RRHE - Remote Replication of Human Emotions

Joaquim Guerra
Instituto Superior Técnico
Universidade de Lisboa
Lisboa, Portugal
joaquimguerra@ist.utl.pt

ABSTRACT

Software-based human emotion detection is an issue that has been debated for a long time. Several solutions have been proposed in the literature, but there are still flaws that impair the effective commercial exploitation of such solutions. Users still do not trust this kind of systems due to the high percentage of classification errors, opting by physical interaction or video-conference communication for visually (and possibly using as well audio clues) transmitting their emotions. One possibility for improving current systems accuracy could be exploiting multimodal sources of emotional content. This will require the integration of multiple techniques of emotion extraction from different sensing modalities. Furthermore, current emotional interfaces are usually bulky. Indeed, emotional algorithms output words corresponding to the detected emotion. We believe that smart user interfaces for emotional detection systems can drastically augment the number of use-cases for this technology, increasing very significantly such systems usability. The proposed multimodal system merges two of the most used modalities for emotion extraction, namely facial expressions and voice properties. With such an algorithm we were able to significantly reduce the error induction from irony, i.e facial expressions that contradict the simultaneously expressed vocal tone. This work was comparatively evaluated with respect to two baseline scenarios, consisting of individually evaluating each of the facial and voice emotion detection algorithms. The results show that this implementation of a multimodal algorithm allows an increase of classification hits, which in turn makes software-based classifications much closer to user-made manual classifications.

Keywords

Emotion Detection, Face Expressions, Voice Emotional Recognition, Classifier, Support Vector Machines, Facial Action Coding System.

1. INTRODUCTION

The concept of Human-Computer Interaction (HCI) emerged with the necessity systems' functionality and usability [47]. The level of functionality is measured with the quantity and efficiency of services in the system [44]. The meaning of usability is the level by which a system can be used efficiently and how it adequates to accomplish some goals for specific users. One of the techniques used to improve accuracy in HCI is precisely the recognition of human emotions, which can be used in a large number of systems. However, according to the Nass and Brave [29] studies on HCI emotions, these kind of stimulus can bring problems since it tends to examine photographs and voices with deliberately performed emotions as opposed to emotions experienced naturally.

Interpersonal communications is dominated by non-verbal expressions [3]. This means that the interaction between human and machines could be richer if machines could perceive and respond to human non-verbal communication, such as emotions. This document places a special focus on emotions recognized from face expressions and speech. However there are more non-verbal gestures very important to detect human emotions, such as posture and hand signals [10].

But which emotions should a system recognize? To address this question, lets introduce briefly the concept of basic emotions, based on Ortony and Turner study [31]. These basic emotions (e.g. fear) are the building blocks for more complex emotions (e.g. jealousy). Plutchik, around 2001, has demonstrated an important property of basic emotions; he argues that these emotions are innate and universal across all cultures [40], which adds universality to this kind of systems (that uses emotional parameters). Defining the set of basic emotions, Ekman and Friesen [13] around 1975 limited the list to the following six: Happiness, Surprise, Fear, Disgust, Anger, Sadness.

These basic emotions are the ones recognized by the majority of recent Emotions Recognition Software's, such as the framework for the facial classifications developed by Ekman and Friesen [12]. Hence, this paper aims to recognize the six previous emotions and the neutral state as a seventh emotion.

For the past 50 years, social scientists community worked hard on facial expressions analysis. They believed that facial expressions are a portal to one's internal mental state [14], [19] and, when an emotion occurs, a series of biological events follow it producing changes in a person (e.g. facial muscles movements). Other authors presented the idea that facial expressions can be used as a strategic tool to accomplish elicit behaviours or social goals in an interaction [17].

In the field of emotion recognition from speech the analysis of Prosody has been the main focus of research. Features like pitch and energy with their meanings, medians, standard deviations, minimum and maximum values [8] are normally combined with some higher level features, such as speaking rate, phone or word duration. A sad agent, for example, typically displays slower, with little high-frequency energy and lower pitched speech. In the other hand, an agent experiencing anger will speak faster and louder, with strong high-frequency energy and more explicit enunciation [39].

Indeed, software for recognizing human emotions has been in use for a long time, but in fact, people are still feeling limited by computers. Indeed, humans communicate through means that are not typically perceived by machines, in particular spatial relations. Even though we do not realize it consciously, we interact with each other through very simple spatial cues that everyone understands. For instance, something as simple as walking in the direction of someone indicates a wish to speak with that person.

Because the software is not usually capable of perceiving this implicit communication, it forces users to inform the system through explicit interactions, like the press of a button. This results in people feeling like the computer is in their way, instead of supporting their tasks as it was intended, because they are forced to repeat something they have already communicated. Hence, users tend to ignore the system, using older methods and not effectively adopting the solutions developed for them. As such, there is the need for software that deals with this implicit protocol.

This paper addresses the aforementioned problems. It proposes a multimodal emotion detection approach that uses both facial expressions and speech when compared with systems that have just one of these modalities. Exploiting multimodal features, the proposed solution is also able to detect irony, as conflicting emotions expressed by visual and speech content.

Since the emotion replication is done remotely, the client application and the actuator do not need to be in the same physical space or in the same sub-network as the detection system. Everything is connected to a server that can be hosted in the Internet. The detection algorithm runs on remote backend servers (e.g. on the cloud), and the information is then presented to the user through emotional agents, such as simple emoticons or more complex avatars, robotic interfaces that may remotely mimic the emotional expressions, or through socialWebs.

This paper is organized as follows. First we discuss the related work in section 2, where we explore the recurring problems of recognizing human emotions and how using a multimodal technique might provide a solution to those problems. In section 3 we explain how we have structured our solution, the **Remote Replication of Human Emotions (RRHE)**, to fulfil all the solution requirements. The implementation decisions and details are explained in section 4. The solution experimental evaluation is described in section 5. Finallly, section 6 concludes this document, discussing the paper main contributions, and directions for further improvements in future work.

2. RELATED WORK

Building a system to remotely recognize emotions during human-machine interactions requires the analysis of previous research works. Hence, the main techniques used to detect and classify facial expressions will be first reviewed. Afterwards, we will focus on previous work for recognizing emotions from speech analysis, to extract the main features given by Prosody. The most relevant previous work that addresses the combination of multiple modalities for emotion recognition is surveyed afterwards.

