tilde{\boldsymbol{H}}_{\boldsymbol{2}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{r}} \beta_i^{(\gamma)} \tilde{H}_{2_i}^{(\gamma)},\quad \boldsymbol{H}_{\boldsymbol{1}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{s}} \beta_{j+r}^{(\gamma)} H_{1_{cj}}^{(\gamma)}$$

$$\boldsymbol{\Omega}_{\mathrm{OptKLC}}^{(\gamma+1)} = \tilde{\boldsymbol{H}}_{\boldsymbol{2}}^{(\gamma)} + \boldsymbol{H}_{\boldsymbol{1}}^{(\gamma)}. \tag{26}$$

*Step 6:* Eigenvector decomposition is performed on $\boldsymbol{\Omega}_{\mathrm{OptKLC}}^{(\gamma+1)}$ to give the next weighted cluster membership matrix, $A^{(\gamma)}$.

*Step 7:* Equation (27) is used to calculate the error of the clustering assignment matrix $\Delta A$ on $A^{(\gamma)}$

$$\Delta A = \frac{\left\|A^{(\gamma+1)} - A^{(\gamma)}\right\|^2}{\left\|A^{(\gamma+1)}\right\|^2}. \tag{27}$$

*Step 8:* The optimization is complete if the condition ($\Delta A < \varepsilon$) is met, where $\varepsilon$ is the threshold error. In this case, the optimal Laplacian coefficients are in $\beta_1^{(\gamma)},\ \ldots,\ \beta_r^{(\gamma)}$, and the optimal kernels coefficients are in $\beta_{r+1}^{(\gamma)},\ \ldots,\ \beta_{r+s}^{(\gamma)}$. If the condition is not met, the Steps 2–7 are repeated for a next iteration ($\gamma$+1).

*Step 9:* The optimal kernel-Laplacian matrix $\boldsymbol{\Omega}_{\mathrm{OptKLC}}^{(\gamma+1)}$ is used to train the RBF neural network. Equation (28) calculates the activation of the RBF hidden layer as the distance $d_i$ between input vector $\boldsymbol{\Omega}_{\mathrm{OptKLC(i)}}$ and RBF neuron center $\mu_j$.

$$d_i = \boldsymbol{\Omega}_{\mathrm{OptKLC}(i)} - \mu_j \tag{28}$$

where *i* is the number of samples, and *j* is the number of hidden units.

*Step 10:* Equation (29) is used to calculate the RBF $\phi_j$ when the distance $d_i$ becomes smaller than the spread width $\sigma_j$.

$$\emptyset_j\,(x) = \exp\left(-\frac{\boldsymbol{\Omega}_{\mathrm{OptKLC}(i)} - \mu_j}{2\sigma_i^2}\right). \tag{29}$$

*Step 11:* Linear regression is performed using the following equation to calculate the target outputs *y*

$$\text{output} = \sum_{i=1}^{j} w_j \phi_j(\mathrm{x}) \tag{30}$$

where $w_j$ are the weights between the hidden and output layers.

For the testing phase, the matrices of the test Laplacians $\boldsymbol{H}_{1-\boldsymbol{test}}$ and test kernels $\widetilde{\boldsymbol{H}}_{2-\boldsymbol{test}}$ are mapped from the testing samples, and combined using (22) to give the OptKLC testing matrix $\boldsymbol{\Omega}_{\mathrm{OptKLC\_test}}$. The trained RBF network is used to perform the classification on $\boldsymbol{\Omega}_{\mathrm{OptKLC\_test}}$.

### C. Audio–Visual Fusion

Fig. 9 shows the audio–visual fusion module that obtains the outputs from both the audio and visual paths and performs the fusion. Our design allows the audio–visual emotion recognition to operate in a dynamic mode for the audio and visual paths. The audio path used a dynamic sliding window to continuously


Fig. 9. Audio–visual fusion module.

