An Introduction to Recent Mobile Affective Inference Techniques: Methods, Applications and Challenges

Ou Changkun

Abstract— Affective Computing has been considered as one of the essential aspects of massive human-computer interaction related projects. This paper provides a broad introduction to recent advances in affective computing for emotion inference based on mobile techniques. We expand the emotion inference principles for different mobile commodity sensors, application context and their possible combinations in recent researches. Then we compare recent popular machine learning methods and models for these sensors, and highlight the most useful techniques and models for their performance. Our comparisons are not limited to traditional machine learning algorithm; they also include the representation learning models. In the end, we discuss few novel applications based on mobile affective computing techniques, such as how adaptive user interface and usability testing work in an emotion-aware system, as well as current limitations and open challenges of this research area.

Index Terms—Mobile Emotion Inference, Deep Networks, Support Vector Machine, Dialogue System, Adaptive User Interfaces

1 Introduction

Affective computing is an emerging interdisciplinary research field ranging from cognitive and social sciences to human-computer interaction (HCI) researchers with techniques like computer vision, machine learning, natural language understanding, etc. With the long-term research on emotion theory from psychology and neuroscience [38, 87], emotion has been confirmed to be a significant effect [39] on human communication, decision making and perception.

On the perspective of human-computer interaction, Picard [65] pointed out that affective computing (involved projects) can be used for *reducing user frustration*, enabling comfortable communication of user emotion, developing infrastructure and applications to *handle affective information*, as well as building tools that *help formulate social-emotional skills*.

Recently, the ubiquitous computing [89] and wearable computing [79], which are strictly related to affective computing, have achieved the pervasive attention of scientists. Ubiquitous computing and wearable computing are the necessary products of the combination of mobile computing technology and computer individualization.

Hence, it creates new research opportunities for affective computing that inferring user emotions with the combination of smart mobile wearable devices. We simply call it *Mobile Affective Inference*. Such devices have been widely used over the world. The key feature of smart devices is the abundant sensors that enable unobtrusive monitoring of various affect-related signals. Exploring the possibility of using smart mobile devices for affective computing will benefit at least three perspectives by the potential of long-term unobtrusive monitoring users affective states: First, it influences the affect-related research literature by the wild, natural and unobtrusive study; Second, it establishes the spontaneous affect databases efficiently to evaluate new effective methods, models, systems more accurately; and third, it enhances the user-centered HCI design for the future ubiquitous computing environments.

In this paper, we present an introduction to the mobile affective computing techniques. Our next section discusses existing data sources from mobile devices, and Section 3 illustrate the recent advances for each different type of data and present the state-of-the-art model and method. Next Section 4, based on the previous information (we reasonably assume we have finished user emotion inference stage), then gives two typical HCI applications in this field. At last, Section 5 and 6 discuss the current challenges of mobile affective inference and our conclusion of this introduction. Figure 1 shows the hierarchically-structured taxonomy of this paper.

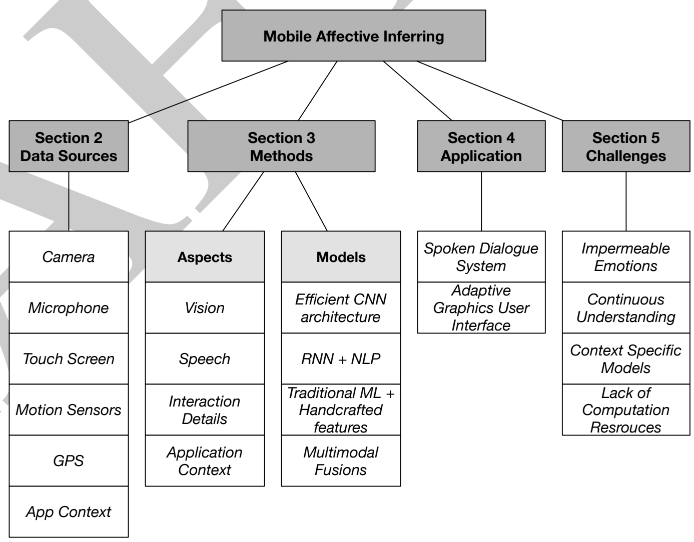

Figure 1. Hierarchically-structured taxonomy of this paper.

2 Data Sources

Previous surveys [1, 92, 69, 24] put forward that different data sources should be applied to various modeling methods in multimodal affective computing. Poria et al. [69] also give the argument that 90% literature consider visual, audio and text information as multimodal affect analysis instead of other dimensions by their extensive literature review.

In this section, we discuss the commonly existing data sources in a typical smartphone and prepare for the modelling method in the next section. Figure 2 illustrates the data flow from sensors to user emotion state.

2.1 Camera

We emphasise vision sensors in the first place since face and facial expressions are undoubtedly one of the most critical nonverbal channels used by humans to convey internal emotion [39, 69]. This part mainly discusses vision sensors, which includes RGB cameras and depth cameras, and illustrates how vision sensors can be used for affective emotion inference.

- *Ou Changkun is studying Human-Computer Interaction at the University of Munich, Germany, E-mail:* `hi@changkun.us`
- *This research paper was written for the Media Informatics Advanced Seminar "Human Computer Interaction", 2017*

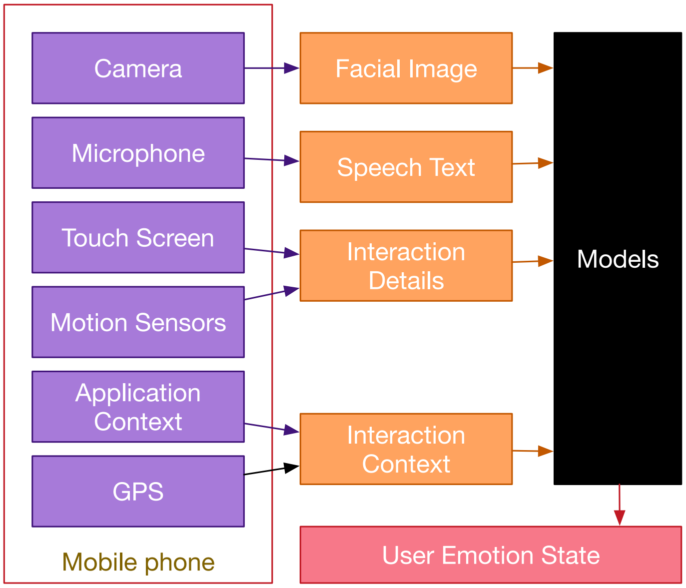

Figure 2. Data sources can be used in emotion inference which provided by the most of commercial mobile phone devices.

