Architecture of cross-platform videoconferencing system with automatic recognition of user emotions

Ilya Duboyskii¹, Aleksandra Shabanova², Oleg Sivchenko², Elizaveta Usina²

¹Saint-Petersburg State University of Aerospace Instrumentation, 190000, Saint-Petersburg, Russia
²Saint-Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, 199178, Saint-Petersburg, Russia

E-mail: lizzzi96@mail.ru

Abstract. This paper considers implementation of the technology of automated detection of emotional condition and video conferencing technology for remote content delivery, such as transport communication systems, polls, lectures, psychotherapy sessions, etc. To establish remote communication sessions a platform-agnostic peer-to-peer architecture was developed. Convolutional neural networks are used for stream processing at the operator end, being utilized to estimate the emotional feedback of the customer. To define the emotional condition, three modalities (video, audio, text), as well multimodal recognition were used. Experiments for 10 pairs of humans were performed, where one of them acted as an operator and asked closed questions, whereas another answered these questions. The neural network shows the following average accuracy values for the individual modalities: video 76 %, audio 57 %. The best output is ensured by multimodal recognition (average accuracy of 80 %). These findings confirm the efficiency of multimodal recognition in videoconferencing systems for classification of human emotions.

1. Introduction

Due to the global epidemiological situation, present nowadays, videoconferencing systems (VS) became highly demanded alternative to business travel and duty trips. This technology enables communication sessions among two or more remote subscribers using technical solutions for transfer of audio-visual information in real-time mode [1]. This technology is used in online learning, online patient consulting in healthcare [2], as well in social polling. So, in [2] successful application of VS is described, where VS is instrumental in session management among branches of a healthcare corporation, as well for online learning and staff performance assessment. In the applied study [3] authors also confirmed, that the quality of knowledge, learned by a small group of medical students with the aid of VS is no less than in face-to-face learning. During final assessment the students from the research group were equally successful as the students of the same specialty in the face-to-face fashion. VS systems can be deployed not only on desktop and mobile computers, but also with the aid of mobile robots [4-6].

In healthcare such mobile robots can be utilized, which act as technical support layer for VS establishment between doctor and patient [4]. In [5] it was confirmed, that utilization of VS-enabled robotic systems in urology meets the same standards of diagnostic quality, comfort for doctor and patient, safety and decision-making quality, as the conventional face-to-face approaches. Thereby the
remote medical rounds ensure more efficient healthcare logistics and optimization of accompanying costs. Notwithstanding that, VS systems, used in healthcare, are criticized for a number of reasons: usage of proprietary technologies, incompatible with alternative devices or software, expensiveness as of the solution itself, as its maintenance. For operation of such system enough high-qualified employees are required [7].

During human communication via VS often such situations arise, where human emotions have to be recognized in real-time mode to detect one’s actual emotional condition. Implementation of automatic recognition of emotional condition within the remote attendance session enables to reveal early depression symptoms in patients [8]. In the servicing domain, where the VS is used for order checkout and customer polling, provides for conflict prevention, improve quality of customer service, as well optimize communication processes and product delivery [9].

In [10] to simplify and ease human-robot interactions, a real-time emotion recognition system was proposed. Authors of [11] to optimize the possibilities of human-robot interactions, presented a hybrid approach to increase emotion recognition accuracy in texts via merging of linguistic, pragmatic and keyword spotting techniques. The paper [12] puts emphasis on neural network-based intelligent recognition of facial emotions; thereby the accuracy of recognition of the seven basic emotions reached 71.3%. The presented paper is very promising for development of personified intelligent agents, i.e. robots with emotional and social intelligence.

Despite that, the VS systems, employed in these workflows, lack an embedded module for human emotion recognition. The majority of available peer-to-peer VS systems have proprietary architectures, what precludes the extension of existing data analysis architectures by the means for processing of different modalities. Therefore, it is relevant to design an architecture for a platform-agnostic VS system, which would allow to recognize human emotions in real-time mode, based on data, processed by neural networks for different modalities.

2. Methods of emotional condition detection

Emotions reflect subjective significance of some phenomenon, event or thing in real world for a certain individual. Different approaches to emotion categorization exist. One of the most common concepts is the framework of basic emotions [13], which assumes retrieval from the whole range of emotions a small subset (usually, 6-7) basic atomic emotions. Thereby more complex emotions within this framework are composed of basic ones [14]. It is also noteworthy, that different frameworks for classification of emotions rely on different premises, such as generic culturally-independent manifestations [15-17], unconditional reflex [18-21], correspondence to evolutionally conditioned adaptation processes [14], universal physiological manifestations [22-24] etc. The following emotions are considered basic: neutral mode, joy, sadness, anger, surprise and disgust. These emotions can be divided into two broad subsets: positive and negative ones, and this distinction allows to characterize human emotional condition. So, in [25] authors treat joy as a positive emotion, whereas anger (aggression), disgust, sadness as negative ones. If a human exhibits positive emotion, it can be treated as satisfaction, otherwise as discontent.

