Received March 1, 2018, accepted April 4, 2018, date of publication April 13, 2018, date of current version May 16, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2826525

Monitoring System for Patients Using Multimedia for Smart Healthcare

ATIF ALAMRI

Chair of Pervasive and Mobile Computing, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia
Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

e-mail: atif@ksu.edu.sa

This work was supported by the Deanship of Scientific Research, King Saud University, through the Vice Deanship of Scientific Research Chairs.

ABSTRACT The use of multimodal inputs in a smart healthcare framework is promising due to the increase in accuracy of the systems involved in the framework. In this paper, we propose a user satisfaction detection system using two multimedia contents, namely, speech and image. The three classes of satisfaction are satisfied, not satisfied, and indifferent. In the proposed system, speech and facial image of the user are captured, transmitted to a cloud, and then analyzed. A decision on the satisfaction is then delivered to the appropriate stakeholders. Several features from these two inputs are extracted from the cloud. For speech, directional derivatives of a spectrogram are used as features, whereas for image, a local binary pattern of the image is used to extract features. These features are combined and input to a support vector machine-based classifier. It is shown that the proposed system achieves up to 93% accuracy in detecting satisfaction.

INDEX TERMS Smart healthcare, user satisfaction detection, emotion detection, cloud computing.

I. INTRODUCTION

With the invention of low-cost processing and storage, several smart solutions are gaining attraction in improving the quality of human life. Particularly, smart healthcare is in great demand because of the increase in population and decrease in doctor-to-people ratio, and some people become are busy to travel to a specialized hospital for treatment. The smart healthcare business is estimated to be more than several billion dollars in next few years [1].

A successful smart healthcare framework requires several parameters, including ease of use of the medical sensors, low cost, high accuracy, ubiquitous nature of the framework, and less delay in making decision. These parameters may not be achieved in a single framework, although efforts have been made for the last several years [2]–[5]. The ease of use of sensors depends on their invasiveness; low-cost depends on the complexity of the devices in acquiring signals and installation; high accuracy depends on the precision of the sensors and the algorithms embedded in the software; and delay depends on the number of features of the signals.

Numerous smart healthcare frameworks have been proposed in the literature from different perspectives. Some frameworks attempt to solve the problems in electrocardiogram signal monitoring [2], smart cities [5], voice pathology assessment [6], emotion recognition [8], and patient's state recognition [9]. A software-defined network was proposed to improve the performance of smart healthcare frameworks in [10].

Customer satisfaction (of users and patients) is an important goal for smart healthcare business. A service provider can obtain feedback on customer satisfaction by using a survey conducted electronically or paper-based. The issue in conducting this type of survey is that sometimes the users do not want to participate. An automatic satisfaction reading from the face, speech, or gesture of the users can greatly solve this problem.

In this study, we propose a user satisfaction detection system as part of a smart healthcare framework. A multimedia-based technique is utilized to capture the signals from the users. Speech and image are the two signals captured for the accuracy of the proposed system. These signals are processed in a cloud server. A cloud manager then sends the result to the stakeholder. Figure 1 shows the smart healthcare framework. A smart home is equipped with multimedia sensors that can capture different signals. These signals come from the expressions of the user. Subsequently, the signals are transmitted to the cloud for processing. The result is then sent to the hospital, doctors, and caregivers, who analyze the satisfaction result for future quality improvement.

The organization of the remainder of the paper is as follows. Section II introduces the proposed user satisfaction

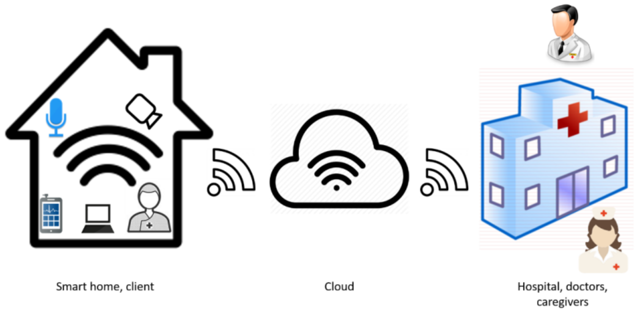

FIGURE 1. Smart healthcare framework for user satisfaction monitoring.

detection system. Section III provides the experiments with results and discussion. Finally, Section IV concludes the paper.

