EmoChat: Bringing Multimodal Emotion Detection to Mobile Conversation

Conference Paper · August 2019

DOI: 10.1109/BIGCOM.2019.00037

3 authors, including:

Meng Jin
Peking University
71 PUBLICATIONS **654** CITATIONS
SEE PROFILE

Yuan He
Tsinghua University
165 PUBLICATIONS **3,743** CITATIONS
SEE PROFILE

EmoChat: Bringing multimodal emotion detection to mobile conversation

Luyao Chong
School of Software and BNRist
Tsinghua University
Beijing, China
chongluyao@gmail.com

Meng Jin
School of Software and BNRist
Tsinghua University
Beijing, China
mengj@mail.tsinghua.edu.cn

Yuan He
School of Software and BNRist
Tsinghua University
Beijing, China
heyuan@tsinghua.edu.cn

***Abstract*—Online chatting is very popular nowadays. However, most of the chatting softwares are based on pure text messages, which cannot completely convey users' emotions, causing information asymmetry. In this paper, we propose a new online chatting system, named EmoChat, which automatically identifies the emotions of the users and attaches the identification result to the messages sent by the user, allowing users to know the emotions of each other during online chatting. EmoChat analyzes the real-time emotions of users based on a joint consideration of facial expressions and text messages. Specifically, we propose an information entropy based method to fuse the multimodal information of these two pieces of complementary information. Furthermore, by realizing the context-sensitive property of the emotion information, a Hidden Markov Model based method is proposed to improve the emotion recognition accuracy with the context information. We implement EmoChat and evaluate its performance through a series of experiments. The experimental results show that EmoChat achieves an accuracy of 76.25% for emotion polarity recognition and an accuracy of 51.64% for emotion category recognition. Moreover, the delay when sending a message with emotions attached is within 50ms on the mobile devices.**

***Index Terms*—online chatting, emotional status, multimodal fusion, context-aware**

I. Introduction

With the development of network communication technologies, more people communicate with each other preferably online rather than face to face. Online communication, however, has an apparent limitation that people cannot get to know the emotion of each other, because in online communication, typically there are no facial expressions, voice intonations, or body gestures. For example, when someone says "I scored 85 points", he/she may mean "I am very satisfied with this achievement", or "I think it was awful", which causes ambiguity when one tries to infer the emotion of the other person solely based on the text information. Therefore, emotion information is demanded in online chatting scenarios.

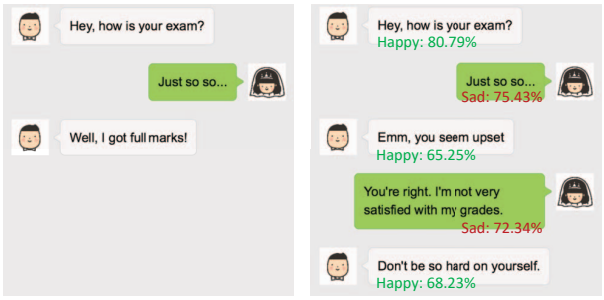
Indeed, emotion information can also benefit human-machine interactions. For example, in the intelligent customer service system, emotion information of the customer is an important indicator for the system to infer the customer's satisfaction and assess the user experience. Similarly, emotion information is also indispensable for the chatbot systems like Siri [1] and Xiao Ice [2]. In these systems, emotion recognition technologies can be helpful for detecting negative emotional status of the user like depression, anxiety, sadness, etc., allowing appropriate response in such conditions.

This motivates a simple vision: Can we build machines that automatically sense our emotions during online communication?

Existing approaches for emotion recognition either rely on audiovisual cues, such as video and audio clips, or require the person to wear physiological sensors like an ECG monitor. Both approaches have their limitations in applying to the online communication scenario. Audiovisual techniques usually need users communicate by voice and a clip of video for the user will be captured, which is infeasible in many scenarios, like in quiet environments (e.g., libraries and classrooms) or when the network bandwidth is limited (e.g., on the train or underground transport). Moreover, audiovisual techniques usually rely on heavy video processing techniques which cannot be applied on the resource-limited mobile devices. The second approach recognizes emotion by monitoring the physiological signals (e.g., the heartbeats) of human beings. These techniques use wearable sensors - e.g., ECG monitors - to measure specific signals and correlate their changes with people's emotional status. However, the use of dedicated hardware will impose additional costs and difficulties for users. What's worse, users' activities can interfere the physiological signals, making this approach unsuitable for regular usage.

We in this paper propose EmoChat to recognize users' emotion during online chatting. Our objective is to enable a universal solution that is applicable on mobile devices without any additional hardware. EmoChat achieves this by fusing the information that extracted from the text messages sent by the user and the facial expressions of the user captured by the front camera of the smartphone, two pieces of complementary information, for accurate and reliable emotion recognition across various scenarios. Specifically, we use deep learning methods to process facial expressions and text messages independently, and skillfully combine the results of these two methods based on their confidence. Furthermore, by realizing the context-sensitive property of the emotion information, an HMM (Hidden Markov Model) based method is further proposed to improve the emotion recognition accuracy with the context information. With EmoChat providing emotion status

978-1-7281-4024-7/19/$31.00 ©2019 IEEE
DOI 10.1109/BIGCOM.2019.00037



(a) traditional application (b) EmoChat

Fig. 1. Contrast between the traditional application and EmoChat.

information for users, the trend of the conversation may be quite different although having the same beginning. Fig. 1 gives a contrast between the traditional online chatting application and the emotion enhanced online chatting application. Firstly, the boy (the user on the left) asks the girl (the user on the right) about the exam, and the girl replys that her grades are just so so. While using the traditional online chatting application, the boy shows off his full marks and the girl never responds to him again, as shown in Fig. 1a. Howerver, with the application displaying the user's real-time emotional status, the boy knows that the girl is sad, so that they talk about the girl's grades and her feelings, narrowing the distance between them, as shown in Fig. 1b.