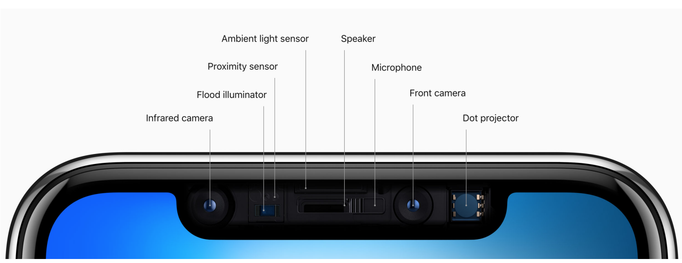

Figure 3. TrueDepth Camera in iPhone X

Pure RGB cameras have been widely used in a commercial smartphone as an image sensor. For the camera with depth information on mobile devices (recently introduced TrueDepth Camera in iPhone X, see Figure 2.1[1]) combines infrared camera, flood illuminator, proximity sensor, ambient light sensor, front-facing camera and dot projector to provide depth images of facial information of a user.

2.2 Microphone

Audio sensors usually refer to built-in microphones; they collect voice information from current environments, which can infer user emotions based on their speech content.

Before recognizing user speech, a system usually should take care and preprocess the environmental noise and detect acoustic fingerprint [6] (i.e., voiceprint) for the current user, isolate their speech from mixed audio information.

Inferring emotions from user speech can be split into two parts; The first part is recognizing the speech text from the user [56, 88], then understanding or inference from the text, namely sentiment analysis [71].

2.3 Touch Screen

Human emotions can be expressed in different ways. Emotional communication has focused predominantly on the facial and vocal channels but has ignored the tactile channel [32], researchers investigated the possible expressions of user emotion in detail while they are using mobile devices with touch screen.

A capacitive touch screen provides touch position, touch pressure, touch angle through time. Among the subsequent researches [23, 76, 60, 3], researchers explored how human emotions can be inferred by capacitive touch channel in a specific application context based on these features. This is typically done only with touch screen interface.

[1]https://www.apple.com/iphone-x/#truedepth-camera

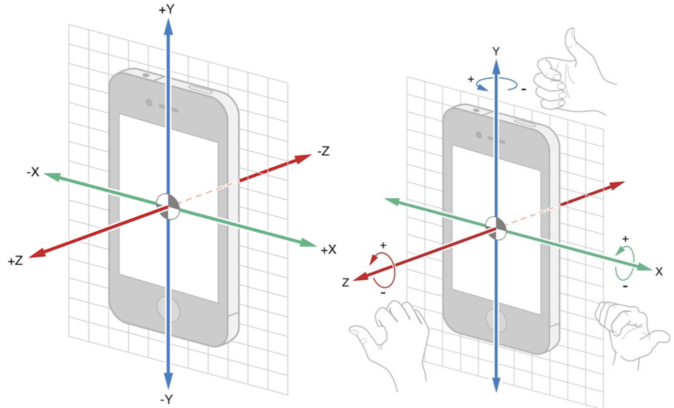

Figure 4. Coordinates information of Accelerometer (left) and gyroscope (right) as motion sensors in iOS CoreMotion framework.

Interestingly, 3D touch screen was largely introduced in commercial devices a few years ago. Some of the researches investigated the possibility of haptic based application [18]. Mazzoni [52] and Lentini [49] showed how a system with haptic touch response essentially can express and influence user emotions. Bhattacharya [4] concludes that haptic-based affect detection remains an understudied topic.

2.4 Motion Sensors

Motion sensors typically combine gyroscope and accelerometer, which are yet another interaction detail information [92]. An accelerometer measures proper acceleration (acceleration it experiences relative to free fall), felt by people or objects. Most smartphone accelerometers trade large value range for high precision. The gyroscope can be a handy tool in peculiar ways and defines gravity. Gyroscopes have been around for a century, and they are now used everywhere from airplanes, toy helicopters to smartphones. A gyroscope allows a smartphone to measure and maintain orientation. Gyroscopic sensors can monitor and control device positions, orientation, direction, angular motion, and rotation. Figure 4 shows the coordinates information of accelerometer and gyroscope sensors.

With these motion sensors, interaction details information such as device holding posture, device moving trajectory can be inferred from these sensor data [60, 69].

2.5 GPS

GPS sensors provide geographical information of a user, and detect the location of the smartphone using 1) GPS [84]; 2) Laceration/Triangulation of cell towers or wifi networks (with a database of known locations for towers and networks) [72]; 3) Location of an associated cell tower or WiFi networks [68].

However, GPS will not work indoors and can quickly kill the battery. Smartphones can try to automatically select best-suited alternative location provider (GPS, cell towers, WiFi), mostly based on desired precision. With the location, we can study the relationship between life patterns and affective states. For example, most people in playground feel happy while most feel sad in a cemetery.

The location provides additional information to verify the subjective report from participants of affective studies. It may also help to build a confidence mechanism [84] for subjective reports. Attaching the location to the subjective report would produce confident weights to measure the significance of collected subjective reports. For example, a participant reported that he was happy in a cemetery. But, in common sense, people in the cemetery would be sad. Thus, we could set a low weight as a confidence value to the report, or ask the user is indeed either choosed a right option or not.

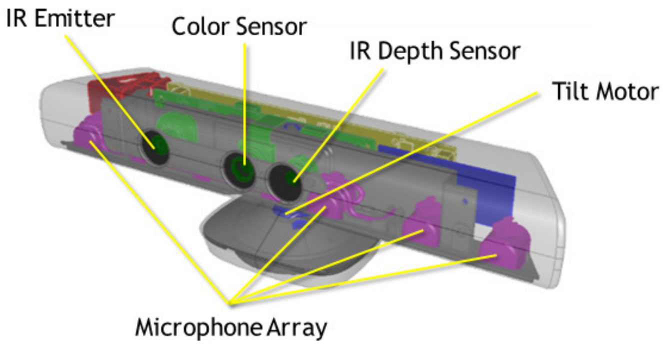

Figure 5. Principle of Microsoft Kinect.