TABLE I
AUDIO VISUAL EMOTION RECOGNITION DATABASES

| Database name | Number of subjects | Audio sampling rate (Hz) | Video frame rate (f/s) |
|---|---|---|---|
| eNTERFACE'05 | 43 | 48 000 | 25 |
| RML | 8 | 22 050 | 30 |

segment and classify the speech signal to give a set of emotion speech outputs. The visual path continuously captured the image frames to give a set of emotion visual outputs. For a certain length of video sequence, two sets of emotion (audio and visual) outputs will be produced and summed to give an audio emotion score set and a visual emotion score set. The audio and visual emotion scores are multiplied with the $W_{\mathrm{Audio}}$ and $W_{\mathrm{Visual}}$ weights. The $W_{\mathrm{Audio}}$ and $W_{\mathrm{Visual}}$ weights can be preset or can be adaptively adjusted based on the channel noise [48]. The emotion with the maximum score is the final decision over a certain period. The proposed I-AVER can dynamically detect and recognize emotion over a continuous video sequence. Unlike many other conventional emotion recognition systems, which can only perform static emotion recognition on the full speech length and/or the selected facial image, the proposed system can dynamically recognize emotion via video chat or conferencing.

## IV. Experimental Results on Audio–Visual Emotion Recognition

The proposed I-AVER was tested on the audio–visual emotion recognition databases, eNTERFACE'05 [49] and RML [28]. Both the eNTERFACE'05 and RML databases contained the six basic emotions (happy, angry, disgust, sad, surprise, fear) of interest in call centers. Table I shows a summary of some characteristics for the two databases in terms of the number of subjects, number of video samples, and number of languages. The eNTERFACE'05 database contains video samples from 43 subjects in one language (English) from 14 different nationalities with an audio sampling rate of 48 000 Hz and a frame rate of 25 f/s. The size of the image frame is 720 × 576 pixels, and the average size of the face region is around 260 × 300 pixels. The RML database contains video samples from eight subjects in six different languages (English, Mandarin, Urdu, Punjabi, Persian, Italian) and accents with an audio sampling rate of 22 050 Hz and a frame rate of 30 f/s.

### A. Recognition Accuracy for Audio, Visual, and Combined Audio–Visual Modalities

Fig. 10(a) and (b) shows the recognition accuracy for the proposed audio–visual emotion recognition system broken down into the audio, visual, and combined audio–visual modalities. The error bars show one standard deviation of uncertainty. The experiments used 1279 and 400 video samples from the eNTERFACE'05 and RML databases, respectively, following the experimental settings as in [28]. The RML video samples were truncated to 2 s. For cross-validation evaluation, the samples were selected in random from ten subjects, and were split into training sets (75% of samples) and testing sets (25% of samples). The average recognition rates were noted for ten experiments. Some parameters used were as follows (audio path: $\mathrm{k_{row}} = 11$, $\mathrm{k_{col}} = 5$, visual path: $\mathrm{k_{row}} = 11$, $\mathrm{k_{col}} = 4$). The values for $r$ and $s$ for kernel and Laplacian mappings were both set to 5.

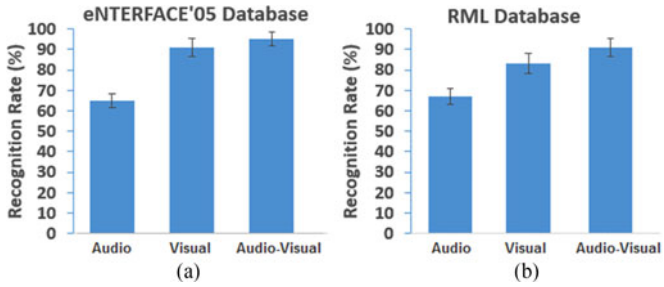


Fig. 10. Recognition accuracies (audio, visual, and audio–visual) of the proposed system for (a) eNTERFACE'05 and (b) RML databases.

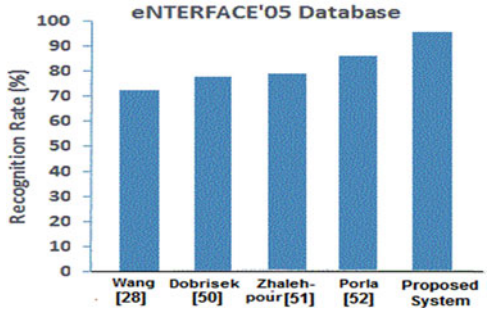


Fig. 11. Comparison–recognition accuracies for eNTERFACE'05.