Using various communication devices, it is possible to transmit information in several ways, also denoted as «modalities» [16]: video, text and audio. Every modality can be used to recognize human emotions. The most successful results of emotion recognition, compared to other machine learning approaches is pertinent to neural networks [27]. Particularly, convolutional neural networks show the best accuracy and speed among all algorithms for object retrieval in the scene [28]. As the main objective of this research is the development of a platform-agnostic VS system, it was decided to rely on out-of-the box solutions in the development of the module for human emotion recognition. By analysis of the existing emotion detection methods, the availability of a free pretrained neural network model was of primary concern.

Emotion recognition based on video modality is considered in papers [29-31]. Within the approach of [29] and [30] the emotion recognition accuracy, based on JAFFE dataset, makes up 85.7% and 93.8% respectively, in [31] the recognition accuracy, based on dataset Cohn-Kanade, makes up
95.1\%. In [32] a method for speech-based emotion recognition is proposed, which uses audio modality. It assumes utilization of deep neural networks, which ensure emotion recognition accuracy of 84.3\%. Authors noted, that the outcomes were unacceptable, and further method refinement was necessary. Textual modality is required for more elaborated classification of ambiguous emotions.

Based on the performed review, the following publicly available pre-trained models were selected: for video modality, Face Expression Recognition Model [33] a convolutional neural network, trained on the JAFFE dataset. This network uses deeply separated convolutions and condensed blocks, thereby recognizing 7 basic emotions. For emotion recognition in audio modality, the model Speech emotion analyzer [34] is used, trained on dataset RAVDESS and SAVEE of 2000 audio files, suitable to detect 8 different male and female emotion manifestations in voice analysis. For recognition of words, extracted from audio files, Google Cloud Speech API was used.

To increase accuracy of emotion recognition, based on single modality, it is proposed to enhance it with subsequent evaluation of emotional condition across several modalities, according to the following scenario: when modalities overlap, emotions are defined according to the analysis of correlation, which respects facial expression and speech spectrogram. The recognition model assumes, that the obtained data have to be processed separately: convolutional neural network for voice-based emotion recognition with attention mechanism outlines the fragments of the spectrogram, that correspond to the manifestation of the emotion; in video, specifically, in the image sequence without audio stream, features are revealed (also with the aid of neural network) features are extracted, which reflect the emotional condition. Further the emotional-related features, revealed from speech, are compared with the features, obtained by voice recognition (Factorised Bilinear Pooling – FBP) to reveal correlations among them for final recognition of emotions.

3. Architecture of application for videoconferencing

Provided, that the platform-agnostic operability of the VS system has to be ensured, the VS architecture was designed for portability and fast deployment of system parts on the devices, used as by customer, as by operator. Because system malfunction cannot be excluded due to inconsistency of hardware architectures of the devices, on which it runs, it was decided to develop a web application architecture.

Currently one of the main technologies for in-browser voice calls is the WebRTC protocol (Web Real Time Communication). The WebRTC protocol enables connections among two or more clients for multimedia data transfer without proxy server. Figure 1 presents the architecture of the developed application.

Before using the application, registration or login is necessary. To achieve this, a registration and authentication module is provisioned in the architecture; this module sends user credentials to the authorization and registration server. This server, in turn, accesses the database. Depending on the success of the validation, the access into the system is granted or denied. To obtain audio-visual data via video camera and microphone, on the client side the module of data capture from device is used. This data is further fed into peer-peer data exchange module, which enables direct traffic transfer between two nodes. The connectivity management module is needed to detect the “caller” and the “callee”. After the modules, managing the connections between client and operator, have performed the handshake, the peer-to-peer connection is established according to the WebRTC protocol. Then the multimedia data exchange proceeds between the peer-to-peer modules directly.
The operator also owns a data processing module, aided by neural networks, which receives streaming audio and video from the peer-to-peer data exchange modules, processes these streams and establishes the resulting emotional response across different modalities, which then goes to the data output module.

