Naas, Si Ahmed; Sigg, Stephan

Real-time emotion recognition for sales

Published in:
Proceedings of 16th International Conference on Mobility, Sensing and Networking (MSN 2020)

DOI:
10.1109/MSN50589.2020.00096

Published: 01/04/2021

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Naas, S. A., & Sigg, S. (2021). Real-time emotion recognition for sales. In *Proceedings of 16th International Conference on Mobility, Sensing and Networking (MSN 2020)* (pp. 584-591). Article 9394292 IEEE. https://doi.org/10.1109/MSN50589.2020.00096

Real-time Emotion Recognition for Sales

Si-Ahmed Naas
Dept. Communication and Networking
Aalto University
firstname.lastname@aalto.fi

Stephan Sigg
Dept. Communication and Networking
Aalto University
firstname.lastname@aalto.fi

***Abstract*—Positive emotion is a pre-condition to any sales contract. Likewise, the ability to perceive the emotions of a customer impacts sales performance.**

To support emotional perception in buyer-seller interactions, we propose an audio-visual emotion recognition system that can recognize eight emotions: neutral, calm, sad, happy, angry, fearful, surprised, and disgusted. We reduced noise in audio samples and we applied transfer learning for image feature extraction based on a pre-trained deep neural network VGG16. For emotion recognition, we successfully obtained an audio emotion-recognition accuracy of 62.51% and 68% and video emotion-recognition accuracy of 97.13% and 97.77% on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Surrey Audio-Visual Expressed Emotion (SAVEE) datasets respectively. For the combination of the two models, our proposed merging mechanism without re-training achieved an accuracy of close to 100% on both datasets. Finally, we demonstrated our system for a customer satisfaction use case in a real customer-to-salesperson interaction using audio and video models, achieving an average accuracy of 78%.

***Index Terms*—Emotion recognition, Internet of Things, deep learning, transfer learning, Customer satisfaction.**

I. Introduction

CUSTOMER satisfaction is economically important for any business [1]. It means that a customer can express enjoyment and excitement about a service [2]. Dissatisfaction demands reaction, while calm or neutral reactions indicate the need for improvements.

The ability to accurately appraise the emotions of others moderates the practice of adaptive selling and customer-oriented selling [3]. It is not only the spoken language, but also facial expressions, tone of voice, or physiological responses that may indicate emotion [4]. Emotion recognition is possible from visual data despite the richness and the complexity of facial emotions [5]. Likewise, speech also carries emotion information, such as tone [6]. Indeed, measuring customer satisfaction in stores is becoming more popular. For instance, the Roads and Transport Authority (RTA) of Dubai has launched AI-enabled cameras that measure customer satisfaction [7]. Canon also presented new cameras for both surveillance and customer satisfaction [8]. As many companies use surveillance cameras with a built-in microphone in stores, customer satisfaction via facial expression and conversation analysis is practical.

We present an audio-visual emotion recognition system to predict customer satisfaction in real-time from visual and audio cues. The system is capable of recognizing the emotions neutral, calm, sad, happy, angry, fearful, surprised, and disgusted. For emotion recognition on two datasets, we derived models that achieve an accuracy of 62.51% and 68% on audio data as well as 97.13% and 97.77% on video data. By relying on confident predictions only, the system did not make incorrect classifications under the premise to provide a prediction at least every 10 seconds (near real-time). We further evaluated the models based on 45 real-world conversations and obtained an average accuracy of 78%.

II. Related Work

Prominent modalities for emotion recognition are audio, video, tactile, or multimodal. *Audio-based* emotion recognition exploits features such as pitch, intensity, energy, and MFCCs (Mel-frequency cepstrum coefficients). Challenges in speech-based emotion recognition are that expression differs among subjects [9] and is directly influenced by age, culture, and externals factors such as the environment [10]. *Video-based* systems extract emotions from features such as facial expression, mouth, or eye shape [4]. However, during speech, the mouth shape does not necessarily reflect the real emotion. Likewise, illumination, background, and pose also impact video-based emotion recognition [11]. *Tactile-based methods* extract features from tactile sensors and are used in human-robot interaction where emotion is estimated based on the types of physical interaction [12]. *Multimodal methods* aim to improve the overall accuracy of emotion recognition using multiple data sources [13]. A good overview is provided in [14]. Most work in emotion recognition is unimodal though [15]. We propose an audio and video-based multimodal emotion recognition system. Audio and video are chosen due to their relevance and availability in customer-salesclerk interaction. Speech and facial expression analysis to measure customer satisfaction have previously been considered in [16] for a customer satisfaction video analytics system

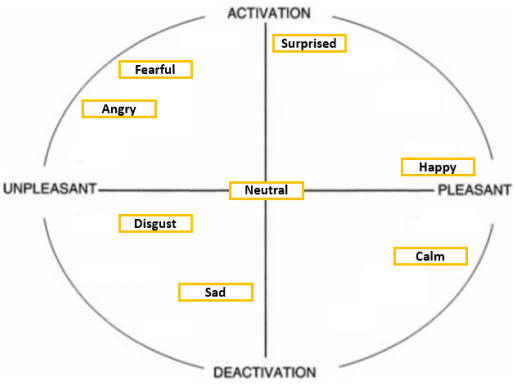

Fig. 1: The eight emotions recognized by our system, arranged according to the circumplex model [22]