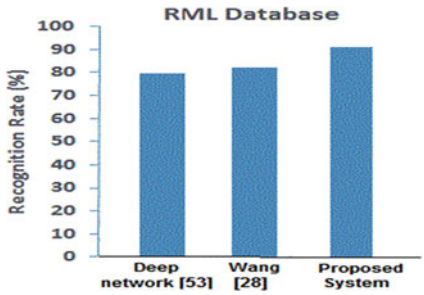


Fig. 12. Comparison–recognition accuracies for RML.

### B. Comparison of Recognition Accuracy With Other Methods

Comparisons were made with other existing techniques and methods in the literature. Figs. 11 and 12 give some comparisons with other reported results on the eNTERFACE'05 and RML databases, respectively. The proposed system achieved better

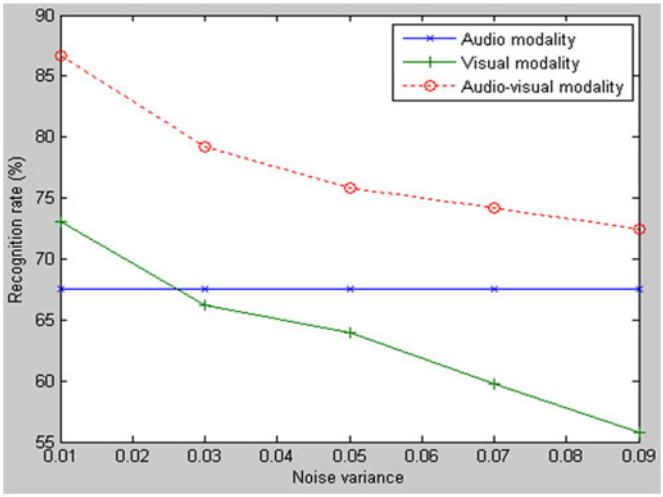


Fig. 13. Recognition accuracies for varying levels of white Gaussian noise for RML database.

recognition rates of 95% and 91%, respectively, compared with other methods.

### C. Recognition Accuracy for Poor Visual Sensors Quality

Further experiments were performed to investigate the recognition accuracy of the proposed system toward noisy and poor quality visual data, which may occur in contact centers. The work by Banda and Robinson in [70] gives a noise analysis for audio–visual emotion recognition. Following the work in [70], we added white Gaussian noise to corrupt the visual data. The audio data were left uncorrupted. Fig. 13 shows the performance of the proposed system on the RML database, respectively, as the Gaussian noise variance was increased from 0.01 to 0.09. The recognition accuracy of the visual modality decreased significantly as the noise variance increased. However, the recognition rate of the combined audio–visual modalities maintained a recognition rate of about 73%, even at a noise variance of 0.09, demonstrating the robustness of the multimodal system toward noisy visual data.

## V. Multimodal Customer Satisfaction Video Monitoring and Analytics

It is important to note that the proposed system in Section III is designed to dynamically recognize the customer's emotions. The system needs to be modified to relate the emotion and derive the customer satisfaction. Based on the studies in [54] and [55], the connection or relationship between customer satisfaction and customer emotion has been found to be relevant and reliable. Some previous studies have attempted to model the relationships between customer satisfaction and design attributes or elements that are based on market or customer survey data. For example, Park and Han [56] developed models to relate the customer satisfaction to design attributes using a fuzzy rule-based approach, and Liu *et al.* [57] proposed fuzzy techniques to calculate an e-commerce customer satisfaction index. In our proposed system, the emotion components contained within the

