The Role of Facial Emotions in Usability Evaluation

Alexandre Antunes Rodrigues

Instituto Superior Técnico
University of Lisbon, Portugal,
Av. Rovisco Pais, 1 1049-001 Lisboa, Portugal,
alexandre.rodrigues@tecnico.ulisboa.pt

Abstract. The usability evaluation of a system is not a trivial task. Over time, there has been much progress in this area despite the difficulties and biases presented in several studies in the literature. A usability evaluation of a given system performed by expert evaluators cannot express all the needs and difficulties of the target users. The difficulty in performing face-to-face tests with users increased significantly due to the rules imposed by the COVID-19 pandemic. Thus, we intend to research the usability evaluation relying on user interactions with the system and the analysis of facial emotions collected on video during user interactions, analyzing in particular the effects of considering emotions during usability evaluation. Emotions expressed by users when interacting with a target system are a valuable source for verifying their satisfaction. In this work, we will study how considering these emotions can influence the results of expert evaluations.

Keywords: Facial expressions, Facial emotions, HCI, Usability, UX evaluation, Usability Evaluation, Cognitive Walkthrough, Usability Smells

Table of Contents

1 Introduction

Nowadays, the importance of evaluating the usability of a system before it is released to the market is crucial. Avoiding unnecessary expenses by analysing how users interact with the system, their needs and having a system capable of being effective, efficient, memorable, error tolerant and easy to learn (Nielsen and Landauer, 1993) are needs that organisations seek. Usability evaluation, although not a trivial task, is a critical part of the product development process and can take place at any point in it. It is performed by expert evaluators to conclude how usable a system can be has drawbacks. The expert evaluator's experience influences the usability evaluation in the domain of the system being evaluated; that is, the reliability and quality of the findings depend on his/her knowledge and experience; usability issues may be missed; false alarms may occur; external factors may distract the expert evaluator's attention when performing usability tests with users; user selection methods may not be appropriate; and the use of expert evaluators is expensive.

All the methods used by expert evaluators have drawbacks. For example, methods like "think-aloud" have the disadvantage that the user does not use the system in a natural way; interviews do not demonstrate the efficiency of use of a system; a cognitive walkthrough used by expert evaluators does not demonstrate the satisfaction of the user when using the system. So, there is a need to study a new method that is non-intrusive, powerful, nonverbal, universal, and intrinsic to all human beings. Thus arises the interest in capturing the emotions expressed by the user when interacting with the system.

Over the years there has been a growing interest in the study of facial emotions and several studies have been done in several areas due to its importance in marketing (PS and Aithal, 2022), medical treatment (Bani et al., 2021), sociable robotics (Webb et al., 2020), human-computer interaction (HCI) (Chowdary et al., 2021) and others. In parallel with the interest in the study of facial emotions, several studies have been carried out with the aim of developing techniques capable of measuring facial emotions.

The motivation of this work is to analyze how the effects of using facial emotions during the process of usability evaluation of a system can influence the results obtained by expert evaluations. The facial expressions of the selected users during their interaction with the system will be captured. The output will be a video that will be used to determine facial emotions during the whole process. Finally, the results that will be compared with evaluations that do not consider the users' emotions will be analysed in order to propose a usability evaluation approach that uses the facial emotions identified during interaction to be used in the usability evaluation process.

1.1 Objectives

As mentioned above, the goal of this work is **to analyze how the effects of using facial emotions during the process of usability evaluation of a system can influence the results obtained by expert evaluations**. In order to achieve this desirable goal, there are essential steps that must be followed, such as

- Collect information available in the literature about facial expressions, facial emotions, usability evaluation and facial emotion recognition algorithms;
- Select users to participate in the testing;
- Create a recording tool to record the users' facial expressions and screen monitor;

– Create a website with and without evident usability problems
– Create a facial emotion recognition algorithm;
– Develop a support website in order to help the expert evaluators to analyze the effects of considering emotions during usability evaluation;
– Qualitative analysis;
– Final conclusions about the effects of using facial emotions during the usability evaluation process

At the end of all the stages are concluded and described in the dissertation writing the final objective is concluded.

1.2 Document Structure

This document is divided into five sections. Section 2 refers to the background of all concepts covered by this study. Section 3 describes an analysis of relevant procedures and contributions from previous work in the literature regarding usability evaluation and the FER algorithm. Section 4 introduces the proposed solution and describes the study that will be conducted. Section 5 indicates the work plan for the proposed study, describing all phases as well as start and end dates. Finally, Section 6 describes, in a general way, the whole document and the expected conclusions to be drawn for the present study.

2 Background

This section introduces theoretical information on several important topics related to the study and is divided into subsections, with each topic considered fundamental to be analysed and described.

2.1 Facial Expressions and Facial Emotions

Facial expressions represent an important and powerful form of non-verbal communication, universal and intrinsic to all human beings (Darwin and Prodger, 1998). They are one of the pillars in social interaction, cognitive processes, are essential in the study of emotions and are part of nonverbal communication that represents two-thirds of human communication, and the other one-third represents verbal communication (Ko, 2018).

Facial expression is a multi-message system and a multi-signal system. It is capable of transmitting three types of signals such as: static signals like skin color, slow signals like permanent wrinkles and rapid signals like eyebrow raising, the first two of which are more difficult to change (Revina and Emmanuel, 2021).

Several studies have been carried out with the aim of developing techniques capable of measuring facial emotions. In 1872, Charles Darwin argued that all humans and animals display their emotions through similar behaviours and that facial movements, in an evolutionary context, were universal signs of emotions (Hess and Thibault, 2009). In 1978, Ekman and Friesen created a system to taxonomize visible facial expressions named Facial Action Coding System (FACS). This system originally consists of 46 action units (AUs) consisting of contraction or relaxation movements of one or more facial muscles that can be combined. The AUs focus on certain facial areas, mostly related to the mouth, eyes and eyebrows. Ekman and Friesen (1978) also defined a set of six basic emotions: anger, disgust, happiness, sadness, fear, and surprise, regardless of culture. In 1995, Feldman argued that emotions can be represented in two orthogonal dimensions: valence and arousal ranging from -1 to 1. Valence is a scale representing the negative or positive affectivity that refers to something unpleasant or pleasant. Arousal defines how calming or exciting the information is (Fasel et al., 2002).

Besides the use of Facial Action Coding System proposed by Ekman, other methods have been created. Over thirty years, the use of the electromyography (EMG) method has been developed in order to recognise the activation of facial muscles through the use of electrodes. Although it is a method that can be used to find an empirically based theory on emotion, it is technically complex to use and does not allow to examine facial expressions in social context. Finally automatic facial emotion recognition (FER) has been frequently used and is the most widely used method because it allows the analysis of facial expressions of emotion in a natural context.