with six emotions [17]. The authors employed prosodic features such as zero-crossing rate (ZCR), Teager energy operator (TEO), pitch, log energy, and spectral features, such as MFCC and RASTA [18]. The audio-based system includes three neural classifiers for {angry, happy}, {sad, disgust}, and {fear, surprise}, and achieved an accuracy of 64%. The visual features are extracted using incremental bidirectional principal component analysis and incremental least-square linear discriminant analysis. The authors classified visual and audio data with an average accuracy of 91% and 94% after merging audio and video-based classification. In [19], an android application to recognize seven emotions in real-time is presented. An SVM algorithm is employed to classify Hausdorff Distance, and Facial Landmark features for facial expression and MFCC features for speech recognition. The system is evaluated based on a random set from the RAVDESS database where the average accuracy is 96.3% for facial expression and 95.43% for speech. The overall accuracy reaches 97%. [20] focuses on cross-corpus evaluation and proposed multimodel emotion recognition from features such as MFCC and AlexNet for facial expression. However, the evaluation study is limited and the proposed approach showed worst performance in terms of recognition accuracy on cross-corpus evaluation. Another work [21] proposed a multimodal temporal deep learning framework to analyze emotion variation through time. It constructs discriminative embeddings based on audio-visual diachronically. We propose an emotion recognition system for the eight emotions neutral, calm, happy, angry, sad, surprise, and disgust. As depicted in Fig. 1, these emotions represent distinct dimensions and angles according to the circumplex model [22]. In contrast to previous work [16], [19]–[21], we perform harmonic/percussive separation to reduce noise in audio samples [23]. Furthermore, since facial landmark localization is sensitive to resolution, occlusion, illumination, and background [24], [25], we propose robust visual features using a Convolutional Neural Network (CNN). Our contributions are as follows (1) we present an audio emotion recognition module using pitch, energy, zero-crossing rates (ZCR), entropy and Mel Frequency cepstral coefficients (MFCCs) based on percussive and harmonic separation; (2) we create a video emotion recognition module extracting robust image features through transfer learning based on VGG16; (3) we apply gain ratio to select important features, balance audio, and video dataset; (4) we propose an efficient technique to merge both audio and video models; (5) finally, we present a real-time customer satisfaction system evaluated on real data.

III. Audio-Visual Emotion Recognition

We assume a salesclerk-customer relationship, where both the customer and the salesclerk are presented the emotion recognized from the respective communication partner (salesclerk or customer). The system is composed of modules to recognize emotion from audio as well as video. We further present a technique to combine the two modalities. Fig. 2 depicts the audio and video emotion-recognition process.

After separating the visual and vocal path from video, each path is preprocess separately. Audio preprocessing (section III-A) includes harmonic/percussive separation, windowing, then extraction of MFCC, zero-crossing rate, energy, entropy, and pitch. The visual path preprocessing (section III-B) includes image resizing, cropping, flopping, normalizing, followed by VGG16 feature extraction. The most representative features of the audio and visual feature vectors are selected using a gain ratio (GA) based feature extraction algorithm. For classification, we employ a CNN with two layers.

We propose a fusion mechanism for audio and video classifiers in section III-C.

A. Emotion recognition from audio

Most audio signals constitute a combination of harmonic and percussive sounds. We separate harmonic and percussive signals, which results in an automatic description of pitched signals and contributes towards better recognition results [26]. We apply median filtering on successive frames and frequency bins to suppress transients [23]. The two obtained median filtered streams are then employed for harmonic and percussive separation. Our overall audio recognition architecture is illustrated in Fig. 4. The architecture separates pitch, energy, zero crossing rate, entropy, as well as MFCCs.

1) Pitch: Pitch is widely used for stress evaluation [27] and to identify speaker tension. Each voice segment P_i is associated to a frequency F_i [28] :

$$\{(P_0, F_0), (P_1, F_1), ..., (P_n, F_n)\}\, n \in \mathbb{N}^* \qquad (1)$$

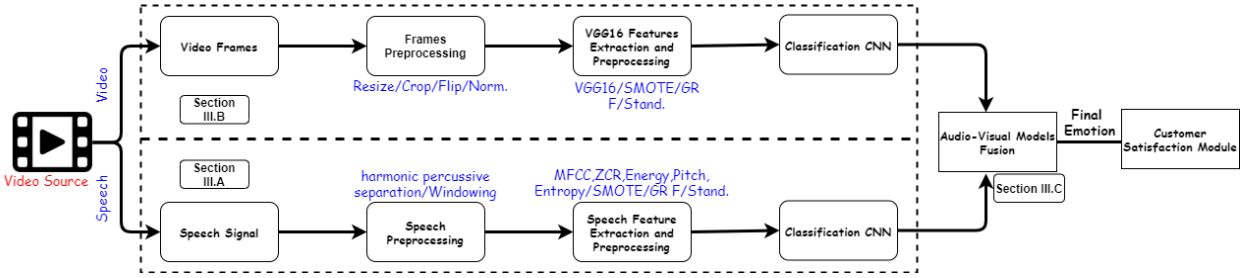

Fig. 2: Overview of Audio and Video Emotion Recognition System

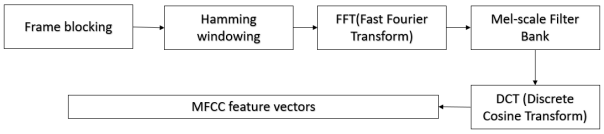

Fig. 3: MFCC feature extraction

2) Speech Signal Energy: The speech signal energy represents the intensity of vocal signals. The energy is influenced by the conditions in which the voice was recorded.

$$E = \frac{1}{N}\sum_{i=1}^{N}\left[x(i)^2\right] \tag{2}$$

$x(i)$ is the sequence of audio samples of the i^{th} frame. N is the length of the audio frame. Energy may vary even for the same emotion, however, it may also be used to separate ambient sound and speech.