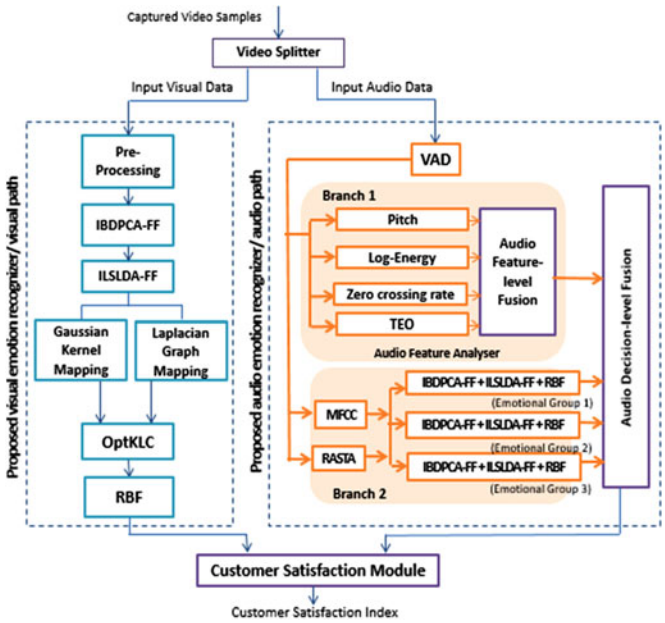


Fig. 14. Architecture of customer satisfaction video analytics system.


Fig. 15. Customer satisfaction module.

visual and audio data in the video can be used as the representations for the customer satisfaction. This section presents a system model for the multimodal customer satisfaction monitoring and analytics. An advanced multichannel customer satisfaction analytics model is also suggested later.

Fig. 14 shows the proposed customer satisfaction monitoring and analytics system. During the conversation or meeting via video conferencing, the audio and visual sensors in computers or mobile devices capture the video. The video is then transmitted to the customer services center. The video consists of customer's faces (visual data) and voices or speech (audio data). At the contact center, the visual and audio data are separated from the video frames using the video splitter and fed into their respective paths of the proposed system. The final audio–visual fusion model is redesigned to allow the transformation from emotions to the customer satisfaction score. The system also allows for continuous performance monitoring. Besides the performance monitoring, the data can also be collected for advanced offline analysis.

The outcomes from the visual and the audio paths are passed into a module called the customer satisfaction module (shown in Fig. 15) that contains two main components: customer satisfaction score conversion and monitoring mechanism. The first module calculates the customer satisfaction score based on the outputs from the visual and audio paths. The two sets of

TABLE II
ASSIGNED INDEX ON EACH EMOTION

| Emotion | Emotion Type, (Positive or Negative) | Index Assignment |
|---|---|---|
| Happy | Positive | $I_1$ |
| Surprise | Positive | $I_2$ |
| Sad | Negative | $I_3$ |
| Fear | Negative | $I_4$ |
| Disgust | Negative | $I_5$ |
| Angry | Negative | $I_6$ |

outputs from the audio and visual paths contain their respective outcomes or recognized emotions. For the customer satisfaction score conversion, the outputs from the audio and visual paths are converted into the customer satisfaction score. Besides $W_{\mathrm{Visual}}$ and $W_{\mathrm{Audio}}$, each emotion score is also multiplied with its respective index. There are six indices $I_1, I_2, \ldots, I_6$ in total. The final customer satisfaction score $CS_{\mathrm{score}}$ is calculated as

$$CS_{\mathrm{score}} = W_{\mathrm{visual}} \sum_{i=1}^{N} y_{i\ \mathrm{visual},} I_i + W_{\mathrm{audio}} \sum_{i=1}^{M} y_{i\ \mathrm{audio},} I_i \tag{31}$$

where $y_i$ denotes the outputs from the audio and visual paths and $I_i$ denotes the assigned indices and carries either a positive $+1$ or negative $-1$ sign depending on the type as shown in Table II.

The six emotions are assigned with different indices. Each index is assigned a value in the range from 0 to 1 (where the high value indicates a high satisfaction level or vice versa). The emotions can be categorized into two types (positive and negative emotions). The positive emotion contributes a positive sign to its term in (31), while the negative emotion gives a negative sign to its term in the same equation. The "happy" and "surprise" emotions tend to represent positive emotions, while the remaining "sad," "disgust," "fear," and "sad" emotions tend to represent negative emotions. The value for each emotion index can be changed or re-defined if it is needed. In our case, the index values for each emotion is designed and set based on the research works in [58]–[61].