2.2 Facial Emotion Recognition Algorithm

Facial emotion recognition (FER) has been the object of study and interest since 1990s. Its rapid development is due to the strong growth of artificial intelligence as well as the fields of virtual reality (VR) (Hickson et al., 2019), augment reality (AR) (Chen et al., 2015a), advanced driver assistant systems (ADASs) (Wilhelm, 2019) and human-computer interaction (HCI) (Nayak et al., 2021).

HCI is part of our daily lives. Facial emotion recognition during HCI has been a challenge due to the contactless user interface, quality of communication between the user and the computer, and real-time recommendation system. The ability of humans to react emotionally to certain situations when interacting with the computer allows for the extraction of facial expressions using static image-based methods and dynamic image sequences (e.g. video) based methods (Nayak et al., 2021; Ko, 2018).

FER systems used in real world using video analysis has been challenging due to head position, environment variations, light intensity, gender, age, background, occlusion caused by scarf, sunglasses, masks, skin diseases etc. (Li et al., 2022; Mellouk and Handouzi, 2020) which has led to rapid advances in emotional artificial intelligence (EAI), computer vision, as well as developments of automatic systems capable of assessing and measuring facial emotions.

In order to analyse the emotions expressed by the user, his/her interaction with the system is recorded on video. From that video the following steps are carried out (Matos et al., 2016):

Facial detection In order to detect the user's face in the extracted frame a search is performed with the face classifier based on Haar-like features (Viola-Jones method (Viola and Jones, 2001)) which divides the input image into windows and the location and detection of the size in the input image of the face resorts to the Adaboost learning algorithm and haar like features. The face region is marked and used in the next phase (Oliveira and Jaques, 2013; Mahersia and Hamrouni, 2015).

Preprocessing Preprocessing consists of removing the noise present in each frame of the video. It is the initial step and has a direct impact on the performance of the system and can be challenging due to head position, environment variations, light intensity, gender, age, background, occlusion caused by scarf, sunglasses, masks, skin diseases, etc. (Li et al., 2022; Mellouk and Handouzi, 2020) which has led to rapid advances in emotional artificial intelligence (EAI), computer vision, as well as developments of automated systems capable of assessing and measuring facial emotions.

In order to have better expression of the frames (Siddiqi et al., 2014), initial steps such as scaling, contrast adjustment, image clarity, and others should be taken into account (Bashyal and Venayagamoorthy, 2008).

Meng et al. (2017) through cropping and scaling processes had the nose as the main feature of identification, being the other features constituent parts of the face. The reduction of the size of the original image can be done by Bessel down sampling without losing quality (Owusu et al., 2014) and skin using the Gaussian filter that maintains its smoothness (Biswas and Sil, 2015).

Regarding the normalisation, using the median filter one can reduce the illumination and variations present in the image (Ji and Idrissi, 2012).

The use of Scale Invariant Feature Transform (SIFT) allows to perform face alignment that identifies the geometric structure of the faces in the input image (Dahmane and Meunier, 2014). Hernandez-Matamoros et al. (2015) defined ROI (Region of Interest) segmentation that allows creating a region of interest in the input image defined through a binary mask in which the pixels that belong to the ROI get the value one and those that are outside get the value zero. It is important to define the features of the face.

Other methods like histogram equalization allow detecting variations of the illumination in the input image and detecting through it the features (Uçar et al., 2016; Cossetin et al., 2016). According to Revina and Emmanuel (2021) ROI segmentation and histogram equalization are two of the most used methods in image processing.

Feature Extraction Feature extraction or selection consists of the extraction of facial features without loss of information from the image input. Generally its application consists of feature-based (Zebari et al., 2019) and model-based (Abduallah and Zeebaree, 2017) techniques or both.

Firstly, texture feature-based techniques consist of the use of Local Binary Pattern (LBP) that uses a threshold between the central pixel and the pixels located around it, can address rotation variation and grayscale (Happy and Routray, 2014), Gabor filter through the feature magnitude allows to know how the face features are organized (Kefal et al., 2016), Resource to spatiotemporal planes as Vertical Time Backward (VTB) and moments descriptor more focused on the stratification of features related are used (Ji and Idrissi, 2012), Discrete Contourlet Transform (DCT) uses two stages such as Laplacian Pyramid (LP) that divides the image in low pass, band pass and all discontinuous positions are confined and Directional Filter Bank (DFB) uses band pass (Biswas and Sil, 2015).

Methods like edge based techniques resort to the use of Line Edge Map (LEM) uses the Dyn2S algorithm that allows to improve geometrical structural features (Gao et al., 2003) and use of Histogram of Oriented Gradients (HOG) uses gradient filter allowing to extract visual features (Dahmane and Meunier, 2014).

Appearance feature-based techniques involve the whole user face who interacts with the system and use dimension reduction that allow extraction of local and global features (Ounachad et al., 2020). Principal Component Analysis (PCA) extracts low and global dimensional features, linear discriminant analysis (LDA) and Component Analysis (ICA) allows using multichannel observations to extract local features (Siddiqi et al., 2014) and the use of Linear Discriminant Analysis (SWLDA) allows through regression models to localize features (Siddiqi et al., 2015).

Resorting to geometric feature-based techniques allows extracting geometric position from the face. They resort to Local Curvelet Transform (LCT) which extracts geometric features like standard deviation, entropy and mean (Uçar et al., 2016), Facial Landmarks that predicts points on the face representing regions of specific muscles of the human face (Farkhod et al., 2022), Fuzzy Hamming Distance based approach resorts to a backnote to obtain specific information (Ionescu and Ralescu, 2005) as well as other measures like harmonic mean, contraharmonic mean, quadratic mean, Hamming distance, Degree of difference, Cardinality of a fuzzy set are also used (Ounachad et al., 2020).

According to Revina and Emmanuel (2021), there are features with high dimensional vectors that resort to techniques like PCA, Whitened Principle Component Analysis, Linear Discriminant Analysis and algorithms like Adaboost.

Classification Final stage of the FER system in which a facial emotion is assigned after preprocessing and extracting distinctively features from the image input such as those mentioned by Ekman as mentioned above.

Traditional methods such as the unsupervised clustering algorithm as Learning Vector Quantisation (LVQ) (Bashyal and Venayagamoorthy, 2008), Classification and Regression Tree (CART) (Yaacoub et al., 2019), ID3 Decision Tree (DT) (Revina and Emmanuel, 2021), statistical model as Hidden Markov Model (HMM) (Rahul et al., 2019), Support Vector Machine (SVM) (Dagher et al., 2019), Random Forest (RF) (Liu et al., 2018) and K-Nearest Neighbour (KNN)(Abdullah, 2019) were the first to be used in classification.