Contributions: We make the following contributions:

- We propose EmoChat, a novel emotion enhanced online chatting application for mobile devices, which combines facial expressions and text messages, two types of complementary information for accurate emotion recognition across different scenarios.
- We design a confidence-aware method to fuse the information extracted from facial expressions and text messages. In addition, an HMM based method is proposed to further enhance the recognition accuracy with the context information.
- We collect a multimodal online chatting dataset and evaluate the performance of EmoChat with extensive experiments across a wide variety of scenarios.

II. Related Work

It is widely believed that emotion information can benefit online communication. Many efforts have been made to improve chatting experience using emotion information. For example, Hua Wang et al. [3] propose a chatting system that presents users with emotion-related animated text messages. Here the emotion information is obtained based on the physiological sensor attached on the user's body and the manually defined emotion categories. The results of their experiments prove that such an enhanced online chatting system makes the users interact with each other more efficiently. Similarly, Chunling Ma et al. [4] and Dey et al. [5] propose enhanced chatting systems which infer the user's emotion based on text messages. Dey et al. evaluate their system using the SemEval dataset, and get an overall precision 56.37% for simple sentences, 42.71% for compound sentences and 27.68% for complex sentences. Note that the sentences in SemEval dataset are news headlines extracted from news websites and newspapers, which is quite different from daily language used in chatting.

Although the above methods are able to enhance the chatting efficiency, they either require users to wear additional hardware or rely only on text messages (resulting in poor recognition accuracy), thus they cannot be directly used in practical chatting scenarios. Considering that emotion recognition techniques have attracted much interest from research community, one may ask: can we directly use the existing emotion recognition methods in chatting systems? The answer is unfortunately no.

Today's emotion recognition methods typically infer users' emotions using various modalities [6], [7], including speech, gesture, physiological signal, facial expression and text.

The first group of approaches identify the emotion of the user based on the spoken utterance [8]. For example, OhWook Kwon et al. [9] extract 5 features from an audio clip, and feed these features to the QDA (quadratic discriminant analysis) classifier and the SVM (support vector machine) for emotion recognition, achieving a 42.3% accuracy for 5-class emotion recognition. Kun Han et al. [10] extract utterance-level features using DNN (Deep Neural Networks). Then based on the extracted features, they identify the emotion of the user using ELM (Extreme Learning Machine). This approach achieves an accuracy of 54.3%, significant outperforming the state of-the-art approaches with accuracy of only 45.1%. The advantage of the above approaches is they do not require users to wear any sensors on their bodies. However, the disadvantage is their limited usage scenarios and high processing overhead.

The second type of emotion recognition systems involves extracting emotion-related features from physiological signals [11], [12]. For example, Dana Kulic et al. [13] recognize user's emotion based on his/her heart rate, skin conductance, and corrugator muscle activity. Wei-Long Zheng et al. [14] train a DBN (deep belief network) with differential entropy features extracted from multichannel EEG as input. Then a Hidden Markov Model (HMM) is integrated to accurately capture a more reliable emotional stage switching. They achieve the accuracy of 87.62% on classifying two emotional categories (positive and negative) from EEG data. These methods, however, require users to wear specific devices such as heart rate belts or skin electrical sensors, which adds additional burden to the users.

Thirdly, another group of emotion recognition techniques rely on body movement and gesture expressivity of the user [15]. In these designs, non-propositional movement qualities, such as amplitude, speed and fluidity of movement, are extracted to infer emotions. For example, Stefano Piana et al. [16] extract a set of postural, kinematic, and geometrical features from a sequence of 3D skeletons of the users and feed them to a multi-class SVM classifier for classifying six emotions. The achieved overall recognition rate (61.3%) is very close to that achieved by human observers (61.9%).

This technique, however, is not applicable in mobile chatting scenarios where people usually hold the mobile phone without any special body gestures.

Different from these three methods, EmoChat makes use of facial expressions and text messages, which can be easily achieved in the chatting scenarios, for emotion recognition.

Benefiting from deep learning technologies, emotion recognition with facial expressions and texts has made significant progress recently [17]–[21]. Traditional facial expression recognition is usually based on handcraft features such as Gabor, LBP (local binary pattern), LGBP (local Gabor binary pattern), HOG (histogram of oriented gradient) and SIFT (scale invariant feature transform). Caifeng Shan et al. [22] illustrate the effectiveness and efficiency of facial expression recognition with LBP feature, with accuracy of 92.6% on Cohn–Kanade database achieved. Handcraft features, however, lack generalizability when applied to unseen images or those that are captured in wild setting. Ali Mollahosseini et al. [17] propose a deep neural network architecture to address the facial expression recognition problem across multiple well-known standard face datasets and show significant improvement compared with traditional recognition methods, proving the generalizability and effectiveness of deep neural networks on facial expression recognition.

As for emotion recognition with text messages, machine learning models with various methods of text vectorization, such as one-hot, TFIDF (term frequency inverse document frequency) are often taken into consideration. For example, Bo Pang et al. [23] apply three machine learning methods (Naive Bayes, maximum entropy classification, and support vector machines) to determine whether a movie review is positive or negative, finding that standard machine learning techniques definitively outperform human-produced baselines, while SVM has the best classification performance with an accuracy rate of 87.5%. The semantics of words and the structure of sentences, however, are ignored within these traditional methods. Deep learning methods, such as [24]–[26] take this information into consideration with Word2Vec [27], TextCNN [28] and LSTM [29] models.