The "happy" emotion is most likely stimulated by customers when they are satisfied with the delivered products or services. Thus, the maximum index value (i.e., 1.0) is assigned to this emotion when it is recognized from the captured data. The "surprise" emotion is normally stimulated by customers due to "unexpected" or "mistakenly expected" attributes for the delivered products and services. This emotion could be in the form of pleasant or unpleasant. However, according to the studies in [58] and [59] that relates the "surprise" emotion to the customer satisfaction, it is more likely to indicate a positive effect and bring positive impact to the business activities. As a result, the "surprise" emotion in this research is assigned with a slightly higher index value (0.6) than the average index value (0.5). The remaining emotions such as "sad," "fear," "disgust," and "angry" are negative emotions. According to [60] and [61], the "sad" emotion is very close to the neutral emotion (or no emotion) on the facial image and speech. Thus, it is assigned with a negative index value of 0.4. This index number is closest to the average customer satisfaction index compared with other negative emotions.

The "fear" emotion on the other hand is more likely to indicate "worry" or "disbelief" from the customer perspective. Thus, it is assigned with the negative index value of 0.7. The "disgust" and "angry" emotions both have the highest characteristics for dissatisfaction. The "disgust" emotion from a customer possibly shows the indications of "disliking" or "unacceptable" to the delivered products or services. Thus, an index value of 0.8 is assigned to this emotion. The "angry" emotion is assigned with the highest negative index value (i.e., 1.0). This is due to its obvious signs of dissatisfaction on the delivered products or services. If a new set of data or real-world data becomes available from the contact centers, then modeling can be performed to identify the right values to be used for the indices $I_i$. After obtaining the $CS_{\mathrm{score}}$, the score is then passed to the second module, which has the monitoring mechanism to allow the supervisors to know which of the conversations between the center agents or the customer services staff members are going smoothly and which ones may require an intervention.

Today, customers contact companies by multiple channels such as telephone, e-mail, online chat, social media, instant messaging, etc. A multichannel contact center management is a new challenge and allows for various kinds of customer interaction services that should be integrated into a contact center strategy. To evaluate the customer satisfaction, the multimodal data from multichannels need to be taken into consideration. The data collected from different channels from a customer can be used for more advanced analysis and suggests a future multimodal analytics system to give an overall customer satisfaction profile. For example, the text analytics for the customer from text channels (e.g., e-mail, social media), speech or voice analytics over the phone channel, and video analytics over video conferencing can be passed to an advanced analytics system to indicate the overall customer satisfaction from multichannels over a certain period. The customer satisfaction models can be designed and obtained from the data obtained from those channels.

## VI. CONCLUSION

In this paper, we investigated the challenging issues pertaining to performing video analytics for customer satisfaction in contact centers. Our investigation suggested that utilizing the audio and visual components in video to detect and recognize human emotions is necessary and feasible to perform the analytics. While the use of audio–visual modalities can result in a higher recognition accuracy compared with using only audio-only or visual-only modalities, they lead to a number of additional challenges such as determining the suitable feature extraction techniques and optimal data fusion techniques prior to classification. To address these challenges, we proposed an I-AVER method that can operate in a dynamic mode to perform human emotion recognition over video conferencing or video chat. Finally, the outcomes of the audio and the visual paths are mapped and translated to give customer satisfaction scores retrieved from a continuous video stream for a customer satisfaction video

analytics system. Numerical results were presented to verify the effectiveness of our proposal.

## REFERENCES


[1] J. J. Cristiano, J. K. Liker, and C. C. White, III, "Key factors in the successful application of quality function deployment (QFD)," *IEEE Trans. Eng. Manage.*, vol. 48, no. 1, pp. 81–95, Feb. 2001.

[2] O. C. Hansemark and M. Albinsoon, "Customer satisfaction and retention:The experiences of individual employees," *Manag. Service Quality*, vol. 14, no. 1, pp. 40–57, 2004.

[3] C. Baumann, G. Elliott, and S. Burton, "Modeling customer satisfaction and loyalty: Survey data versus data mining," *J. Services Market.*, vol. 26, no. 3, pp. 148–157, 2012.