2.6 Application Context

As we slightly mentioned in the previous section, most of the mobile affective inference techniques based on a touch screen and motion sensors[23, 76, 60, 3] are based on specific application context, for example, application user interface. This kind of information can be another confidence mechanism in a inference system. Vice versa, the inference systems are only modeling for a specific context. However, this could be a drawback of multimodal emotion inference since a complete system will integrate more models for emotion inference.

3 Methods

Emotion inference problems are mostly considered as a classification problem, which classifies three states of user emotion: Happy, Unhappy, Neutral. The reason for this consideration lies in technical constraints: the more types of emotion we need to classify, the more data we need (to prepare). In this section, we will see the most accurate specialized method, models and their datasets in different data type aspects mainly all based on machine learning methods.

3.1 Vision Aspect

The standard RGB camera focusing on how to conduct emotion recognition with RGB images. Through depth camera was recently introduced on a commercial mobile phone, its principle is as same as Microsoft Kinect (see Figure 5). Considering these two different sensor aspects, we dive into two distinct research area on vision sensors.

3.1.1 Plain recognition

Recently, convolutional neural networks (CNN) method has successfully made break-through contributions to computer vision as well as its application to emotion inference. AlexNet [47] popularized deep convolutional neural networks by winning the ImageNet Challenge (a large-scale image classification challenge). Subsequently, other powerful CNN architectures were proposed such as VGG [77], Inception series [81, 82, 83], ResNet [31], DenseNet [36] and CapsNet [74]. However, the most accurate CNN's usually have hundreds of layers and thousands of parameters, which is (entirely) not possible to deploy on a mobile system. The increasing of mobile emotion inference needs of running high quality deep neural networks on embedded devices encourage the study of efficient model designs [30].

SqueezeNet [37] is the first model that reduces parameters and computation significantly while maintaining accuracy. MobileNet [35], ShuffleNet [93] and Xception [10] utilize the depthwise separable convolutions among lightweight models. Table 1 shows the complexity comparation of these CNNs.

The classification of RGB images is just the final step of feedforward propagation. To determine the facial information inside an image essentially become more difficult than only a classification. Then the problem refers to Landmark detection. The recent break-through contributions in this area are the Region-based CNN (R-CNN) approach [26] that propose bounding-box object detection, which is attend to a manageable number of candidate object regions. The state-of-the-art contribution is Mask-R-CNN [29], which is a conceptually simple, flexible, and general framework for object instance segmentation.

Table 1. Complexity comparison of CNN models, smaller number indicates good performance

Model	Cls. Error (%)	Complexity (MFLOPs)
AlexNet [47]	42.8	720
VGG-16 [77]	28.5	15300
SqueezeNet [37]	42.5	833
MobileNet [35]	31.6	325
ShuffleNet [93]	31.0	292
Xception [10]	21.0	486

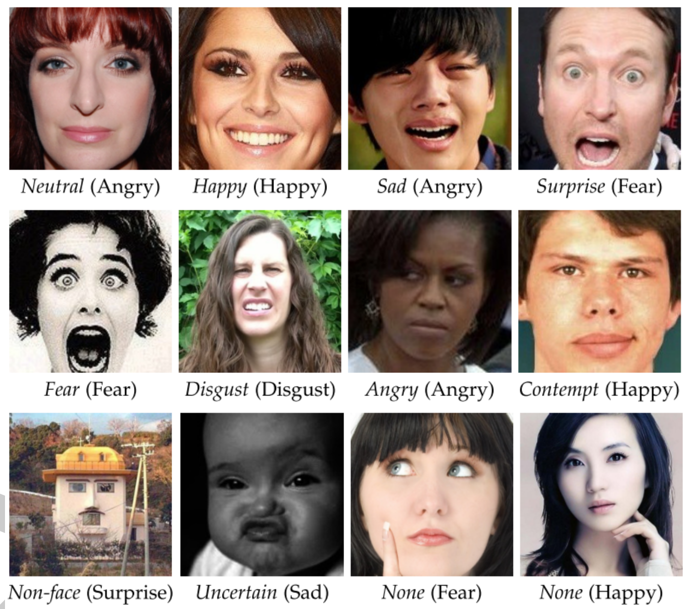

Figure 6. Samples of AffectNet database and classification results. The emotion expression labels is written in parentheses. Image source [57].

With all above computer vision methods, the ICML 2013 Challenges in Representation Learning introduced the Facial Expression Recognition 2013 (FER-2013) database [27]. Fortunately, human actors/subjects databases portraying the basic emotions of external human emotion has been created, which solves the problem of training data.

Benitez-Quiroz et al. [20] proposed EmotioNet database that extracts features by using Gabor filters. Their database is subject-independent and indicates cross-database experiments. Mollahosseini et al. [58] uses FER-Wild database, and trained them on AlexNet with noise estimation methods and archived 82.12% accuracy (on FER-Wild). AffectNet [57] is the state-of-the-art database that proposed gives the most extensive database of facial expression, valence, and arousal in the wild (see Figure 6). In their paper, various evaluation metrics show that their deep neural network gives the state-of-the-art performance in facial emotion recognition.

3.1.2 Depth recognition

As we discussed in the previous section, depth camera principle is the same as Microsoft Kinect, the main difference between depth camera and an RGB camera is it provides 3D facial information, which leads the model difference in this field. Unlike Kinect, depth camera in most cases can only offer facial reconstruction model information instead of body gesture. Thus, depth recognition mainly focuses on modeling 3D facial points.

Chen et al. [9] was the recent research that considering 3D modelling. They propose a real-time 3D model-based method that continuously recognizes dimensional emotions from facial expressions in natural communications. The most challenge parts of their research cover the 3D facial information reconstructed from 2D images. Zhang et al. [94] proposed their exploration on 3D facial points modeling of emotion recognition that directly gets depth information from Kinect. However, their recognition only gives three different emotion states.

Despite the already existing depth cameras in mobile phones, research in this area is rare and not in popular demand. We believe that primary reason is 3D modeling requires extensive computation which is not possible from mobile devices at the moment.