2.1 Facial Expressions

According to Mehrabian [27], 55% of the effect conveyed by a human communication message is reflected by facial expressions. It is extremely important to have an effective representation of the human face to successfully recognize facial expression. Nowadays, there are two common methods used to obtain facial features: geometric and appearance features [21].

The locations and shape of facial components that represent the face geometry are represented by geometric features. Valstar et al. [48] demonstrated that geometric feature-based methods have an identical or superior performance than appearance-based approaches in Action Unit recognition. In appearance-based methods, the idea is to apply image filters to specific face portions as well as to the whole face, to extract appearance changes over the time.

There are different methodologies studied in the literature for developing classifiers for emotion recognition [43], [11]. In a static approach, the classifiers evaluate each frame in videos to one of the facial expression category. Bayesian network classifiers and Naïve Bayes classifiers were often used on these approaches. In dynamic approaches, the classifiers watch for temporal patterns to recognize facial expressions [4]. Classifiers based on Hidden Markov Models (HMM) and multi-level HMM [24] are often used in dynamic approaches.

Facial Action Coding System (FACS) was developed by Ekman and Friesen [14] to represent movements on the face as facial expressions codes. They described a set of action units (AUs). An action unit has a direct link to a muscle movement (e.g. blinking) and they proposed 44 AUs to mask all the possible movement combinations. FACS does not contain any system to classify facial expressions, it needs to be done in an independent system which is preconfigured, manually, with a set of rules.

Black and Yacoob [6] used another classification technique, employing local parametrized samples of image motion to retrieve non rigid motion. Once estimated, these parameters were used as entry to a rule-based classifier to identify the six essential emotions. Yacoob and Davis [51] developed another optical flow technique applying identical rules to execute the classification of the six basic emotions. Rosenblum, Yacoob and Davis [41] also developed optical flow for face fractions and later implemented a function to classify expressions.

Ohya and Otsuka [32] developed yet another optical flow approach. However, they additionally introduced 2D Fourier transform coefficients that were used as feature collections for hidden Markov model (HMM) to classify facial expressions present on each frame. Finally, tracked motions were

employed to command the facial expression of an animated Kabuki system [33]. For each one of the six basic expressions it was obtained a detection.

Martinez [26] brought in an indexing approach based on the recognition of frontal face images beneath distinct facial expressions, occlusions and illuminations conditions. A Bayesian approach was implemented to get the right combination between learned features model and local observations. Furthermore, since new conditions could be different from the previous ones, an Hidden Markov Model was applied to increase recognition rates.

Oliver et al. [30] used lower face tracking as a strategy to select mouth features, using the obtained values as information to an Hidden Markov Model based system. The mentioned techniques are akin because they initially extract a few features from each frame, which will then be used as input to a classification system. In addition, the outcome of these techniques is one of the emotion categories previously picked.

2.2 Prosody

The computer speech community, has traditionally focused on "what was said" and "who said it", instead of "how it was said". Languages cannot be considered equally. The large variety of languages, and the correspondent number and variability of features in each one, makes it difficult to predict how to connect these features to obtain better results on the recognition rate. [38].

The recent studies for emotion recognition in speech have been using diverse classification algorithms like HMM (Hidden Markov Models), GMM (Gaussian Mixture Model), MLB (Maximum-Likelihood Bayes), KR (Kernel Regression), k-Nearest Neighbour and NN (Neural Network) [16].

Affective applications are being developed and gradually appearing in the market. However, the development of efective solutions depends strongly on resources like affective stimuli databases, either for recognition of emotions or for synthesis. The information is normally recorded by the affective databases, by means of sounds, psychophysiological values, speech, etc. and actually there is a great amount of effort on increasing and improving its applications [18]. Some other important resources include libraries of machine learning algorithms such as classification via artificial neural networks (ANN); Hidden Markov Models (HMM); genetic algorithms; etc.

By analysing speech patterns the user's emotions are identified by emotional speech. Parameters extracted from voice and Prosody features such as intensity, fundamental frequency and speaking rate are deeper correlated with the emotion expressed in speech. Fundamental frequency (F_0), normally known as pitch (since it represents the perceived fundamental frequency of a sound) is one of the most important attributes for determining emotions in speech [28].

One possible way to extract and analyse features from human speech is statistical analysis. Using this method, the features connected with the pitch, Formants of speech and Mel Frequency Cepstral Coefficients, can be chosen as inputs to the classification algorithms. Bäzinger et al. said that statistics related to pitch carry important information about emotional status [9]. Nevertheless, pitch was also considered to be the most gender dependent feature [1].

According to Kostoulas et al. [20], the emotional state of an individual is much related to energy and pitch. From these features of the speech signal, it may be easier to understand happiness or anger, but not so easy to detect, for instance, sadness.

Besides pitch, there are some other important features that are linked to speaking: rate, formants, energy and spectral features, such as MFCCs. The spectrum peaks of the sound spectrum |P(f)|' of the voice can be defined as formants; this term is a polysemic word and it also refers to an acoustic resonance of the human vocal tract. It is usually calculated as an amplitude peak in the frequency spectrum of the sound. It is useful to distinguish between genders and to predict ages.

Wang & Guan [50] used MFCCs, formant frequency and prosodic features to represent the characteristics of the emotional speech.

MFCCs are an universal way to make a spectral representation of speech. They are used in many areas, such as speech and speaker recognition. Kim et al. [35] referred that statistical assumptions with MFCCs also brings emotional information. MFCCs are generated with a Fast Fourier Transform followed by a non-linear warp of the frequency axis. Afterwards it is calculated the power spectrum, to have frequencies logarithmically spaced. In the end, MFCCs result from the appreciation with the cosine basis functions of the first N coefficients of this strained power spectrum.

2.3 Multimodal Implementations

Emotion recognition from multimodal techniques is still an open challenge. Pantic and Rothkrantz [36] presented a survey where the focus was on audiovisual affect recognition. Since then, an increasing number of studies were made on this matter. As evidenced by the state-of-the-art for Prosody and facial expressions implementations (single-modal techniques), most of the existing studies focus on the recognition of the six basic emotions.

Pal et al. [34] presented a system for detecting hunger, pain, sadness, anger and fear, extracted from child facial expressions and screams. Petridis and Pantic [37] investigated the separation of speech from laughter episodes taking into account facial expressions and Prosody features.