[4] D. J. Flint, C. P. Blocker, and P. J. Boutin, Jr., "Customer value anticipation,customer satisfaction and loyalty: An empirical examination," *Ind. Market. Manage.*, vol. 40, no. 2, pp. 219–230, 2011.

[5] A. Gilmore and L. Moreland, "Call centers: How can service quality be managed?" *Irish Market. Rev*., vol. 13, no. 1, pp. 3–11, 2000.

[6] A. Feinberg, I. S. Kim, L. Hokama, K. De Ruyter, and C. Keen, "Operational determinants of caller satisfaction in the call center," *Int. J. Service Ind. Manage.*, vol. 11, no. 2, pp. 131–141, 2000.

[7] S. Ben-David, A. Roytman, R. Hoory, and Z. Sivan, "Using voice servers for speech analytics," in *Proc. Int. Conf. Digit. Telecommun*., Aug. 29–31, 2006, p. 61.

[8] D. Melamed and M. Gilbert, "Samsa: Speech analytics," in *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. Hoboken, NJ, USA: Wiley, 2011, pp. 397–416.

[9] R. Polig *et al.*, "Giving text analytics a boost," *IEEE Micro*, vol. 34, no. 4, pp. 6–14, Jul./Aug. 2014.

[10] R. Polig, K. Atasu, and C. Hagleitner, "Token-based dictionary pattern matching for text analytics," in *Proc. 23rd Int. Conf. Field Programmable Logic Appl.*, 2013, pp. 1–6.

[11] A. Gilmore and L. Moreland, "Call centers: How can service quality be managed?" *Irish Market. Rev*., vol. 13, no. 1, pp. 3–11 2000.

[12] A. Feinberg, I. S. Kim, L. Hokama, K. De Ruyter, and C. Keen, "Operational determinants of caller satisfaction in the call center," *Int. J. Service Ind. Manage.*, vol. 11, no. 2, pp. 131–141, 2000.

[13] *Hype Cycle for Contact Center Infrastructure 2014, Gartner*, published on Jul. 24, 2014. [Online]. Available: https://www.academia.edu/8426323/Hype_Cycle_for_Contact_Center_Infrastructure_2014

[14] T. Sobol-Shikler and P. Robinson, "Classification of complex information: Inference of co-occuring affective states from their expressions in speech," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1284–1297, Jul. 2010.

[15] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifier using acoustic-prosodic information & semantic labels," *IEEE Trans. Affective Comput.*, vol. 2, no. 1, pp. 10–21, Jan.-Jun. 2011.

[16] A. Chakraborty, A. Konar, U. K. Chakraborty, and A. Chatterjee, "Emotion recognition from facial expressions and its control using fuzzy logic," *IEEE Trans. Syst., Man Cybern. A, Syst. Humans*, vol. 39, no. 4, pp. 726–743, Jul. 2009.

[17] D. Arumugam and S. Purushothaman, "Emotion classification facial expression," *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, pp. 92–98, 2011.

[18] M. Song, J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition—A new approach," in *Proc. 2004 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2004, vol. 2, pp. II-1020–II-1025.

[19] 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 Technol.*, vol. 22, no. 6, pp. 869–874, Jun. 2012.

[20] 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.

[21] C. Cheng-Yao, H. Yue-Kai, and P. Cook, "Visual/Acoustic emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2005, pp. 1468–1471.

[22] L. C. De Silva and N. Pei Chi, "Bimodal emotion recognition," in *Proc. 4th IEEE Int. Conf. Automat. Face Gesture Recognit.*, 2000, pp. 332–335.

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

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

[25] Z. Zeng, Y. Hu, G. Roisman, Z. Wen, Y. Fu, and T. Huang, "Audio-visual spontaneous emotion recognition," in *Artificial Intelligence for Human Computing*, vol. 4451, T. Huang, A. Nijholt, M. Pantic, and A. Pentland, Eds. Berlin, Germany: Springer-Verlag, 2007, pp. 72–90.

[26] Z. Zeng *et al.*, "Audio-visual affect recognition," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 424–428, Feb. 2007.

[27] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 936–946, Aug. 2008.

[28] 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.