Currently, neural networks are the most widely used in the classification process. Multilayer Feed Forward Neural Network (MFFNN) (Revina and Emmanuel, 2021), Bayesian Networks (He et al., 2019), Deep Neural Network (DNN) (Jain et al., 2019) and Convolution Neural Network (CNN) (Mehendale, 2020). Recently deep learning approaches are the most widely used for their ability to significantly improve the results of different problems of recognition of facial expressions (Abdullah and Abdulazeez, 2021). Thus, currently the use of CNNs has been high due to the powerful capacity of hierarchical representations. Examples are AlexNet CNN (Sekaran et al., 2021), Convolution Neural Network with attention mechanism (ACNN) (Li et al., 2020), VGG16 (Dubey and Jain, 2020), Xception (Kalwad et al., 2022), ResNet50 (Li and Lima, 2021), MobileNet (Nan et al., 2022), Google Net (or Inception V1) (Caroppo et al., 2020), etc.

Figure 1 represents the steps required to classify the emotions extracted from the facial emotions expressed by the user when interacting with the system.

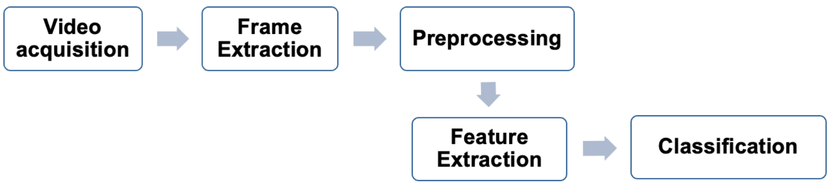


Fig. 1: Diagram of facial expression recognition algorithm.

2.2.1 FER Performance Evaluation

According to Ko (2018), the evaluation of a FER system is essential as it is possible to have a quantitative compaction. There are several ways to evaluate the performance of a FER system. One of the methods is the K-fold cross-validation capable of validating the generalization ability of a model and consists in dividing the dataset K-1 folds to train the model and evaluate the same with the rest. This model allows avoiding overfitting. Another method is the use of evaluation metrics such as precision, recall, accuracy, and F1-score.

2.2.2 Datasets

According to Kanade et al. (2000), the problem space of facial expression is constituted by levels of description, transitions between facial expressions, existence of training data and test data, head orientation, scene complexity, differences between facial features of different individuals.

Regarding the datasets used to train a FER system, a variety of specimens exist in the literature that differ mainly according to the number of emotions, the type of environment where they were recorded, which can be in a laboratory environment where conditions are controlled or in a field environment where conditions are wild, as mentioned in Section 2.2, and the number of users that participated. For example, in an environment with wild conditions, there are AFEW (Kossaifi et al., 2017), FER-2013 (Giannopoulos et al., 2018), EmotionNet (Fabian Benitez-Quiroz et al., 2016), FER-Wild (Mollahosseini et al., 2016) and AffectNet (Handrich et al., 2020), and in an environment with controlled conditions there are CK+ (Lucey et al., 2010), Multipie (Gross et al., 2010), DISFA (Mavadati et al., 2013), and AM-FED (McDuff et al., 2013).

Table 1 briefly describes the characteristics of each of the aforementioned datasets.

Table 1: Description of the characteristics of the analyzed datasets.

Dataset	# Emotions	Environment	# of Users	Description
AFEW	7	Wild	330	Videos
FER-2013	7	Wild	35.900	Images from web
EmotionNet	23 based on AUs	Wild	100.000	Images from web
FER-Wild	7	Wild	24.000	Images from web
AffectNet	8 with Valence and Arousal	Wild	450.000	1M images with facial landmarks 450.000 images annotated manually
CK+	7	Controlled and Posed	123	Frontal and 30º images
MultiPie	7	Controlled and Posed	337	Various angles of illuminations
DISFA	6	Controlled and Posed	27	Video and clips from stereo camera
AM-FED	14 AUs	Spontaneous	242	Facial videos

2.3 Facial emotion representation

There are several models capable of quantifying facial emotions. According to Mollahosseini et al. (2017) they can be quantified directly using a categorical model that uses the six basic emotions defended by Ekman (Ekman and Friesen, 1978), they can be quantified using a dimensional model (Kollias and Zafeiriou, 2021) that uses valence and arousal that as said above their values are between -1 and 1 and can be defined using FACS in which facial emotions are described in terms of AUs that contrary to the two models mentioned above this does not describe the state of facial emotion directly but rather explains the facial movements that occur (Álvarez et al., 2018).

2.4 Usability

Usability is a multidisciplinary concept defined in different ways and varies according to different models and standards where it is applied.

Shackel (1981) was one of the first authors to define usability. He defined the concept as the ability of a system to be used easily and effectively by certain users to perform certain tasks in a concrete environmental scenario.

Nielsen (1994b) defines usability as "a quality attribute that assesses how easy user interfaces are to use". The word "usability" also refers to methods for improving ease-of-use during the design process. ISO 9241-210 describes usability as the ability of specific users to use a product to achieve certain goals with efficiency, effectiveness and satisfaction (for Standardization, 2010).

Usability has been applied in various contexts, such as mobile apps, interfaces, software, and so forth. It is a quality attribute used to assess the ease with which the user uses a given interface. Its focus has to do with how the human is able to finish a task by interacting with the computer (Li and Jagadish, 2012).

According to Nielsen and Landauer (1993), usability can be defined in five modules that influence product acceptance: learnability (ease with which users have to perform basic tasks the first time they encounter the design), efficiency (after users have learned the design, how fast they are at performing tasks), memorability (when users return to use the design after a period of time, how easy it is to use the system again), low error rate (amount of errors that the user makes, impact of these errors and how they recover from mistakes) and satisfaction (ability of the user to use the system with pleasure) (Dourado, 2018).

A system must be designed in such a way that it fulfills all the modules mentioned above. Usability does not boil down to the sum of the scores of these measures but to achieving a certain level for each one. There are two usability measures that usually conflict. Among them, learnability and efficiency can influence each other negatively, so a system should have high learnability and efficiency (Nilsson and Følstad, 2012).

2.4.1 Usability evaluation methods

When talking about evaluation in terms of usability, two ways of reasoning can be considered. Seek to find out, with the support of the user, how easy it is to learn, remember, and use the system and to know its effectiveness, efficiency, and safety in relation to the system. In addition to these, emotional and fun measures can be considered (Gonçalves et al., 2017).

Usability evaluation is an integral part of design and should occur throughout the software development life cycle (SDLC) (Jindal, 2016). It is essential to identify and rectify any problems that may exist.

Fernandez et al. (2011) defines usability evaluation as a procedure that encompasses a set of well-elaborated activities with the purpose of collecting usage data related to the end-user's interaction with the software product or how the properties of this product allow reaching a certain level of usability.

According to Moradian et al. (2018), evaluation is a multi-dimensional methodology that can investigate user perceptions such as satisfaction, design, and learnability.