3) Zero-crossing rate: It is used in speech/music classification and indicates when a sign change happens between successive audio samples $s(n)$.

$$Sign[s(n)] = \begin{cases} 1, & \text{if } s(n) \geq 0 \\ -1, & \text{if } s(n) < 0 \end{cases} \tag{3}$$

4) Entropy: It indicates the meaningfulness of the information. Extracting signal speech entropy is a technique that aims to detect voiced and unvoiced segmentations in speech.

5) Mel-Frequency Cepstral Coefficients (MFCCs): MFCCs are the most employed features in speech recognition. The MFCC can be of high accuracy especially in noiseless environments [29].

$$MFCC_i = \sum_{k=1}^{20} X_k cos\left[n(k-\frac{1}{20})\frac{\pi}{20})\right], \tag{4}$$

$$n = 1, ..., M \tag{5}$$

$$X_k = log(E(k)) \tag{6}$$

X_k is the log-energy output of the k^{th} filter, while M represents the number of cepstrum coefficients. Fig. 3 summarizes the feature extraction process.

B. Emotion recognition from video

Several feature descriptors have been proposed to enhance the recognition accuracy from video, such as facial landmarks, grey level dependency statistics [30], first-order moment to extract the texture information [31], rotation, and scale-invariant Gabor filters [32]. However, handcrafted image features are proven to be inefficient when the dataset is large or the classes are diverse [33].

1) Transfer learning: Deep learning has shown great success in computer vision and is efficient in complex problems and high dimensional datasets and known for the ability to extract good representation from data [34]. However, it requires long training time and large amounts of data, which makes it impractical for small datasets [35].

A solution to overcome the need for large datasets and to reduce resource consumption is transfer learning [36], which reuses the knowledge of a pre-trained network on a large dataset and transfers it to another domain [35]. Examples of pre-trained networks are AlexNet and VGG16 [37]. VGG16 is trained on over 1000 classes so universal features for facial expressions are expected. We then employ a CNN for the classification of the obtained feature vectors (as illustrated in Fig. 5).

VGG16 was initially proposed for object recognition and features a deeper CNN architecture than AlexNet [38]. It contains 16 layers designed for image classification and achieves 92.7% top-5 test accuracy based on the ImageNet dataset. It contains more than 14 million images and 1000 classes. We employ transfer learning to extract robust video features. The front layers of the pre-trained VGG16 are used as feature extractor (as illustrated in Fig. 5). The constructed image dataset using transfer learning resulted in 2200 features and lasted for approximately seventeen hours.

Let D_{vgg16} be the ImageNet dataset, and T_{vgg16} the object recognition task. Transfer learning improves the predictive function [39] $f(T_e)$ for the learning task T_e (emotion recognition) by transferring the knowledge from $\{D_{vgg}, T_{vgg}\}$, where: $D_{vgg} >> D_e$.

2) Image processing: We perform image resizing, crop to 224x224, horizontal flip based on the original