Zeng et al. [54] introduced a method to fuse multi-streams using HMMs. The goal is to form, according to the maximum common information, an ideal link between several streams extracted from audio and visual channels. Afterwards, Zeng et al. [53] further evolved this technique, presenting a middle-level training approach. Under this layer, several learning schemes can be used to combine multiple component HMMs. Song et al. [45] introduced a solution where upper face, lower face and prosodic behaviours are modelled into individual HMMs to model the correlation features of these elements. Fragopanagos and Taylor [16] proposed an artificial neural networks (NN) based approach. This proposal also incorporates a feedback loop, named ANNA, to assimilate the data extracted from facial expressions, lexical content and prosody

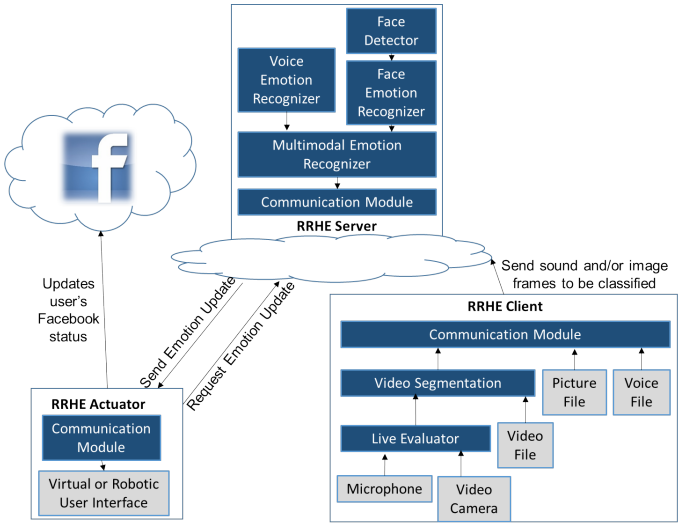

Figure 1: RRHE Architecture.

analysis. Sebe et al. [42] utilized a Bayesian network (BN) to combine features from facial expressions and prosody analysis.

2.4 Irony Detection

Human expressions are often employed to express irony. For instance, bad news (like "you are fired") may turn someone's face with a sad emotional expression, while the person, with a happy voice, states a positive sentiment such as "but these are good news". Indeed, irony is an important instrument in human communication, both verbally as well as written. Indeed, irony is quite often used in literature, website, blogs, theatrical performances, etc.

To the best of our knowledge, no previous work addressed irony detection from multimodal sensing modalities. However, there are some works addressing irony detection in written texts. But even such work is very recent, as demonstrated by Filatova [15] first corpus including annotated ironies in texts. Buschmeier et al. [7] analyzed the impact of several features, as well as combinations of them, used for irony detection in written product reviews. They evaluated different classifiers, reaching an F1-measure of up to 74% using logistic regression. Another work [46] used a sentiment phrase dictionary combined method to address multiple semantic recognition problems, such as text irony. Machine learning methods were also employed for satire detection in Web Documents [2].

3. ARCHITECTURE

RRHE's architecture is based on a Client-Server-Actuator model as shown in Figure 1. The internal architecture of RRHE-Client module is composed by 3 major logical components: communication module, video segmentation, and live evaluator. The communication module is responsible for the communication with the server, sending the captured data - images and audio files. The video segmentation component incorporates the algorithm which splits a video into several segments containing still images and an audio file. This data is subsequently analyzed by the server. The live evaluator module in the RRHE-Client allows the overall system to run in real-time, using the microphone and video camera available at the hosting device.

The RRHE-Server module is the brain of the system. Similar to the client module, the server also has a communication module. The communication module is listening for requests, either from the client - with data for classification - or from the actuator - with requests for users' status updates. The components responsible for the capture of emotions are a layer above the communication module. For the facial expressions we have a component capable of detecting and extracting faces in images. After running this process, the Face Emotion Recognizer component evaluates the extracted face and outputs an emotional category. In the vocal expression side we have the Voice Emotion Recognizer that incorporates algorithms capable of extracting properties from the audio signal for further analysis in the classification phase. This paper also proposes another module performing the multimodal integration of facial and audio emotional content for better classifying emotions and for enabling the detection of ironies.

Finally, the RRHE-Actuator module replicates the emotions detected by the server and remotely transmitted through the communication channel. The actuator starts a cycle of requests to the server where it requests the most recent emotional state of a user. RRHE-Actuator can replicate the detected emotions in several ways, from the display of an emoticon to the status update in a social network or mimic interpretation by a robotic agent.

The next subsections present in detail each module of the RRHE architecture.

3.1 Client

The client module (RRHE-Client) is an application that can be installed in any PC or mobile device. The purpose of the RRHE-Client is the collection of sound and image data (from the microphone and video camera devices, respectively), and their transmission to the server for proper classification. However, some processing needs to be carried out at the client. Hence, RRHE-Client's interface is mainly composed by the following features:

1. Send image - allows to select an image file to be sent to the server for evaluation and classification as an emotion detected in the face transmitted - if there is one. This is a very important function since it allows to separately test the recognition of facial emotions.
2. Send sound - allows to select a sound file to be sent to the server, for evaluation and classification as an emotion detected on the voice transmitted - if there is one. It is equally important function since it allows to separately test the recognition of voice emotions.
3. Send video - allows to select a video file for testing the system as a whole, combining facial and voice emotions. Image and sound segment pairs are extracted from the provided video.
4. Begin Live Evaluation - enters the Live Evaluation module.

5. Log - shows a list of error messages that help the user understand the system's behavior and fix any problem.

6. Settings - enters the Settings module to configure the system.

3.2 Actuator

The actuator module (RRHE-Actuator) is an application that can be installed on any PC or mobile device. The purpose of RRHE-Actuator is to represent emotions detected in the data sent by RRHE-Client. RRHE-Actuator establishes a request-on-demand connection with RRHE-Server and periodically asks for the update of the emotional status of the user. After receiving the notification from RRHE-Server, RRHE-Actuator updates the emotional state. The representation of the emotional status is made through two different aspects:

1. Emoticon - An emoticon representing the detected emotion is displayed in the user interface.

2. Facebook integration - To demonstrate a possible way for RRHE to be integrated with external systems, RRHE-Actuator can also update the Facebook status of the user using a 'Feeling' emoticon according to the detected emotion. The Facebook integration can be enabled or disabled in the RRHE-Actuator user interface and the login is requested in the first status update.

