Multimodal Emotion Recognition for Human-Computer Interaction: A Survey

Michele Mukeshimana, Xiaojuan Ban, Nelson Karani, Ruoyi Liu
School of Computer and Communication Engineering, University of Science and Technology Beijing, 100083 Beijing, China.

Abstract— Today, the computer and its applications has invaded our daily life, Ubiquitous Computing. The interaction between the users and computing devices is becoming similar to human-human interactions. The integration of emotion recognition in Human-Computer Interaction aims at making the interaction easier and smarter, natural interaction. This paper, presents a brief survey on current state of the multimodal emotion recognition in Human-Computer Interaction (HCI) research. We explain Emotion recognition as a research trend, its application in Human-Computer Interaction, its challenges and opportunities. The Human-Computer Interaction and Information Fusion Researches are also presented, and conclude with a proposition of steps to design a multimodal emotion recognition analyzer.

Index Terms— Multimodal emotion recognition, information fusion, affective computing, Human-Computer Interaction, machine learning, affect states, Data mining, feature extraction.

1 Introduction

The Ubiquitous Computing (UC) is a new computing wave aiming to make computer so imbedded, so fitting and so natural that man will use it without focusing on it rather focus on the user's need [1, 2, 3]. Among the underlying promising technologies for the success of UC are Nanotechnology, wireless computing, context-aware computing and Natural interaction. The first two technologies are more computer-oriented to provide the required infrastructure, while the last two are more user communication/interaction- oriented to make the infrastructure easy and helpful to use.

The context-awareness aims to endow computers with the ability of understanding the user's situation and intentions in order to supply proper and adequate services, resources, or other relevant information. Meanwhile, the natural interaction permits the user to be able to get services, resources and information from the computer without needing to know or think about the rules of use of the computers [4]. Thus, the current Human-Computer Interaction(HCI) or Man-Machine Interaction (MMI) design needs to consider many aspects of human behaviours and needs to be useful, adaptive and more intelligent. For example, in an Intelligent Tutoring System the integration of an automatic emotion recognition module, improves the feedback of the system [5, 6, 7, 8] or in Healthcare System [9, 10].

The HCI research is concerned with the study, conception, design, and implementation of the ways in which humans make use of computational artefact, systems and infrastructures. It aims to make the interaction between man (user) and the computer or information systems/technologies more effective, more comfortable, easier, safer, and more faultless for the users [11]. Human natural interactions are emotionally driven and multi-modally expressed [12, 13, 14].

Emotion plays an important role in human social and intelligent interactions [15, 16, 17]. The study done by Reeves and Nass [18] has shown that people react to computers as if they are social actors with emotions, feelings and intelligence though people were well sentient they are interacting with only computers. The absence of the emotions can make one feel uncomfortable in the interaction process even towards non human agent. Therefore, to integrate the emotion detection, recognition and responses in the user interface is one of the ways of designing useful, effortless and meaningful intelligent interaction experiences [19, 12, 20].

The HCI research is oriented to design intelligent, adaptive interfaces to make the interaction more effortless and more natural-like [21, 22, 23, 24]. Computers ought to understand both the meaning and the context of the user's message in order to provide the user with an appropriate service, resources or information to his/her particular situation [25, 26, 27].

Humans interact with the environment using multiple means [4, 26, 14] namely, face, speech, voice, gesture, etc. To endow computers with such abilities is the intention of the Multimodal Human-Computer Interaction research [28, 29]. One of the first practical examples is the "Put That There" system, as the first system to demonstrate the feasibility of the multimodal interfaces [30]. The following table (Table 1) shows some of the multimodal systems:

Table 1. Some Multimodal based systems

Systems	Year of publication	Modalities engaged	Applications
Put-That-There[31]	1980	Voice, Gesture	Graphical Interface
Quickset [32]	1997	Pen, voice	Military training Interface
Vision-based Tabletop Interface [33]	2007	Face	Human Computer Interface
Emotional Healthcare Systems [10]	2014	Face Expression	Healthcare
Emotional Avatar [34]	2015	Visual, Audio, Text	Customer Care

Numerous works have been done, for more information we refer the reader to [35, 22, 36, 37, 30]. The emergence of the new

intelligent devices based on basic human modality recognition [38, 39] has stimulated the automatic emotion recognition research to add naturalistic property to the interaction [40, 41, 22, 6, 42].

Information fusion is utilized during the multimodal interface design. Multimodality in computing is the capacity of a system to communicate with the user through different types of communication channels and to extract and convey meaning automatically. The multi-modality is used to enhance the accuracy and bring the spontaneity in the interaction. Our focus in this paper is on multimodality in emotion recognition.

In this paper, we present integration of multimodal emotion recognition in human computer interaction. The section 2 portrays the emotion recognition research and related challenges. The section 3 presents the Human-Computer Interaction as a research direction. In the section 4, we present the information fusion in Multimodal Emotion Recognition.

2 Multimodal Emotion Recognition research

2.1 Emotion Recognition outline

The word "emotion" was introduced in English in 17th century and widely employed in the 18th century English referring to mental experiences and in the 19th century it became a theoretical term [43]. Since the question of William James, "What is emotion?" [44], there have been many debates and attempts striving to answer this question and many can be found in literature [45, 46, 47, 48]. Numerous definitions have been proposed, but we retain that the seat of the emotion is the brain. It is a reaction of an individual to environmental events according to his needs, goals and concerns. Emotion involves physiological, affective, and cognitive parts [28, 49, and 50]. In 1997, Picard [15] introduced the Affective computing aiming to endow computers to recognize and express affect. So the terms "emotion" and "affective states" will be equally used throughout the text.

Emotions play an essential role both in human cognition and perception [51, 52, 53]. Therefore, for a natural and intelligent interaction between man and computer, it is a must for a computer to acquire the ability to at least recognize and express affect [54, 55]. Yet human emotion has to be well modelled [56, 57, 58], to be efficiently integrated in current human- computer interaction.

Human emotions are very vast and complex entity which needs careful and adequate conceptualization for the design of useful and meaningful interface [4, 13]. In reality, it is not all the computing systems that will need that ability of sensing and expressing emotion. Thus, to endow computers with that ability, there is need to consider the usefulness, to make the product more pleasant and likable [59, 26].

The experimental works in human emotion modelling have been mostly based on one modality like facial expression [40, 60,61]; speech and voice expression [14, 62], body gesture [63], physiological signals [64], etc. Other techniques are based on combination of multiple modalities [65, 36, 66, 26, 42, 34, 67]. The physiological-based emotion recognition is technically easier to collect than others and less affected by human's will. But, they are more uncomfortable to use in the real life because data acquisition will require more device on the body of the subject; and the daily use will cause an extra expense to the user. The study of one single modality is less efficient to produce an accurate model of the complete situation, thus the need for combining two or more modalities.

2.2 Emotion Recognition Multimodality

The use of multiple modalities is in agreement with the fact that the human emotion is produced simultaneously by different channels likewise its interpretation. So the multimodal emotion recognition in human computer interaction makes it closer to the natural interaction. The modality can be visual (the face, body, hand…), audio (speech, voice), physiological signal, and Input dynamic sequence (Fig. 5). Note that multimodal systems are able to communicate through more than one mode and extract a meaning; but the one which doesn't extract any meaning from them, is a multimedia system [68]. So a multimodal emotion recognition interface can recognise/express emotion by multiple channels [25, 69, 70]. The main objective is the enhancement of robustness due to combining different partial information sources; and a flexible personalization based on user and context. This was proved in the works of [71, 72, 39, 34].

Emotion modelling systems has captured more attention in the last decades, but among all works the multimodal based emotion recognition has been demonstrated to be the most accurate and robust. The following illustration, Fig.1 represents results from nine references as an illustration:

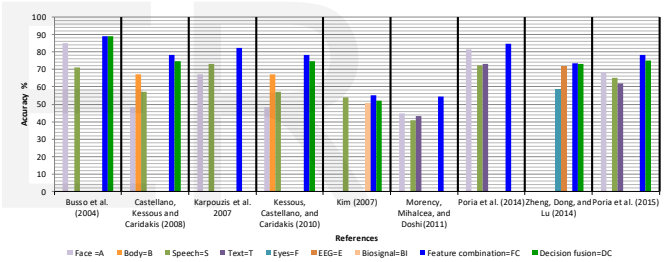

Fig. 1. Comparison between unimodal and multimodal emotion recognition

Fig.1 above graphically represents the results of nine (09) different works on multimodal emotion recognition. The multimodal emotion recognition was more accurate than individual; because the increase of the variables for the training permit a better discrimination of relevant features, thus the decision function generalizes better [73]. The fusion is mostly done on two levels: feature and decision; and the fusion on the features-level (FC variable in Fig.1) is better than the fusion on the decision-level (DC variable in Fig. 1). The main reason is that the later fusion (decision) has lost some information in earlier steps, and uses synthesised information resulting from another approximation. However, the main issue is to know which combination is better to be considered for on optimised prediction and which best cue that convey the information.