4. Approbation of the designed architecture
For approbation of the designed architecture a web application was developed, which ensures connection between two users. The graphical user interface of this web application is presented in figure 2.
The developed web application has an intuitive and simple graphical interface, accessible for anybody, even not involved in IT domain. To initiate the streaming of local media content, the user has to click the “Start” button. Then in the left window, shown with dashed line, a local video stream from the webcam of the operator will be displayed. To establish the connection with the remote user, the respective nickname should be clicked, after which the image from the webcam of the client appears.

To estimate the accuracy of emotion recognition with the aid of the developed videoconferencing system an experiment was performed, which simulates the survey of restaurant guests, performed with a mobile robot [35]. For this a list of closed questions was composed, assuming unambiguous answers «yes» or «no». The operator asked these questions to the guests remotely through the developed videoconferencing application, run on the mobile robot. Here goes the list of questions, addressed to the guests: «Did you like your order?» «Was everything okay?» «Did you like the service?» «Did you like the music?» «Was the order served timely to you?» «Did you like the décor of the place?» «Did you like the prices?» «Would you like to recommend this place to your friends?» «Would you come here again?» «Would you like to celebrate some events here?» After survey, performed by the operator, the users are also asked to leave a textual feedback on service quality. The answers of users on the proposed questions, asked by the operator, as well parsing of the textual feedback allow to deduce, whether the emotion recognition results, pulled from the developed system, correspond to the actual human responses on their level of satisfaction.

To simulate the restaurant service process, an environment was established with places for operator and guest. 10 respondents were invited, whereby half of them received poor service and another half – good one. After service session, a robot approached to humans, and the operator asked the aforementioned questions in real-time mode. The answers, that followed, were processed by a neural network, and the result was given. For the sake of experiment validity, the respondents were asked to reply in natural fashion. The examples of emotion recognition, based on video modality, are given in figure 3.
Figure 3 illustrates different emotions, expressed by guests, after receiving service of different quality. In figure 4 the values of recognition accuracy are given across different modalities. To detect accuracy, the frames from video sequences and instants of audio recordings were considered, where the respondents replied the questions. If the answer was positive, and the system treated the detected emotion as positive one, it was deemed, that the system works correctly, otherwise incorrectly. Concerning the relatively mild emotions (surprise, neutral mood) any reply was deemed to be correctly processed by default. The classification accuracy outcomes are presented in figure 4.

Figure 4. Accuracy of emotion classification across different modalities.

As follows from figure 4, video and audio modalities are the best choice to recognize neutral mood, sadness and joy, with the respective accuracy levels over 83\%, whereas the accuracy of recognition of the other emotions does not exceed 73\% for video modality and 45\% for audio modality. The respondents experienced no fear, so in all experiments the share of fear equals zero. The emotion classification accuracy in multimodal recognition exceeds the respective accuracy levels for individual modalities from 2\% to 66\%, ignoring the accuracy of recognition of neutral mood. So, the accuracy of multimodal surprise recognition is 5 times higher, compared to audio modality alone. The average accuracy of emotion recognition for audio and video modalities was 57\% and 76\% respectively, whereas the multimodal recognition resulted in accuracy level of 80\%. In figure 5 the percentage of emotions, expressed by respondents, is given.

Figure 5. The pie chart of the expressed emotions.
As evident from figure 5, the most respondents replied the questions neutrally, what complicates the interpretation of their answers in the view of level of guest satisfaction with the delivered service. This is because the neutral mood replies, as well the ones with surprise, can be loaded as positively, as negatively, i.e., it precludes unambiguous estimation of customer satisfaction. In contrast, all the other emotions are easier to estimate the satisfaction level, because they are expressively loaded. Hence, if customer satisfaction has to be estimated based on the single modality, the definitive conclusion is almost impossible to draw.

Figure 6 presents the accuracy of classification of positive and negative replies. The reply classification accuracy across the whole range of modalities is 8.5% higher in average, than for negative modalities only. This is because disgust and anger detection accuracy is two times poorer in average, as more facial muscles are involved in the expression of these emotions, what impairs the overall recognition accuracy.

Figure 6. Classification accuracy of positive and negative replies across modalities.

As follows from figure 6, the classification accuracy of positive and negative replies in multimodal recognition is higher, than for individual modalities, and it varies in the range between 2% and 12%.

Hence, as per the obtained results, by multimodal emotion recognition the classification accuracy is higher, than in individual modalities (audio and video). This evidence support the efficiency of multimodal techniques of emotion recognition in videoconferencing systems.