3. SmartLamp - SmartLamp is a desktop lamp with robotic behaviors and personality. This product is being developed by YDreams Robotics and its main features are: face tracking during video calls; video surveillance with motion detection; play games; express emotions. SmartLamp incorporates a smartphone/tablet and it is compatible with Android and iOS. RRHE was developed aiming its integration into the SmartLamp, by including the RRHE-Client and RRHE-Actuator as part of SmartLamp's applications. We will have multiple SmartLamps communicating with one RRHE-Server, sending data to be classified or requesting emotions updates. The representation of each emotion can be modelled and/or mimicked in robotic movements as well as using its screen.

3.3 Server

The server module (RRHE-Server) is a console application dedicated to the treatment of the information captured and sent by RRHE-Client. RRHE-Server is the core module of RRHE since it is responsible for processing audio and image data to recognize the respective emotion. RRHE-Server consists of:

1. TCP server - a typical TCP server listening to requests.

2. Server Manager - maintains the execution context of the server (e.g., the emotional state of each active user).

3. Worker Threads - launched (one per core) at the start of the application. They work together with the Server Manager in a Single Producer-Multiple Consumer type of environment, processing the RRHE-Client and RRHE-Actuator requests.

RRHE-Server receives an image file for which it must extract an emotion. This module has been developed purely in C++ with Qt and OpenCV frameworks. The first problem to address is to crop only the significant part - human face - of the entire image. This is achieved using the Haar Cascade methods given by OpenCV. This framework is also useful for rotating the recognized faces to proper positions. Afterwards, we have developed our Gabor Bank implementation, which will filter the image, returning a features vector. We can use our facial classifier, which uses the OpenCV SVM as the learning algorithm, to process such input vector. In the end, this module returns all the confidence values found for each supported emotion.

RRHE-Server also receives an audio file - WAV format - to extract an emotion from the voice in the audio signal. We first need to extract some audio properties used during the classification process. Since it is easy for Matlab to handle sound files, we decided to use this language for such work. In addition, we also decided to implement our vocal classifier using the Matlab SVM as the underlying learning algorithm. This module will return all the confidence values found for each of the supported emotions.

Once both facial and vocal emotions have been estimated, our fusion algorithm is applied to estimate a final multimodal emotion. This module has been developed purely in C++ with Qt-Framework.

4. IMPLEMENTATION

This section will focus on the implementation of RRHE three main modules for emotion recognition. namely: voice emotion recognition; facial emotion recognition; and the proposed emotion fusion technique. Hence, the following subsections describe the techniques and algorithms used to implement each module.

4.1 Voice Emotions Extraction

This module uses the Support Vector Machine (SVM) algorithm from Matlab. SVM uses a binary classification based on statistical learning with data represented in a vectorial space. It finds an hyperplane of maximum margin through internal kernel functions to get the final classification. SMVs have the capacity to generalize new information accurately using trained models, which are created during the learning phase.

As in other problems, this classification is multi-class, since we considered 7 different classes that can be returned. For problems like this, there are various algorithms that can be applied, such as one-against-all (OAA), multi-class ranking and pairwise SVM.

We opted the one-against-all (OAA) algorithm which means that for a given input, all emotion classes are going to classify this parameter and return its degree of confidence. In the end, the chosen class is the one that presents the highest degree of confidence, after comparing all classes.

Using the highest degree of confidence alone as the decision factor, the number of missclassifications is increased if the two confidence values are relatively close. To mitigate this problem, we decided to implement the algorithm of hybrid kernel and thresholding fusion proposed by Yang et al. [52].

During the training phase, for each utterance, 60 attributes and the respective labels are used to feed the models X_i, where $i = 1, ..., 7$ corresponding to the number of emotions (emotional classes) used in this project. The best kernel function between quadratic, linear, polynomial, radial basis function (RBF) and multilayer perceptron (MLP) is calculated for each of the trained models, resulting in an hybrid kernel for the generated classifiers. The average (μ_i) and standard deviation (σ_i) of the confidence values returned during the training phase are calculated for each classifier. On the end, results the construction of the models trained for each class of classification (emotion).

As just mentioned, 60 attributes were used to analyse each utterance, as follows. The algorithm uses 12 features:

- Pitch (1 feature): Defined as the relative lowness or highness with which a tone is perceived by the human hearing. Its value depends on the number of vibrations per second produced by the vocal chords. The pitch values are extracted and represented by cepstrum - the Inverse Fourier Transform (IFT) of the logarithm of the signal frequency spectrum - in the frequency domain.
- Energy (1 feature): The energy represents the speech intensity. It is calculated for each 60ms segment, by adding the amplitude of the squared values of each 1ms sample in the segment.
- Pitch difference and Energy difference (2 features): Is the difference between the pitch values and energy values of two contiguous segments. The higher the fluctuation of these values, the most evident is the presence of emotions.
- Formant: Calculated from the format (frequency and bandwidth) of the vocal channel. In the context of this project the frequency and bandwidth were used for the first four formants of each segment (2 x 4 = 8 features). Each formant was determined by the Linear Predictive Coding (LPC) method.

The average, maximum, minimum, range and standard deviation in each 60ms segment are calculated for each of these 12 features. This way, for each segment we have 12 x 5 = 60 attributes that will be used in the classifier, either for training or for classification.

The concept of speaker-dependent emotion classification was not implemented, meaning that parameters specific to the speaker (e.g., sex) are not considered in the evaluation. This could be a future update to the algorithm which we believe could improve its results.

4.2 Face Emotions Extraction

The approach followed to implement this module is based on two previous research works [5], [23], where several algorithms are combined to achieve the emotional states recognition. The methodology proposed is the following: Face detection and extraction; Facial features extraction through Gabor Filters Bank; Training SVM classifiers with the labelled data (e.g., emotion labels).

The face area represents the region of interest (ROI) in the context of this module. Therefore, we first focused on the detection of faces in an image. This kind of task has been widely discussed in the literature and several algorithm implementations exist. We adopted a Haar feature based cascade classifier [49], which is also part of the OpenCV framework.

This algorithm includes:

- Haar Features: these features are calculated in small windows of the image. In each window a binary mask is virtually applied and the value of the feature is the difference between the sum of the pixel values above the part of the mask with value 1 and the sum of the other part.
- Cascade Classification: this classification approach is based on several classification steps. At each step a different feature is considered and if a feature value does not match the trained model, the process is aborted and further stages are not evaluated.