II. RELATED WORKS

Several studies have been published on patients' emotions and status monitoring system. This section presents some of the important works.

The emotion or human's mental status can be recognized using speech only, image only, and their combination. An emotion recognition system using nonlinear features from speech was proposed in [20]. An optimal set of features was selected by a particle swarm optimization algorithm, which achieved 99.47% accuracy in the Emo-DB database. The same database was used in other works. For example, in [21], a support vector machine (SVM) with spectral and prosody features was used, which achieved 94.9% accuracy. A deep neural network was used in [22]; several selected acoustic features were fed into the network, which obtained accuracy of 81.9%. A hidden Markov model-based classification was utilized in [23], which achieved approximately 73% accuracy. Wavelet packet energy with entropy was used as the input to an extreme learning machine-based classifier in [24], which obtained accuracy of 97.24%.

Human facial expressions have been automatically recognized using images or videos in several studies. The most commonly used database in these works is the Cohn–Kanade (CK) database [25]. SVM-based systems were proposed in [26] and [27]. The most prominent features of images were wavelet and geometric features, and texture pattern. The accuracies of the systems varied between 94% and 97% using the CK database.

Monitoring systems of patients' expressions were proposed in [28]–[30]. In [28], a healthcare framework using big data of emotions and deep learning model was developed. Another framework for smart cities was introduced in [29]. A center-symmetric local binary pattern (LBP) with bandlet transformation was also realized. A combination of speech and facial features were employed to monitor patients' status in [8]. A facial expression recognition for an e-healthcare system was proposed in [30] based on Weber local descriptor. The recognition accuracy reached up to 98%. A biologically-inspired multimedia management system was proposed in [31].

Multimedia-based human expression recognition systems were proposed in [4], [8], and [28]. The multimedia inputs consisted of speech and image signals. These systems achieved higher recognition rate than that of systems using one type of signal only.

III. PROPOSED SYSTEM

Figure 2 displays the block diagram of the proposed user satisfaction detection system for a smart healthcare framework. In the proposed system, multimodal input signals are processed, namely, speech and image signals. A microphone records the speech from the user while a video camera captures the facial expressions.

A. PROCESSING OF SPEECH SIGNAL

Speech signal is transmitted to the cloud, where a server extracts and classifies the features. In the server, the speech signal is framed and is calculated using Hamming window. The frame length is 40 ms, and the frame shift is 20 ms. Each windowed frame is transformed into a frequency domain representation (spectrum) using Fourier transform, such that the time domain signal is converted into a frequency domain signal. Twenty-four band-pass filters are passed through the frequency domain signal to mimic the hearing perception of the user. The center frequencies of the filters are distributed on a mel scale. The bandwidths of the filters correspond to the critical bandwidth. The result of this step is a mel spectrogram [11].

The mel spectrogram is passed through directional derivatives in four directions to obtain the relative progress of the signal along four directions, that is, 0°, 45°, 90°, and 135°, which correspond to time, increasing time frequency, frequency, and decreasing time frequency, respectively. The derivatives used are a linear regression, where the window size is three frames before and three frames after the current frame [12]. The following equation shows the calculation of the linear regression along time (0°), where $S_{\mathrm{n,f}}$ corresponds to the mel spectrogram at frame n and filter f.

$$0^\circ d_{n,f} = \frac{\sum_{j=1}^{3} j\left(S_{n,f+j} - S_{n,f-j}\right)}{\sum_{j=1}^{3} j^2} \tag{1}$$

The reason of using the directional derivatives is to capture the relative dependency of the features in time, frequency, and time-frequency directions. The satisfied mood speech may have slow transition in time and frequency directions, while the unsatisfied mood speech may contain fluctuations very often in time and frequency directions. The directional derivatives are then processed via a discrete cosine transform for compression and de-correlation. Subsequently, 48 features per frame are available for the speech signal.