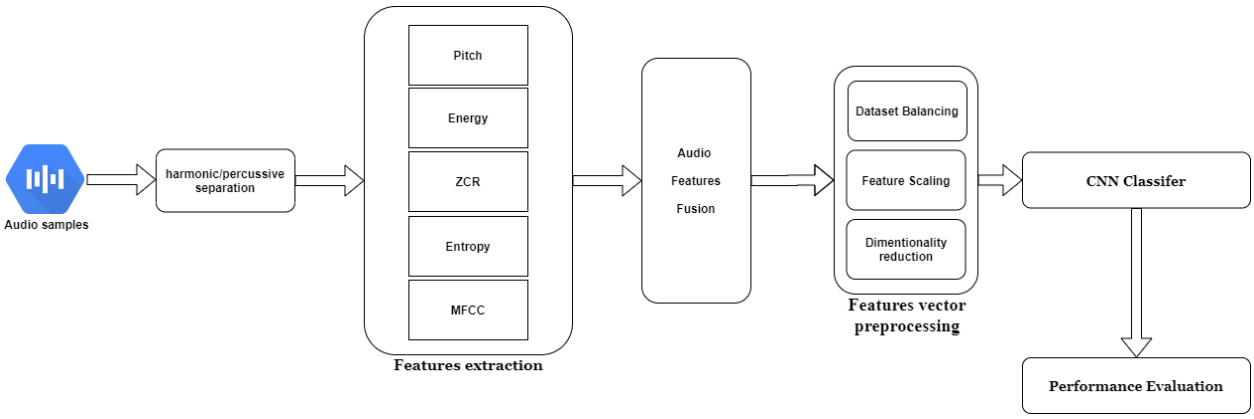

Fig. 4: Audio emotion recognition architecture

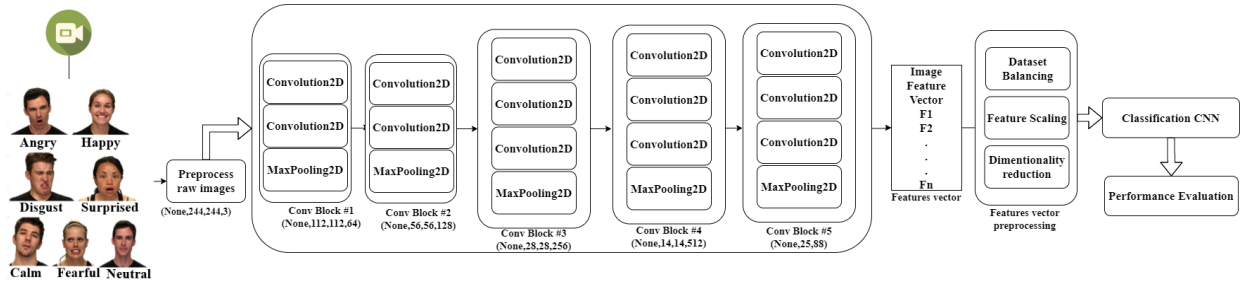

Fig. 5: Video emotion recognition architecture

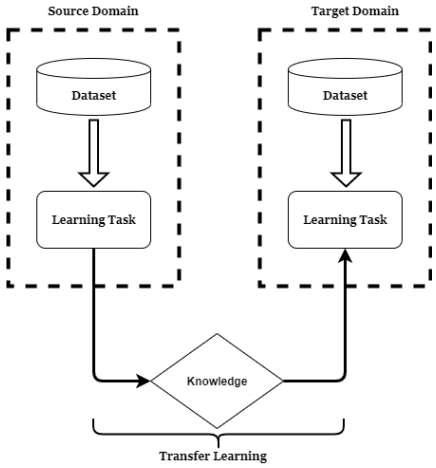

Fig. 6: Transfer learning architecture

training samples to enrich the dataset with more training samples. The employed techniques can reduce overfitting. We take 3-channel images (3 x H x W) as the input which we further normalize using **mean = [0.485, 0.456, 0.406]** and **std = [0.229, 0.224, 0.225]**, to scale the images with respect to the pre-trained model.

C. Combined audio-visual emotion recognition

Emotion in our system is obtained from the combined audio and visual recognition paths. We propose a technique to merge both, as described in Alg. 1. We consider two parameters to optimize α and β, where α is the number of samples that are wrongly predicted, and β is the number of samples that are ignored. To obtain these two parameters, we define a prediction score threshold to select the final emotion class. We ignore samples if the audio and video prediction score are below the defined threshold and also if the audio score is equal to video score and they refer to different classes. Otherwise, we assign the class with the highest score among the audio and the video prediction.

The threshold is defined based on α and β by experiments, where we aim to maximize α and minimize β. We obtain the overall prediction accuracy as

$$\textbf{Prediction Accuracy=}\ \frac{N-(\alpha+\beta)}{N-\beta} \tag{7}$$

where N is the total number of samples.

IV. System Evaluation

We present the preprocessing steps, as well as the performance of the emotion recognition system. We use Haar feature-based cascade classifiers to detect the

Algorithm 1 Merging audio video models

Output: α , β **Input:** M_{Audio}, M_{Video}, X_{Test}, γ, M
M_{Audio} : trained audio recognition model
M_{Video}: trained video recognition model
M: Maximum number of instances
$\alpha = 0$: wrongly classified instances.
$\beta = 0$: ignored instances.
Test set : $X_{Test} = \{x_1, ..., x_m\}$, $Y_{Test} = \{y_1, ..., y_m\}$
γ : prediction score threshold
Begin
For Each (x_i, y_i) in (X_{Test}, Y_{Test})
$\hat{Y}_{audio}$= M_{Audio}.predict(x_i)
$\hat{Y}_{video}$= M_{Video}.predict(x_i);
$Score_{Audio} = M_{Audio}$.predict($x_i$).score($\hat{Y}_{audio}$,$y_i$)
$Score_{Video} = M_{Video}$.predict($x_i$).score($\hat{Y}_{video}$,$y_i$)
IF $\hat{Y}_{audio} <> y_i$ and $\hat{Y}_{video} <> y_i$ and $Score_{Audio} < \gamma$ and $Score_{Video} < \gamma$
β ++
continue;
End if
IF $\hat{Y}_{audio} <> y_i$ and $\hat{Y}_{video} <> y_i$
α ++
continue;
End if
IF $Score_{Audio} < \gamma$ and $Score_{Video} < \gamma$
$\beta++$
continue;
End if
IF $Score_{Audio} >= Score_{Video}$ and $\hat{Y}_{audio} == y_i$
AssignClass(x_i)= $\hat{Y}_{audio}$
continue;
End if
IF $Score_{Video} > Score_{Audio}$ and $\hat{Y}_{video} == y_i$
AssignClass(x_i)= $\hat{Y}_{Video}$
continue;
End if
$\alpha++$;
End FOR
End =0

face region. Also, we employ Python scikit-learn, and Weka [40] for dataset preprocessing as well as OpenCV for image preprocessing. The processing was conducted using a laptop of 32 GB of RAM, i7-7700HQ CPU @ 2.80 GHz, and Nvidia GTX 1050 Ti.

A. Datasets

We employ in our study the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [41] and Surrey Audio-Visual Expressed Emotion (SAVEE) database [42]. RAVDESS contains an audio-visual recording of 24 performers (12 female). The actors sing and speak the same two sentences with repetition ("Kids are talking by the door", "Dogs are sitting by the door") while expressing the emotions neutral, calm, sadness, happiness, anger, fear, surprise, and disgust. Each file is 3 seconds long with an audio sampling rate of 48000 Hz and a frame rate of 25 f/s (7356 files in total). The SAVEE dataset comprises males recording 7 emotions (no calm class). The actors perform a total of 480 sentences. The video images are 1280x720 pixels and we extract the faces with a size of 224 x 224 pixels.