A simple face rotation correction algorithm was implemented, which is based on the position of the eyes detected via specifically trained cascade classifiers. Once the eyes position is obtained, a simple trigonometry calculation is performed to get the value of the angle between the eye-line and the x-axis. This value is then used for the definition of the rotation matrix which is applied to the whole image.

Gabor filters are, roughly speaking, linear filters obtained by modulating a complex sinusoid with a Gaussian. These filters are typically used in image processing for tasks like edge detection, texture classification and face recognition. They are particularly effective in case of a time-frequency analysis, which is an analysis technique that aims to simultaneously study a signal in both time and frequency domain. This is due to multi resolution and multi-orientation properties. Multi-resolution is a method for orthonormal base creation by slicing the signal space into subspaces at different scales. One reason behind the great success of this kind of filters is the discovery that simple cells in the human visual cortex can be modelled with this particular filter. The following Gabor function formula is employed:

$$g(x, y, \lambda, \theta, \psi, \sigma, \gamma) = exp(-\frac{x^{'2} + \gamma^2 y^{'2}}{2\sigma^2})exp(i(2\pi\frac{x^{'}}{\lambda} + \psi)) \quad (1)$$

where $x' = xcos\theta + ysin\theta$ and $y' = -xsin\theta + ycos\theta$ represent the rotated component of the complex sinusoid, λ is the wavelength of the sinusoid, θ is the spatial orientation of the filter, ψ is the phase offset, σ is the standard deviation of the Gaussian support and γ is the aspect ratio factor (e.g. 1.0 for a circular shape).

Derived parameters can also be considered. For instance, the spatial frequency bandwidth of the filter is defined as:

$$b = log_2(\frac{\frac{\sigma}{\lambda}\pi + \sqrt{\frac{ln(2)}{2}}}{\frac{\sigma}{\lambda}\pi - \sqrt{\frac{ln(2)}{2}}}), \frac{\sigma}{\lambda} = \frac{1}{\pi}\sqrt{\frac{ln(2)}{2}}\frac{2^b + 1}{2^b - 1} \quad (2)$$

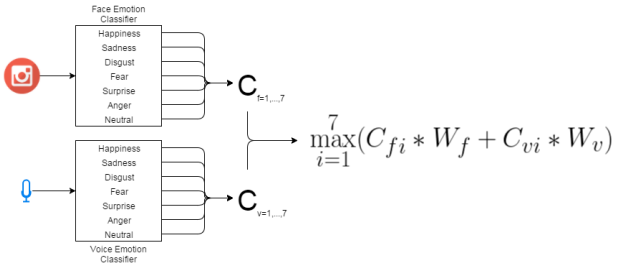

Figure 2: RRHE fusion algorithm.

These additional relationship between λ, θ and b is useful for generating Gabor filter banks.

Gabor Filter banks are one of the main methods for selection of Gabor filters, typically adopted in texture segmentation problems. Families of filters are typically obtained by generating Gabor kernels with spatial frequencies λ, sinusoid orientation θ and bandwidth in ad hoc intervals, while scaling parameters are sometimes selected intuitively and assumed to be constant.

We used the OpenCV implementation's of linear SVM, which realizes a C-Support linear SVM, this is, a linear SVM with Soft Margin, as follows:

$$min\frac{1}{2}\omega^T\omega + C * \sum_{i=1}^{l}\xi_i \tag{3}$$

$$y_i * (\omega^T\phi(x_i) + b) \geq 1 - \xi_i, \tag{4}$$

$$\xi_i \geq 0, i = 1, ..., l \tag{5}$$

where $\omega \cdot x - b = 0$ is the hyperplane: ω is the normal vector to the hyperplane and $x_i \in \mathbb{R}^n, i = 1, ...l$ are the training vector; $y \in \mathbb{R}^l, y_i \in \{1, -1\}$ is a class indicator, ξ_i is a non-negative slack variable measuring the degree of missclassification on x_i. The parameter C gives a weight to these missclassification variables. In other words, C is a trade-off between margin maximization and error minimization.

4.3 Emotion Fusion Technique

Figure 2 illustrates our idea to combine both facial and vocal emotions. C_{fi} is the confidence degree for facial emotion i, W_f is the weight for face classifier, C_{vi} is the confidence degree for vocal emotion i and W_v is the weight for voice classifier, with $i = 1, ..., 7$ representing the 7 emotions supported by RRHE.

The idea behind this algorithm was as simple as to weigh up each one of the classifiers, and apply this weight to their emotional confidences degrees. Afterwards, the weighted emotional classes are summed together, and the final emotion will be the one with the maximum value.

We considered that face expressions are the most relevant element when evaluating an emotional scenario. Humans tend to reflect what is on their mind by actively issuing facial expressions. After several experiences we decided to weigh up the facial classifier with 60% and voice classifier with 40%.

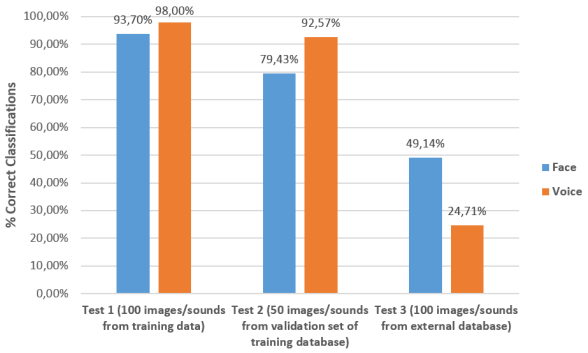

Figure 3: Percentage of correct classifications.

5. EVALUATION

To validate this solution, which presents a new approach for detecting and replicating human emotions, multiple evaluations were conducted that allow the comparison between results obtained with our system and other annotations (human and computerized).

Regarding system evaluation the extracted data aims to demonstrate the accuracy and performance of our solution. To train the SVM applied for face emotion recognition, it was used a set of images from the Cohn-Kanade Expression Database [25]. Likewise, to train the SVM applied for voice emotion recognition, it was used a set of recordings from LDC database [22].

The first tests had the objective to individually determine performance of the voice and face classifiers. Then, in order to have a comparison between RRHE and human annotations, we decided to make a questionnaire where 30 participants classify 30 videos regarding face emotion, voice emotion and overall emotion. To finalize, we compared the results obtained with the new Kinect V2 with our results.