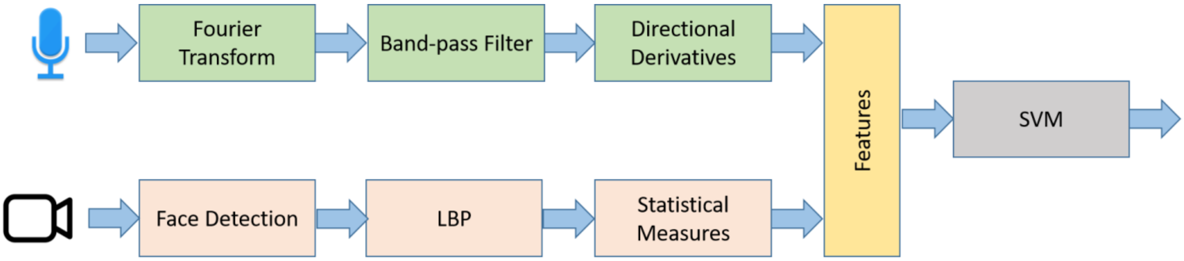

FIGURE 2. Block diagram of the proposed user satisfaction detection system.

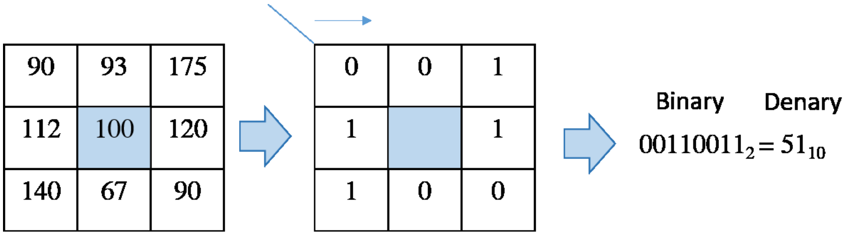

FIGURE 3. Graphical representation of LBP calculation.

B. PROCESSING OF IMAGE SIGNAL

First, a key image frame per one-second video is selected. The key frame is determined by a histogram comparison of the frames. It is selected when the minimal distance between a frame and its previous and next frames is achieved. This key frame is termed as the image signal. A face detection algorithm extracts the face area of the image signal. This process is performed in the local processor to decrease the transmission cost of the video.

An LBP is applied to the face image once it is transmitted to the cloud server to obtain an LBP image. LBP is a powerful texture descriptor and is computationally efficient [13]. In a rectangular LBP, a window size of 3 x 3 pixels is selected. The intensity of the middle pixel is set as a threshold of the window. If the intensity value of a neighboring pixel is larger than the threshold, then "1" is assigned to the location of this neighboring pixel; otherwise, "0" is assigned. The arrangement of "1" and "0" of location of the eight neighboring pixels is concatenated to form an 8-bit binary number, which is then transformed to a denary value. The denary value is the LBP of the middle pixel. The window is slid by one pixel, and the process is repeated. Figure 3 shows an illustration of the LBP calculation. A histogram is formed from the LBP image. Several features are calculated from the histogram to describe the facial image. The extracted features are the average (mean), standard deviation, skewness, and kurtosis. These are well-known statistical features, which are successfully used in many applications.

The LBP has several variants designed to make them robust against noise and other distortions. These variants include circular LBP, multivariate LBP, center symmetric LBP, and magnitude-sign LBP. All these variants have their own advantages and disadvantages. The rectangular LBP is the basic LBP, and it has low computational complexity and comparable accuracy. Other texture descriptors are used in [14] and [15]; however, the LBP is used in the present work for its simplicity.