2.3 Challenges in Emotion Recognition

The first challenge is sensing and the recognition of emotion (affects state); in other words, the definition of what emotion is, and how it influences the representation of the human affective. Some sustain the theory of discrete basic emotion [72, 74] and admit that these basic emotions are universally displayed and cross-culturally recognized. Others say emotions can be represented in terms of small number of latent dimensions [75,76] with various representations [77, 78]. The

social constructivists argue that emotions are culturally constructed and there is no universality [79, 26]. However, the observation that, some emotional expressions universally come from infants, blind and sighted independent of race and culture [80, 81] has supported the evolutionary perspective of the discrete basic emotion. Some theoretical models have attempted to account for both universality and cultural variations by specifying which particular emotional aspect show similarities and differences across cultural boundaries [72, 81, 82]. Currently, this problem hardens the collection of necessary data for training. In order to overcome this issue one can treat affective states as being correlated not only to discrete emotions but also to behavioural cues identifying attitudinal states and to social signaling [37, 26]. In attempting to solve this problem, Sellers [83] used the definition of emotion in arousal and valence spaces coupled with an adaptation of Maslow's hierarchy [84] to give a comprehensive theory of emotions. Larue et al. [85] worked on the Biological Inspired Cognitive Architecture introduced by Samsonovich [86] and combined the psychological approach [87, 88] and biological approaches [89, 90] to add emotions from cognitive and neurological processes to a multi-agent.

The second challenge relates to how the emotion information is conveyed. There is a problem of the exactness of the human behavioural cues which convey information about affective states [91, 74]. In different works, the combination including face and body gesture simultaneously has given better accuracy than other combinations [92, 93, 66]. It sustains the Dr. Albert Mehrabian's 7-38-55% Rule [94,95], man communication is 55% by the facial expression. However, the problem of knowing which channel to consider, is still open. The attempting solution is to study techniques used by the actors or artists or psychologists for the facial expression interpretation.

The third challenge is the choice of optimal combination. The emotion recognition is improved when using multiple modalities (Fig.1). However, the problem of knowing which modalities to combine to optimize the accuracy still preoccupies researchers. In solving it, some authors proposed the integration of the context (who, where, what, when, why, and how) and the consideration of interdependency between the different feature to increase the accuracy [96, 37, 97]. The new learning paradigm using privileged information [98, 99] in analogy of human learning, supports the recognition increase in a multi-modal combination [100].

The fourth challenge is about the data collection. Most of the existing datasets are well prepared in good condition of extreme emotion expression which won't accurately reflect the reality. This impacts the data collection because human-being try to moderate their emotion according to the environment, none likes one's internal condition to be completely known beyond their expectation. Even spontaneous emotion recording can be affected by the fact that the person knows she is being recorded and will alter the spontaneous expression. In such cases, the record on a longer period can stabilize the subject expression. Likewise, the works on the dataset built from the public websites [66, 34] have given promising results.

The fifth challenge is relating to the ethical and safety or privacy aspects of human emotion. Emotions are more personal and very private. This ethic affects the incorporation of emotion in the Human-Computer Interaction, because a computer with ability of detecting, recognizing and even manipulating user's emotion is subject to a potential rejection [97]. Design should aim at consistency with more discretion and privacy. Instead of targeting an emotional interaction, it is preferable to target an effective and useful one, especially in Human-Computer Interaction application.

3 HUMAN-COMPUTER INTERACTION (HCI) RESEARCH.

3.1 Human Computer Interaction description

The Human-Computer Interaction (HCI) term is newer than its original well reputable disciplines [101]. The study of Human performance began with Second World War, motivated by the need of producing more effective weapon systems. In the development of Information Science and Technology, the need for effective and efficient management of the Information has influenced the design of user and systems interaction [102].

HCI research aims to understand and create different interfaces (software and other technology) between humans (users) and Computers. The resulting systems have to be enjoyable to use, engaged and accessible [24]. As illustrated in the Fig. 2 there are three main components: human as the user, Computer or system as the information auto-matic processor and the interaction to link the two ends.

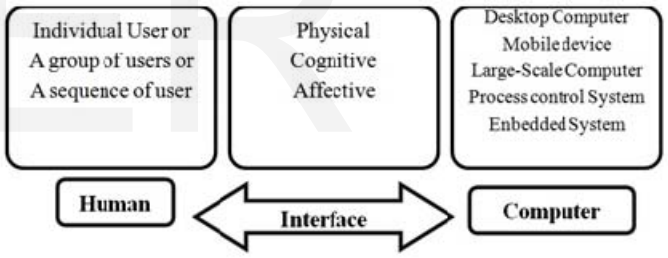

Fig.2. Human-Computer Interaction main components

The user inputs his/her request via an interface (commands, graphic user interfaces or virtual reality) and the computer processes it and gives back an output to the user via the interface. The user can be involved at a physical, cognitive or affective level. The physical level is when the user is interacting with the machine mechanically. The cognitive level is the way permitting the user to understand and interpret the system and expect from the system in order to interact with it. The affective aspect is when the interaction becomes intelligent and active; the computer can react as having emotion, thinking [15]. The most important fact in human computer interaction is functionality and usability of the system [103].

Functionality of the system concerns the actions or services provided to the user; and it is measured by its usability [104]. The human tasks are complex to be efficiently modeled therefore the human computer interaction design involves many expertise [102]. Thus the HCI research engages many areas as illustrated in the Fig.3.

Fig. 3. Disciplines involved in HCI and their contribution.

Fig.3 is inspired by the book edited by Dix et al. [102], the center is the Human-Computer Interaction domain, and the surrounding are the expertise domains and their contribution. There are tributary researches to these main domains which are also related to HCI research area, for example: Artificial Intelligence, Pattern Recognition, Computer Vision, etc in Computer Science and Engineering. The functionality and usability of the interface have made its progress along with the computer and its usability evolution with an increasing focus on the user.

3.2 HCI Evolution

HCI has evolved through time according to the technological change. From the middle of the 20th Century, there have been three important waves of technological change that has determined the place of technology in our daily life as well as interactions with it [21]. The first one is the mainframe wave (or mainframe era) where one computer is shared by many users. The second wave is the Personal Computer wave, where one person uses one computer; interact with it directly in a deep way. The transition step is marked by the emergence of internet and distributed computing that interconnects many people from different place and using different resources. Along with the new innovation in technical computing, it has created a new computing wave changing our daily relationship with the technology. The third wave is the Ubiquitous Computing era [2, 3], when the computer completely pervades the user's life; many computers are shared among many users [1]. Thus, the Interaction between the user and the computers has become intelligent [105,106], where the computer is supposed to understand the meaning of the user's message and the context.

In the beginning the interfaces, were lines of commands [107]. They were used till the venue of the interfaces based on desktop referred to as WIMP-based (Windows, Icons, Menus, Pointers- based) Graphical User Interfaces (GUI) which has dominated the interface style since 80s. In the last decades, there have been the introduction of the new interaction techniques that go beyond the traditional desktop [108, 82, 103], more intelligent and active multimodal interaction [24, 30], embedded with user's natural environment.

3.3 HCI Current Research

The ubiquitous and pervasive computing has introduced the need for a new design of user interface leveraging the human natural capacity of communication. Thus, the interaction has become multimodal, perceptual and multimedia. The multimodal interaction systems intend to deliver natural and efficient interaction by the use of recognition-based technologies [30]. Research in the HCI, has moved from being machine-centered to being human-centered. The multimodality is introduced to build interfaces in the matching of user's communicative method. There are many surveys and reviews about the multimodal interaction in the literature, and many relating works. We refer the reader to [41, 24, 104, 109, 110, 20, 30].

The multimodal human-computer interaction is a new and ongoing innovation and the final result is the creation of powerful, efficient, natural and persuasive multimodal interfaces. The emerging of new human recognition based devices has enhanced the modeling of human natural interaction.