3.2 Voice Aspect

Voice aspect as we discussed in the previous section, emotion inference from user speech is primarily processing user speech. The first part is to deducing user speech text from their voice, and the second part of recognition is calculated sentiment from these documents.

3.2.1 Speech recognition

Speech recognition is a board research area, and there exist broad approaches to achieve this goal. Previous years commercial systems modeling speech recognition by using Hiden Markov model, which achieved good performance. However, with the rise of deep learning methods, recurrent neural network (RNN) [56], Long-short term memory cell [34] and attention mechanism [88] becomes the dominant methodology. It is laborious to compare which RNN model is the state-of-the-art model since speech recognition is much more complicated than a typical vision task when training a RNN. Consider there exists very successful commercial system such as Google Speech API [2]can performs stream speech recognition with returning the speech text information. We don't consider this area in detail for the primary goal of mobile affective computing.

3.2.2 Sentiment Analysis

Sentiment analysis requires text understanding, and it is not an easy problem to solve. Some machine learning techniques, including various supervised and unsupervised algorithms, are being utilized. Some algorithms rank the importance of sentences within the text and then construct a summary out of essential sentences, others are end-to-end generative models. After we have the speech text from the user, sentiment analysis can be performed to evaluate user emotions for each speech sentence or a chunk of speech contents during a period.

Rajalakshmi et al. [71] provides board and comprehensively survey on sentiment analysis. We conclude here for the general steps of sentiment analysis. The first step to for calculating sentiment value from text is tokenize each word via a public sentiment calculation dataset; then for each word, compare it with positive sentiments and negative sentiments word embedding in the dictionary, and increment positive count or negative count; finally, based on the positive count and negative count, one can get result percentage of sentiment to decide the polarity.

Sentiment calculation for text is essentially clear defined in engineering, and the final sentiment value suits of user speech can be a training feature for emotion inference.

3.3 Touch Aspect

Touch interaction modality in previous research all considered using a handcrafted feature for touch behavior and using kernel Support Vector Machine (SVM) to train linear models for classification. Gao et al. [23] is the first application specific in game, which the recognition rates are very robust even in naturalistic settings in the context of smartphone-based computer games. Shah et al. [76] proposed a reasonable handcrafted features, for the three classes (happy, unhappy, neutral); The recent studies in [3] has 7 proposed features, for four classes (Excited, Relaxed, Frustrated, Bored) classification; and Tikadar [86] compares four discriminative models, namely the Naive

[2]https://cloud.google.com/speech/

Table 2. Most commonly used handcrafted features for a touch model.

Feature	Unit (%)	
Deviation in number of strikes		
Deviation in number of taps		
Mode of strike length	Millimeter	
Average strike length	Millimeter	
Mode of strike speed	Meter/second	
Average strike speed	Meter/second Mode of delay	Millisecond
Average delay	Millisecond	
Total delay	Second	
Turnaround time	Second	

Table 3. Comparison of classifiers.

Classifier	Accuracy (%)
Nave Bayes	86.13%
kNN	92.82%
kernel SVM	96.75%

Bayes, K-Nearest Neighbor (KNN), Decision Tree and kernel SVM were explored. SVM gives the highest accuracy of 96.75%. Table 2 shows a feature set of touch interaction information and Table 3 shows the performance comparison for different classifiers. However readers should be aware that these provide solution doesn't provide any stability analysis of their classification model, it is possible can be counted as overfitting if readers cannot recur the accuracy.

Touch interaction is typical interaction information on mobile which outputs from users. One can conclude that kernel SVM is the most accurate model for a handcrafted model in this research direction.

3.4 Sensors Fusion

Multimodal fusions of previous sensors is necessarily a complex challenge. Much research has focused on this analysis, in particular, Zeng et al. [91] gave bimodal affective recognition via facial emotion recognition and combined audio with visual modalities so that the final affect recognition accuracy is greatly improved to almost 90%. Dermody [16] proposed a multimodal system with real-time feedback for public speaking. The system has been developed within the paradigm of positive computing which focuses on designing for user wellbeing.

The recent surveys of multimodal affect analysis on mobile [17, 70, 69], focuses mainly on state of the art in collecting sample data, and reports performance comparison of selected multimodal and unimodal systems, as opposed to comprehensively reviewing key individual systems and approaches, from the growing literature in the field.

Table 4 from [17] shows the most commonly used experiments result of fusion modalities.

On the same training and test sets, the classification experiment using SVM, NN (Neural network) and ELM (Extreme learning machine). ELM outperformed NN by 12% in terms of accuracy (see Ta-

Table 4. Results of feature-level fusion, results from [17]

Combination of modalities	Precision
Accuracy of the experiment carried out on Textual Modality	0.619
Accuracy of the experiment carried out on Audio Modality	0.652
Accuracy of the experiment carried out on Vision Modality	0.681
Experiment using only visual and text-based features	0.7245
Result obtained using visual and audio-based features	0.7321
Result obtained using audio and text-based features	0.7115
Accuracy of feature-level fusion of three modalities	0.782

Table 5. Comparison of classifiers, results from [17]

Classifier	Recall (%)	Training time
SVM	77.03	2.7 min
ELM	77.10	25 s
NN	57.81	2.9 min

ble 5). Regarding training time, the ELM outperformed SVM and NN by a considerable margin (Table 7). A real-time multimodal sentiment analysis engine, the NN as a classifier which provided the best performance regarding both accuracy.

In conclusion, we discussed the techniques for different mobile modalities. We conclude here of this section for each type of data:

- *Vision data*: CNN models perform the state-of-the-art performance of facial analysis;
- *Speech data*: RNN models performs the state-of-the-art performance of speech recognition and sentiment analysis;
- *Interaction data*: handcrafted features are the most commonly used feature and neural networks for this kind of unstructured data are unminded;
- *Context data*: application context and geographic location are normally used as a validation dimension for emotion inference, researchers don't consider them how to use as a training feature properly.

4 Applications Case Study

Assuming that user emotions have inferred by models or systems, Conati et al. [11] address a variety of issues related to the development of affective loop as well as a synthesis of the appropriate affective expressions. In subsequent studies, emotional systems were suggested in [13] since emotional-sensitive systems make our interactions with machines like human to human communications [65]. In this section, we illustrate the most popular applications of emotion-sensitive HCI system in the mobile affective research area.