B. Class balancing

Since the training set was not balanced (classes not equally distributed), SMOTE (Synthetic Minority Over-sampling Technique) balancing [43] was performed, which reduces the imbalance by adding new samples to the minority class [44]. It randomly selects one from k (we choose k=5) neighbors of each sample of the minority class, then produces new samples using linear interpolation (8).

$$y = x_i + \delta \times (x^{'} - x_i) \tag{8}$$

where x_i is the training set sample, $x^{'}$ is the neighbour sample, and δ is a random value between 0 and 1.

C. Feature Selection

After obtaining the audio and video features, we perform feature selection using a gain ratio based feature technique. We applied C4.5 which is a successor of ID3 [45] to measure the information gain and select the attributes with the highest gain ratio. IG is the information gain, H is the entropy, H_{attr} calculates the entropy of the attribute $attr$ contributing to class C.

$$GainRatio(attr) = \frac{IG(attr)}{H(attr)} \tag{9}$$

where

$$IG(attr) = H(C) - H_{attr}(C) \tag{10}$$

where

$$H(attr) = \sum_{x \in values(attr)} -P(x) log_2 P(x) \tag{11}$$

D. Feature scaling

Every sample in the dataset is normalized between -1 and 1 ($-1 \leq x^i \leq +1$). All the values are replaced by their Z scores where each feature is redistributed with a mean $\mu = 0$ and divided by its standard deviation σ.

$$x^{'} = \frac{x - mean\,(x)}{\sigma} \tag{12}$$

Here, x is the original feature vector.

E. Training and testing set

We divide the dataset into training and testing sets and train our models with a split 7:3 for both datasets. For true positive (TP), true negative (TN), false positive (FP), and false negative (FN), we compute

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$\tag{15}$$

	Neutral	Calm	Happy	Sad	Angry	Fearful	Surprised	Disgust	Recall
Neutral	**28/39**	3/39	3/39	2/39	1/39	0	1/39	1/39	**71,79 %**
Calm	2/37	**28/37**	0	5/37	0	0	2/37	0	**75,68 %**
Happy	0	0	**25/36**	2/36	3/36	3/36	1/36	2/36	**69,44 %**
Sad	5/56	8/56	4/56	**27/56**	2/56	5/56	1/56	4/56	**48,21 %**
Angry	2/45	1/45	7/45	2/45	**26/45**	2/45	4/45	1/45	**57,78 %**
Fearful	1/50	1/50	7/50	2/50	3/50	**29/50**	2/50	5/50	**58,00 %**
Surprised	5/62	2/62	6/62	3/62	4/62	1/62	**37/62**	4/62	**59,68 %**
Disgust	2/56	0	5/59	0	4/59	7/59	2/59	**39/59**	**66,10 %**
Precision	**62,22 %**	**65,12 %**	**43,86 %**	**62,79 %**	**60,47 %**	**61,70 %**	**74,00 %**	**69,64 %**	

Audio confusion matrix

	Neutral	Calm	Happy	Sad	Angry	Fearful	Surprised	Disgust	Recall
Neutral	**104/117**	8/117	0	0	1/117	4/117	0	0	**88,89 %**
Calm	0	**101/101**	0	0	0	0	0	0	**100,00 %**
Happy	0	3/169	**164/169**	0	2/169	0	0	0	**97,04 %**
Sad	0	0	0	**100/102**	0	2/102	0	0	**98,04 %**
Angry	0	0	0	1/135	**134/135**	0	0	0	**99,26 %**
Fearful	0	0	0	0	0	**127/127**	0	0	**100,00 %**
Surprised	0	0	0	2	0	0	**133/135**	0	**98,52 %**
Disgust	0	0	1/96	0	0	0	0	**95/96**	**98,96 %**
Precision	**100,00 %**	**90,18 %**	**99,39 %**	**97,09 %**	**97,81 %**	**95,49 %**	**100,00 %**	**100,00 %**	

Video confusion matrix

Fig. 7: Emotion recognition confusion matrices for the RAVDESS dataset

	Neutral	Happy	Sad	Angry	Fearful	Surprised	Disgust	Recall
Neutral	**33/40**	1/40	3/40	0	0	3/40	0	**82,50 %**
Happy	0	**11/14**	0	2/14	0	1/14	0	**78,57 %**
Sad	3/14	0	**10/14**	0	0	1/14	0	**71,43 %**
Angry	1/12	1/12	0	**8**	2/12	0	0	**66,67 %**
Fearful	1/16	2/16	0	0	**10/16**	1/16	2/16	**62,50 %**
Surprised	5/25	3/25	1/25	1/25	4/25	**11/25**	0	**44,00 %**
Disgust	0	3/23	0	0	3/23	1/23	**16/23**	**69,57 %**
Precision	**76,74 %**	**52,38 %**	**71,43 %**	**72,73 %**	**52,63 %**	**61,11 %**	**88,89 %**	