5. Conclusion
Within the presented research a web application was developed for customer satisfaction estimation, which uses neural networks for processing of three content modalities: audio, video, text, as well for multimodal recognition. The application ensures human communication via Internet-enabled devices. It has a peer-to-peer architecture, which ensures direct exchange of multimedia data between clients. The approbation of the developed application showed, that, in multimodal emotion recognition, the accuracy was 2-64% higher, than by involvement of individual modalities only (audio or video). The classification accuracy of positive and negative emotions in multimodal context is 7% higher in average, than for individual modalities.

The developed system can be applied in different domains, such as service delivery and healthcare, where human emotion recognition in real-time mode is required. The application still has some downsides, though; currently it supports only one-to-one user connections. Further it is expected to extend the application functionality to increase the number of simultaneous user connections.

References
[1] Karasev E, Vatamaniuk I, Saveliev A and Ronzhin A 2018 Management Information Systems 1
[2] Gerasimov A E, Bessoltsev M A and Korsunov L C 2017 J. Clinical practice 1 18
[3] Allen M, Sargeant J, Mann K, Fleming M and Premi J 2003 J. of Continuing Education in the Health Professions 23(1) 38
[4] Akhmedov R R and Ivanyugin V M 2017 Scientific journal “Commonwealth” 14 201
[5] Sivkov A V 2015 Experimental and Clinical Urology 4 128
[6] Vatamaniuk I and Saveliev A 2017 Extreme Robotics 1 37
[7] Jang-Jaccard J, Nepal S, Celler B and Yan B 2016 Computing 98(1-2) 169
[8] Vrijen C, Hartman C A and Oldehinkel A J 2016 European child & adolescent psychiatry 25 1 1255
[9] Varavva M Yu and Radchenko A V 2015 New Science: from idea to result 6(1) 39
[10] Liu Z, Wu M, Cao W, Chen L, Xu J, Zhang R, Zhou M and Junwei J 2017 EEE/CAA Journal of Automatica Sinica 4(1) 37
[11] Jang-Jaccard J, Nepal S, Celler B and Yan B 2016 Computing 98(1-2) 169
[12] Zhang L, Jiang M, Farid D and Hussain M A 2013 J. Expert Systems with Applications 40(13) 5160
[13] Ortony A, Turner T J 1990 Psychological review 97(3) 315
[14] Plutchik R 1980 Theories of emotion (Academic press) p 3
[15] Ekman P 1982 Emotions in the human face p 39
[16] Frijda N H 1986 Cambridge University Press
[17] Scherer K R, Johnstone T and Klasmeyer G 2003 Handbook of affective sciences p 433
[18] Gray J A 1982 Behavioral and Brain Sciences 5(3) 469
[19] Izard C E 1971 Appleton-Century-Crofts
[20] Panksepp J 1982 Behavioral and Brain sciences 5(3) 407
[21] Watson J B 1930 University of Chicago Press
[22] James W 1884 Oxford Journals 9 188
[23] Tomkins S S 1984 Approaches to emotion 163 163
[24] Lange C G and Kurella H G 1911 F. Alcan
[25] Kwon D S et al. 2007 RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication (IEEE) p 351
[26] Levonevskii D, Shumskaya O, Velichko A, Uzdiaev M and Malov D 2020 Proceedings of 14th International Conf. on Electromechanics and Robotics “Zavalishin's Readings” (Springer, Singapore) p 511
[27] Levonevskii D, Shumskaya O, Velichko A, Uzdiaev M and Malov D 2020 Proceedings of 14th International Conf. on Electromechanics and Robotics “Zavalishin's Readings” (Springer, Singapore) p 511
[28] Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou M A, Schuller B and Zafeiriou S 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
[29] Verma K and Khunteta A 2017 International Conf. on Information, Communication, Instrumentation and Control (ICICIC) (IEEE) p 1
[30] Salmam F Z, Madani A and Kissi 2018 International J. of Electrical and Computer Engineering 8(1) 52
[31] Kotsia I and Pitas I 2006 IEEE transactions on image processing 16(1) 172
[32] Badshah A M, Ahmad J, Rahim N and Baik S W 2017 International Conf. on Platform Technology and Service (PlatCon) (IEEE) p 1
[33] Face-api.js https://github.com/justadudewhohacks/face-api.js/
[34] Speech-Emotion-Analyzer https://github.com/MITESHPUTHRANNEU/Speech-Emotion-Analyzer
[35] Pavliuk N, Smirnov P and Kovalev A 2019 Izvestiya TulGU Technical science 10 181