In accord with Asarbakhsh and Sandars (2013), usability in HCI aims to determine how satisfied users are when using computers. HCI is important in several scientific areas and is indirectly associated with the system design, evaluation, and implementation of a system designed for human use (Hollender et al., 2010).

A system's usability is evaluated by users or expert evaluators and should take place throughout the SDLC. Its realization is expensive in terms of financial costs and time-consuming and may compromise its acceptance by users due to a lack of analysis of critical aspects of usability. Financial costs for hiring expert evaluators Time-consumingly due to the stoppage of the development process, availability of expert evaluators and users, data collection, and generation of an expert report by the expert evaluators (de Souza Santos et al., 2022).

In both HCI and usability evaluation, attention is paid to what users do. Usability evaluation is an important part of HCI work. The difference is that the main focus of HCI is on designing new techniques and tools that improve usability, while usability evaluation tries to solve these problems (Li and Jagadish, 2012).

The performance of tests in order to evaluate the usability of a system can be in a controlled environment such as a laboratory or in a field environment that is performed in the user's work environment based on real-life situations and variables (Gonçalves et al., 2017).

According to Gonçalves et al. (2017), the evaluation of a system's usability can be performed using empirical and analytical methods. The empirical methods involve the use of users in the evaluation of a system and have as their purpose to evaluate the performance and satisfaction of the users in the accomplishment of a certain task. In order to collect information from them, we use interviews and questionnaires. An analytical evaluation uses expert evaluators, does not require the involvement of users, and uses predictive models or inspection methods. Predictive models are essential in an analysis of the mental and physical actions that the user needs to complete a given task that has been assigned to him, being able to conclude, for example, the time needed to complete the task. Analytical methods have the advantage of being quick to perform and do not depend on the user's availability.

2.4.2 Inspection methods

Inspection methods aim to find usability problems in the design, correct them, and improve the usability of the design. It also allows the user interface specifications not to be fully implemented, so they can be used at any stage of the SDLC cycle (Nielsen, 1994b).

Cognitive walkthrough (CW) It is a group-based expert method developed by (Polson and Lewis, 1990; Polson et al., 1992) and was based on exploratory cognitive learning theories analysing the users' ability to learn from their actions. For its implementation there are three aspects that must be taken into account. Initially, the characteristics of the users that will use the system are identified and documented, there is a detailed description of the prototype system that will be used in the usability evaluation, a list of all the tasks that will be performed and for each task a list of the necessary actions to complete it (Harms and Grabowski, 2014). The expert evaluators attempt to emulate the problem-solving capabilities of users by initially placing a sequence of actions to be performed by users in order to achieve a specific goal. Next, four questions are asked (Table 2) that, when answered for each sub-task action, allow the identification of usability problems in the system. Finally, when performing the documentation, the expert evaluator will write down the answers as well as personal notes for each action (Georgsson et al., 2019).

CW has the advantage of generating fast answers, having low cost, it can be applied in one phase in the SDLC as the design phase Wang and Caldwell (2002). The disadvantages are that it does not use real users, it is a cumbersome and invasive evaluation method and it generates a lot of data on low-priority interface design problems (Georgsson et al., 2019).

Table 2: Questions for the Cognitive Walkthrough (Wharton et al., 1994)

Q1	Will the user try to achieve the effect that the sub-task has?
Q2	Will the user notice that the correct action is available?
Q3	Will the user understand that the wanted subtask can be achieved by the action?
Q4	Does the user get feedback? Will the user know that they have done the right thing after performing the action?

According to Mahatody et al. (2010), there are several variants of CW such as Streamlined Cognitive Walkthrough, Cognitive Walkthrough for the Web, Cognitive Walkthrough with Users, Extended Cognitive Walkthrough, Distributed Cognitive Walkthrough and others.

Cognitive Walkthrough With Users (CWU) The introduction of the CW variant named CWU was initially proposed by Granollers and Lorés (2006). The need to introduce users was necessary in order to overcome some CW drawbacks. The proposed solution is defined in three main steps (Granollers and Lorés, 2005):

1. The completion of CW is performed in the traditional way;
2. After its completion users will be incorporated so that:
   (a) Users with the sought profile should be selected and recruited;
   (b) An explanation is shared on how the test will be performed, all the objectives, and the prototype system. Then, all users are invited to perform all actions in order to express their thoughts, feelings, and opinions as they interact with the system prototype out loud (the "think-aloud method");
   (c) The realization of the tasks by the users is done autonomously without any help or intervention during the test;
   (d) In addition, the users must comment on all the usability problems mentioned.
3. The expert evaluators analyse the doubts expressed by the users during the stage (2).

The main advantage of this method is its ability to significantly increase the number of usability problems found since it uses the advantages of CW and uses the participation of real users to perform tests on the system's prototype so that they note all the problems found (Granollers and Lorés, 2005).

Heuristic Evaluation (HE) , developed by Jakob Nielsen Nielsen (1994a) and Rolf Molich Nielsen and Molich (1990), is another flexible usability evaluation method capable of finding usability problems. Through the use of expert evaluators, they verify if the interface is in accordance with a certain set of usability principles known as heuristics (Gonçalves et al., 2017). This aims to find heuristic problems from the interface review taking into account can be used during development and at the end of the SDLC (Vieira et al., 2019).

This evaluation method is a good solution when there is difficulty finding users since the expert evaluator puts himself in the user's place during the evaluation process and allows completing the evaluation with users due to the experience that expert evaluators have in finding heuristic problems. Other advantages are that it is quick to perform, inexpensive as it does not require users, laboratories, or adequate and expensive equipment, and it is easy to teach (Gonçalves et al., 2017). Its disadvantages are the lack of validation after the use of a given heuristic, the expert evaluator not being the typical user of the evaluated system, and the lack of rigor and robustness of the method used (Vieira et al., 2019).

This method consists of several phases of heuristic evaluation. In the evaluation phase, the expert evaluator analyzes the interface at least twice, recording the usability problems using a set of ten heuristics (Gonçalves et al., 2017). At the end of the evaluation, a report on all problems found should be produced. In the consolidation phase, all the expert evaluators who participated in the system evaluation should compile all the problems found into a list of problems, thus identifying duplicates and similarities. In this phase, a severity grade is assigned to estimate and manage the effort needed to solve the problems encountered on a scale of 0 to 4. The balance phase brings together all the expert evaluators in order to find solutions to the usability problems and discuss the general characteristics of the interface. Finally, the design team analyzes the final report produced by the expert evaluators and, according to the degree of severity, corrects the interface's usability problems (Gonçalves et al., 2017).

Usability Smells According to de Souza Santos et al. (2022), the use of usability inspection is influenced by the experience of the expert evaluator regarding the domain of the application, the system to be evaluated, and the knowledge of the users used to perform it. Its request is not always possible, and the subjectivity factor of each expert evaluator has an impact on the final results.