In this paper, we leverage the light-weighted deep learning models for facial expression recognition and text emotion analysis, along with the multimodal fusion and context-aware algorithm, to achieve the inference of the current emotional status of online chatting users.

III. Background and Motivation

In this section, we first show the results of the survey about people's willingness in sharing their emotions. Then we conduct a set of experiments to show how different information (i.e., text message and facial expression) contribute to the emotion recognition process.

A. People's willingness to share emotions in online chatting scenarios

Although emotion information can enhance communication efficiency during online chatting, many people also consider their inner feelings as privacy. We design a questionnaire to investigate whether people are willing to share emotions in online chatting scenarios. In this questionnaire, we introduce the concept and design of emotion enhanced online chatting application, and tell that the application will capture the facial expression of users to analyze their real-time emotional status. Then we ask whom surveyed if they would like to share such emotion information with their families, lovers, intimate friends, ordinary friends and colleagues during chatting. At the end of the survey, we recycle 100 questionnaires, and the statistical results of the survey is shown in Fig. 2. It can be seen that most people are willing to share their true emotions with their lovers and intimate friends, and many people are also willing to share their emotions with their families. For ordinary friends and colleagues, people are less willing to share their emotions. This result indicates that people do have the need and willingness to share emotions in certain chatting relationships, especially in more intimate relationships.

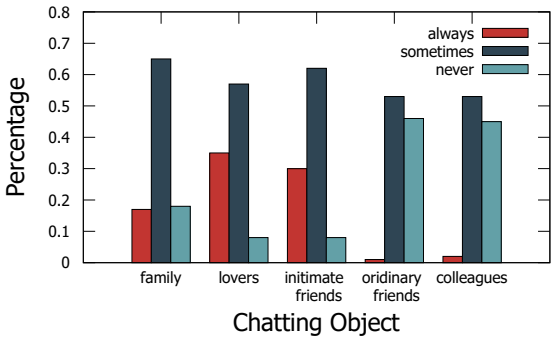
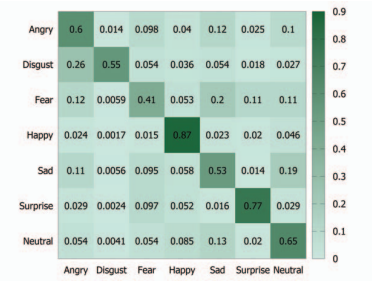


Fig. 2. Statistical results of the questionnaires.

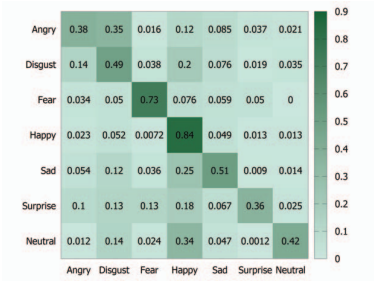
B. Complementarity of facial expressions and text messages in emotion recognition

There have been plenty of works using facial expressions to recognize emotions. The facial expressions under each emotion, are similar even for different persons. For example, when a person is happy, he/she will smile with corners of his/her mouth upward. With the rapid development of convolutional neural network recently, the facial expression recognition has reached a high accuracy. There are also lots of works using text information to recognize emotions. People often use similar words or sentence patterns to express the same emotion, thus the classification results can benefit a lot with machine learning methods and large sentence corpus.

However, although both these two pieces of information (facial expression and text message) provide important hints for emotion recognition, using either of them alone cannot provide satisfactory result. This is because both of them can cause certain ambiguity. As the example we showed in the beginning of this paper, when someone says "I scored 85 points", we can hardly infer whether he/she is happy or sad solely based on such text message. As for the facial expression, it is easy to confuse angry and sad since both these two emotions incur mow in facial expression. We further conduct a series of experiments to verify such ambiguousness.



(a) facial expression based method



(b) text message based method

Fig. 3. Confusion matrix of facial expression and text message based emotion category recognition methods.

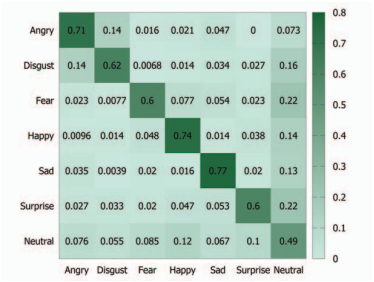


Fig. 4. Transition probability matrix of each emotion category.

Specifically, we build two emotion classification models based on facial expressions and text messages, respectively:

For facial expression based model, we use a light-weighted CNN [30], mini-Xception [31], for feature extraction and emotion classification. For the text message based model, we use TextCNN [28] model for classification. As for the dataset, we use the fer2013 dataset [32] to train/test the mini-Xception and use the public microblog benchmark corpus provided by NLP&CC 2013 to train/test the TextCNN. The images and sentences are labeled as one of these seven categories: Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral. Note that the first six emotion categories is in accordance with Ekman's six basic emotions theory [33], and neutral means the emotional status of the user is quite calm and not belong to the six basic emotions. In this paper, all of the "emotion category recognition" refers to classifying the emotion into one of these seven categories. On the fer2013 test set, a classification accuracy of 65.51% is achieved. The confusion matrix is shown in Fig. 3a. And on the NLP&CC 2013 test dataset, we achieve a classification accuracy of 62.61%. The confusion matrix is shown in Fig. 3b.

By comparing Fig. 3a and 3b, we find that facial expression based model exhibits higher accuracy in distinguishing positive and negative emotions since these two groups of emotions have quite different facial features. However, facial expression more likely to confuse the emotion in the same group. For example, facial expression based model seldom confuses "happy" and "sad", but always confuse "angry" and "sad". While, compared with facial expression based model text message based model is more likely to confuse emotions on different groups due to the language ambiguity. For example, it may confuse "disgust" and "happy", but it can accurately distinguish between "sad" and "angry".