4 Multimodal Emotion Recognition System Conception

4.1 Information Fusion

The multimodal interaction brings in the need of fusing information in many levels. In literature there are various definitions of information fusion process, for example in Boström et al. [111]. The following Table 2 represents a summary of different terms used in the process of information fusion

Table 2. Information fusion definition summary

Taxonomy	Process	Objects	Sources	Objective
Data fusion, Multi sensor fusion, Sensor fusion, Information multimodal fusion, Information fusion, Data integration	Association, Joint, Integration, Combination, Manage complementary, information uncertainty, Bring together, Correlation, Merge, Multilevel	Data, Information, Sequences of observations, Measurements, Unknown objects, Data knowledge, Tools	Single or multiple Sources or sensors	Refinement, estimates, complete and timely assessment of a po-sition, an identity or a state of network, a threat, Specific and unified data about entity, activity or event, Specific vector of an observed system

By taxonomy we do refer to different names given to information fusion process but the most used word found is "Data fusion". Thus, data level is the basic target and the most reliable because the raw data are richer in information than post-processed features. In total have distinguished six different terms defined as core to the whole process; Fig.4 gives a global representation based on that taxonomy. By process we mean the principle actions done in fusion and though there are ten terms they differ from one another in literature but in similarity they join or combine or integrate the data/information. Objects are the subjects of the fusion and relate to the type of the data like sequences of observations in case of a multimedia resource. By sources we mean the origin of the data. Multiple sources can be many sensors of the same type that are differently used or different sensors used for same situation like a camera for visual recording and a microphone for audio recordings. By objective we mean the target of

the fusion. There are many terms used in literature to describe objective but the general meaning is a specific structure of data of an entity (classification, regression). The corresponding references were not inserted in the table to save the space, but they are in the global list of the references.

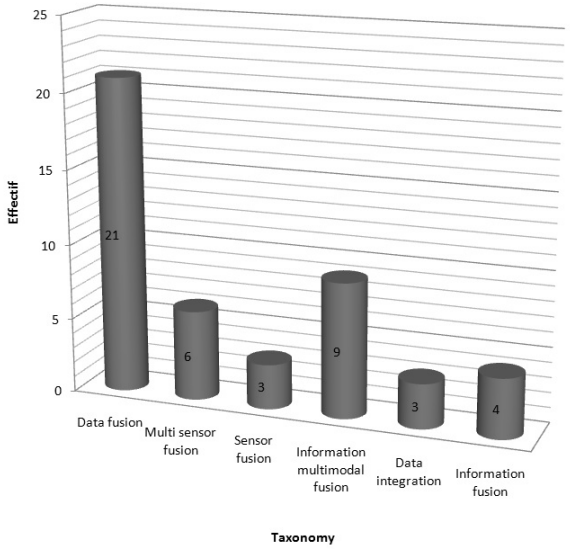

Fig. 4. Distribution of the different concept in the literature (46 references in total)

Fig.4 gives a representation of the nomenclature employed in 46 references. We have chosen the nomenclature representation because it is the controller of the subsequent steps. The variation in their distribution depends on the focus of the research: "data fusion" definition is more used when the research or the work is done on raw data (low level), "the multisensory and multimodal fusion" definition is used when the research focus is on the inputs; and the "data integration" mostly when the research focuses on the method. "Data fusion" distribution in the figure shows that most of the work focuses on combining the raw data though in reality most of the work especially in the emotion recognition is done on the features and/or high level.

As the current research and benefit of multimodality in emotion recognition are de-tailed in the section 2 here we go on by summarizing the important steps to be taken in a multimodal emotion analyser design process.

4.2 Data acquisition

Data acquisition consists of collecting raw data from different sensors according to the modality to be combined and the targeted emotional states: discrete versus continuous. This step is crucial for the design and the implementation automatic emotion recognition; the data have to reflect almost all the possible cases in real life. In many works, the data are acquired from own made dataset or from the datasets made by other researchers and made public for research. The dataset design setup is beyond the scope of this paper so we will limit on a brief description of the already existing multimodal emotional datasets in literature.

These datasets can be subdivided into three groups according to the accessibility towards the researchers: full access for free download and are especially designed for researches [112], there are others which are free for download after an End User Licence Agreement signature [113], and the last group is the commercial da-tasets which have to be purchased. However, the self made dataset is the best because it is supposed to contain all the information needed for the work in view. The only handicap is that it needs more time and special settings. Table 3 gives a summarized view of some examples of the emotion expression based datasets in the literature.

Table 3. Examples of existing emotion-based datasets

Dataset	Modalities	Naturalness	Brief description/References
Youtube dataset	Audio-visual and text	Natural	Collected from social website [65]
Audio Visual Vera Am Mittag (VAM)	Audio-Visual	Spontaneous	Audio-visual recording taken from a popular authentic and unscripted German TV talk show [142]
eNTERFACE'05 [112]	Audio-Visual	Acted	benchmark database for evaluation of multimodal emotion recognition algorithms.
FAU Aibo Emotion Corpus	Audio	Spontaneous	speeches of Children interact with a robot dog (Aibo) in a Wizard of Oz scenario [113].
SEMAINE corpus	Audio visual	Induced	A person converse with a limited agent [143]
SAL Data Set	Audio Visual	Induced	People react to the Sensitive Artificial Listener [35]
SmartKom dataset	Audio	Spontaneous	Spontaneous speech and natural emotions for German and English via wizard Of Oz
SAFE Corpus	Audio	Spontaneous	focus on fear type emotion [144]
Biwi 3D Corpus	Audio Visual	Acted	Recognition and simulation of emotion state [145]
IEMOCAP Dataset	Audio visual	Acted	Multimodal, multispeaker database [146].
GEMEP corpus	Audio, visual motion	Acted	10 actor portraying 18 affective states with different utterance content [147].
Cohen2003 dataset	Audio-visual	Acted	a comprehensive test-bed for the facial expression analysis algorithm [149].
AVIC Corpus	Audio Visual	Spontaneous	database created in order to deal with interest[148].
MMI da-	Facial expres-	Acted Sponta-	database with com-

taset [96]	sion	neous	mon training and testing forms
Cohn-Kanade FACS database	Facial expression	Acted/ Spontaneous	research in automatic facial image analysis and synthesis and for perceptual studies [149]
MAHNOB HCI [150]	Audio video and gaze	Spontaneous	a database with the goal of emotion recognition and implicit tagging research [151, 154]
DEAP database [152]	Physiological	Acted	A multimodal dataset for analysis of emotional states.
MIT Dataset [153]	Audio Visual	Spontaneous/Natural	For the detection of stress during the car driving by physiological measurements
FEEDTUM Database	Facial Expression	Spontaneous	For emotional states analysis from the facial expression [147,155,156]
PIT Corpus	Audio-visual	Spontaneous	Conceived for the multi-party conversation technology research [157]

In this Table 3, the datasets are arranged according to two main facts namely, recording context (spontaneous, induced and acted) and the modalities (face, voice, speech, eyes, body, hand, bio-physiological signal…) concerned.

Based on the recording context of the actors, we have categorized different datasets into three classes: acted, induced and spontaneous. Acted datasets are the sample in which the subjects' emotion is predefined. They are the most available and the easiest to acquire in research [128]. However, that facility is one of the shortcomings of such datasets, because the predefined conditions of the trained samples fail to cope with the real-life situations. However, in the raw data the use of perception tests to filter exaggerations and unnatural behaviour can help in handling some stereotypical expressions [114, 62].

In the induced emotion datasets, the actors are presented to some situations and are supposed to naturally express their emotions according to the situation. They are better than the acted in the sense that they permit the subject to express their emotion without the designer's special indication. However, they still have weakness resulting in the human inhibition/exaggeration of emotion due to the awareness of recording process [92].

The Natural (spontaneous) datasets are the best in view of the real life application, but they are the hardest to obtain because of the security and privacy of the actors. In the literature they are mostly collected from the internet social and/or multimedia websites (YouTube, Facebook, Twitter, etc), TV show or live news broadcasts. Their limitation is that they are difficult to accurately annotate and very expansive to record [115].

The modalities of interest can be audio, visual, textual, physiological signal and/or Key-strokes. They can be one mode (unimodal), two modes (bimodal) or more modes (multi-modal). The choice of the modality depends on the work at hand and the material availa-ble. The most existing modality is the facial expression datasets because of the availability of the recorders (camera, webcam, etc); and the least is the bio-physiological based dataset because of the difficult accessibility to the recording material and their obstruction (need sensors on the body).