Audio confusion matrix

	Neutral	Happy	Sad	Angry	Fearful	Surprised	Disgust	Recall
Neutral	**5783/5832**	21/5832	1/5832	1/5832	7/5832	3/5832	16/5832	**99,16 %**
Happy	6/2332	**2293//2332**	0	0	13/2332	0	20/2332	**98,33 %**
Sad	1/4492	0	**4406/4492**	0	78/4492	3/4492	4/4492	**98,09 %**
Angry	23/4053	7/4053	1/4053	**3952/4053**	2/4053	56/4053	12/4053	**97,51 %**
Fearful	11/3703	7/3703	68/3703	11/3703	**3549/3703**	23/3703	34/3703	**95,84 %**
Surprised	5/4310	5/4310	7/4310	21/4310	1/4310	**4269/4310**	2/4310	**99,05 %**
Disgust	34/3150	30/3150	16/3150	1/3150	21/3150	5/3150	**3043/3150**	**96,60 %**
Precision	**98,64 %**	**97,04 %**	**97,93 %**	**99,15 %**	**96,68 %**	**97,94 %**	**97,19 %**	

Video confusion matrix

Fig. 8: Emotion recognition confusion matrices for SAVEE dataset

F. Model Architecture

We use a CNN for audio and video emotion recognition. The model was implemented using Keras and TensorFlow. It consists of two 1-Dimensional convolutional layers. The first has 300 neurons with ReLU activation function followed by a dropout layer of rate 0.05 to reduce overfitting. The second convolutional layer was trained with 128 neurons with ReLU activation function followed by a dropout layer of rate 0.05. To convert the output into a single layer, we added a flatten layer. Finally, we use a Softmax activation function for the last dense layer. We use Adam optimizer with a learning rate of 0.001, and $\epsilon = 1e\text{-}07$. We train our network with 5000 epochs and a batch size of 32.

G. Emotion recognition system evaluation

We analyze the performance of audio and video recognition for all emotions: {calm (only RAVDESS), neutral, sad, happy, angry, fearful, surprised, disgusted}. Figure 7 and 8 show the respective confusion matrices. The algorithm obtained good results for emotions {neutral, calm, sad, happy, angry, surprised, and disgusted}. Average accuracies (audio, video) on the RAVDESS dataset are 62.51% and 97.13% in comparison to 68% and 97.77% for SAVEE. We combine the audio and video emotion recognition models using Alg. 1. For each dataset, we split the training and testing set with a ratio of 5:5. Fig. 9 and Fig. 10 show the performance after merging. Without ignoring any sample, our system achieved an accuracy of 98.14% on the RAVDESS dataset, and 99.68% on the SAVEE dataset. Furthermore, we run the algorithm for all possible values of the threshold [0.1,0.99] to obtain optimal parameters and to minimize the amount of ignored samples. Table I depicts the accuracy achieved for different values of α and β when varying the prediction score threshold from 0.1 to 0.99.

TABLE I: Accuracy vs. prediction score thresholds

	Score threshold	0.10	...	0.97	0.98	**0.99**
SAVEE	β (%)	0	...	0	0	**3.85**
Dataset	α (%)	99.65	...	99.65	99.65	**100**
RYVDESS	β (%)	0	...	0	0	**2.58**
Dataset	α (%)	89.28	...	96.15	98.04	**100**

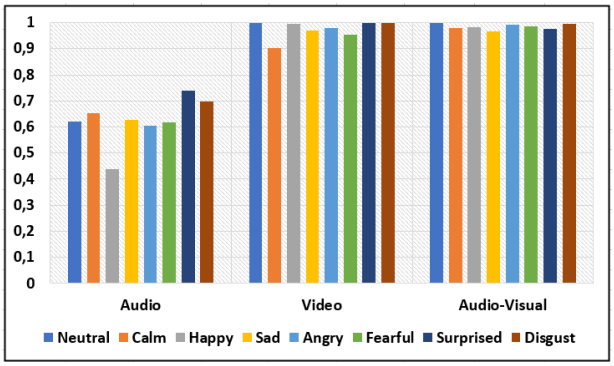

Fig. 9: Speech emotion, visual emotion, and audio-visual emotion recognition system on the RAVDESS dataset

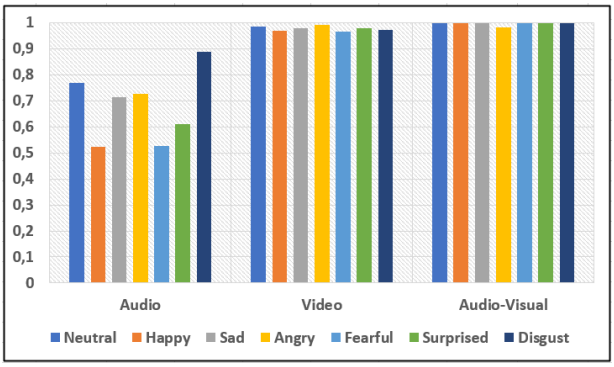

Fig. 10: Speech emotion, visual emotion, and audio-visual emotion recognition system on the SAVEE dataset

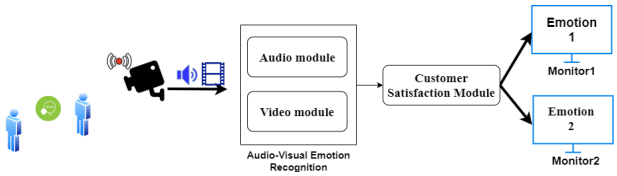

Fig. 11: Customer satisfaction module

Participant	nbr conv	conv. duration	accuracy	recog. time	S_Score
P1	9	4s	89% (8/9)	1,953s	0,22 (Partialy satisfied)
P2	6	4s	67% (4/6)	2,02s	0 (Partialy satisfied)
P3	5	4s	60% (3/5)	1,98s	0,2 (Partialy satisfied)
P4	25	4s	80% (20/25)	1,973s	-0,04 (not satisfied)
Overall acc.	77,78 %				