In summary, text message and facial expression exhibit different accuracy in distinguishing different emotions. **So, these two pieces of information indeed complement each other and should be considered together for emotion recognition.**

C. Context information in emotion recognition

Existing emotion recognition method usually consider people's emotion at each time points independently. In this work, we find that context information is indeed helpful in emotion recognition. Specifically, people's inner emotion usually changes gradually. So, people usually keep a certain emotion state for several minutes to several hours before changing to another emotion state. In addition, the transition probability between different emotion is quite different. For example, the emotion sequence {disgust, angry, sad} obviously exhibit higher likelihood than the emotion sequence {disgust, happy, sad}. So, instead of considering each emotion in isolation, **we should treat each emotion in the context of its predecessors in the time series.**

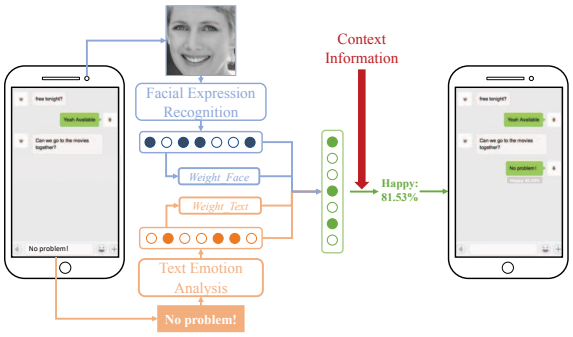


Fig. 5. Overview of the system architecture.

We further conduct a series of experiments to verify the above idea. Specifically, we invite 20 volunteers and ask them to select 10 dialogs from their recent conversation logs and label the sentences in the dialogs with one of the seven emotion categories. Then we calculate the transition probability between these emotion categories. For the transition probability matrix $A = [a_{ij}]_{N*N}$, a_{ij} represents the probability of transferring from the i-th emotion category to the j-th emotion category, $N = 7$. The transition probability matrix is shown in Fig. 4. As can be seen from the figure, the transition probability between different emotions are quite different. Therefore, using context information can help to improve the accuracy of emotion classification in online chatting scenarios.

IV. Overview

In this section, we introduce EmoChat, a new type of mobile online chatting application which can display the emotional status of the chatting object in real time during the chatting process. Fig. 5 shows the architecture of EmoChat. The

interface of EmoChat is very similar with a common online chatting application, but for each message, a hint of emotional status is attached below the message, as the example shown in Fig. 5 (the mobile phone interface on the right).

We achieve this using the combination of facial expressions of the user and text messages sent by the user. When the user is entering messages in the input box, the entering event will be captured automatically by the application. Then the application will call the front camera to capture the user's current facial expression. After getting the face area, the front camera will be closed to save energy. And then the achieved face area will be fed to the mini-Xception model to infer the probabilities of each emotion category or emotion polarity. We process the facial expression during the period when the user entering the text message, which usually takes a few seconds. When the user clicks the button to send the message, the message content will be achieved and submitted to the TextCNN model for emotional status inference. After the inference using the facial expressions and the text messages both complete, the fused probabilities will be calculated based on these two results. Then we fed a series of fused results to a Hidden Markov model to get the final results of the user's emotional status by considering the context information. Finally, the final results will be sent to the other side along with the text message, including the emotional status and the confidence of the inference.

V. EmoChat Design

A. Problem definition

Suppose there is a dialog $D = \{d_1, d_2, ..., d_T\}$, where d_i is a sentence belong to D. Select all of the sentences that are sent by the same user as $D_a = \{d_{a_1}, d_{a_2}, ..., d_{a_T}\}$. For each d_{a_i}, a facial expression f_i and a text message t_i can be observed, thus we can observe a sequence of facial expressions and text messages. In our system, we need to recognize the current emotional status of user using the observed face expressions and text messages. Thus, our problem is: given a sequence $S = \{(f_1, t_1), (f_2, t_2), ..., (f_T, t_T)\}$ and the possible emotional status set $C = \{c_1, c_2, ..., c_N\}$, find the most possible emotional status c_j for (f_T, t_T). There are two different standards to classify emotions. The one is to classify emotions into positive and negative, thus we call it "emotion polarity recognition". The other one is to classify emotions into detailed categories, such as angry, disgust, happy, surprise and so on, thus we call it "emotion category recognition", as mentioned in section III.B. In this paper, we take these two classification standards into consideration. So that C_1 = {positive, negative} corresponds to the emotion polarity recognition problem, and C_2 = {angry, disgust, fear, happy, sad, surprise, neutral} corresponds to the emotion category recognition problem. Below we use a general expression C to represent the emotional status set.

We solve this problem in three stages. Firstly, we analyze the facial expression f_T and text message t_T separately, and get $P(c_j|f_T)$ and $P(c_j|t_T)$ for each $c_j \in C$. Secondly, we fuse the $P(c_j|f_T)$ and $P(c_j|t_T)$ to get the fused probabilities $P(c_j|f_T, t_T)$. And finally, we take the context $P(c_j|f_1, t_1)$, $P(c_j|f_2, t_2)$ ... $P(c_j|f_{T-1}, t_{T-1})$ into consideration using Hidden Markov Model, and get the most possible emotional status c_j for d_T. Below we show the details of the three stages.

B. Emotion classification based on facial expressions and text messages

The development of hardware and deep neural network compression technology have made it possible to leverage convolutional neural networks on mobile devices. After obtaining the face area from the front camera using android API, we feed the face area into the mini-Xception model to get the inference. We take the mini-Xception model trained with the training set of the fer2013 dataset as basic model, and finetune the model with our own collected data. As for text messages, we use TextCNN to classify the utterances, and train the model with NLP&CC 2013 dataset. However, the NLP&CC 2013 dataset has few utterances labeled with fear and neutral, so that we add some fear and neutral instances until the number of instances of each category is almost the same. We also finetune the TextCNN model trained with the NLP&CC 2013 dataset using our own collected data.