The recording material's quality matters because some issues like sensor failure can add errors and noise in the collected data. The smaller amount of training data can facilitate the training process but the test hardly generalizes [116]. So these facts are to be well known before the subsequent steps can be addressed accordingly. The data collected contain some unnecessary data, irrelevant and erroneous data, or sometimes missing values. Thus, they need to be prepared in the data pre-processing step for later processing.

4.3 Data Pre-Processing

Data pre-processing is the process related to preparing the data for further processing in view of satisfying the intended use. Major steps involved in data pre-processing are data cleaning, data integration, data reduction and data transformation [117]. For further information about the above steps, we refer the reader to Han et al. [117] and the book by Theodoridis and Kourtroumbas, Pattern Recognition [118].

The data pre-processing is mostly about:

Cleaning the raw data by removing noise or outliers, filling in missing values, resolving inconsistency using different methods. Outlier removal techniques insensitive to outliers can be found in the book by Huber and Ronchetti, Robust statistics [116]. In case of outliers resulting from the sensor (measurement) it is very difficult because there is no replacement of a good sensor, thus it is better to examine the information content of the sensor data and intelligently select algorithms for sensor pre-processing [119].

Dealing with missing data: when the training samples in complete value are in high number, the small number of the incomplete samples can be discarded otherwise missing values will be replaced by approximated value like the mean calculated from the available values of the respective feature or by interpolation.

Uncertainty and inconsistent handling approaches can be enriched by the details given in the work of Khaleghi et al. [120] and Abdulhafiz and Khamis [121].

There are other data transformation strategies and we refer the reader to Han et al. [117] and to Lyons [122] The data pre-processing step result in a dataset clean and crude, it needs to be put in a good representation for a straightforward modeling during feature extraction step.

4.4 Features Extraction

Up to this stage the data collected contain examples or samples with same variables or attributes but different values characterizing each examples by class or category. These variables are called Features or Input variables or attributes [123]. They are the representation of the data; they can be binary, categorical or continuous and differ from one modality to an-

other (Table 4). The feature extraction consists of drawing out the feature relating to the modality. In multimodal emotion recognition, the feature extraction often defines basic features which can encompass other intermediate features as shown in Table 4. Multiple feature values will constitute the vector feature of the modality [124, 125]. Table 4 gives a summary of some different feature extracted in view of emotion recognition:

Table 4. Modality and extracted features

Modality	Feature		
	Basic	Intermediate	
Face expression	eyes, eyebrows, nose, mouth	Action Units	
Speech	linguistic	Words, multiwords, phrases, sentences, documents	
	Paralinguistic	Pitch, bandwidth, duration, prosodic, voice quality, Mel frequency Cepstral coefficients (MFCC)	
Body	Head gestures	Head position	
		Head movement	
	Hand Gestures	Shape	
		Motion	
	Body motion	Spinal column	Neck, chest and Abdomen
		DOF body	Symmetrical arms
		Body center mass	Movement of body center of mass
		Joints	Degree of joint rotation
Physiologic	Hearth, Brain, Limbs, Blood	Electrocardiogram(ECG); Electro dermal activity (EDA); Electromyogram (EMG)	

The exact determination of the most relevant feature to extract in view of the affective states is an open research topic [126, 127]. The most common modalities in studies are; the face expression, speech, body motion, hand gesture, physiological signal; it is mostly dependent on the datasets availability and accessibility. The features extraction can be subdivided into two steps: Features construction and Features selection.

The feature construction consists of determining the good data representation according to the domain specifications and measurements availability [123]. The features can be manually built and complemented by automatic feature construction methods.

Recently, some automatic feature Extraction Tools have been proposed to be used in the extraction of some specific features. Some examples of the automatic features extraction tools are summarized in the Table5.

Table 5. Feature Extraction tools/algorithms

Toolkit	Modality	Feature extracted/ Function	Brief description
PRAAT [129]	Audio	Duration, F0, Range, Movement, Slope, Energy features	PRAAT (a system for doing phonetics) [130]
FEELTRACE [71, 26]	Audio	Labelling	FEELTRACE, allowing the emotional dynamics of speech episodes to be examined.
OpenEAR [131, 61, 34]	Audio	Signal Energy, Loudness, Mel-/Bark-/Octave-Spectra, MFCC, PLP-CC	openEAR provides effi-cient (audio) feature ex-traction.
OpenSMILE [134, 135]	Audio	Signal Energy, Loudness, Formants, Mel-/Bark-/Octave-Spectra, MFCC, PLP-CC, Pitch, Voice quality (Jitter, Shimmer), LPC, Line Spectral, Pairs(LSP), Spectral, Shape description	It is an open source toolkit, for feature extraction in machine learning and data mining by researchers in voice processing, vision signal processing and Music Information Retrieval.
EyesWeb [124]	Body	Quantity of motion, cue, Contraction index of the body, velocity, Acceleration, fluidity of the hand's, barycentre	Open software for extended Multimodal Interaction. University of Genoa - Italy
Luxand FSDK 1.7	Face	Action Units	Facial recognition software [61]
ANVIL [132, 133]	Audio	Annotation tool in a multi-modal dialogue	Free for research purposes. [63, 136]
AAM [137]	Face	Texture and shape	It uses a statistical approach [60]
GAVAM [61]	Face	Face component detection	Constrained local models

In Table 5 above, these tools are, mostly open source and can be downloaded from the internet. They are mostly compatible with the most popular platforms such as Windows, Macintosh, and Linux. The features construction is followed the selection of the useful and significant features.

The features selection mainly aims at selecting features which are more informative and relevant to the task at hand. In general, the features construction can build up to thousands of features which requires an important amount of storage and slow down the training process. In order to solve these inconveniencies, the feature selection uses a general data reduction method by eliminating irrelevant, redundant information to a sufficient minimum dimension. Features selection helps to understand the data and improve the performance. The main idea is to have features with large distance between classes and small variance in the same class. For more details, we refer the reader to Theodoridis and Kourtroumbas [118].

4.5 Information Fusion

The information fusion is done in three levels: Data level Fusion, Feature level fusion and Decision (or semantic) level

fusion. However, the feature and decision level fusions are mostly used techniques in multimodal emotion recognition. The following are brief descriptions of these fusion levels.

4.5.1 Data-Level Fusion.

Data-level fusion is carried out just on the raw data. It is better done for the data provided by a similar kit for same type of data such as multiple cameras recording the same scene from different point of view. The raw data is smeared by noise but the data-fusion level is hypothetical to give better accuracy because there is less loss of basic information [41]. Data level fusion is more accurate with expensive computation.

4.5.2 Feature-Level Fusion.

It is the fusion done by combining the features from each modality to form one vector which will be presented to the classifier for training [41, 138]. Feature fusion is more adequate when the modalities are closely coupled and synchronized to get an easy generalization. It has to be noted that at this stage, the dimensionality of the input features will increase and necessitate collecting a large amount of the training samples [126].

4.5.3 Decision-Level Fusion.

The decision-level fusion is the highest level of combination. Initially, each modality is trained on its own classifier and the output will be combined and filled to one final classifier for the final decision. It concerns the integration of asynchronous and temporally correlated modalities. In practice, it is more favored above the previous ones because each modality is trained apart and the integration doesn't need retraining [11]. In the literature, the techniques used for the final decision vary from the weighted sum rule, Fisher Discrimant Analysis, decision trees, etc.

The following figure (Fig. 5) summarizes the different steps for the multimodal emotion recognition design.

Fig.5 represents the four main steps of multimodal emotion recognition design namely, Data collection, Data preprocessing, Feature extraction and Fusion. The data collection step represents the different ways of collecting the data. Though each modality has its own recording symbol, there can be recordings that can combine two or three modality like in the case of a webcam with integrated microphone, which can record visual and audio content. The data flow continues to the pre-processing step. We represent them in features extraction because the steps of data pre-processing and feature construction are related. There is no data level fusion in multimodal emotion recognition because the multi-modal emotion there must be at least two modalities and the data level fusion is held in. The Feature extraction, we represent some of the features for the either the feature level combination of feature classification in order to the further classification; the decision-level fusion.

The main issue is the evaluation of the algorithms because up to now there is no standard evaluation framework to measure the performance of data fusion algorithms [120]. Most of the learning cases, the solution is evaluated under conditions similar to the training conditions [139]; the research should focus on building system that can be educative, and increase knowledge, skills and plans through experience [37]. The work of Janssen et al. [140], machines outperformed human in emotion recognition gives an example of how to set benchmark dataset.