[29] C. S. Ooi, K. P. Seng, and L.-M. Ang, "Enhanced incremental bi-directional principal component analysis with forgetting factors," in *Proc. Int. Conf. Adv. Comput. Inf. Technol.*, Kuala Lumpur, Malaysia, 2012, pp. 46–50.

[30] 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.

[31] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[32] M. F. Wilkins, L. Boddy, C. W. Morris, and R. Jonker, "A comparison of some neural and non-neural methods for identification of phytoplankton from flow cytometry data," *Comput. Appl. Biosci.*, vol. 12, no. 1, pp. 9–18, 1996.

[33] M. F. Wilkins, L. Boddy, C. W. Morris, and R. R. Jonker, "Identification of phytoplankton from flow cytometry data by using radial basis function neural networks," *Appl. Environ. Microbiol.*, vol. 65, no. 10, pp. 4404–4410, Oct. 1999.

[34] 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.

[35] X. Yang, B. Tan, J. Ding, J. Zhang, and J. Gong, "Comparative study on voice activity detection algorithm," in *Proc. Int. Conf. Elect. Control Eng.*, 2010, pp. 599–602.

[36] S. Zhalehpour, Z. Akhtar, and C. E. Erdem, "Multimodal emotion recognition with automatic peak frame selection," in *Proc. IEEE Int. Symp. Innov. Intell. Syst. Appl.*, Jun. 2014, pp. 116–121.

[37] 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.

[38] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Detection of stress and emotion in speech using traditional & FFT based log energy features," in *Proc. Joint Conf. 4th Int. Conf. Inf., Commun. Signal Process., 4th Pacific Rim Conf. Multimedia*, 2003, vol. 3, pp. 1619–1623.

[39] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate & energy of the speech signal," in *Proc. Amer. Soc. Eng. Edu. Zone Conf.*, 2008, pp. 1–7.

[40] 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.

[41] M. Jones and P. Viola, "Fast multi-view face detection," Mitsubishi Elect. Res. Lab, Cambridge, MA, USA, TR-20003-96, vol. 3, 2003.

[42] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, pp. 71–86, 1991.

[43] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.

[44] 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. B*, vol. 36, no. 4, pp. 946–953, Aug. 2006.

[45] Y.-R. Yeh and Y.-C. F. Wang, "A rank-one update method for least squares linear discriminant analysis with concept drift," *Pattern Recognit.*, vol. 46, pp. 1267–1276, 2013.

[46] S. Yu, L. C. Tranchevent, W. Glänzel, J. A. Suykens, B. De Moor, and Y. Moreau, "Optimized data fusion for K-means Laplacian clustering," *Bioinformatics*, vol. 27, pp. 118–126, 2011.

[47] S. Yu, L.-C. Tranchevent, B. Moor, and Y. Moreau, "Rayleigh quotient-type problems in machine learning," in *Kernel-Based Data Fusion for Machine Learning*, vol. 345. Berlin, Germany: Springer-Verlag, 2011, pp. 27–37.