Fig. 12: Real-time emotion recognition results

	Sad+Fearful	Calm	Happy	Angry	Surprised	Disgust	Recall
Sad+Fearful	**3/4**	0	1/4	0	0	0	**75,00 %**
Calm	2/17	**9/17**	1/17	0	5/17	0	**52,94 %**
Happy	0	0	**8/9**	0	1/9	0	**88,89 %**
Angry	0	0	0	**6/6**	0	0	**100,00 %**
Surprised	1/9	0	0	0	**8/9**	0	**88,89 %**
Disgust	0	0	0	0	0	**0**	/
Precision	**50,00 %**	**100,00 %**	**80,00 %**	**100,00 %**	**57,14 %**	/	

Fig. 13: Real data confusion matrix

The prediction accuracy reaches 100% already for both datasets when only less than 4% of lowest confidence predictions are ignored.

Comparing our system with [19]–[21], our system achieves accuracies of 99.65% for SAVEE (77.40% [19], 90.38% [20]) and 98.04% for RAVDESS (67.70% [21], 95.43% [20]).

V. Customer Satisfaction Case Study

We conducted a customer satisfaction study to evaluate our system in a real-world instrumentation. In particular, we modelled customer to salesperson interaction while both the audio and video of the interaction was captured and analyzed by our system (cf. Fig. 11). During a conversation, the interaction is streamed to a server that extracts audio and video and predicts the final emotion. Based on the predicted emotion, it further calculates the customer satisfaction to display it on the screen so that the appropriate reaction can be taken by the staff. The obtained customer satisfaction is monitored during the whole conversation. The level of satisfaction is distinguished as (Neutral, 0), (Calm, 0), (Surprised, 0), (Happy, 1), (Sad, -1), (Angry, -1), (Fearful, -1), (Disgust, -1). Happy mood is considered as an only positive emotion (cf. [2]), and Neutral, Calm, Surprised as partially satisfied while the other emotions are negative. The overall satisfaction score is calculated as the average level of satisfaction during the entire conversation (equation 16 and equation 17).

$$S = \frac{1}{N}\sum_{i=1}^{N} e_i \tag{16}$$

N is the number of detected emotions and e_i is the satisfaction level for a each emotion.

$$CuS = \begin{cases} \text{Satisfied} & \text{IF S=1} \\ \text{Partially satisfied} & \text{IF S} \in [0,1[\\ \text{Not satisfied} & \text{ELSE} \end{cases} \tag{17}$$

For further evaluation, 45 conversations of 4 volunteers are involved in our experiment. As hardware, we utilize one smartphone (Samsung Galaxy M30) and use an IP Webcam to record the conversation between the volunteers. The server processes the received data from the IP Webcam and predicts the final emotion to calculates the satisfaction score. The predicted emotions are displayed on the monitor. Each volunteer is facing the smartphone for recording purposes and is asked to freely perform a reaction in which we consider the intended emotion as our ground truth. We display the predicted emotion on the screen, and we calculate the satisfaction score using (Eq. 17) The conversations are distributed as follows where the participants act naturally: P1(angry=3, happy=4, fearful and sad=1, calm=1), P2(angry=1, happy=1,surprised=1, calm=2, fearful and sad=1), P3(happy=1, calm=2, surprised=2), and P4(angry=2, fearful and sad=2, happy=3, surprised=6, calm=12). Figs. 12 and 13 illustrate the experiment results with an average emotion detection accuracy of 77.78%.

A. Discussions & Conclusions

We proposed an emotion recognition system for customer satisfaction. We used a smartphone as a major IoT device [46] to collect evaluation data. Our work is based on both audio and video data to obtain robust emotion recognition based on diverse and challenging datasets RAVDESS and SAVEE. We discussed feature selection for audio and video, and the combination of the obtained models at decision level. We employed a CNN for the classification of each data source separately and proposed a merging technique for the two models that achieves an accuracy close to 100%. The emotion recognition system is able to work with any video conferencing, meetings, emotion aware IoT service, or even social gathering applications.

Acknowledgment

The authors appreciate partial funding from NII International Internship Program and from Nokia Solutions and Networks Oy.

REFERENCES

[1] S. N. Mansor, S. A. Mostafa, A. Mustapha, and R. Darman, "An emotional agent for the analysis of customer satisfaction surveys," in *Int. Symp. on Agent, Multi-Agent Systems and Robotics*, 2018.

[2] J. Wang, "From customer satisfaction to emotions: Alternative framework to understand customer's post-consumption behavior," in *Int. Conf. on Service Sciences*, 2012.

[3] B. Kidwell, R. G. McFarland, and R. A. Avila, "Perceiving emotion in the buyer–seller interchange: the moderated impact on performance," *Journal of Personal Selling & Sales Management*, vol. 27, no. 2, pp. 119–132, 2007.

[4] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expr., speech and multimodal inform." in *Int. Conf. on Multimodal interfaces*, 2004.

[5] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Video emotion recognition with transferred deep feature encodings," in *Int. Conf. on Multimedia Retrieval*, 2016.

[6] L. Yu, B. Wu, and T. Gong, "A hierarchical support vector machine based on feature-driven method for speech emotion recognition," in *ECAL*, 2013.