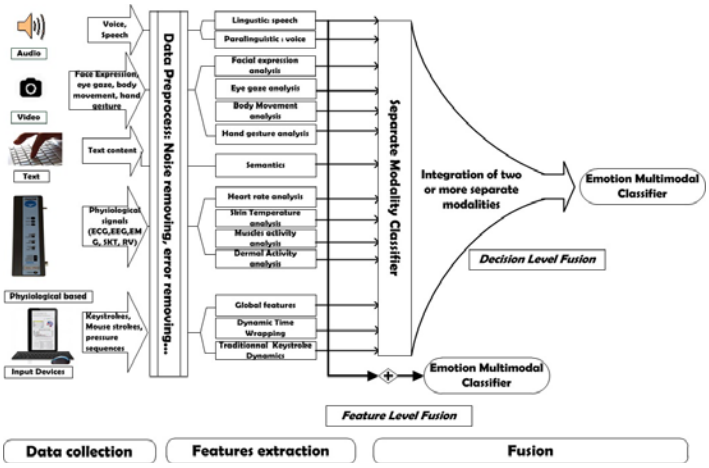

Fig. 5 Multimodal Emotion Recognition Design System conception summary

Multimodal emotion recognition is still at its infancy especially in some areas such as the input devices for emotion recognition, but as Picard [15], suggests, in a difficult situation partial solution can still be of value. The progress in machine learning methods [16, 141, 98] demonstrates that the real-time and effective applications' manifestation is at hand.

5 Conclusion and Recommendation

This paper gives a summary of three research topic: emotion recognition (affective computing) research, human computer interaction and information fusion research, focusing on the definition of the emotion to the automation of emotion recognition. An emotion study is an interdisciplinary field.

Human emotion expression is multimodal. An automatic multimodal emotion recogniser implies information fusion techniques. We have shown that through different studies done on information fusion the main idea is about fusing the raw data but the semantic level is the more performed because it is computationally easier to implement than data and features level fusions. However, the two first levels yield better accuracy because of wealth of information. In addition, we have done the analysis of main steps taken through the design of multimodal emotion classification and pointed out the more important issue and probable remedy.

Though much work has already been done, the natural and real time affective user-machine interaction is still a desire. The progress already done in the automatic feature extraction, and the new efficient and robust machine learning algorithms such as Support Vector Machine and its derivatives, Neural Networks and its derivatives; have given hope that the natural and real-time affective-based software, is not for long.

Emotion recognition integration in human computer interaction has added utility in daily human life such as in Intelligent Tutoring System and Healthcare Systems. Thus, it is one of the emerging and growing research trend that needs more support. Although it is difficult to deal with instability in human behaviour, the intended improvement will enhance ubiquitous computing. The same way airplanes are getting in highest without flapping wings; computers with ability of rec-

ognizing, communicating and expressing emotions, aim at promoting intelligent interaction and decision making.

ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China under Grant No.61300074 and No.61272357, the new century personnel plan for the Ministry of Education (NCET-10-0221).

REFERENCES

[1] Friedewald, M., Oliver, R., 2010. Ubiquitous computing: An overview of technology impacts. Telematics and Informatics 28 (2), 55-65

[2] Weiser, M., 1991. The Computer for the 21st Century. Scientific American 265, pp. 94-104

[3] Weiss, R. J., Craiger, J.P., 2002. Ubiquitous Computing. Industrial-Organizational Psychologist 39 (4), 44-52

[4] Palen, L., Bødker, S., 2008. Don't Get Emotional. In Peter, C., and Beale, R. (Eds), Affect and Emotion in HCI of the series LNCS 4868, 12–22.

[5] Banda, N., Robinson, P., 2011. Multimodal Affect Recognition in Intelligent Tutoring Systems. In S. D'Mello et al.(Eds.): ACII, Part II, LCNS 6975, pp.200-207.

[6] Petrovica S., Ekenel H. K., 2016. Emotion Recognition for Intelligent Tutoring. In Joint Proceedings of the BIR 2016 workshops and Doctora; Vol.1684.

[7] Wu, Y.W., Liu, W., Wang, J.B., 2008. Application of Emotion Recognition in Intelligent Tutoring System. In Workshop on Knowledge Discovery and Data Mining, 449-452.

[8] Zatarain-Cabada, R., Barron-Estrada, M. L., Alor-Hernaneez, G., Reyes-Garcia, C. A., 2014. Emotion Recognition in Intelligent Tutoring Systems for Android-Based Mobile Devices. A. Gelbukh et al. (Eds.): MICAI 2014, Part I, LNAI 8856, pp. 494-504.

[9] Tokuno, S., Tsumatori, G., Shono, S., Takei, E., Yamamoto, T., Suzuki, G., Mituyoshi, S., Shimura, M., 2011. Usage of Emotion Recognition in Military Health Care: Detecting Emotional Change under Stress. Defense Science Research Conference and Expo, pp. 1-5.

[10] Tivatansakul, S. Ohkura, M., Puangpontip, S., Achalakul, T., 2014. Emotional healthcare system: Emotion detection by facial expression using Japanese database. 6th Conference on Computer Science and Electronic Engineering. CEEC2014, pp41-46.

[11] Corradini, A., Mehta, M., Bernsen, N. O., Martin, J.C., Abrilian, S.,2003. Multimodal Input Fusion in Human-Computer Interaction - On the Example of the NICE Project." In NATO-Advanced Studies Institute (ASI) conference on Data Fusion for Situation Monitoring, Incident Detection, Alert, and Response management. Science Series, III: Computer and Systems Sciences, 223-234.

[12] Pantic, M., Pentland, A., Nijholt, A., Huang, T. S.,2007. Human Computing and Machine Understanding of Human Behavior: A Survey. Human Computing of the series LNAI 4451, edited by Huang, T. S., Nijholt A., Pantic M., Pentland A., 47–71.

[13] Peter, C., Beale, R., 2008. The Role of Affect and Emotion in HCI. In Peter, C., and Beale, R. (Eds), Affect and Emotion in HCI of the series LNCS 4868, 23–34.

[14] Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X. F., Kirbas, C., McCullough, K. E., Ansari, R., 2002. Multimodal human discourse: gesture and speech. Journal of ACM Transactions on Computer - Human Interaction (TOCHI) 9 (3), 171–193.

[15] Picard, R.W., 1997. Affective Computing. MIT Press1997, MIT Media Laboratory Perceptual Computing Section Technical Report No 321Press.

[16] Deng, L., Yu D., 2013. Deep Learning: Methods and Applications. In Foundations and Trends in Signal Processing, Vol. 7, Nos. 3-4, 197-387.

[17] Bunt, H., Beun, R. J., Borghuis, T., 1998. Multimodal human–computer communication systems, techniques, and experiments. Series LNCS (LNAI) 1374.

[18] Reeves, B., Nass, C., 1996. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.University of Chicago Press

[19] Nigay, L., Coutaz, J., 1993. A design space for multimodal systems: Concurrent processing and data fusion. In Proceedings of the Conference on Human Factors in Computing Systems, INTERCHI'93, 172-178.

[20] Sebe, N., 2009. Multimodal interfaces: Challenges and perspectives", Journal of Ambient Intelligence and Smart Environment 1, 19-26.

[21] Weiser, M., Brown, J. S., 1996. The Coming Age of Calm Technology." Bits flowing through the wires of computer network are invisible; a "network monitor", 1-8

[22] Lew, M., Bakker, E. M., Sebe, N., Huang, T. S., 2007. Human-Computer Intelligent Interaction: A Survey. In Proceedings of International Workshop on Human-Computer Interaction (HCI 2007), 1–5.

[23] Shen, L. P., Wang, M.J., Shen, and R.M., 2009. Affective e-Learning: Using "Emotional" Data to Improve Learning in Pervasive Learning Environment. Educational Technology & Society, 12 (2), 176-189.

[24] Gupta, R., 2012. Human Computer Interaction - A Modern Overview. International Journal of Computer Technology and Applications (IJCTA), 3 (5), 1736-1740.

[25] Busso, C., Deng, Z-G., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S., 2004. Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. In Proceedings of 6th International Conference on Multimodal Interfaces (ICMI'04), 205-211.

[26] Pantic, M., Caridakis, G., André, E., Kim, J., Karpouzis, K., Kollias, S., 2011. Multimodal Emotion Recognition from Low-Level Cues. In Petta, P., Pelachaud, C., Cowie, R., (Eds), Emotion-Oriented Sytems: The HUMAINE Handbook with 104 Figures and 35 Tables, pp.115-132.