C. CLASSIFICATION

An SVM-based classifier is used in the cloud server for classification. The SVM is a simple yet powerful binary classifier, and it has successfully been applied in many signal processing applications [16] such as speaker recognition, image classification, image forgery detection, and electrocardiogram signal classification. The main idea of the SVM is to maximize the distance of a linear separator from two classes of samples. Normally, real data of two classes cannot be separated by a line in a two-dimensional space. Therefore, a kernel function is applied to project the data in a high-dimensional space so that the data of two classes are separated by a hyperplane. Many kernels are proposed in the literature; each having its own advantages and disadvantages. Polynomial kernel and a radial basis function (RBF) kernel are the two most common kernels used in many applications. In image and speech processing applications, these kernels achieved high classification accuracies, and so we investigated these kernels in the proposed system. The famous library for the SVM, LibSVM is used in our experiments [17].

IV. EXPERIMENTS

This section discusses the experimental setup, database creation, and experimental results.

A. EXPERIMENTAL SETUP

The proposed system detects three classes of emotions, namely, satisfied, unsatisfied, and indifferent. In the SVM, we adopted the one versus the rest approach. Several experiments were conducted. In different sets of experiments,

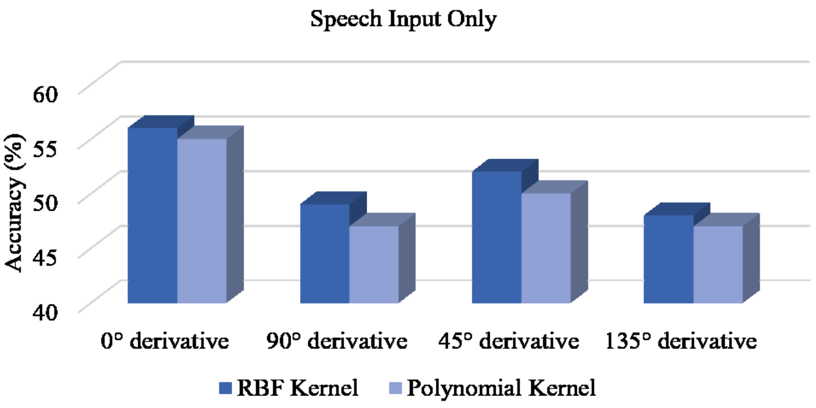

FIGURE 4. Accuracy of the proposed system with speech input and two types of SVM kernels.

we evaluated the system by using the speech signal only, the image signal only, and the combined signals. The parameters of the SVM, namely, optimization and kernel parameters, were fixed during system training. We investigated two SVM kernels, namely, RBF and polynomial kernel.

B. DATABASE

We created a database consisting of speech and image signals to evaluate the proposed system. The 40 participants were all male students. The mean age of the students was 22 years with a variance of 4.9. Before recording the database, all the participants were trained to act as either satisfied, unsatisfied, or indifferent. For a sample training session, which was not considered in the database, the participants delivered their speech and video signals as instructed. The actual session started after reaching the level of our expectation.

Each participant had three instances of each class, thereby obtaining nine instances for all the classes. For each instance, the speech and video signals were recorded. After the completion of all the recordings, they were replayed in front of a set of students to evaluate the validity of the signals based on the classes. The mean opinion score was 4.2, which indicated excellent recording.

The recording was taken place in a quiet office environment. The office was not populated other than the recorder and the participant. The noise level was very low. There was enough light to ensure a uniform illumination in the office.

For the experiments, we implemented a five-fold cross-validation tactic. In this tactic, the entire dataset was separated into five equal groups. The experiments were performed in five instances, where four groups in each instance were used in the training, and the other group was used in the testing. The concluding accuracy was achieved by averaging the accuracies of the five instances.