[7] A. Frangoul, "Dubai introduces cameras that use ai to measure people's happiness," Mar 2019. [Online]. Available: https://www.cnbc.com/2019/03/12/dubai-introduces-\ \cameras-that-use-ai-to-measure-peoples-happiness.html

[8] S. Writer, "Canon security cameras keep close eye on customer satisfaction," Jul 2018. [Online]. Available: https://asia.nikkei.com/Business/Business-trends/Canon-security-cameras-keep-close-eye-on-customer-satisfaction

[9] M. Kotti and Y. Stylianou, "Effective emotion recognition in movie audio tracks," in *Acoustics, Speech and Signal Proc.*, 2017.

[10] M. S. Sinith, E. Aswathi, T. M. Deepa, C. P. Shameema, and S. Rajan, "Emotion recognition from audio signals using support vector machine," in *Recent Adv. in Intell. Comp. Sys.*, 2015.

[11] Mingli Song, Jiajun Bu, Chun Chen, and Nan Li, "Audio-visual based emotion recognition - a new approach," in *Computer Vision and Pattern Recognition*, 2004.

[12] H. Yan, M. H. Ang, and A. N. Poo, "A survey on perception methods for human–robot interaction in social robots," *Int. Journal of Social Robotics*, vol. 6, no. 1, pp. 85–119, Jan 2014.

[13] J. K. P. Seng and K. L. Ang, "Multimodal emotion and sentiment modeling from unstructured big data: Challenges, architecture, techniques," *IEEE Access*, vol. 7, pp. 90 982–90 998, 2019.

[14] D. Massaro, "Illusions and issues in bimodal speech perception," in *Auditory Visual Speech Perception*, 2002.

[15] A. A. Varghese, J. P. Cherian, and J. J. Kizhakkethottam, "Overview on emotion recognition system," in *Soft-Computing and Networks Security*, 2015.

[16] K. P. Seng and L. Ang, "Video analytics for customer emotion and satisfaction at contact centers," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 3, pp. 266–278, 2018.

[17] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface' 05 audio-visual emotion database," in *Int. Conf. on Data Engineering Workshops*, 2006.

[18] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE T. Speech and Audio Proc.*, vol. 2, no. 4, 1994.

[19] H. Alshamsi, V. Kepuska, H. Alshamsi, and H. Meng, "Automated facial expression and speech emotion recognition app development on smart phones using cloud computing," in *Inf. Techn., Electr. and Mobile Comm.*, 2018.

[20] E. Avots, T. Sapiundefinedski, M. Bachmann, and D. Kamiundefinedska, "Audiovisual emotion recognition in wild," *Mach. Vision Appl.*, vol. 30, no. 5, p. 975–985, 2019.

[21] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *Affective Computing and Intelligent Interaction*, 2019.

[22] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[23] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Digital Audio Effects*, 01 2010.

[24] H. Ouanan, M. Ouanan, and B. Aksasse, "Facial landmark localization: Past, present and future," in *Colloq. Inf. Sci. and Techn.*, 2016.

[25] S. Minaee and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *CoRR*, vol. abs/1902.01019, 2019.

[26] W. Lim and T. Lee, "Harmonic and percussive source separation using a convolutional auto encoder," in *EUSIPCO*, 2017.

[27] A. A. Khulage, "Extraction of pitch, duration and formant frequencies for emotion recognition system," in *Adv. in Recent Techn. in Comm. and Comp.*, 2012.

[28] S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, "Speech emotion recognition," in *Adv. in Electr. Comp. and Comm.*, 2014.

[29] N. Zheng, X. Li, H. Cao, T. Lee, and P. C. Ching, "Deriving mfcc parameters from the dynamic spectrum for robust speech recognition," in *Symp. on Chin. Spoken Lang. Proc.*, 2008.

[30] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans Syst Man Cybern*, vol. SMC-3, pp. 610–621, 1973.

[31] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE T. Pattern Analysis and Machine Intell.*, vol. 18, no. 8, pp. 837–842, 1996.

[32] J. Han and K.-K. Ma, "Rotation-invariant and scale-invariant gabor features for texture image retrieval," *Image and Vision Computing*, vol. 25, no. 9, pp. 1474 – 1481, 2007.

[33] M. Shaha and M. Pawar, "Transfer learning for image classification," in *Int. Conf. Electr., Comm. and Aerospace Techn.*, 2018.

[34] H. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *ICMI '15*, 2015.

[35] F. R. da Silva Oliveira and F. C. Farias, "Comparing transfer learning approaches applied to distracted driver detection," in *Lat. Amer. Conf. Comp. Intell.*, 2018.

[36] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE T. Knowl. and Data Engin.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[37] S. Liu and W. Deng, "Very deep cnn based image classif. using small training sample size," in *Asian Conf. Patt. Recogn.*, 2015.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classif. with deep cnns," *CACM*, vol. 60, no. 6, pp. 84–90, 2017.

[39] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transf. learn." *CoRR*, vol. abs/1808.01974, 2018.

[40] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[41] S. R. Livingstone and F. A. Russo, "Ryerson audio-visual db of emot. speech and song: A dynamic, multimodal set of facial and vocal expr. in north american engl." in *PloS one*, 2015.

[42] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[43] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[44] Z. Zheng, Y. Cai, and Y. Li, "Oversampling method for imbalanced classif." *Comp. and Inform.*, vol. 34, pp. 1017–1037, 2015.

[45] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, p. 81–106, 1986.

[46] M. ajana el khaddar and M. Boulmalf, *Smartphone: The Ultimate IoT and IoE Device*, 11 2017.