[27] Pitsikalis, V., Katsamanis, A., Theodorakis, S., Maragos, P., 2015. Multimodal Gesture Recognition via multiple Hypotheses Rescoring. Journal of Machine Learning Research 16, 255-284.

[28] Brave, S., Nass, C., 2002. Emotion in human-computer interaction. In Jacko, J.A., Sears, A. (Eds), The Human-Computer Interaction Handbook, pp.81-96.

[29] Sun, Y., Chen, F., Shi, Y. (David) Chung, V., 2006. A novel method for multi-sensory data fusion in multimodal human computer interaction." In Proceedings of the 18th International Conference on Computer-Human Interaction: Design: Activities, Artefacts and Environment. OZCHI'06 206, pp.401-404.

[30] Turk, M. 2013. Multimodal interaction: A review. Pattern Recognition Letters 36, 189-195.

[31] Bolt, R., 1980. "Put That There": voice and gesture at the graphics interface." ACM Computer Graphics 14 (3), 262-270.

[32] Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., ittman, J., Smith, P., Chen, L., Clow , J.,. 1997. Quickset: multimodal interaction for distributed applications. In Proceedings of the Fifth ACM international Conference on Multimedia, MULTIMEDIA'97, 31-40.

[33] Song, P., Winkler, S., Gilani, S. O., Zhou, Z-Y., 2007. Vision-based projected Tabletop Interface for Finger Interactions. IEEE International Workshop, HCI 2007, 49-58.

[34] Poria, S., Cambria, E., Howard, N., Huang, G. B., Hussain, A., 2015. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Journal of Neurocomputing 174 (A-22), 50-59.

[35] Karpouzis, K., Caridakis, G., Kessous, L., Amir, N., Raouzaiou, A., Malatesta, L., Kollias, S., 2007. Modeling Naturalistic Affective States via Facial, Vocal, and Bodily Expressions Recognition. In Huang T.S., Nijholt A., Pantic M., Pentaland A. (Eds), Artificial Intelligence for Human Computing, LNAI 4451, pp. 91–112.

[36] Lin, H. C. K., Wang, C.H., Chao, C.J., Chien, M.K. 2012. Employing Textual and Facial Emotion Recognition to design an Affective Tutoring System. TOJET 11 (4), 418-426.

[37] Pantic, M., Pentland, A., Nijholt, A., Huang, T. S., 2006. Human computing and machine understanding of human behavior: a survey. In Proceedings of the 8th International Conference on Multimodal Interfaces, 239-248.

[38] Lee, H., Choi, Y. S., Lee, S., Park, I. P., 2012. Towards Unobtrusive Emotion Recognition for Affective Social Communication. In the 9th annual IEEE Consumer Communication and Networking Conference – Special Session Affective Computing for Future Consumer Electronics, 360-364.

[39] Microsoft Kinect Sensor and its Effect, www.microsoft.com/research/wp-content/uploads/2016/02/Microsoft Kinect Sensor and its Effect – IEEE MM 2012.pdf.

[40] Akputu, K. O., Seng, K. P., Lee, Y. L., 2013. Facial Emotion Recognition for Intelligent Tutoring Environment. In 2nd International Conference on Machine Learning and Computer Science (IMLCS'2013) pp.9-13.

[41] Dumas, B., Lalanne, D., Oviatt, S., 2009. Multimodal Interfaces: A Survey of Principles, Models and Frameworks. In Lalanne, D., Kohlas, J. (Eds), Human Machine Interaction (5440) of the series LNCS, 3-26.

[42] Poria, S., Cambria, E., Hussain A., Huang, G.B.,2014. Towards an intelligent framework for multimodal affective data analysis. Journal of Neural Networks 63, 104-116.

[43] Dixon, T., 2012. "Emotion": The History of a Keyword in Crisis. Emotion Review vol.4 (4) pp.338-344.

[44] William James, "What is an Emotion?" Mind. Vol.9(34), pp. 188-205, Apr.1884.

[45] Egon L. van den Broek, 2011. AFFECTIVE SIGNAL PROCESSING unravelling the mystery of emotions, PhD Dissertation.

[46] Izard, C. E, 2010. More Meanings and More Questions for the Term "Emotion", Emotion Review, vol. 2(4), pp.383-385.

[47] Kleinginna Jr., P. R., Kleinginna, A. M., 1981. A categorized List of Emotion Definitions, with Suggestions for a Consensual Definition. Motivation and Emotion, Vol.5 (4), 345-379.

[48] Wassmann, C., 2016. Forgotten Origins, Occluded Meanings: Translation of Emotion Terms. Emotion Review. Doi:10.1177/1754073916632879.

[49] Damasio, A.R., Grabowski, T. J., Bechara, A., Damasio, H., Ponto, L. L. B., Parvizi, J., Hichwa, R.D., 2000. Subcortical and cortical brain activity during the feeling of self-generated emotions. Nature Neuroscience 3, 1049–1056.

[50] Panksepp, J., 2003. At the interface of the affective, behavioral, and cognitive neuro-sciences. Brain and Cognition 52, 4–14.

[51] Allen, J. A., 1997. Pragmatic Neuropsychology: A Review of the Neurological Side of Neuropsychology by Richard Cytowic. PSYCHE 3(3). 1-6.

[52] Cytowic, R. E., 1996. The Neurological Side of Neuropsychology. Behavior and Philosophy 24 (2), 191-194.

[53] Goleman, D., 1995. Emotional intelligence: Why It Can Matter More Than IQ. New York, USA.

[54] Ishii, H., Wisneski, C., Brave, S., Dahley, A., Gorbet, M., Ullmer, B., Yarin, P., 1998. AmbientROOM: integrating ambient media with architectural space. In the Conference Summary on Human Factors in Computing Systems (CHI'98), 173–174.

[55] Pelachaud, C., Carofiglio, V., De Carolis, B., De Rosis, F., Poggi, I., 2002. Embodied Contextual Agent in Information Delivering Application. In Proceedings of the 1st International conference on autonomous agents and multi-agent systems: part 2, 758-765.

[56] Balomenos, T., Raouzaiou, A., Ioannou, S., Drosopoulos, A., Karpouzis, K. Kollias, S., 2005. Emotion Analysis in Man-Machine Interaction Systems. In S. Bengio and H. Bourlard (Eds.): MLMI 2004, LNCS 3361, pp.318-328.

[57] Fragopanagos, N., Taylor, J. G., 2005. Emotion Recognition in Human-Computer Interaction, Neural Networks, 18, pp.389-405.

[58] Scherer, R. K., 2009. Emotions are emergent processes: they require a dynamic computational architecture. Phil. Trans. R. soc. B 364, 3459-3474.

[59] Ekman, P., Huang, T. S., Sejnowski, T. J., Hager, J.C., 1993. Understanding the face. A Human Face eStore.

[60] Edwards, G.J., Taylor, C.J., Cootes, T.F., 1998. Interpreting face images using active appearance models." In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition (FG'98), 300-305

[61] Saragih, J.M., Lucey, S., Cohn, J.F., 2009. Face alignment through subspace constrained mean-shifts. IEEE 12th International Conference of Computer Vision (ICCV), 1034 – 1041.

[62] Scherer, R. K., Johnstone, T., Klasmeyer, G.,2003. Vocal expression of emotion. In Davidson, R. J., Scherer, K. R., Goldsmith, H. H. (Eds), Handbook of affective sciences, Chap.23,433-456.

[63] Navarretta, C., 2011. Individuality in Communicative Bodily Behaviors. In Esposito, A., Esposito, A., Vinciarelli, A., Hoffmann, R., Müller, V. C. (Eds), Cognitive Behavioral Systems, LNCS 7403, 417–423.

[64] Knapp, B. R., Kim, J., André, E.,2011. Physiological Signals and Their Use in Augmenting Emotion Recognition for Human–Machine Interaction. In Petta, P., Pelachaud, C., Cowie, R., (Eds), Emotion-Oriented Sytems: The HUMAINE Handbook with 104 Figures and 35 Tables, pp.133-161.

[65] Kim, J.,2007. Bimodal Emotion Recognition using Speech and Physiological Changes. In Grimm, M., Kroschel, K.,(Eds), Robust Speech Recognition and Understanding, pp.265-280.

[66] Morency, L. P., Mihalcea, R., Doshi, P.,2011. : Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. In Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI'11), 169-176. (2011).