C. RESULTS AND DISCUSSION

Figure 4 shows the accuracy of the proposed system using the speech signal only. Four directional derivatives were used in the calculation of speech features. The figure shows the accuracy of the system using one directional derivative at a time. The highest accuracy was obtained by the 0° directional derivative, followed by the 45° directional derivative. The RBF kernel performed superior to the polynomial kernel. This result indicated that the temporal derivative is the most significant in user satisfaction detection when speech signal is the input.

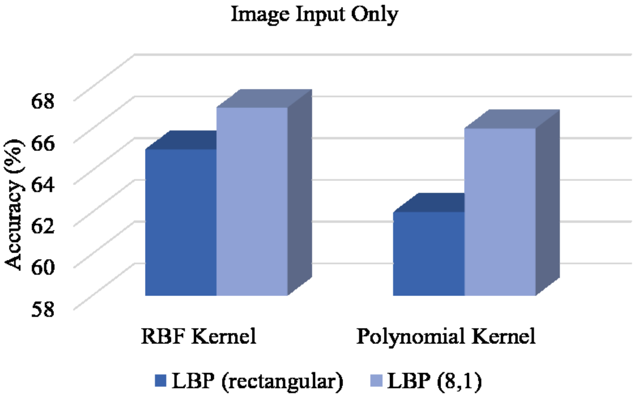

FIGURE 5. Accuracy of the proposed system with image input and two types of SVM kernels.

Figure 5 shows the accuracy of the system when the input was the image signal. We examined two variants of the LBP, namely rectangular LBP and circular LBP with radius of 1 and 8 neighbors. The figure shows that the rectangular LBP performed better than the circular LBP. Moreover, the RBF kernel of the SVM performed better than the polynomial kernel.

Figure 6 displays the accuracy of the system when the input was the speech signal only, image signal only, and both signals. In the case of speech signals, all the directional derivatives were combined. The figure shows that the system achieved 64% accuracy while using the speech signal, 67% while using the image signal, and 78% accuracy while using both of the signals. These results, which were obtained by using the RBF kernel in the SVM, indicate that the image signal is slightly more important than the speech signal

TABLE 1. Comparison of accuracies obtained by the systems with different feature extraction techniques.

Speech		Image	
Technique	Accuracy (%)	Technique	Accuracy (%)
Proposed	64	Proposed (with LBP)	67
MFCC	61	HoG	65
LPC	57		

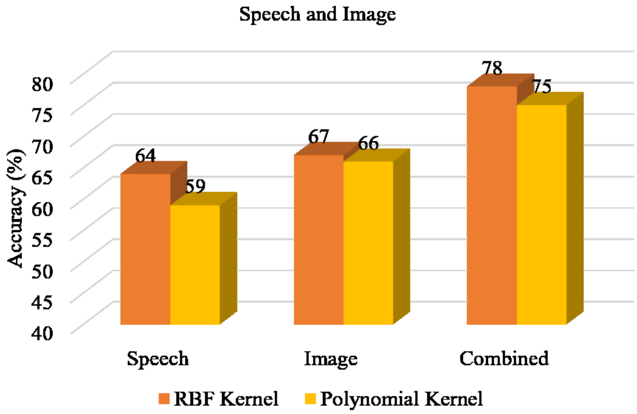

FIGURE 6. Accuracy of the proposed system with single and bi-modal inputs and two types of SVM kernels.

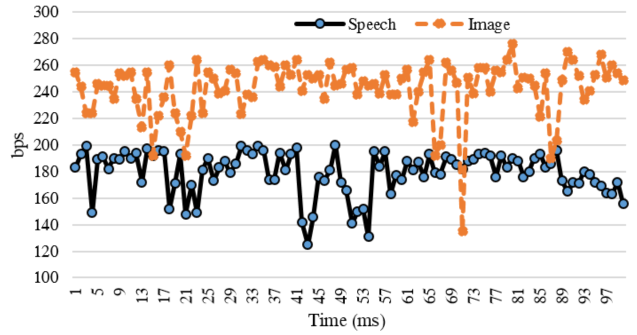

FIGURE 7. Bandwidth consumption (in bps) by the proposed system for speech and image transmission.

in terms of accuracy; however, both signals are needed to achieve good accuracy.