4.1 Case 1: Spoken Dialogue Systems

A voice interaction system capable of sending the users emotional messages can improve the intelligence and interaction experience of a voice user interface [80].

In the early research stage, [46] proposed a personalized voice-emotion user interface in the desktop system regardless of speaker's age, sex or language is presented. They experimented with participants, and the results showed that voice emotion sensitive agents are feasible.

The most recent papers in emotion-sensitive voice user interface consider emotional voice tones to caring users [8], as well as the voice control system in an in-vehicle infotainment system [43].

Due to the intangible property of voice interaction system, it is almost endless of how we integrate user emotions to a spoken dialogue system, McTear et al. [54] addressed emotional spoken system by a markup language, which means developers can easily incorporate user emotion state to adjust the system voice.

With the raising of voice assistant, there already exists successful commercial voice system such as Apple Siri[3], Google Assistant[4], Amazon Alexa[5] and Microsoft Cortana[6]. Voice user interfaces design has become the part of user experience design. Even though the commercial system provides the above described markup language, there is still huge unmined research into an open problem of how to evaluate this kind of emotional voice system.

[3] `https://www.apple.com/ios/siri/`
[4] `https://assistant.google.com/`
[5] `https://developer.amazon.com/alexa`
[6] `https://www.microsoft.com/en-us/windows/cortana`

4.2 Case 2: Adaptive Graphics User Interfaces

Adaptive user interfaces have been researched for years [75, 48] and addressed in many ways. User emotion is one of the aspects that considered in adaptive user interface.

Dalvand et al. [14] introduced an adaptive user interface, the colors of user interface change according to the emotional states and mood state of users. Emotional states of a user are specified according to his/her interactions with the keyboard. After detection of emotions, the user mood reflects appropriate colors, and [40] as another example of this adaption. However, they even didn't have an evaluation of such kind of system. A recent paper [21] also addresses UI adaptation by user emotions (positive, negative and neutral) at run-time. Their prototype was tested successfully of how it reacts to emotions (negative).

In conclusion, the adaptive graphics user interfaces system suggests covered components with inference engine (integrate with techniques we discussed in the previous section), the adaptation engine (with common personalized system rules) and the interactive system (for standard graphical user interfaces). It also evidenced that GUI changes follows emotion changes at run-time, which was mainly beneficial for most users.

5 Challenges

Politou et al. [68] discussed most of the challenges in smartphone affective research which covers *privacy*, *data misuse*, *trust and engagement*, *multimodal fusion*, *resource constraints*, *affect modelling and representation*, *cultural differences* and *system building costs*. From the technique prespective, challenges and limitations can be summarized to the following sections.

5.1 Impermeable Emotions

The primary limitation of traditional affective computing research refers to as impermeable emotions [64]. Impermeable emotions are broad, with many of these modalities being inaccessible (e.g., blood chemistry, brain activity, neurotransmitters), and many others being too non-differentiated. It makes it unlikely that collecting the necessary data will be possible or feasible in the near future.

There is a time period to express emotion, and a time period to forbear; a time period to sense what others are feeling and a time period to ignore feelings. In every time, human need a balance when they express their emotions, and this balance is missing in computing. Figure 7 illustrates a map of human emotions. The largest text represents the most expressible emotions and the arrow in the graph indicates how emotions related to each other, which is can be consider as a analogy of Markov chain. The all recent classification method doesn't provides this consideration. A close look to the smaller text in this picture, almost all of them are impermerable emotions that can never be discovered by emotion inference model since can not be labeled in a dataset.

In most cases, researchers feel positive and argue that impermeable emotions can be expressed explicitly in other expressed Emotions [62] and implicit emotions are trivial and not crucial for most of the case of affective computing application when we need an emotion state [92]. However, impermeable emotion is still an open problem and challenging research area in affective computing.

5.2 Continious Understanding

Emotions are not states. Continuous emotion understanding is much more challenging since it's not related to external emotion but also related to internal emotions; various emotions can be expressed as the map *(see Figure 7)* we discussed above.

Most of the researchers we introduced in the previous section treated emotion inference problem as a classification problem, whereas the human emotions are always passive and instantaneous. People's expression of emotion is so idiosyncratic and variable, that there is little hope of accurately recognizing an individuals emotional state from the available data sometimes.

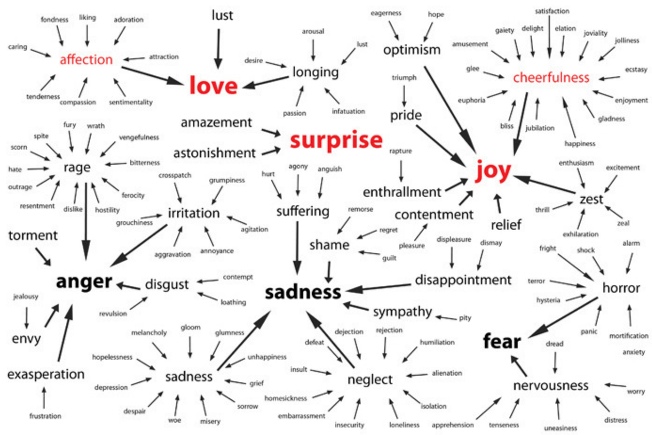

Figure 7. A map of human emotions. Image from [78].

5.3 Context Specific Models

A potential ethical limitation in research studies comes from the fact that perceptions of emotions and personality are not universal, but they are highly dependent on the mobile application context [23, 76, 3, 86] as well as the cultural differences between humans [55, 51, 25]. According to these perspectives, a reasonable challenge is how an affective recognition model should be designed and built in an application and transparent cultural way.

5.4 Lack of Computation Resources

As we discussed before, CNN and depth camera methods require large computation resource. Apparently, computation resources are limited in a mobile device. Consequently, an issue that impacts the effectiveness of emotion inference system is the fusion of all kinds of multimodal.

To handling such informations in run-time, for mobile devices with limited operation time and processing power is not possible. Cloud computing is a way of sending processing data from mobile device to powerful server, it significantly mitigates the computation stress of a mobile device but also increases the delay of emotion inference due to communication delay between server and mobile phone.

6 Conclusions

In this paper, we investigated the recent advances in mobile affective computing related to human-computer interaction projects and inference techniques.