[67] Zheng, W.L., Dong, B.N., Lu, B.L., 2014. Multimodal Emotion Recognition using EEG and Eye Tracking Data. In Proceedings of 36th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (EMBC), 5040-5043.

[68] Fazekas, A., 2006. Multi-modal Human-Computer Interaction. Image Processing Group of Debrecen.

[69] Kessous, L., Castellano, G., Caridakis, G., 2010. Multimodal Emotion Recognition in Speech-based Interaction Using Facial Expression, Body Gesture and Acoustic Analysis. Journal on Multimodal User Interfaces 3(1), 33-48.

[70] Lefter, I., Burghouts, G. J., Rothkrantz, L.J. M., 2014. An audio-visual dataset of human–human interactions in stressful situations. Journal on Multimodal User Interfaces 8 (1), 29–41.

[71] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M.,

Schröder, M.,2000. FEELTRACE: an instrument for recording perceived emotion in real time. In Proceedings of the ISCA workshop on speech and emotion: A Conceptual Framework for research, 19-24.

[72] Ekman, P., Friesen, W. V.,1969. The repertoire of nonverbal behavioral categories – origins, usage, and coding. Semiotica 1, pp. 49–98.

[73] Zhu, X., Ramanan, D.,2012. Face detection, pose estimation and landmark localization in the wild. In Proceedings/CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[74] Keltner, D., Ekman, P., 2000. Facial expression of emotion. In Lewis, M., Haviland-Jones, J. M.(Eds), Handbook of emotions, 2nd edition, pp. 236–249.

[75] Ekman, P., 1994. Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. Psychological Bulletin 115 (2), 268–287.

[76] Watson, D., Weber, K., Smith-Assenheimer, J., Clark, L. A., Strauss, M. E., McCormick, R. A., 1995. Testing a tri-partite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptom scales. Journal of Abnormal Psychology 104 (1), 3–14.

[77] Larsen, R. J., Diener, E., 1992. Promises and problems with the circumflex model of emotion. In Clark, M.S.(Eds), Personality and Social Psychology Review 13, 25-59.

[78] Levenson, R. W., 1994. Human emotion: a functional view. In Ekman, P., Davidson, R. J. (Eds), The Nature of Emotion: Fundamental Questions, 123–126.

[79] Fridlund, J. A.,1997. The new ethology of human facial expression. In Russell, J.A., Fernandez-Dols, J.M. (Eds.), The psychology of facial expression, pp.103-129.

[80] Esposito, A., Marinaro, M.: Some notes on nonlinearities of speech. In Chollet, G., Esposito, A., Faundez-Zanuy, M., Marinaro, M.(Eds), Nonlinear Speech Modeling and Applications, LNCS 3445, 1–14.

[81] Jack, R. E., Garrod, O. G. B., Yub, H., Caldara, R., Schyns, P. G., 2012. Facial expressions of emotion are not culturally universal. Proceedings of the National Academy of Science 109 (19), 7241–7244.

[82] Taylor, J. G., Fragopanagos, N. F., 2005. The interaction of attention and emotion. Journal of Neural Networks 18 (4), 353-369.

[83] Sellers, M., 2013. Towards a comprehensive theory for biological and artificial agent. In Biological Inspired Cognitive Architecture 4, 3-26.

[84] Maslow, H.A., 1954. Motivation and personality. New York, NY: Harper.

[85] Larue, O., Poirier, P., Nkambou, R.,2012. The emergence of (artificial) emotions from cognitive and neurological process. In Biological Inspired Cognitive Architecture 4, 54-68.

[86] Samsonovich, V. A., 2012: On a roadmap for the BICA Challenge. In Biological Inspired Cognitive Architecture 1, 100-107.

[87] Lindsquit, A. K., Wager, T. D., Kober, H., Bliss-Moreau, E., Barrett, L. F., 2012. The brain basis of emotion: A meta-analytic review. Behavior and Brain Sciences 35 (03), 121-143.

[88] Stanovich, E. K., 2010. Rationality and reflexive mind. Oxford University Press.

[89] Fellous, J.-M., 1999. Neuromodulatory Basis of Emotion. In Neuroscientist 5(5), 283-294.

[90] Lövheim, H., 2012. A new three-dimensional model for emotions and monoamine neurotransmitters. In Medical Hypotheses 78 (2), 341-348.

[91] Jaimes, A., Sebe, N., 2005. Multimodal Human Computer Interaction: A Survey. IEEE International Workshop on Human Computer Interaction in conjunction with ICCV.

[92] Ambady, N., Rosenthal, R.,1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A Meta-analysis. Psychological Bulletin 111(2), 256–274.

[93] De Gelder, B., 2009. Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. Philosophical transactions of The Royal Society: Biological Sciences 364, 3475-3484.

[94] Mehrabian, A., 1968. Communication without words. Psychology Today 2 (4), 52–55.

[95] Van Vliet, V., 2012. Communication Model by Albert Mehrabian. http://www.toolshero.com/communication-skills/communication-model-mehrabian.

[96] Pantic, M., Rothkrantz, L. J. M., 2003. Towards an affect-sensitive multimodal human–computer interaction. Proceedings of the IEEE Special Issue on Multimodal Human-Computer Interaction (HCI) 91(9), 1370-1390.

[97] Picard, R. W., 2003. Affective computing: Challenges. International Journal of Human-Computer Studies 59 (1), 55–64.

[98] Vapnik, V., Izmailov, R., 2015. Learning Using Privileged Information: Similarity Control and Knowledge Transfer. Journal of Machine Learning Research 16, 2023-2049.

[99] Vapnik, V., Vashist, A., 2009. A new learning paradigm: Learning using privileged information. Neural Networks 22, 544-557.

[100] Wang, S.-F., Zhu, Y.-C., Yue, L.-H., Ji Q., 2015. Emotion Recognition with the Help of Privileged Information. In IEEE Transactions on Autonomous Mental development, 189-200.

[101] Card, K. S., Moran, T. P., Newell, A.,1980. The Keystroke-Level Model User Performance Time with Interactive Systems. Communications of the ACM 23 (7), 396-410.

[102] Dix, A., Finlay, J., Abowd, G., Beale, R., 2003. Human-Computer Interaction: Second Edition.

[103] Te'eni, D., Carey, J. M., Zhang, P., 2007. Human Computer Interaction: Developing Effective Organizational Information Systems. Hoboken.

[104] Karray, F., Alemzahed, M., Saleh, J. A., Arab, M. N., 2008. Human-Computer Interaction: Overview on State of the Art. International Journal on Smart Sensing and Intelligent Systems 1 (01), 137-159.

[105] Turk, M., 1999. Perceptual User Interfaces. in Workshop report.

[106] Turk, M., Robertson, G., 2000. Perceptual User Interfaces. Communication of ACM, 33-34.

[107] Myers, B. A., 1998. A Brief History of Human Computer Interaction Technology. ACM interactions 5 (2), 44-54.

[108] Srivastava, L., 2005. Mobile phones and the evolution of social behavior. Behavior & Information Technology 24(2), 111-129.

[109] Lalanne, D., Nigay, L., Palanque, P., Robinson, P., Vanderdonckt, J., Ladry, J. F., 2009. Fusion Engines for Multimodal Input: A Survey. ACM International Conference on Multimodal Interface, pp.153-160.

[110] Pleari, M., Lisetti, C. L.2006. Toward multimodal fusion of affective cues. HCM '06 Proceedings of the 1st ACM international workshop on Human-centered multimedia, pp. 99-108.

[111] Boström, H., Andeler, S. F., Brohede, M., Johansson, R., Karlsson, A., van Laere J., Niklasson, L., Nilsson, M., Persson, A., Ziemke, T., 2007. On the Definition of Information Fusion as a Field of Research. Technical report.

[112] Martin, O., Kotsia, I., Macq, B., Pitas, I., 2006. The eNTERFACE'05 Audio-Visual Emotion Database. Proceedings of 22nd International Conference on Data Engineering Workshops (ICDEW'06).

[113] Steidl, S., 2009. Automatic Classification of Emotion-related user states in Spontaneous Children's Speech. PhD Dissertation.

[114] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., 2005. A database of german emotional speech. In Proceedings Interspeech'2005-Eurospeech, 9th European Conference on Speech Communication and Technology, 1517–1520.

[115] Campbell, N., 2007. On the use of non verbal speech sounds in human communication. In Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (Eds), Verbal and Nonverbal Communication Behaviors of the series LNCS 4775, Chap. 11, 117–128.