In response to this issue, usability smells have been used to indicate the use of poor interface design that makes it difficult for the end-user to complete a given task and thus needs to be corrected. (Souza et al., 2021; de Souza Santos et al., 2022).

According to Harms and Grabowski (2014), usability smells are related to user behaviors and that there are strategies such as user testing or feedback, inspection methods and the use of access logs capable of identifying them.

For Souza et al. (2021), bad smells are symptoms capable of identifying that something is wrong in the design and are an important factor capable of affecting the maintainability of a software system. There are a variety of bad smells that can be identified in a system (Paternò et al., 2017).

Almeida et al. (2015) argues that "bad smells" are indicators capable of identifying the use of inadequate design in an application that harm maintenance and its evolution. There is an extensive list of "bad smells" that can be analysed and identified in a system in order to be corrected. Their identification and subsequent classification can be done using event logging (Grigera et al., 2016).

2.4.3 Predictive Evaluation

Predictive evaluation allows you to evaluate an interface before it exists. It does not involve the use of users and is useful for evaluating systems with predictable tasks. This type of evaluation is based on psychological principles as well as on experiments. It consists of two phases. To begin, the task's sequence and steps are identified. Then the steps are analyzed in order to determine usability measures such as the steps in which errors may occur and the time required to complete each step (Gonçalves et al., 2017; Ribeiro et al., 2019).

GOMS Developed by Card, Moran, and Newell John (1995), the GOMS (Goals, Operators, Methods and Selection Rules) model focuses on analyzing the complexity of users' interaction with the system. It has a "divide and conquer" approach in which users achieve goals by solving sub-goals. The term GOMS is an acronym for: Goals - Goals that users want to achieve; Operators - Cognitive processes and physical actions that users have to do in order to achieve the goals; Methods - Sequences of steps needed to achieve a goal or subgoals; Selection Rules - Used to determine which method to use when there are several available that allow achieving the goals (Gonçalves et al., 2017; Rosyidah et al., 2019).

KLM Keystroke Level Model, developed by the same authors as the aforementioned GOMS. It is a model that allows users to determine how much time they spend on specific tasks. It considers expert evaluators and decomposes the execution phase into five physical-motor operators, a mental operator, and an operator that relates to the response that the system is capable of giving (Gonçalves et al., 2017).

2.4.4 Empirical Usability Evaluation

Conducting an evaluation with real users requires careful planning. It is necessary to have an experimental plan and an experimental script. The experimental plan is a document that should indicate the objective, location, date, duration, equipment needed for the tests, software status, tasks, response time on the system used, users, data that will need to be collected and analyzed, help from manuals or other information resources that will be available for consultation by the user during the test, and criteria that will dictate whether an interface is successful or not. The experimental script is developed during the performance of the tests and is based on the experimental plan. It should include the introduction and objectives of the testing session, consent forms to be handed out and filled out by the user, a pre-test questionnaire, tasks, a post-test questionnaire, and an interview (Dix et al., 2003; Gonçalves et al., 2017).

Usability evaluation methods with users Can be carried out in several ways. Without interfering with their work, passive observation allows you to understand where they have more difficulty performing certain tasks in the laboratory and in the field. Using the think-aloud method helps the observer understand the users' thoughts as they speak aloud about what they perceive and think about the system they use during the tests. Indirect observation through automatic data collection both in the field and in the lab allows for detailed study of the users' use of the system and identifies usability problems more accurately. Finally, group interviews are conducted in order to obtain comments and suggestions from the users regarding the version of the system they used during the test (Dix et al., 2003; Gonçalves et al., 2017).

Usability Questionnaires Usability questionnaires are used in usability evaluation in order to obtain information on how easy the interface is to use. The SUS (System Usability Scale) Brooke et al. (1996) consists of a 10 item questionnaire with five response options for respondents from strongly agree to strongly disagree. SEQ is a direct question asked after users finish tasks in order to measure task performance satisfaction (Gonçalves et al., 2017). TAM (Technology Acceptance Model) is a questionnaire to measure technology acceptance consisting of 10 items to measure usefulness and 10 items to measure ease of use (Silva, 2015). ASQ (After-Scenario Questionnaire) consists of three statements that allow the user to express how easy it was for them to complete a scenario, state the time spent as well as the support they received to complete it (Lewis, 1991). CSUQ (Computer System Usability Questionnaire) is performed at the end of usability testing and evaluates the system overall from 19 statements with answers on a seven-point scale (Lewis, 1995). UEQ (User Experience Questionnaire) allows a quick evaluation on interactive systems with the elaboration of 26 semantic questions with a seven-point scale (Laugwitz et al., 2008). SUPR-Q (Standardized User Experience Percentile Rank Questionnaire) allows measuring trust, credibility, perception of usability and appearance of websites from 10 items distributed by the four factors that allows obtaining an overall value and partial values for each one (Sauro, 2015). Developed by Hart (2006), NASA-TLX is a multidimensional assessment tool capable of evaluating the cognitive load performed during a task. It contains six subjective subscales such as mental demand, physical demand, temporal demand, performance, effort and frustration. Other instruments such as the LEM-Tool (Huisman et al., 2013) cartoon representations that define eight discrete emotions through facial expressions and body postures containing a total of four positive and four negative emotions, the PrEmo (Product Emotion)(Desmet, 2003; Laurans and Desmet, 2012) is a self-report that uses non-verbal measures, various user emotions that include satisfactory and unpleasant emotions. Finally, AttackDiff evaluates users' feelings towards the evaluated system (Gonçalves et al., 2017).

According to Gonçalves et al. (2017), usability evaluation should not be something done only at the end of the design process, as is the case with summative evaluation, but should be performed during the SDLC in order to modify and improve the interface design. These can be formative or summative. On the one hand, formative evaluation is carried out throughout the iterative design cycle. Its objective is to identify specific aspects of the interface that give rise to usability problems so that they can be resolved and a new version of the prototype created. The prototype can be on paper or be a final functional prototype. This type of evaluation method includes heuristic evaluation and the "think aloud method." On the other hand, summative evaluation is carried out after the final design, so it focuses on evaluating the success of the final product in order to compare it with the usability criteria established at its initialization. It also allows for statistical calculations to compare with existing solutions on the market.

3 Related Work

This section presents relevant studies related to this work, namely studies on usability evaluation and facial emotion recognition (FER) algorithm. For each subsection a table is presented in order to summarise similarities and differences of all the analysed studies found in the literature.