Section 2 addresses different data sources in various mobile commodity sensors for emotion inference in previous studies. These include camera, touch screen, motion sensors, microphone, GPS, and application context.

Next, in the Section 3, we first carried out the review of emotion inference methods based on different type of data source, and compared the tested methods and inference models from previous researchers. In these comparisons, we first reviewed various models for user emotion inference, researchers usually transfers emotion inference problem into a classification problem. As a classification problem, most researchers consider user emotions can be inferred to three different states (Happy, Unhappy, Neutral). In each subsection, we highlighted the most useful methods for different type of emotion inference that concluded by the most recent research papers, such as CNN as the best way of vision type, RNN as the best way of auditory type, and handcrafted feature with SVM as the best way of interaction details. At the end of this section, we considered the combinations of these types of data. According to our investigation, the most commonly used data type are suggested to *vision and audio* data in mobile affective computing; However, the combination of *vision, audio, interaction details, and context for aspects fusion are unmined open topics*.

In Section 4, we survey two novel applications in human-computer interaction related projects driven by emotion inference. Voice user interfaces consider user emotion as inputs and suggest to please users by adjusting system voice tone; Adaptive graphics user interfaces then considers emotion state as the context of UI theme.

Despite we researched the scientific approaches of mobile emotion inference in human-computer interaction related topics, there still apparent challenges in this area. Section 5 pointed out the current problems of this research area. The main challenges of this area are *impermeable emotions* and *continuous understanding*. Moreover, the generalisability of mobile affective computing applications are subject to certain limitations. For instance, most of *inference models are context specific* and multimodal inference then *requires large computation resources*.

Nowadays, new technologies and methods provide us new opportunities of affect emotion inference in an unobtrusive mobile device. Since the complexity of the interpretation of human behavior at a very deep level is tremendous and requires a highly interdisciplinary collaboration, we believe the true break-throughs application in this field can be established by precisely modeling and new sensing technologies in the future.

Acknowledgement

The author would like to thank reviewers for their comments and suggestions, and appriciate the topic advisor Daniel Buschek for his patient on massive grammar error while reading, useful guidance of paper structuring and proofread suggestions.

References

[1] "Affective computing: A review". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 3784 LNCS (2005), pp. 981–995.

[2] Jeremy N Bailenson, Nick Yee, Scott Brave, Dan Merget, and David Koslow. "Virtual interpersonal touch: expressing and recognizing emotions through haptic devices". In: *Human-Computer Interact.* 22.November (2007), pp. 325–353.

[3] S Bhattacharya. "A predictive linear regression model for affective state detection of mobile touch screen users". In: *International Journal of Mobile Human Computer Interaction* 9.1 (2017), pp. 30–44.

[4] Samit Bhattacharya. "A Predictive Linear Regression Model for Affective State Detection of Mobile Touch Screen Users". In: *Int. J. Mob. Hum. Comput. Interact.* 9.1 (2017), pp. 30–44.

[5] Jeffrey P Bigham and Michael S Bernstein. "Human-Computer Interaction and Collective Intelligence". In: *The handbook of collective intelligence* (2014).

[6] A. Boles and P. Rad. "Voice biometrics: Deep learning-based voiceprint authentication system". In: *2017 12th System of Systems Engineering Conference (SoSE)*. IEEE. 2017, pp. 1–6.

[7] Jessica R. Cauchard, Kevin Y. Zhai, Marco Spadafora, and James A. Landay. "Emotion encoding in human-drone interaction". In: *ACM/IEEE International Conference on Human-Robot Interaction* 2016-April (2016), pp. 263–270.

[8] Woosuk Chang, Angel Camille Lang, Miki Nobumori, Luca Rigazio, Gregory Senay, and Akihiko Sugiura. *Intelligent caring user interface*. US Patent App. 15/085,761. 2016.

[9] Hui Chen, Jiangdong Li, Fengjun Zhang, Yang Li, and Hongan Wang. "3D model-based continuous emotion recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1836–1845.

[10] François Chollet. "Xception: Deep Learning with Depthwise Separable Convolutions". In: *arXiv preprint arXiv:1610.02357* (2016).

[11] Cristina Conati, Stacy Marsella, and Ana Paiva. “Affective interactions”. In: *Proc. 10th Int. Conf. Intell. user interfaces - IUI ’05* (2005), p. 7.

[12] Roddy Cowie. “Ethical issues in affective computing”. In: *The Oxford Handbook of Affective Computing* (2015), p. 334.

[13] Elizabeth a Crane, N Sadat Shami, and Christian Peter. “Let’s get emotional: emotion research in human computer interaction”. In: *Proceedings of ACM CHI 2007 Conference on Human Factors in Computing Systems* 2 (2007), pp. 2101–2104.

[14] K. Dalvand and M. Kazemifard. “An Adaptive User-Interface Based on User’s Emotion”. In: *2012 2nd International eConference on Computer and Knowledge Engineering (ICCKE)*. 2012, pp. 161–166.

[15] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[16] Fiona Dermody. “Multimodal positive computing system for public speaking with real-time feedback”. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016* (2016), pp. 541–545.

[17] Sidney K D’mello and Jacqueline Kory. “A review and meta-analysis of multimodal affect detection systems”. In: *ACM Computing Surveys (CSUR)* 47.3 (2015), p. 43.

[18] M. A. Eid and H. Al Osman. “Affective Haptics: Current Research and Future Directions”. In: *IEEE Access* 4 (2016).

[19] “Emotion recognition in human-computer interaction”. In: *Neural Networks* 18.4 (2005), pp. 389–405.

[20] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. “Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 5562–5570.

[21] J Galindo, Sophie Dupuy-Chessa, and Éric Céret. “Toward a User Interface Adaptation Approach Driven by User Emotions”. In: *ACHI07, Nice, France*. 2017, pp. 12–17. ISBN: 9781612085388.

[22] Raghu K Ganti, Fan Ye, and Hui Lei. “Mobile crowdsensing: current state and future challenges”. In: *IEEE Communications Magazine* 49.11 (2011).

[23] Yuan Gao, Nadia Bianchi-Berthouze, and Hongying Meng. “What Does Touch Tell Us about Emotions in Touchscreen-Based Gameplay?” In: *ACM Trans. Comput. Interact.* 19.4 (2012), pp. 1–30.