5.1 Experiences with Individual Classifiers

The goal of this experiment was to individually test each emotion classifier. For each classifier:

1. 100 images and sounds for each emotion were extracted as validation data from the DB used during the classifier's training phase, with the goal of testing if the algorithms recognize well the data used for their training (no generalization);
2. 50 images and sounds for each emotion were extracted from the validation set of the same DB used to train the classifier thus already requiring some generalization capability;
3. 100 images and sounds for each emotion were extracted from a different DB than the one used to train the classifier, which corresponds to a more demanding test concerning the generalization capability.

Figure 3 compares the percentage of matches from both classifiers in all tests. As we can see, both the facial emotion classifier and the vocal emotion classifier had excellent results in the first test: above 90%. This proves that when the

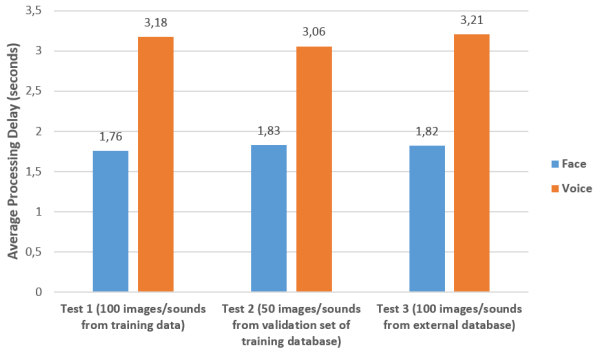

Figure 4: Average processing delay, in seconds.

data used for evaluation is the same data that was used for training the accuracy is very high. It is important to note that, albeit by a small margin, the vocal classifier obtained a better result than the facial classifier, which shows that this classifier has a deeper connection with the training data, being able to extract relevant properties during the learning process. As expected, although the percentage of matches is still very high, the second test shows a reduction in the match percentage of both classifiers. In this case the voice classifier has even better results when comparing with the facial classifier, which shows once again that the sound properties used for training are very useful for the classifier when evaluating and training data have similar properties. The bad results of the third test are debatable. Starting with the facial classifier, we believe that the bad results are mainly due to the noticeable difference between the images used in the training and evaluation phases. For the vocal classifier, the explanation resides on the presence of a distinguishable background noise. This is a problem identified by our algorithm that could be alleviated by a background noise removal phase before the classification of the sound.

Figure 4 shows the analysis to the performance of RRHE-Server and the classifiers. Firstly, it is important to note that the Processing Delays are relatively close to each other in all tests, which shows that the performance of the classifiers is not affected by the data. We can see that the facial classifier is faster than the vocal classifier by a factor of 2 which shows that the analysis of a sound file involves a lot more operations than the analysis of an image file. Finally, it is important to state that the timings of RRHE-Server are acceptable in a system with the capacity to work remotely. Considering that the client-server and server-actuator exchange of messages was made over the Internet, a maximum of 3 seconds of delay until the emotion replication is a good result.

5.2 Manual Annotation With Questionnaire

This experiment involved a total of 30 people whom individually answered the questionnaire. The number and profile of people was chosen so that there are as many answers of people coming from distinct professional areas as possible.

The procedure of this experiment was the same for all participants so that answers are not influenced by non-controllable variables. The age of the respondents varied between 19 and 40 years. Out of all users, 63.33% were male and 36.67% were female. Regarding nationality, all users were Portuguese and all understand English quite well. All sessions happened in

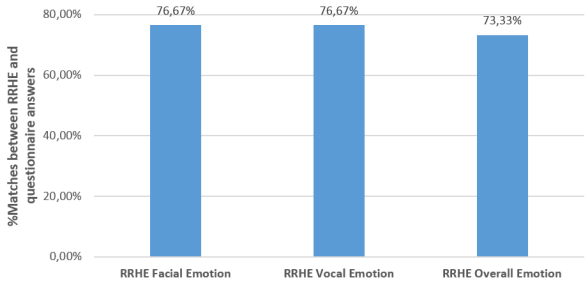

Figure 5: Total percentage of matches between RRHE and questionnaire results.

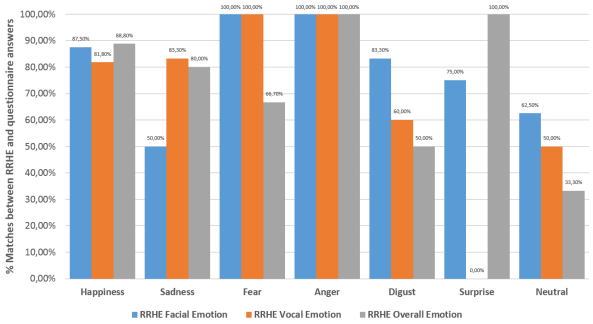

Figure 6: Percentage of matches, by emotion, between RRHE and questionnaire results.

the same room and using the same computer. Each session consisted of: 10 minutes of presentation of the work and the goal of the questionnaires - this presentation was verbal with the help of a set of slides to illustrate the most important concepts; 5 minutes to fill out a pre-session questionnaire; 35 minutes for the session itself and filling the questionnaire; 10 minutes for demonstration and interaction with RRHE.

The pre-session questionnaire had the goal of collecting information about the respondent, understand their level of knowledge about the matter and understand their decision making process in the following questionnaire.

During the session the respondents faced a set of 30 videos for observation. Each of these videos had an image with a facial expression and a human voice also expressing an emotion. These videos were compiled from the validation sets in the database used for the classifier training and had the duration of 1 second. Each video was repeated 5 times with a time interval of 10 seconds, so that the respondents had time to correctly understand the transmitted facial and vocal emotions. In each video the users were asked to classify the interpreted emotions in 3 categories: Facial, Vocal and Overall video emotions.

Figure 5 shows a global view over the results obtained in the questionnaire, comparing them with RRHE results. The percentages show the proportion between the number of correct guesses of RRHE and the respondents. Curiously, the results reveal that the percentage of matches obtained by the facial emotion classifier is exactly the same as the vocal classifier. This data shows that both classifiers have the same level of accuracy which lends credibility to both algo-

rithms. The overall classification has in some cases a lower match percentage, which can be explained by the difference in criteria of RRHE and the respondents when choosing the final classification. Nevertheless, the percentage of matches reveals that most of the times RRHE is correct according to the classification of the respondents. Given the high number of ironies shown in the video set, we can say that RRHE was able to correctly identify them, assigning the same classification as the respondents in the majority of the videos. This constitutes a very promising result for future improvements on irony recognition systems.