From this stage, we get $P(c_j|f_T)$ and $P(c_j|t_T)$ for each $c_j \in C$ separately. We select to use the mini-Xception and TextCNN, because they are effective, efficient and light-weighted.

C. Multimodal information fusion

We have shown that, classification results got from facial expressions and text messages have advantages in different aspects in section III, thus we can get a better performance by combining these two results. One intuitive method to combine these two results is weighted averaging. However, how to determine the weights of these two results becomes a challenge. Compared to giving fixed weights to these two methods, giving weights to each piece of results is more meticulous and effective. To carefully measure the weight of each piece of results, we use the concept of *Information Entropy*. Information entropy quantifies the uncertainty of a random variable. Larger information entropy means greater uncertainty of the random variable and less confidence of the output value. The information entropy $H(X)$ of a discrete random variable X is defined as:

$$\begin{aligned} H(X) =& E[I(X)] = E[\log \frac{1}{P(X)}] \\ =& -\sum_{i=1}^{n} p(x_i) \log p(x_i) \end{aligned} \tag{1}$$

Here x_i represented one of the possible values of X, $p(x_i)$ represented the output probability of x_i.

In our emotional status classification problem, the possible value set of X is the emotional status set C mentioned above. We have already achieved $P(c_j|f_T)$ and $P(c_j|t_T)$ using facial expressions and text messages in stage one, thus the information entropy of the facial expression based result is:

$$H_f(X) = -\sum_{j=1}^{N} P(c_j|f_T) \log P(c_j|f_T) \tag{2}$$

and the information entropy of the text message based result is:

$$H_t(X) = -\sum_{j=1}^{N} P(c_j|t_T) \log P(c_j|t_T) \quad (3)$$

Since larger information entropy means less confidence, we use the reciprocal of the information entropy as the weight of the classification result. Thus, the normalized weight of facial expression based result is:

$$W_f = \frac{H_t(X)}{H_f(X) + H_t(X)} \quad (4)$$

and the weight of text message based result is :

$$W_t = \frac{H_f(X)}{H_f(X) + H_t(X)} \quad (5)$$

Finally, the confusion probability is:

$$P(c_j|f_T, t_T) = W_f * P(c_j|f_T) + W_t * P(c_j|t_T) \quad (6)$$

D. Context information

We have shown that context information can be very helpful for emotion classification of the current sentence in section III. Here we leverage the HMM (Hidden Markov Model) to take the context information into consideration. We first give a brief introduction of HMM, and then describe how we calculate the results using this model.

For the HMM model, firstly we assume that $Q = \{q_1, q_2, ..., q_N\}$ is the set of all of the possible hidden states, and $V = \{v_1, v_2, ...v_M\}$ is the set of all of the possible observed states, where N is the number of possible hidden states and M is the number of all possible observed states. For a sequence of length T, $H = \{h_1, h_2, ..., h_T\}$ is the hidden states sequence, $O = \{o_1, o_2, ..., o_T\}$ is the corresponding observation sequence. The HMM model makes two assumptions as follows. Firstly, HMM assumes that the hidden state at any time only depends on its previous hidden state. If the hidden state at time t is $h_t = q_i$, and the hidden state at time $t+1$ is $h_{t+1} = q_j$, the HMM state transition probability a_{ij} from time t to time $t+1$ can be expressed as: $a_{ij} = P(h_{t+1} = q_j|h_t = q_i)$. Thus a_{ij} forms the state transition probability matrix:

$$A = [a_{ij}]_{N*N},\ a_{ij} = P(h_{t+1} = q_j|h_t = q_i) \quad (7)$$

Secondly, the observation state at any time depends only on the hidden state of the current moment. If the hidden state at time t is $h_t = q_j$ and the corresponding observation state is $o_t = v_k$, then the probability that the observed state v_k is generated under the hidden state q_j is $b_{jk} = P(o_t = v_k|h_t = q_j)$. Thus b_{jk} forms the emission probability matrix:

$$B = [b_{jk}]_{N*M},\ b_{jk} = P(o_t = v_k|h_t = q_j) \quad (8)$$

In addition, we need to know about the probability distribution of the possible hidden states at time $t = 1$, that is, the initial hidden state probability distribution matrix:

$$\Pi = [\pi_i]_N,\ \pi_i = P(h_1 = q_i) \quad (9)$$

In summary, an HMM model can be determined by the state transition probability matrix A, the emission probability matrix B and the initial hidden state probability distribution matrix Π, . The HMM is used to solve three kinds of problems. Here we concern the most famous one, that is, given HMM models A, B, Π and a sequence of observations $O = \{o_1, o_2, ..., o_T\}$, find the optimal hidden state sequence. This problem can be resolved by the Viterbi decoding.

Inspired by HMM, we treat the emotional status of users as the hidden states and the facial expressions and text messages as the observations. Then, the emotion classification problem is transformed to finding the optimal sequence of emotional status given a sequence of facial expressions and text messages. Thus, the possible hidden states set Q is the possible emotional status set $C = \{c_1, c_2, ..., c_N\}$. The possible hidden states set V is the set that contains all of the combination of the facial expressions and text messages. The observation sequence O is $S = \{(f_1, t_1), (f_2, t_2), ..., (f_T, t_T)\}$. Note that we empirically select 30 minutes as a threshold to pick out the context information of the current sentence. That is, if the time interval between (f_i, t_i) and (f_T, t_T) is within 30 minutes, then (f_i, t_i) is considered as the context information of (f_T, t_T). The hidden states sequence H is exactly which to be found. And the crucial task is how to represent our problem in form of HMM ($i.e.(A, B, \Pi)$). Below we describe the modeling process.