[24] Jose Maria Garcia-Garcia, Victor M. R. Penichet, and Maria D. Lozano. “Emotion Detection: A Technology Review”. In: *Proceedings of the XVIII International Conference on Human Computer Interaction*. ACM, 2017.

[25] Maria Gendron, Debi Roberson, Jacoba Marietta van der Vyver, and Lisa Feldman Barrett. “Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture.” In: *Emotion* 14.2 (2014), p. 251.

[26] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[27] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. “Challenges in representation learning: A report on three machine learning contests”. In: *International Conference on Neural Information Processing*. Springer. 2013, pp. 117–124.

[28] Mariam Hassib, Daniel Buschek, PawelW W Wozniak, and Florian Alt. “HeartChat: Heart Rate Augmented Mobile Chat to Support Empathy and Awareness”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), pp. 2239–2251.

[29] K. He, G. Gkioxari, P. Dollár, and R. Girshick. “Mask R-CNN”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. arXiv preprint arXiv: 1703.06870, 2017, pp. 2980–2988.

[30] Kaiming He and Jian Sun. “Convolutional neural networks at constrained time cost”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 5353–5360.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conf. Comput. Vis. Pattern Recognit.* (2016), pp. 770–778.

[32] Matthew J Hertenstein, Rachel Holmes, Margaret McCullough, and Dacher Keltner. “The communication of emotion via touch.” In: *Emotion* 9.4 (2009), p. 566.

[33] Heinke Hihn, Sascha Meudt, and Friedhelm Schwenker. “Inferring mental overload based on postural behavior and gestures”. In: *Proceedings of the 2nd workshop on Emotion Representations and Modelling for Companion Systems - ERM4CT ’16* (2016), pp. 1–4.

[34] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[35] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. “Mobilenets: Efficient convolutional neural networks for mobile vision applications”. In: *arXiv preprint arXiv:1704.04861* (2017).

[36] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. “Densenet: Implementing efficient convnet descriptor pyramids”. In: *arXiv preprint arXiv:1404.1869* (2014).

[37] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size”. In: *arXiv preprint arXiv:1602.07360* (2016).

[38] William James. “What is an emotion?” In: *Mind* 9.34 (1884), pp. 188–205.

[39] William James. *What is an Emotion?* Simon and Schuster, 2013.

[40] Robin Kaiser and Karina Oertel. “Emotions in HCI – An Affective E-Learning System”. In: *Computer (Long. Beach. Calif).* (2006), pp. 105–106.

[41] Herman Kamper, Aren Jansen, and Sharon Goldwater. “A segmental framework for fully-unsupervised large-vocabulary speech recognition”. In: *Computer Speech and Language* 46 (2017), pp. 154–174.

[42] Wazir Zada Khan, Yang Xiang, Mohammed Y Aalsalem, and Quratulain Arshad. “Mobile phone sensing systems: A survey”. In: *IEEE Communications Surveys & Tutorials* 15.1 (2013), pp. 402–427.

[43] Dong-hyu Kim and Heejin Lee. “Effects of user experience on user resistance to change to the voice user interface of an in-vehicle infotainment system: Implications for platform and standards competition”. In: *International Journal of Information Management* 36.4 (2016), pp. 653–667.

[44] Hyun-Jun Kim and Young Sang Choi. “Exploring emotional preference for smartphone applications”. In: *2012 IEEE Consumer Communications and Networking Conference (CCNC)*. 2012, pp. 245–249. DOI: 10.1109/CCNC.2012.6181095.

[45] Joohee Kim, Na Hyeon Lee, Byung-Chull Bae, and Jun Dong Cho. “A Feedback System for the Prevention of Forward Head Posture in Sedentary Work Environments”. In: *Proceedings of the 2016 ACM Conference Companion Publication on Designing Interactive Systems - DIS ’16 Companion* (2016), pp. 161–164.

[46] V Kostov and S Fukuda. “Emotion in user interface, voice interaction system”. In: *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*. Vol. 2. IEEE. 2000, pp. 798–803.

[47] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in neural information processing systems (NIPS)*. 2012.

[48] Pat Langley. “Machine learning for adaptive user interfaces”. In: *Annual Conference on Artificial Intelligence*. Springer. 1997, pp. 53–62.

[49] Rodrigo C Lentini, Beatrice Ionascu, Friederike A Eyssel, Scandar Copti, and Mohamad Eid. “Authoring Tactile Gestures : Case Study for Emotion Stimulation”. In: *International Journal of Computer, Electrical, Automation, Control and Information Engineering* (2017).

[50] Daniel Lopez-Martinez and Rosalind Picard. “Multi-task Neural Networks for Personalized Pain Recognition from Physiological Signals”. In: *arXiv preprint arXiv:1708.08755* (2017), pp. 3–6.

[51] Takahiko Masuda, Phoebe C Ellsworth, Batja Mesquita, Janxin Leu, Shigehito Tanida, and Ellen Van de Veerdonk. “Placing the face in context: cultural differences in the perception of facial emotion.” In: *Journal of personality and social psychology* 94.3 (2008), p. 365.

[52] Antonella Mazzoni and Nick Bryan-kinns. “Mood Glove : A haptic wearable prototype system to enhance mood music in film”. In: *Entertainment Computing* 17 (2016), pp. 9–17.

[53] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. “AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit”. In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA ’16* (2016), pp. 3723–3726.

[54] Michael F McTear. “The Rise of the Conversational Interface: A New Kid on the Block?” In: *International Workshop on Future and Emerging Trends in Language Technology*. Springer. 2016, pp. 38–49.

[55] Batja Mesquita and Nico H Frijda. “Cultural variations in emotions: a review.” In: *Psychological bulletin* 112.2 (1992), p. 179.

[56] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. “Recurrent neural network based language model.” In: *Interspeech*. Vol. 2. 2010, p. 3.

[57] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild”. In: *IEEE Transactions on Affective Computing* (2017). arXiv: `1708.03985`.

[58] Ali Mollahosseini, Behzad Hasani, Michelle J Salvador, Hojjat Abdollahi, David Chan, and Mohammad H Mahoor. “Facial expression recognition from world wild web”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016, pp. 58–65.

[59] “Motion, Emotion, and Form: Exploring Affective Dimensions of Shape”. In: *Late-Breaking Work: Designing Interactive Systems* (2016), pp. 1430–1437.