The usability evaluation of a system is made by users or expert evaluators and should occur throughout the SDLC. Its implementation is expensive in terms of financial costs and time-consumingly, and may compromise its acceptance by users due to the lack of analysis of critical aspects of usability. Time-consumingly due to the stoppage of the development process, availability of expert evaluators and users, data collection and generation of an expert report by the expert evaluators (de Souza Santos et al., 2022). Another important aspect influencing the evaluation of the system's usability is the explicit feedback (e.g. user comments, bug reports, etc.) which is not always accurate and complete so that for example unsatisfied users tend to give more explicit feedback than satisfied users (de Souza Santos et al., 2022; Johanssen et al., 2019). The need to obtain valuable information with low cost and in a short time from usability evaluation is important for organisations. The use of facial emotions in the detection of usability problems can be the solution. The obtaining of implicit feedback (Johanssen et al., 2019) using facial emotions in an automatic way through the analysis in Human-Computer Interaction can help in obtaining valuable information about user experience in a less invasive way, in individuals with disability and less accustomed to use a new system or that never had contact with technology (de Souza Santos et al., 2022). The analysis of the methods and techniques developed, studied and present in the literature capable of usability evaluation, facial emotion recognition algorithms capable of classifying emotions accurately and how to measure the emotions involved in the interactions of humans with the machine system are an important object of study in HCI.

3.1 Usability evaluation

Aballay et al. (2021) during his study, conducted by experts in affective technologies and education, he proposed a method to obtain the emotions expressed and felt by students when using the functionalities of a virtual learning environment (VLE). The results obtained will allow, on the one hand, to improve the design of future mechanisms capable of collecting emotions and, on the other hand, to select specific and generic emotions from a VLE. The authors consider the following features when evaluating a VLE: pedagogical facility (functionalities that the environment provides to teachers and students in the learning process), support (focused on task completion by providing a list of specific steps), content (material used for learning), user interface (aspects of the system with which the user has contact), error handling (error detection, visible and clear error messages, etc.), tools (usability). After the characteristics of the VLE have been identified, it is necessary to know the methods used to evaluate these characteristics from the affective aspect. An online questionnaire was created on Google Forms and contained a demographic section and eight multiple-choice questions, one for each VLE characteristic mentioned above. In order to select the emotions associated with each VLE characteristic, participants mentioned a set of emotions initially pre-defined by experts. Next, scales (Magallanes et al., 2012) were defined for the emotions based on categorical approach (Josephs, 2005). In conclusion, the UX evaluation is necessary to know the preferences of the VLE students in order to know their satisfaction with the use of the VLE. The list of emotions created for each VLE feature validated by experts is essential to define the emotions of each feature that students could experience when using the VLE, although no definitive conclusions have been drawn about the emotions obtained.

Although e-commerce is increasing sales, stationary retail is decreasing customers. Meyer et al. (2021) to address this issue, they examined the benefits of stationary retail providing new customer services. Following the Design Science Research approach, it presented an emotion-based information system (IS) with the objective of supporting the interaction between customers and sales personnel. The evaluation of the interface prototype created was presented to experts to evaluate its use in stationary retail. This evaluation considered the influence of emotions, which, according to the experts, is a crucial factor mainly in private purchases. The participants who took part in this study had to make purchases both online and offline, ensuring the situation of a customer in stationary retail. During the use of the interface, users' emotions were captured using the Circumplex Model of Affect (CMoA) (Russell, 1980). The CMoA classifies a variety of emotions using the two dimensions valence and arousal. After using the prototype interface, all users completed a post-test questionnaire using a five-point Likert scale in which primarily online shop is coded with 1 or stationary retail is coded with 5. Next, a System Usability Scale (SUS) was used to test the usability of their instantiation and in order to select emotional situations general statements were made and evaluated with a five-point Likert scale (5 = strongly agree to 1 = strongly disagree). Finally, this study allowed them to conclude that the use of emotions is a way to support and reinforce the interactions between customers and sales personnel in stationary retail as from the creation of an interface.

Liu et al. (2021) studied usability and emotions in mental health assessment tools comparing mobile app and paper-and-pencil modalities. Although there are studies of psychiatric mobile apps focused on diagnostic accuracy and perceived usability, they do not focus on the impact of user emotional experiences. Thus, using EarlyDetect, which is a mental health assessment tool composed of several clinical questionnaires available in mobile app and paper-and-pencil formats. In a total of 191 participants who used the mental health clinic, they completed the EarlyDetect questionnaires for 10 to 15 minutes using either paper-and-pencil or the mobile app. In the assessment of emotion (enjoyment, boredom, frustration, and anxiety) questionnaires from Harley et al. (2020) and Poitras et al. (2019) were modified to include experiences of enjoyment, boredom, frustration, and anxiety. Each emotion was measured only once and the questionnaires had a five-point Likert scale (Allen and Seaman (2007)) where 1 corresponds to completely disagree and 5 to completely agree. Multivariate analysis of covariance (MANCOVA) was performed to avoid the Type-1 error rate that occurs when a null hypothesis is rejected when it is true.

Mental disorders are present all over the world and their study and treatment using mental health services is essential both to understand and help individuals that have them. Denecke et al. (2020) in order to help due to the limiting resources that exist in this area, they created SERMO, a mobile application with an integrated chat capable of implementing methods from cognitive behaviour therapy (CBT) in order to support people with mental illness in regulating emotions. Using SERMO, the user can interact with SERMO on a daily basis to automatically determine emotion on certain key events. The emotion determination is based on natural language input using natural language processing and a lexicon-based approach. From the emotion, activities or mindfulness exercises are determined at the moment in order to help the user. Furthermore, an emotion diary is created as well as a list of pleasant activities. This study had the participation of 21 subjects and used the User Experience Questionnaire (UEQ) in order to describe the user experience. The methodology used in the usability test consisted of each user performing six tasks. Each task included defining a goal, enter a mood, enter a current event, choose a pleasurable activity, run a mindfulness exercise and chat with SERMO for at least one minute. In the second part, as mentioned, each user had to judge concrete user experience aspects through the user experience questionnaire (UEQ) provided by Schrepp et al. The UEQ consisted of 26 items in which each item can be rated on a 7-point Likert scale and was grouped into six scales (attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty). In conclusion, the study showed that the application is well understood by psychologists but still needs to deal with variability of chatbots and unexpected user input and responses as well as improvement of system stability.

Schmidt et al. (2020) investigated the benefit of using the emotions given by emotion recognition software in predicting usability similar to traditional usability metrics. Three modalities were analysed such as speech, text and face so I will only mention what was mentioned regarding facial emotions. In a total of 125 users with an average age of 28.48, three tasks were performed on two different E-shop websites at various locations. As usability metrics were used task completion rate and System Usability Scale score (SUS) (Brooke et al. (1996)), Guerilla Usability Testing (Nielsen (1994a)) was used and all tests were recorded by desktop webcam. According to the author, facial emotion recognition did not present as good results as expected since some emotions may not resemble what they really mean so no significant correlations were found with facial emotion recognition and SUS scores, the tools used for emotion recognition presented some limitations for specific usability tests during a given task. In conclusion, the study performed was inconclusive and no positive points were mentioned about its accomplishment.