[116] Huber, P. J., Ronchetti, E. M.: Robust statistics (2nd Edition). Wiley. (2009).

[117] Han, J., Kamber, M., Pei, J. 2012. Data Preprocessing. Chap.3 in Data Mining: Concepts and Techniques, 84-124.

[118] Theodoridis, S, Kourtroumbas, K., 2009. Feature Selection." Chap. 5 in Pattern recognition, 4th edition.

[119] Hall, D. L., Garga, A. K., 1999. Pitfalls in data Fusion (and How to Avoid Them). In Proceedings of the 2nd International Conference on Information Fusion – FUSION'99, 1, 429-436.

[120] Khaleghi, B., Khamis, A., Karray, F. O., Ravazi, S. N., 2013. Multisensor data fusion: a review of the state-of-the-art." Information Fusion14 (1), 28–44.

[121] Abdulhafiz, W. A., Khamis, A., 2013. Handling Data Uncertainty and Inconsistency Using Multisensor Data Fusion." Advances in Artificial Intelligence Volume 2013.

[122] Lyons, R. G., 2004: Understanding Digital Signal Processing. Prentice Hall.

[123] Guyon, I., Elisseeff, A., 2006. An Introduction to Feature Extraction. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. (Eds), Feature Extraction: Foundations and Applications, Volume 207 of the series Studies in Fuzziness and Soft Computing, Chap 0, 1-25.

[124] Castellano, G., Kessous, L., Caridakis, G., 2008. Emotion Recognition through Multiple Modalities: Face, Body Gesture, and Speech. In Affect and Emotion in HCI, edited by Christian P., Beale R., of the series LNCS 4868, 92-103.

[125] Graves, A., Schmidhuber, J., Mayer, C., Wimmer, M., Radig, B., 2008. Facial Expression Recognition with Recurrent Neural Networks.

[126] Karpouzis, K., 2011. Editorial: Signals to Signs – Feature Extraction, Recognition, and Multimodal Fusion. In Petta, P., Pelachaud, C., Cowie, R., (Eds), Emotion-Oriented Sytems: The HUMAINE Handbook with 104 Figures and 35 Tables, pp. 65-70.

[127] Vertegaal, R., Slagter, R., van der Veer, G., Nijholt, A., 2001. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In Proceedings of the SIGCHI conference on human factors in computing systems, 301–308.

[128] Scherer, S., Glodek, M., Layher, G., Schels, M., Schmidt, M., Brosch, T., Stephan T., Schwenker, F., Neumann, H., Palm, G., 2012. A generic framework for the inference of user states in human computer interaction: how patterns of low level communicational cues support complex affective states. Journal on Multimodal User Interfaces 6, 117- 141.

[129] Huang, Z.Q., Chen, L., Harper, M., 2006. An Open Source Prosodic Feature Extraction Tool. In Proceedings of the Language Resources and Evaluation Conference (LREC'2006).

[130] Pokorny, F., 2011. Extraction of Prosodic Features from Speech Signals. Toningenieur-Projekt, UniverstätfürMusik und darstellende.

[131] Eyben, F., Wöllmer, M., Schuller, B., 2009. openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009), 1-6.

[132] Kipp, M., 2001. Anvil - a generic annotation tool for multimodal dialogue. In Proceedings of Eurospeech 2001, 1367-1370.

[133] Hoffmann, R., 2006. Recognition of non-speech acoustic signals. In Kacic, Z.(Eds), Proceedings of the International Workshop on Advances in Speech Technology Advances, AST 2006.

[134] Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In Proceedings of the 21st ACM International Conference on Multimedia (MM), 835-838.

[135] Eyben, F., Wöllmer, M., Schuller, B., 2010. openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In Proceedings of the 18th ACM International Conference on Multimedia (MM'10), 1459-1462.

[136] Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., Konosu, H., 2009. Being bored? Recognizing natural interest by extensive audiovisual integration for real-life application. Journal of Image and Vision Computing 27 (12), 1760–1774.

[137] Cootes, T. F., Edwards, G. J., Taylor, C.J., 1998. Active appearance models. In Burkhardt H., Neumann B. (Eds), Computer Vision - ECCV'98: LNCS 1407, 484-498.

[138] Pantic, M. Valstar, M. F., Rademaker, R., Maat, L., 2005. Web-based Database for Facial Expression Analysis. Proceedings of IEEE International Conference on Multimedia and Expo (ICME'05).

[139] Van Laere, J.2009. Challenges for IF performance evaluation in practice. In Proceedings of 12th International Conference on Information Fusion, 866-873.

[140] Janssen, J. H., Tacken, P., de Vries, J.J.G(Gert-Jan), van den Broek, E. L., Westerink, J. H.D.M., Haselager, P., Ijsselsteijn, W. A., 2012. Machines Outperfom Laypersons in recognizing Emotions Elicited by Autobiographical Recollection. Human-Computer Interaction 28(6), 479-571.

[141] Huang, G. B., 2015. What are Extreme Learning Machines? Filling the Gap between Frank Rosenblatt's Dream and John von Neumann's Puzzle. Cognitive Computation, 7 (3), 263-278.

[142] Grimm, M., Kroschel, K., Narayanan, S., 2008. The Vera am Mittag German Audio-Visual Emotional Speech Database. Proceeding of the IEEE International Conference on Multimedia and Expo (ICME).

[143] McKeown, G., Valstar, M. F., Cowie, R., Pantic, M., Schroeder, M., 2012. The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Transactions on Affective Computing 6 (1), 5-17.

[144] Clavel, C., Vasilescu, I., Devillers, L., Richard, G., Ehrette, T., Sedogbo, C., 2006. The SAFE corpus: illustrating extreme emotions in dynamic situations. Language Resources Evaluation Conference (LREC'06).

[145] Fanelli, G., Gall, J., Romsdorfer H., Weise, T., Van Gool, L., 2010. A 3-D Audio Visual Corpus of Affective Communication. IEEE Transactions on Multimedia, Special Issue Multimodal Affective Interaction 12 (6), 591-598.

[146] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., Narayanan, S. S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Journal of Language Resources and Evaluation, 42 (4): 335-359.

[147] Wallhoff, F., Schuller, B., Hawellek, M., Rigoll, G., 2006. Efficient Recognition of Authentic Dynamic Facial Expressions on the Feedtum Database. Proceeding of 2006 IEEE International Conference on Multimodal and Expo.

[148] Schuller, B., Müller, R., Hörnler, B., Höthker, A., Konosu, H., Rigoll, G., 2007. Audio-visual recognition of spontaneous interest within conversations. Proceedings of 9th International Conference on Multimodal Interfaces (ICMI), ACM SIGCHI, 30–37.

[149] Kanade, T., Cohn, J. F., Tian, Y-L., 2000. Comprehensive database for facial expression analysis. Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), 46-53

[150] Lichtenauer, J., Soleymani, M., 2012. MAHNOB-HCI-Tagging Data-

base.http://mahnob-db.eu/hcitagging/media/uploads/manual.pdf

[151] Lichtenauer, J., Valstar M. F., Shen J., Pantic, M., 2009. Cost-effective solution to synchronized audio-visual captures using multiple sensors. Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'09), 324 – 329.

[152] Koelstra, S., Mühl, C., Soleymani, M., Lee, J-S, Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I., 2012. DEAP: A Database for Emotion Analysis using Physiological Signals. IEEE Transactions on Affective Computing 3(01), 18-31.

[153] Healey, J.A., Picard, R. W., 2005. Detecting Stress during Real-World Driving Tasks Using Physiological Sen-sors. IEEE Transactions on Intelligent Transportation Systems 6 (2), 156-166.

[154] Mohammad S., Lichtenauer, J., Pun, T., Pantic, M., 2012. A multimodal database for affect recognition and implicit tagging. IEEE Transactions on Affective Computing 3(1), 42 – 55.

[155] Boughrara, H., Chen, L-M., Ben Amar, C., Chtourou, M., 2012. Face recognition under varying facial expres-sion based on Perceived Facial Images and local feature matching. Proceedings of 2012 International Conference on Information Technology and e-Service (ICITeS), 1-6.

[156] Dhall, A., Asthana, A., Goecke, R., 2011. A SSIM-based approach for finding similar facial expressions. Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), 815-820.

[157] Strauss, P-M., Hoffmann, H., Minker, W., Neumann, H., Palm, G., Scherer, S., Traue, H. C., Weidenbacher, U., 2008. The PIT corpus of German multi-party dialogues. Proceedings of the sixth international language resources and evaluation (LREC'08), 2442–2445.