Hidden State and Transition Probability. $C = \{c_1, c_2, ..., c_N\}$ is the possible emotional status set , and N is the number of different emotional status. The state transition matrix $A = [a_{ij}]_{N*N}$, where $a_{ij} = P(i_{t+1} = c_j|i_t = c_i)$, $c_i, c_j \in C$. To calculate the state transition matrix A, we use our own collected data mentioned before in section III. We ask 20 volunteers label their recent conversation logs with emotion category and emotion polarity. We count all of the transition relationships, and get n_{ij} represented the count that from the i-th emotional status to the j-th emotional status. And then we normalize n_{ij} for each i to get a_{ij}:

$$a_{ij} = \frac{n_{ij}}{\sum_{j=1}^{N} n_{ij}} \quad (10)$$

Initial Hidden State Probability Distribution. We count the number of occurrences of each emotional status in the data we use to calculate the transition probability. Take m_i as the number of occurrences of emotional status i,

$$\pi_i = \frac{m_i}{\sum_{i=1}^{N} m_i} \quad (11)$$

Observation and Emission Probability. We consider the facial expressions and text messages as the observation, thus the observation set V contains all of the facial expressions and text messages in the dataset. From the probability distribution achieved from the multimodal fusion results, we have got $P(c_j|v_k = (f_k, t_k))$. According to the Bayesian formula,

$$P(f_k, t_k|c_j) = \frac{P(c_j|f_k, t_k) * P(f_k, t_k)}{P(c_j)} \quad (12)$$

We make no assumption of the occurrence of any facial expressions and text messages, thus the probability of the occurrence of every f_k, c_k is equal. We take the initial hidden state probability as $P(c_j)$, so that

$$b_{jk} = \frac{P(c_j|f_k, t_k)}{\pi_j} \quad (13)$$

With A, B and Π calculated, we can find the optimal emotional status sequence by the Viterbi decoding.

VI. Experiments and Results

A. Data Collection

We develop an android application to collect facial expressions and text messages during online chatting. We find 20 volunteers with age ranging from 19 to 25 and divide them into 10 groups, asking each two volunteers in the same group to communicate with each other. To elicit different emotional status, we design six scenarios for volunteers to act, such as they quarrel because of trivial matters, or they plan to go to the movies. Our data collection process consists of two phases. In the first phase, the 20 volunteers communicate with each other according to the well-designed scenarios, and the application records their facial expressions and text messages during their chatting process. In the second phase, the volunteers label their own messages with emotion categories and emotion polarities, and then export the labels to the storage of the mobile phone. Finally, we obtain a total of 1495 samples, each sample contains one facial expression image and one text message, and is labeled with the emotion category and the emotion polarity.

B. Accuracy

1) Overall Accuracy: We take two different standards to classify emotions into consideration as mentioned in section V, that is, emotion polarity recognition (positive and negative) and emotion category recognition (six basic emotions and neutral). We use a common algorithm for these two standards. We take use of our own collected data to evaluate our algorithm, with 60% as the finetuning set and the other 40% as testing set. Here we show the accuracy of emotion recognition.

Table. I shows the accuracy of the emotion polarity classification and emotion category classification. Here F means facial expression, and T means text message. F + T means the method which fuses the information of both facial expressions and text messages, and F + T + Context means the method considering both the multimodal fusion result and the previous context information. We show that we get the overall accuracy of 76.25% for emotion polarity recognition and 51.64% for emotion category recognition. Taking the method with text messages only as baseline, the accuracy improves about 3% with multimodal fusion method and further improves about 5% with the context aware method for the emotion polarity recognition. And for the emotion category recognition, accuracy improves over 4% with multimodal fusion method and further improves about 7% with context aware method. It proves that taking facial expressions of chatting objects into consideration is effective for emotion recognition.

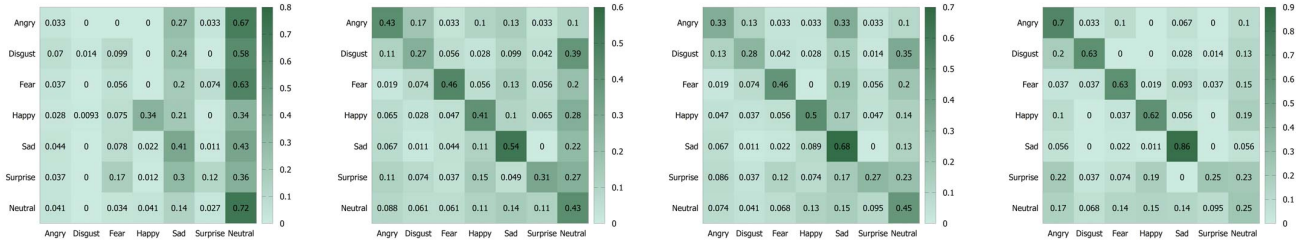
To be more detailed, we show the confusion matrix of emotion category classification in Fig. 6. Fig. 6a shows the confusion matrix of facial expression based method. The accuracy of recognizing angry, disgust, fear and surprise is obviously lower than that of recognizing happy, sad and neutral, while the accuracy of each category is similar on fer2013 dataset. That is because the amplitude of facial expressions is relatively small in online chatting scenarios, thus the facial expressions tend to be recognized to neutral. We can see that the multimodal fusion method truly combines the advantages of facial expressions and text messages (from Fig. 6a, Fig. 6b and Fig. 6c). For example, the accuracy of sad improves to 68% while the accuracy of facial expression and text message based methods is 41% and 54%, separately. From the confusion matrix of context aware method (Fig. 6d), we can also see that angry and sad can be easily recognized, with accuracy of 70% and 86% separately. Surprise is difficult to be recognized, that may be because data labeled with surprise is less. Neutral category also has a low accuracy, the reason may be that neutral can be transferred to or from any other emotion categories, thus the context information bring confusion for its classification. Compared with multimodal fusion method, the accuracy of most categories improves in the context aware method, proving that the context information plays a constructive role in emotion recognition.