Porcu et al. (2019) in order to analyse the Emotional Impact of Video Quality conducted a study with the main objective of investigating the effects of visual degradation on quality perception. With the participation of 19 healthy participants with an average age of 29.5 years exposed to a series of short soccer-related video clips, they had to decide at the end of playback whether a certain event happened in the video. Regarding the study data collected through a camera placed behind the person or beside him, spontaneous facial expressions were captured. In order to analyse the same data two approaches were taken into account. The first approach used a multivariate analysis of variance to determine the effects of visual degradation factors on perceived quality and emotional dimensions, revealing that both were sensitive to degradation intensity. The second, using machine learning, used an automatic Video Quality of Experience (VQoE) prediction system to show the correlation between facial expressions and perceived quality. In conclusion, the results were able to demonstrate that emotional facial expressions were correlated with perceived quality meaning that non-obtrusive VQoE prediction indeed remains a feasible goal and should be further studied.

Carlotta Olivetti et al. (2019) proposed a methodology to evaluate virtual environments through the integration between traditional methods of self-report and emotion recognition methods through face analysis. Using these methods, the author's goal is to be able to draw conclusions that allow him to produce the design of any type of product or service. Design science research (DSR) follows the following steps: problem statement, hypothesis, development and evaluation. Regarding the problem statement it is mentioned that current self-reports are not able to fully evaluate user feedback. The hypothesis states that the FER methodology can help in the analysis of user feedback. The development was done with SVM. Finally, the evaluation proposes the solution during navigation of the virtual environment. An experiment was conducted with the participation of twelve subjects between 20 and 30 years old. Using a computer the user interaction was recorded and using a structured-light depth camera it was possible to record the body posture and facial expressions during the user interaction. During the experience two expert evaluators recorded the user's emotional responses. After the experience the users filled out the UES (user engagement scale) questionnaire (O'Brien and Toms, 2010). At the end, during the analysis of the recordings, the two expert evaluators selected frames of interest according to their notes, in order to classify the facial expressions during that period of interaction. Regarding the study of emotional activation, three classes adopted from Russell's model (Russell, 1980) were used: deactivated, averagely activated and activated. At the end, in order to analyse the results, a Comparison between Questionnaires and SVM Classification was made through the median UES (user engagement scale) and respective engagement classification groups. Finally, the author concluded that the use of facial expressions is a powerful tool to get more spontaneous feedback from users in a non-invasive way.

According to Johanssen et al. (2019), implicit feedback, of which emotions are a part, represents an important source of information in the delivery of software releases during the SDLC process and allows the identification of usability problems. A framework named EmotionKit was developed for use in mobile applications that use EMFACS. This framework uses the device's camera to extract facial expressions and is composed of six static views with usability problems defined by Nielsen. The study was done in a controlled environment with twelve participants—students or academic staff—with no selection rules. All participants ran through the six scenarios in the application, and the user's emotions were recorded. Every user action, such as clicking on buttons, was recorded and timestamped to relate it to the emotion expressed by the user. Finally, a quantitative analysis shows that there is a higher emotional response towards interactive parts (scenarios) than static ones (such as clicking a button) and a qualitative analysis to consider the performance of the framework, studying its applicability in emotion detection based on facial expressions.

Galindo et al. (2018) during his study proposed a method to adapt the interface according to the users' age and gender through the emotions detected at runtime to trigger adaptations. In his experiment, he categorised users' emotions according to the levels of UI usability. Each emotion is measured, using the FaceReader tool to capture the emotions advocated by Ekman, and interpreted in terms of affects that allow a decision to be made. Initially in his study, two hypotheses were defined related to the possibility of the existence of a relationship between emotions and UI aesthetics and usability levels and also if there was an emotional threshold to detect usability and aesthetics problems. During the course of the experiment, four UI versions of a travel website were used that varied according to usability and aesthetics aspects. During user interaction (45 users, 24 male from 19 to 67 years old and 21 female from 23 to 63 years old) with each UI, facial expressions were captured in order to obtain emotions. At the end of each task, in which a set of three tasks were assigned to each user, each of them answered an online questionnaire. Finally, it was possible to categorize the emotions depending on the users' age and gender, and there are emotion thresholds that can be defined to detect usability problems that can improve the system interface. As a future work, the same author defended that it would be interesting to have statistical analyses that would allow them to analyse more precisely usability problems.

Xu et al. (2018) mentioned that in-vehicle information systems (IVISs) are used while driving since they are able to provide timely information during the driver's actions as well as the environment outside the vehicle. The need to discover the type of interface that increases the usability of such a system is crucial. An experiment was therefore carried out to compare the usability of two different IVISs by performing finger tracking and facial expression recognition technologies that will be mentioned, usability performance analysis, and facial expression arousal level scores. A total of 29 participants between the ages of 23 and 45 who did not wear glasses and had no experience in using IVISs were selected to carry out the experiment for two years. The users used an SUV for two years and used two IVISs called "Calink" similar to using a smartphone and "Carlife" representing a traditional vehicle information system, while a camera (Logitech C920) installed in the SUV recorded their facial expressions. The captured video was run though by the FaceReader software that determined the emotions. During the test, the participants were asked to complete three tasks in the two IVISs and could ask for help from the expert evaluator who was standing next to the participant and recorded the times the user asked for help during the tasks. At the end of the experiment they were asked to complete a three-question questionnaire with a seven-point rating scale. The data analysis was carried out along the three dimensions of usability according to the ISO definition. Finally as negative points it concluded that the velocity (mean, peak, and standard deviation) and valence may not reflect the usability effectively and as positive point it was able to prove that Calink (smartphone style) had better performance than Carlife (traditional style) through consistent analysis of subjective scores from facial expression data.

Munim et al. (2017) in his study created a web based user experience evaluation tool that serves to measure user emotions. The tool has as input the facial expressions recorded by the webcam and the screen recording that will be loaded in the system dataset that will allow the automatic generation of a UX evaluation report. The tool essentially consists of recording the facial expressions and the screen from AForge.NET and to detect the emotions Affdex SDK was used, which saves them in a text file. Four participants with average age of 30 years, and three to four years of experience using gaming systems participated in this study. Thus, the game "Zombie Tsunami" was used for three minutes in order to make use of this tool, and also had the participation of two expert evaluators who wrote down the emotions and compiled the evaluation report graphically. In conclusion, through the analysis and combination of the data collected by the tool, it was possible to determine the *effectiveness* and *efficiency* of the system. Regarding *effectiveness*, it was verified that the graph obtained by the tool and by the expert evaluator were identical, so it was concluded that the use of an expert evaluator brought unnecessary costs, since the tool could produce the same results in a single click and had less probability of error and possibility of biasness. *Efficiency* was achieved through the analysis of emotions in a specific time of game use, so when the user felt that he was going to lose he had the fear emotion and when he felt he was going to win he had the happy emotion. Finally, the author mentions that a recommendation of UX evaluation using user emotions in different phases of the game can be generated.