Figure 6 presents the results regarding the different emotional states recognized by RRHE. It is important to take into consideration that we only have a sample of 30 videos, which explains scenarios such as the 0% in the voice classifier for the emotional state of Surprise that was not classified in any video by RRHE or the respondents. It is worthy to notice that, for the emotional states with the higher correct classification rate, the match percentage between the different classifiers is very similar and shows considerable values, reaching 100% on 4 occasions. Another relevant fact is that the facial emotion classifier shows a higher accuracy when compared to the vocal classifier. The facial emotions are usually more expressive and thus having a higher impact in the emotional analysis. This was one of the motives that lead us to give a bigger weight to this classifier. Finally, looking at the overall emotion values, it is important to note the high accuracy in certain emotions. The emotions more easily detected by RRHE are clear: Happiness, sadness, anger and surprise. We believe that these results are directly connected to the drastic facial and vocal changes that these emotional states cause in the human body.

5.3 Experiences with Kinect V2

We decided to test the new emotion recognition feature included in Kinect V2 SDK so that we could compare with the results obtained by RRHE. Since the Kinect only recognizes emotions from facial expressions, we could only compare the results with our facial emotion extraction algorithm.

We could only compare 3 emotions with those existing in our project - Happy, Fear and Surprise - amd thus this experiment is not very conclusive. For this comparative study we decided to use the Kinect camera as a capture source, since this camera is required to provide the inputs to the Kinect SDK. Afterwards we have used the images captured by Kinect in RRHE to obtain its classification.

According to this experiment's evidence, that in 38.33% of the cases the Kinect and RRHE are in accordance to eachother, returning the same classification. However, the Kinect does not present the neutral state, for each most emotions are mapped on RRHE. This may correspond to smaller degrees of happiness, fear or surprise that are very close to neutral. As such, RRHE will map them as neutral, while the Kinect as to map them into one of the other 3 available emotional states. Although in a smaller extent, there are anyway some significant classification differences remaining mapping Kinect emotions to other RRHE different emotional states.

6. CONCLUSIONS

We hypothesized that the application of multimodal approaches for emotion detection could result in classification improvements, while enabling the detection of ironies, hardly possible using individual sensing modalities. RRHE also aimed to exploit the remote replication of such emotions in robots, virtual agents and social webs. This paper showed that it is possible to replicate human emotions remotely with good results, both at the level of the correct classifications and system performance.

The commitment to have a ubiquitous system forced the design of an architecture that supports it, so that RRHE can be integrated in multiple use-cases (for instance robots, call centers and conference systems). Another big commitment in this project was to have a client with a low memory footprint and low CPU usage so that it can be executed in any mobile device. Hence we decided to have only the capture of images and sound in the client side. The data is sent to the server for its immediate classification, corresponding to the heavier processing. In this kind of systems one of the requisites is the small delay between data extraction and emotion representation, which we managed to implement with very good results.

Another problem was the definition of the algorithm for human emotion extraction. After investigation and analysis of the State of the Art algorithms, we concluded that single-modal techniques are more used and explored but still show several deficiencies. A clear example where single-modal algorithms are still lacking is in irony detection. We concluded that it would be preferable to use a multimodal algorithm for extraction of emotions. Among the available options, based on the objective of this work, we decided to use facial and voice expressions as an information source by implementing a function that merges both sources.

With this work we developed a client application, highly portable for any mobile device capable of image and sound capture for later transmission to the central server. The central server was developed on a multi-thread architecture for bigger scalability.

The experimental evaluation demonstrated the adequacy of the solution performance both in terms of accuracy and processing speed. As expected, the tests with data extracted from the database used to train the classifiers show good results whenever the samples are known to the classifiers. On the other hand, when using a different database - not the one used for classifier training, the results show that the system reacted well to inputs captured in controlled environments, even without having been used in classifier training. Relative to the results of the questionnaires we conclude that, in most cases, the classification of the respondents matches that of RRHE. Finally, we compared RRHE with Kinect 2, using the emotion detection features provided by Microsoft's SDK. The results are not relevant due to the big difference between the available labels in both systems, although both systems had similar classification results in the matching labels.

6.1 Future Work

Regarding facial and vocal emotion extraction algorithms, we suggest the implementation of a gender-dependent algorithm with different classifiers depending on the user sex. We also suggest to investigate different algorithms for both facial and vocal classifiers - for example, neural networks, HMM and AdaBoost Classifiers. We also suggest, as a potential

improvement factor, to train both classifiers with multiple databases, containing data recorded with people of different races and different background conditions.

The algorithm to merge both emotion extraction techniques can be significantly improved. Instead of using weights for each of the techniques, some intelligence can be added. Instead of using only the confidence level of each classifier, other algorithm execution parameters can be used such as pitch, energy and some FACS extracted during facial analysis to feed a classifier trained with this kind of data.

Unfortunately, the integration of RRHE into the SmartLamp was not possible (the SmartLamp prototype is not yet ready for such integration) and hence it must be addressed by future work.

7. REFERENCES

[1] W. Abdulla and N. Kasabov. Improving speech recognition performance through gender separation, 2001.

[2] T. Ahmad, H. Akhtar, A. Chopra, and M. Akhtar. Satire detection from web documents using machine learning methods. In *Soft Computing and Machine Intelligence (ISCMI), 2014 International Conference on*, pages 102–105, Sept 2014.

[3] M. Argyle. *Bodily Communication*. University paperbacks. Methuen, 1988.

[4] D. Arumugam and S. Purushothaman. Emotion Classification Using Facial Expression. *International Journal of Advanced Computer Science and Applications*, 2, 2011.

[5] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *in 7th International Conference on Automatic Face and Gesture Recognition*, pages 223–230, 2006.

[6] M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *In ICCV*, pages 374–381, 1995.

[7] K. Buschmeier, P. Cimiano, and R. Klinger. An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[8] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. *Analysis of emotion recognition using facial expressions, speech and multimodal information*. ACM Press, 2004.

[9] T. Bï£¡nziger and K. Scherer. The role of intonation in emotional expressions. *Speech Communication*, 46(3-4):252–267, 2005.

[10] J. Cassell and K. Thï£¡risson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4-5):519–538, 1999.