TABLE I
Accuracy of emotion recognition.

Method	Polarity	Category
Facial expression	64.89%	35.11%
Text message	67.99%	41.14%
F + T	71.08%	45.78%
F + T + Context	**76.25**%	**51.64**%

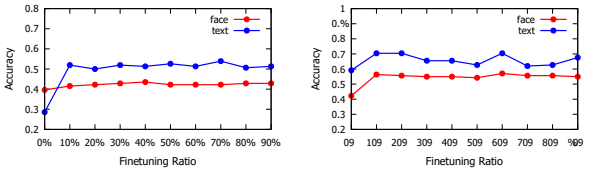
2) The effect of split ratio for finetuning: Severe loss of precision can be observed when transferring the pre-trained mini-Xception and TextCNN models to our own collected dataset since the training datasets and the testing dataset are not in the same data distribution. So that, we split our own collected data into "finetuning set" and "testing set", and finetune the mini-Xception and TextCNN models with the finetuning set. To evaluate the effect of the split ratio for finetuning, we simply take 10% data as testing set, and take the other 0%, 10%, ..., 90% data as finetuning set, separately. Fig. 7 shows the accuracy of emotion recognition with different finetuning ratio. Here 0% means taking no data for finetuning, and 90% means taking all of the rest data except for the testing set for finetuning. We find that the accuracy improves significantly even with only 10% data as finetuning set, for example, from 42.2% to 56.3% for facial expression and 59.1% to 70.4% for text messages on emotion polarity recognition, and the accuracy fluctuates slightly with different finetuning ratio. On average, the performance is the best with taking 60% of data as finetuning set, thus we randomly take 60% of data as finetuning set while the other 40% as testing set in our experiments.

3) The effect of the frequency of capturing the users' faces: Capturing the users' faces and deal with facial expressions are the most energy-intensive and time-consuming operations. Thus, we evaluate the effect of the frequency of capturing the users' faces, and the results are shown in Fig. 8. The x-axis means capturing the users' faces every X sentences, and the y-axis shows the accuracy after fusing both facial expression and text message based results. The accuracy



(a) Facial expression (b) Text message (c) F+T (d) F+T+Context

Fig. 6. Confusion matrix of emotion category recognition.



(a) Emotion category recognition (b) Emotion polarity recognition

Fig. 7. Accuracy of emotion recognition with different split ratio for finetuning.

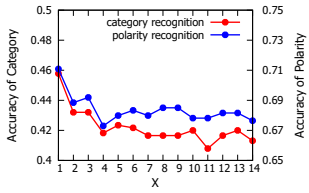


Fig. 8. Accuracy of emotion recognition with different frequency of capturing users' faces.

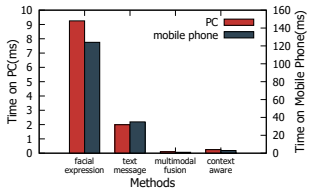
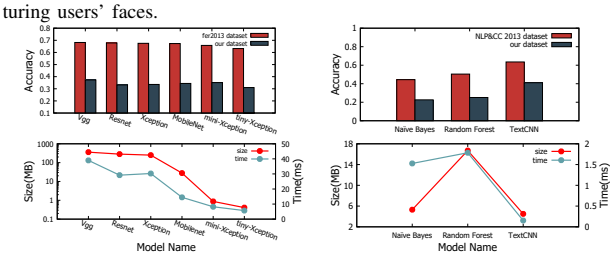


Fig. 9. Time consumption of different methods on PC and mobile devices.

continuously declines with X increasing, that is, taking fewer facial expression makes the accuracy reduce. The accuracy declines faster when X increases from 1 to 4, that is, from 45.8% to 41.8% for category recognition and from 71.1% to 67.3% for polarity recognition, and then the accuracy tends to be stable, close to the accuracy using only text messages. It proves that combining the information of facial expressions is truly effective although bringing extra consumption of time and energy.

C. Time Consumption

We measure the time consumption of our algorithm on both PC and mobile phone. Fig. 9 shows the time consumption of processing one piece of data of each stage. The processor of the laptop device is 3.1 GHz Intel Core i5, and the memory is 16GB. The type of the mobile phone is Xiaomi 5 with processor frequency 2.15GHz and memory 3GB. On the laptop device, processing one facial expression image costs the longest time, 9.25ms. The analysis of one sentence using TextCNN uses 1.99ms. The time cost of multimodal fusion is 0.1ms and the estimation of current emotional status with Hidden Markov Model is 0.25ms. We can see that the most time-consuming operation is image processing, using about 10ms. On the mobile phone, facial expression analysis also costs the most time, 124ms. And the text analysis costs 35ms. The time cost of multimodal confusion is less than 1ms and the estimation of current emotional status with Hidden Markov Model is 3ms. It takes lots of time to deal with the facial expression images, however, we continue trying to take photos using the front camera after the user starts typing the message, making full use of the typing time to process images in our system architecture. So that the time delay only consists of text messages processing, multimodal fusion and final prediction with context information. These three stages take less than 50ms, thus satisfying the real-time requirements of online chatting.



(a) Models for facial expression recognition (b) Models for text message analysis

Fig. 10. Performance of different models on emotion category recognition.