[60] Aske Mottelson and Kasper Hornbæk. “An affect detection technique using mobile commodity sensors in the wild”. In: *ACM Int. Jt. Conf. Pervasive Ubiquitous Comput.* (2016), pp. 781–792.

[61] Ganapreeta R. Naidu. “A Computational Model of Culture-Specific Emotion Detection for Artificial Agents in the Learning Domain”. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI ’15* (2015), pp. 635–639.

[62] Brian Parkinson. *Ideas and realities of emotion*. Psychology Press, 1995.

[63] Phuong Pham and Jingtao Wang. “Understanding Emotional Responses to Mobile Video Advertisements via Physiological Signal Sensing and Facial Expression Analysis”. In: *Proceedings of the 22nd International Conference on Intelligent User Interfaces - IUI ’17* (2017), pp. 67–78.

[64] Rosalind W Picard. “Affective computing: challenges”. In: *International Journal of Human-Computer Studies* 59.1 (2003), pp. 55–64.

[65] Rosalind W Picard. “Affective Computing for HCI”. In: *Human Computer Interaction: Ergonomics and User Interfaces. Proceedings of HCI International ’99 8th International Conference on Human Computer Interaction* (1999).

[66] Rosalind W Picard. “Affective computing: from laughter to IEEE”. In: *IEEE Transactions on Affective Computing* 1.1 (2010), pp. 11–17.

[67] Matthew Pike, Richard Ramchurn, and Max L Wilson. “Two-way Affect Loops in Multimedia Experiences”. In: *Proceedings of the 2015 British HCI Conference* (2015), pp. 117–118.

[68] Eugenia Politou, Efthimios Alepis, and Constantinos Patsakis. “A survey on mobile affective computing”. In: *Computer Science Review* (2017).

[69] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. “A review of affective computing : From unimodal analysis to multimodal fusion”. In: *Information Fusion* 37 (2017), pp. 98–125.

[70] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. “Fusing audio, visual and textual clues for sentiment analysis from multimodal content”. In: *Neurocomputing* 174 (2016), pp. 50–59.

[71] S. Rajalakshmi, S. Asha, and N. Pazhaniraja. “A comprehensive survey on sentiment analysis”. In: *2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN)*. 2017, pp. 1–5.

[72] Rajib Rana, Margee Hume, John Reilly, Raja Jurdak, and Jeffrey Soar. “Opportunistic and context-aware affect sensing on smartphones: the concept, challenges and opportunities”. In: *arXiv preprint arXiv:1502.02796* (2015).

[73] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 779–788.

[74] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. “Dynamic Routing Between Capsules”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 3857–3867.

[75] Matthias Schneider-Hufschmidt, Uwe Malinowski, and Thomas Kuhme. *Adaptive user interfaces: Principles and practice*. Elsevier Science Inc., 1993.

[76] Sachin Shah, J. Narasimha Teja, and Samit Bhattacharya. “Towards affective touch interaction: predicting mobile user emotion from finger strokes”. In: *J. Interact. Sci.* 3.1 (2015), p. 6.

[77] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014), pp. 1–14. arXiv: `arXiv:1409.1556v6`.

[78] stanchew. *A map of human emotions*. https://stanchew.wordpress.com/2012/04/23/a-map-of-human-emotions/. [Online; accessed 20-November-2017]. 2012.

[79] Thad Starner. "Human-powered wearable computing". In: *IBM systems Journal* 35.3.4 (1996), pp. 618–629.

[80] Kevin J Surace, George M White, Byron B Reeves, Clifford I Nass, Mark D Campbell, Roy D Albert, and James P Giangola. *Voice user interface with personality*. US Patent 6,334,103. 2001.

[81] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[82] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. "Rethinking the Inception Architecture for Computer Vision". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 2818–2826. arXiv: 1512.00567.

[83] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning". In: *AAAI*. 2017, pp. 4278–4284. arXiv: 1602.07261.

[84] Guang Tan, Hongbo Jiang, Shengkai Zhang, Zhimeng Yin, and Anne-Marie Kermarrec. "Connectivity-based and anchor-free localization in large-scale 2d/3d sensor networks". In: *ACM Transactions on Sensor Networks (TOSN)* 10.1 (2013), p. 6.

[85] Jordan Tewell, Jon Bird, and George R. Buchanan. "The Heat is On: A Temperature Display for Conveying Affective Feedback". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17* (2017), pp. 1756–1767.

[86] Subrata Tikadar, Sharath Kazipeta, Chandrakanth Ganji, and Samit Bhattacharya. "A Minimalist Approach for Identifying Affective States for Mobile Interaction Design". In: *Human-Computer Interaction - INTERACT 2017*. Ed. by Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler. Cham: Springer International Publishing, 2017, pp. 3–12.

[87] Sherry Turkle. *The second self: Computers and the human spirit*. Mit Press, 2005.

[88] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is All You Need". In: *Advances in Neural Information Processing Systems (NIPS)*. 2017.

[89] Mark Weiser. "The computer for the 21st century". In: *Scientific american* 265.3 (1991), pp. 94–104.

[90] N. B. Wunarso and Y. E. Soelistio. "Towards Indonesian speech-emotion automatic recognition (I-SpEAR)". In: *2017 4th International Conference on New Media Studies (CONMEDIA)*. Vol. 2017. 2017, pp. 98–101.

[91] Zhihong Zeng, Jilin Tu, Ming Liu, Tong Zhang, Nicholas Rizzolo, Zhenqiu Zhang, Thomas S. Huang, Dan Roth, and Stephen Levinson. "Bimodal HCI-related affect recognition". In: *Proceedings of the 6th international conference on Multimodal interfaces - ICMI '04* (2004), p. 137.

[92] Shengkai Zhang and Pan Hui. "A Survey on Mobile Affective Computing". In: *ArXiv Prepr. ArXiv14101648* 1 (2014). arXiv: 1410.1648.

[93] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. "Shufflenet: An extremely efficient convolutional neural network for mobile devices". In: *arXiv preprint arXiv:1707.01083* (2017).

[94] Zhan Zhang, Liqing Cui, Xiaoqian Liu, and Tingshao Zhu. "Emotion detection using Kinect 3D facial points". In: *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*. IEEE. 2016, pp. 407–410.