[11] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. Huang. Facial expression recognition from video sequences: Temporal and static modelling. In *Computer Vision and Image Understanding*, pages 160–187, 2003.

[12] P. Ekman and W. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.

[13] P. Ekman and W. Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Prentice-Hall, Oxford, 1975.

[14] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.

[15] E. Filatova. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *LREC*, pages 392–398. European Language Resources Association (ELRA), 2012.

[16] N. Fragopanagos and J. Taylor. Emotion recognition in human-computer interaction. *Neural Networks*, 18(4):389–405, 2005.

[17] A. Fridlund. *Human Facial Expression: An Evolutionary View*. Acad. Press, 1994.

[18] V. Hozjan and Z. Kacic. Context-independent multilingual emotion recognition from speech signals. 6:311–320, 2003.

[19] C. Izard. *The face of emotion / Carroll E. Izard*. Appleton-Century-Crofts, New York :, 1971.

[20] T. Kostoulas and N. Fakotakis. A speaker dependent emotion recognition framework. In *Proc. 5th International Symposium, Communication Systems, Networks and Digital Signal Processing(CSNDSP)*, pages 305–309, 2006.

[21] S. Li and A. Jain. *Handbook of Face Recognition*. Springer, 2005.

[22] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell. Emotional Prosody Speech and Transcripts, 2002. [Online; accessed 19-July-2015].

[23] G. Littlewort, M. S. Bartlett, I. R. Fasel, J. Susskind, and J. R. Movellan. Dynamics of facial expression extracted automatically from video. *Image Vision Comput.*, 24(6):615–625, 2006.

[24] G. Littlewort, I. Fasel, M. S. Bartlett, and J. Movellan. Fully automatic coding of basic expressions from video. Technical report, Tech. rep.(2002) U of Calif., S.Diego, INC MPLab, 2002.

[25] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression.

[26] A. Martinez. Face image retrieval using hmms, 1999.

[27] A. Mehrabian. *Communication without words*, pages 51–52. 2 edition, 1968.

[28] D. Morrison, R. Wang, and L. D. Silva. Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2):98–112, 2007.

[29] C. Nass and S. Brave. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. MIT Press, Cambridge, MA, 2005.

[30] N. Oliver, A. Pentland, and F. Bï£¡rard. Lafter: Lips and face real time tracker. pages 123–129, 1997.

[31] A. Ortony and T. Turner. What's basic about basic emotions? *Psychological Review*, 97(3):315–331, 1990.

[32] T. Otsuka and J. Ohya. Recognizing multiple persons' facial expressions using hmm based on automatic extraction of significant frames from image sequences. In *In Proc. Int. Conf. on Image Processing (ICIP-97*, pages 546–549, 1997.

[33] T. Otsuka and J. Ohya. A study of transformation of facial expressions based on expression recognition from temporal image sequences. In *Technical report, Institute of Electronic, Information, and Communications Engineers (IEICE)*, 1997.

[34] P. Pal. *Emotion Detection from Infant Facial Expressions and Cries*. Temple University, 2006.

[35] S. N. Pantelis Georgiou, Sungbok Lee. Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. In *Proceedings of IEEE Multimedia Signal Processing Workshop, Chania, Greece*, 2007.

[36] M. Pantic and L. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. In *Proceedings of the IEEE*, pages 1370–1390, 2003.

[37] S. Petridis and M. Pantic. Audiovisual discrimination between laughter and speech. In *ICASSP*, pages 5117–5120. IEEE, 2008.

[38] V. Petrushin. Emotion recognition in speech signal: experimental study, development, and application. In *In: Proc. ICSLP 2000*, pages 222–225, 2000.

[39] R. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997.

[40] R. Plutchik. The Nature of Emotions. *American Scientist*, 89(4):344+, 2001.

[41] M. Rosenblum, Y. Yacoob, and L. Davis. Human emotion recognition from motion using a radial basis function network architecture, 1994.

[42] N. Sebe, I. Cohen, T. Gevers, and T. Huang. Emotion recognition based on joint visual and audio cues. In *Proceedings of the 18th International Conference on Pattern Recognition - Volume 01*, ICPR '06, pages 1136–1139, Washington, DC, USA, 2006. IEEE Computer Society.

[43] N. Sebe, M. Lew, I. Cohen, A. Garg, and T. Huang. Emotion recognition using a cauchy naive bayes classifier. In *ICPR (1)*, pages 17–, 2002.

[44] B. Shneiderman and C. Plaisant. *Designing the User Interface: Strategies for Effective Human-Computer Interaction (4th Edition)*. Pearson Addison Wesley, 2004.

[45] M. Song, J. Bu, C. Chen, and N. Li. Audio-visual based emotion recognition - a new approach. In *CVPR (2)*, pages 1020–1025, 2004.

[46] H. Sui, Y. Jianping, Z. Hongxian, and Z. Wei. Sentiment analysis of chinese micro-blog using semantic sentiment space model. In *Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on*, pages 1443–1447, Dec 2012.

[47] D. Te'eni, J. Carey, and P. Zhang. *Human Computer Interaction: Developing Effective Organizational Information Systems*. John Wiley & Sons, Hoboken, 2007.

[48] M. Valstar, I. Patras, and M. Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *CVPR-V4HCI*, jun 2005.

[49] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. pages 511–518, 2001.

[50] Y. Wang and L. Guan. Recognizing human emotional state from audiovisual signals. *Trans. Multi.*, 10(4):659–668, June 2008.

[51] Y. Yacoob and L. Davis. Recognizing human facial expressions from long image sequences using optical flow. pages 636–642, 1996.

[52] N. Yang, R. Muraleedharan, J. Kohl, I. Demirkol, W. Heinzelman, and M. Sturge-Apple. Speech-based emotion classification using multiclass svm with hybrid kernel and thresholding fusion. In *SLT*, pages 455–460. IEEE, 2012.

[53] Z. Zeng, Y. Hu, M. Liu, Y. Fu, and T. Huang. Training combination strategy of multi-stream fused hidden markov model for audio-visual affect recognition. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, MULTIMEDIA '06, pages 65–68, New York, NY, USA, 2006. ACM.

[54] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T. Huang, and S. Levinson. Audio-visual affect recognition through multi-stream fused hmm for hci. In *CVPR (2)*, pages 967–972. IEEE Computer Society, 2005.