[48] Y. W. Wong, K. P. Seng, and L.-M. Ang, "Audio-visual recognition system in compression domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 5, pp. 637–646, May 2011.
[49] 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, p. 8.
[50] S. Dobrisek, R. Gajsek, F. Mihelic, N. Pavesic, and V. Struc, "Towards efficient multi-modal emotion recognition," *Int. J. Adv. Robotic Syst.*, vol. 10, no. 53, pp. 1–10, 2013.
[51] 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. 827–834, 2016.
[52] 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.
[53] C. Fadil, R. Alvarez, C. Martinez, J. Goddard, and H. Rufiner, "Multimodal emotion recognition using deep networks," in *Proc. Latin Amer. Congr. Biomed. Eng.*, 2014, pp. 813–816.
[54] D. Hill, *About Face: The Secrets of Emotionally Effective Advertising*. London, U.K.: Kogan Page, 2010.
[55] M. Gobe, *Emotional Branding: The New Paradigm for Connecting Brands to People* (Updated and Revised ed.). New York, NY, USA: Allworth, 2010.
[56] J. Park and S. H. Han, "A fuzzy rule-based approach to modeling affective user satisfaction towards office chair design," *Int. J. Ind. Ergon.*, vol. 34, no. 1, pp. 31–47 2004.
[57] X. Liu, X. Zeng, Y. Xu, and L. Koehl, "A fuzzy model for customer satisfaction index in e-commerce," *Math. Comput. Simul.*, vol. 77, pp. 512–521, 2007.
[58] J. Vanhamme, "The link between surprise and satisfaction: An exploratory research on how best to measure surprise," *J. Market. Manage.*, vol. 16, pp. 565–582, 2000.
[59] D. Snelders, "The role of surprise and satisfaction judgements," *J. Customer Satisfaction, Dissatisfaction Complaining Behav.*, vol. 14, pp. 27–45, 2001.
[60] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 582–596, May 2009.
[61] J. M. Leppänen, M. Milders, J. S. Bell, E. Terriere, and J. K. Hietanen, "Depression biases the recognition of emotionally neutral faces," *Psychiatry Res.*, vol. 128, pp. 123–133, 2004.
[62] L. L. Berry, L. P. Carbone, and S. H. Haeckel, "Managing the total customer experience," *MIT Sloan Manage. Rev.*, vol. 43, no. 3, pp. 1–6, 2002.
[63] P. C. Verhoef, K. N. Lemon, A. Parasuraman, A. Roggeveen, M. Tsiros, and L. A. Schlesinger, "Customer experience creation: Determinants, dynamics and management strategies," *J. Retailing*, vol. 85, no. 1, pp. 31–41, 2009.
[64] C. M. Froehle and A. V. Roth, "New measurement scales for evaluating perceptions of the technology-mediated customer service experience," *J. Oper. Manage.*, vol. 22, pp. 1–21, 2004.
[65] K. A. Richards and E. Jones, "Customer relationship management: Finding value drivers," *Ind. Market. Manage.*, vol. 37, pp. 120–130, 2008.
[66] A. Gandomi and H. Murtaza, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, pp. 137–144, 2015.
[67] C. M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Proc. IEEE Workshop Automat. Speech Recognit. Understanding*, 2001, pp. 240–243.
[68] L. Vidrascu and L. Devillers, "Detection of real-life emotions in call centers," in *Proc. Annual Conf. Int. Speech Commun. Assoc.*, 2005, pp. 1841–1844.
[69] D. Pappas, I. Androutsopoulos, and H. Papageorgiou, "Anger detection in call center dialogues," in *Proc. IEEE Int. Conf. Cognitive Infocommun.*, 2015, pp. 139–144.
[70] N. Banda and P. Robinson, "Noise analysis in audio-visual emotion recognition," in *Proc. 11th Int. Conf. Multimodal Interaction*, 2011, pp. 1–4.

**Kah Phooi Seng** (M'02) received the B.Eng. and Ph.D. degrees from the University of Tasmania, Hobart, Australia.

She is currently an Adjunct Professor in the School of Computing and Mathematics, Charles Sturt University, Sydney, Australia. Before returning to Australia, she was a Professor and the Department Head of Computer Science and Networked Systems at Sunway University. Before joining Sunway, she was an Associate Professor in the School of Electrical and Electronic Engineering, Nottingham University. She has published more than 230 papers in journals and international refereed conference proceedings. Her research interests include intelligent visual processing, multimodal signal processing, artificial intelligence, multimedia WSN, affective computing, and the development of intelligent system and multimodal big data analytics.

**Li-Minn Ang** (SM'09) received the B.Eng. and Ph.D. degrees from Edith Cowan University, Joondalup, Australia.

He is currently a meber of Senior Academic Staff in the School of Computing and Mathematics, Charles Sturt University, Sydney, Australia, and the Leader of the Intelligent Analytics and Sensing Research Group. He was previously an Associate Professor at Nottingham University. He has published more than 150 papers in journals and international refereed conference proceedings.

Dr. Ang is a Fellow of the Higher Education Academy (U.K.). His research interests include visual information processing, embedded systems, WSNs, reconfigurable computing, real-world computer systems, large-scale data gathering and big data analytics for sensor networks, and multimedia IoT.