Given the limited literature, we compared the performance of the proposed system with systems composed of different feature extraction techniques of speech and image. Particularly, we selected two well-known speech features, namely, mel-frequency cepstral coefficients (MFCC) and linear predictive coding (LPC) [11], and one popular image texture descriptor, namely, histogram of gradients (HoG) [14]. Table 1 demonstrates the accuracy of the systems with various feature extraction techniques. The table shows that the proposed system performed the best.

The bandwidth requirement of the proposed system is also less. Fig. 7 shows the bandwidth requirement in bps to transmit speech and image, respectively. From the figure, we see that the bps required by the image modality is higher than that by the speech modality.

V. CONCLUSION

A user satisfaction detection system using speech and image for a smart healthcare framework was proposed. For the speech signal, we used the directional derivative features from the mel spectrogram, whereas for the image signal, we used the LBP features. SVM was used as the classifier, and several experiments were performed. The best accuracy (78%) was obtained by combining the features from the speech and image signals.

In future works, we intend to use highly sophisticated classifying approach, such as active learning [18], which has been successfully used in emotion recognition. In [19], MPEG-7 audio features were effectively used in an audio–visual emotion recognition. We may use such features and include other input modalities to enhance the accuracy of the proposed system.

REFERENCES

[1] M. S. Hossain and G. Muhammad, "Cloud-assisted industrial Internet of Things (IIoT)—Enabled framework for health monitoring," *Comput. Netw.*, vol. 101, pp. 192–202, Jun. 2016.

[2] S. Cuomo, A. Galletti, R. Farina, G. De Pietro, and G. Sannino, "A framework for ECG denoising for mobile devices," in *Proc. 8th ACM Int. Conf. Pervas. Technol. Rel. Assist. Environ. (PETRA)*, New York, NY, USA, 2015, Art. no. 48.

[3] H. B. Sta, "Quality and the efficiency of data in 'smart-cities,'" *Future Generat. Comput. Syst.*, vol. 75, pp. 409–416, Sep. 2017.

[4] M. S. Hossain, G. Muhammad, B. Song, M. M. Hassan, A. Alelaiwi, and A. Alamri, "Audio–visual emotion-aware cloud gaming framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2105–2118, Dec. 2015.

[5] U. Aguilera, O. Peña, O. Belmonte, and D. López-de-Ipiña, "Citizen-centric data services for smarter cities," *Future Generat. Comput. Syst.*, vol. 76, pp. 234–247, Nov. 2017.

[6] M. S. Hossain and G. Muhammad, "Healthcare big data voice pathology assessment framework," *IEEE Access*, vol. 4, no. 1, pp. 7806–7815, Dec. 2016.

[7] A. Baseliner, S. Steidi, C. Hacker, and E. Nöth, "Private emotions versus social interaction: A data-driven approach towards analysing emotion in speech," *User Model. User-Adapted Interact.*, vol. 18, nos. 1–2, pp. 175–206, 2008.

[8] M. S. Hossain, "Patient state recognition system for healthcare using speech and facial expressions," *J. Med. Syst.*, vol. 40, no. 12, pp. 272:1–272:8, Dec. 2016.

[9] L. Hu *et al.*, "Software defined healthcare networks," *IEEE Wireless Commun. Mag.*, vol. 22, no. 6, pp. 67–75, Jun. 2015.

[10] G. Muhammad, T. A. Mesallam, K. H. Malki, M. Farahat, M. Alsulaiman, and M. Bukhari, "Formant analysis in dysphonic patients and automatic Arabic digit speech recognition," *BioMed. Eng. OnLine*, vol. 10, p. 41, May 2011.

[11] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[12] G. Muhammad *et al.*, "Spectro-temporal directional derivative based automatic speech recognition for a serious game scenario," *Multimedia Tools Appl.*, vol. 74, no. 14, pp. 5313–5327, 2015.