D. Unimodal Model Selection

We use the mini-Xception model to deal with facial expressions and TextCNN model to deal with text messages in our work, taking both efficiency and accuracy into consideration. We evaluate six models (Vgg, Resnet, MobileNet, Xception, mini-Xception and tiny-Xception) for facial expression recognition and three models (Naive Bayes, Random Forest, TextCNN) for text message analysis with the emotion category recognition task.

For facial expression recognition, we train these models with fer2013 dataset and then finetune them with our own collected finetuning set. The results of evaluation are shown in Fig. 10a. The upper subfigure shows the accuracy on the public dataset and our own collected dataset, and the lower subfigure shows the size and the time consumption for processing one image. It can be observed that mini-Xception achieves a relatively high accuracy with smaller model size and lower time consumption. Thus, we choose mini-Xception to deal with facial expressions.

Similarly, for text messages analysis, we train these models with the NLP&CC 2013 dataset and finetune them with our own data. Fig. 10b shows the performance of each model. The accuracy of TextCNN on NLP&CC 2013 dataset is much higher than Naive Bayes and Random Forest models. The

size of TextCNN is also smaller than the other two models. TextCNN is better than the other two traditional models on all aspects, thus we take TextCNN to deal with text messages in our algorithm.

VII. Conclusions

In this paper, we propose, design and implement EmoChat, a new online chatting mobile application enhanced with emotion information. EmoChat makes users know about each other's emotional status, thus they will feel more intimate with their conversation partners. We combine the output probabilities achieved from both facial expressions and text messages with the reciprocal of information entropy as weights, and use the Hidden Markov Model to take context information into consideration. Finally, we achieve the accuracy of 76.25% for emotion polarity recognition and 51.64% for emotion category recognition, and the time delay is within 50ms. We plan to improve the accuracy of the algorithm and reduce the time consumption in the future work. Besides, the emotion recognition algorithm proposed in this paper is also helpful for intelligent chatbots, making them more emotional and more humane.

Acknowledgment

This work is supported in part by the research fund of Tsinghua - Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] J. Aron, "How innovative is apple's new voice assistant, siri?," 2011.

[2] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of xiaoice, an empathetic social chatbot," *arXiv preprint arXiv:1812.08989*, 2018.

[3] H. Wang, H. Prendinger, and T. Igarashi, "Communicating emotions in online chat using physiological sensors and animated text," in *CHI'04 extended abstracts on Human factors in computing systems*, pp. 1171–1174, ACM, 2004.

[4] C. Ma, H. Prendinger, and M. Ishizuka, "Emotion estimation and reasoning based on affective textual interaction," in *International Conference on Affective Computing and Intelligent Interaction*, pp. 622–628, Springer, 2005.

[5] L. Dey, M.-U. Asad, N. Afroz, and R. P. D. Nath, "Emotion extraction from real time chat messenger," in *Informatics, Electronics & Vision (ICIEV), 2014 International Conference on*, pp. 1–5, IEEE, 2014.

[6] N. Sebe, I. Cohen, and T. S. Huang, "Multimodal emotion recognition," in *Handbook of Pattern Recognition and Computer Vision*, pp. 387–409, World Scientific, 2005.

[7] Y. Zhang, Y. He, J. Wang, Y. Kang, D. Liu, B. Li, and Y. Liu, "Share brings benefits: Towards maximizing revenue for crowdsourced mobile network access," in *2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pp. 1–9, IEEE, 2017.

[8] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[9] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[10] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth annual conference of the international speech communication association*, 2014.

[11] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, pp. 410–415, IEEE, 2011.

[12] J. Zhang, X. Zheng, Z. Tang, T. Xing, X. Chen, D. Fang, R. Li, X. Gong, and F. Chen, "Privacy leakage in mobile sensing: Your unlock passwords can be leaked through wireless hotspot functionality," *Mobile Information Systems*, vol. 2016, 2016.

[13] D. Kulic and E. A. Croft, "Affective state estimation for human–robot interaction," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 991–1000, 2007.

[14] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu, "Eeg-based emotion classification using deep belief networks," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2014.

[15] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *International Conference on Affective Computing and Intelligent Interaction*, pp. 71–82, Springer, 2007.

[16] S. Piana, A. Stagliano, F. Odone, A. Verri, and A. Camurri, "Real-time automatic emotion recognition from body gestures," *arXiv preprint arXiv:1402.5047*, 2014.

[17] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–10, IEEE, 2016.

[18] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 435–442, ACM, 2015.

[19] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2983–2991, 2015.

[20] Y. He, J. Liang, and Y. Liu, "Pervasive floorplan generation based on only inertial sensing: Feasibility, design, and implementation," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1132–1140, 2017.

[21] X. Chen, X. Wu, X.-Y. Li, Y. He, and Y. Liu, "Privacy-preserving high-quality map generation with participatory sensing," in *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pp. 2310–2318, IEEE, 2014.

[22] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[23] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86, Association for Computational Linguistics, 2002.

[24] D. Li and J. Qian, "Text sentiment analysis based on long short-term memory," in *2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI)*, pp. 471–475, IEEE, 2016.

[25] Y. Zhang, J. Fu, D. She, Y. Zhang, S. Wang, and J. Yang, "Text emotion distribution learning via multi-task convolutional neural network.," in *IJCAI*, pp. 4595–4601, 2018.

[26] C. Jiang and Y. He, "Smart-dj: Context-aware personalization for music recommendation on smartphones," in *2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 133–140, IEEE, 2016.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[28] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[31] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," *arXiv preprint arXiv:1710.07557*, 2017.

[32] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*, pp. 117–124, Springer, 2013.

[33] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.