[13] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[14] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, pp. 51–59, Jan. 1996.

[15] G. Muhammad, "Automatic speech recognition using interlaced derivative pattern for cloud based healthcare system," *Cluster Comput.*, vol. 18, no. 2, pp. 795–802, Jun. 2015.

[16] R. O. Duda, P. E. Hart, and H. G. Strork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2000.

[17] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, Jul. 2007, Art. no. 27. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[18] G. Muhammad and M. F. Alhamid, "User emotion recognition from a larger pool of social network data using active learning," *Multimedia Tools Appl.*, vol. 76, no. 8, pp. 10881–10892, Apr. 2017.

[19] M. S. Hossain, M. Moniruzzaman, G. Muhammad, A. Ghoneim, and A. Alamri, "Big data-driven service composition using parallel clustered particle swarm optimization in mobile environment," *IEEE Trans. Services Comput.*, vol. 9, no. 5, pp. 806–817, Sep./Oct. 2016.

[20] C. K. Yogesh *et al.*, "A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal," *Expert Syst. Appl.*, vol. 69, no. 1, pp. 149–158, Mar. 2017.

[21] J. B. Alonso, J. Cabrera, M. Medina, and C. M. Travieso, "New approach in quantification of emotional intensity from the speech signal: Emotional temperature," *Expert Syst. Appl.*, vol. 42, pp. 9554–9564, Apr. 2015.

[22] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2011, pp. 5688–5691.

[23] S. Deb and S. Dandapat, "A novel breathiness feature for analysis and classification of speech under stress," in *Proc. 21st Nat. Conf. Commun. (NCC)*, 2015, pp. 1–5.

[24] H. Muthusamy, K. Polat, and S. Yaacob, "Improved emotion recognition using Gaussian mixture model and extreme learning machine in speech and glottal signals," *Math. Problems Eng.*, vol. 2015, Mar. 2015, Art. no. 394083.

[25] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2000, pp. 46–53.

[26] S. Datta, D. Sen, and R. Balasubramanian, "Integrating geometric and textural features for facial emotion classification using SVM frameworks," in *Proc. Int. Conf. Comput. Vis. Image Process. (CVIP)*, 2016, pp. 619–628.

[27] M. Jampour, V. Lepetit, T. Mauthner, and H. Bischof, "Pose-specific non-linear mappings in feature space towards multiview facial expression recognition," *Image Vis. Comput.*, vol. 58, pp. 38–46, Feb. 2017.

[28] M. S. Hossain and G. Muhammad, "Emotion-aware connected healthcare big data towards 5G," *IEEE Interne Things J.*, 2017, doi: 10.1109/JIOT.2017.2772959.

[29] G. Muhammad, M. Alsulaiman, S. U. Amin, A. Ghoneim, and M. F. Alhamid, "A facial-expression monitoring system for improved healthcare in smart cities," *IEEE Access*, vol. 5, no. 1, pp. 10871–10881, Dec. 2017.

[30] M. Alhussein, "Automatic facial emotion recognition using weber local descriptor for e-healthcare system," *Cluster Comput.*, vol. 19, pp. 99–108, Sep. 2016.

[31] M. S. Hossain, A. Alamri, and A. El Saddik, "A biologically inspired framework for multimedia service management in a ubiquitous environment," *Concurrency Comput., Pract. Exper.*, vol. 21, no. 11, pp. 1450–1466, Aug. 2009.

Atif Alamri is currently an Associate Professor with the Software Engineering Department, College of Computer and Information Sciences, King Saud University. His research interest includes multimedia assisted health systems, ambient intelligence, and service-oriented architecture. He was a Guest Associate Editor of the IEEE Transactions on Instrumentation and Measurement and a Co-Chair of the 10th IEEE International Symposium on Haptic Audio Visual Environments and Games, and serves as a program committee member of many conferences in multimedia, virtual environments, and medical applications.