Kujala and Miron-Shatz (2013) examined emotions in real-life mobile phone experiences over a period of five months. Their goal was to relate emotions and memories to usability evaluation and user experience evaluation. Positive emotions are related to a good user experience and negative emotions to poor usability. In a total of 22 participants who agreed to participate in the experiment for five months, they had different jobs. The participants, during the period of using the mobile phone every six days, reported their emotions on the first five days of using it Kahneman et al. (2004) and on the sixth day mentioned their overall evaluation of the product as well as their satisfaction in a questionnaire. The questionnaire contained five measures related to user experience, behavioural intentions, emotional reactions, perceived usability and the most memorable experience episodes. The same author has also analyzed the results through correlations between the means of daily experienced and recalled emotions after five days of use, as well as means and standard deviations of the product behavioural intentions, evaluation measures, and emotions. Finally, as positive points, he concluded that the user experience is built periodically over time and the role of usability increases with the use of the mobile phone throughout the experience, however, not all experience episodes throughout the experience are equally important and critical for user satisfaction and that usability problems had more impact and attention when they were written in the questionnaires.

According to Staiano et al. (2012), he proposed a non-invasive system called UXMate that relies on users' facial expressions, interprets them based on pre-set rules and relates the occurrence of a particular emotion to an interactive event. Compared to other user experience evaluation methods, it has the advantage of being non-invasive, can be installed on any device with a camera, can be used in a natural environment without the need for lighting adjustments, and is low-cost. The facial emotions are based on anatomical analysis of the human face derived from FACS as mentioned in the "Facial

Expressions and Facial Emotions" section 2.1. Prior to the study, participants signed a consent form stating that they could be recorded during the course of the experiment. Four media-players were tested (iTunes, MusicBee, Songbird and MediaMonkey) that had different levels of usability and functionalities. Each of the media-players was installed on the participants' laptops as well as a program capable of recording video and audio simultaneously. Three tasks were performed on each media player. After they were completed, the participants filled out a UX questionnaire stating which media player they would use, information about the use of all media players and demographic date. The media player was used for one month with the help of an expert in case of technical problems. The study collected performance data (errors and time) obtained by the expert evaluator, questionnaires and facial expressions.

Champney and Stanney (2007) mentioned that emotions are present in the product design discussion and are essential in usability evaluation. More than encompassing user satisfaction, they are able to collect information about the pleasure of using a given product. He defines emotions as brief, acute and intentional episodes that facilitate reaction to certain events that can last seconds or minutes. The authors introduce the Emotional Profiling (EP) method that can show how emotions in HCI can be used by usability experts. EP uses a non-verbal electronic questionnaire with 18 animated negative and positive unlabeled emoticons expected to be selected (can be multiple responses) when a user interacts with a product. This method has three phases consisting of the user interacting with a product, assessment part which uses the questionnaire to note the emotions chosen by the user during his interaction with the product and finally a short interview to obtain notes given by the user on the chosen emotions. Finally, the information is analyzed by an expert to ensure that the user's description and choices match the user's answers. For example, if the user has said that he/she would like to buy the product, this would express desire. In order to represent the emotions through the information collected from the questionnaire and the short interview, a Graphical Emotional Profile (GEP) was created, which represents the distribution of emotions and their intensity, and an Emotion Appraisal Profile (EAP) contains a list of what was said briefly in the short interviews and identifies the respective emotion. The combination of these two graphs allows us to identify the intensity of the emotions as well as their source.

Yammiyavar et al. (2007) conducted a study involving 18 cross-cultural 'Think Aloud Method'. The think-aloud usability tests were conducted in two stages. The first part was performed with interactions through gestures and the second part was performed through facial expressions, which is the part that is interesting to be examined in this study. Focusing on the second part, 10 participants between the ages of 20 and 23 were trained on how to perform "Think Aloud" tests. The users had to perform tasks on three websites for a maximum of 45 minutes, and at the end they were given a qualitative interview in order to know how the interaction, TA behaviour, level of satisfaction as well as the impact of the method used by the participants. During the whole process, a camera captured the users' facial expressions. In the end, people known to the participants and three reviewers who did not know them were asked to evaluate the facial expressions captured by the camera in order to describe what the participant was thinking and saying out loud ('Think Aloud Method'). Finally, the same author stated that conclusions were drawn about the impact of facial expressions on the ability to convey the participant's thoughts when performing tasks, however, the culture of the users influences the way people they know describe the facial expressions observed in the participants' videos.

Zaman and Shrimpton-Smith (2006) in order to determine the value of

FaceReader when conducting usability research, an experiment was conducted on a total of twelve participants between the ages of 20 and 60. The test contained nine tasks performed on the personal computer with the purpose of testing different levels of computer/Internet knowledge and skills. After the test, a questionnaire was given, which contained a first part concerning self-reported “subjective” emotions and another with usability questions based on a QUIS questionnaire appropriate for measuring user interaction satisfaction (Chin et al., 1988). The test was conducted in a room with a computer and a camera positioned towards the user’s face and in such a way that the user was not aware of it, and a glass that allowed the expert evaluator to observe the user and that the user could not see. The collection and analysis of the observed data was done from ’The Observer 6.1’, a Noldus software program. On the one hand, in order to evaluate the usability, the input given by the expert evaluator is crucial to rate the effectiveness and efficiency of the system that will observe the usability problems that the users cannot describe, but the user’s experiences, feelings and thoughts only he can express through the QUIS questionnaire. On the other hand, emotions have the advantage of helping the expert evaluator in his evaluation, since besides the observation, essential verbal cues such as comments, voice intonation, stop words and sighs he has from the user, he cannot observe and write down all the emotions, so the use of FaceReader is essential.

Duh et al. (2006) conducted a study investigating the differences between usability tests on mobile phones in real life situations and in the laboratory. The benefits of having implemented a good design allow for reduced mental and physical stress, increased operability in the use of devices as well as the overall quality of the product. Most of the usability studies performed are laboratory based, which allows for easy data collection, however the fact that it is laboratory based does not counteract uncontrollable facts that influence the use of mobile phones in real life situations. A pre-test questionnaire was used to obtain information about the 10 participants such as demographics, experience in using mobile phones, frequency in using mobiles phones during the day, SMS functions, and a final question to test if the participants knew the procedure and purpose of the study. A post-test questionnaire was also conducted to obtain subjective information from the participants about their achievement in performing the tasks, opinions about the tested features, key usability problems encountered, and the difference they noticed in the use of the device in the two environments. A five point Likert scale was also used. The participants performed the tasks according to “prioritize by frequency” (Duh et al., 2006). For the laboratory usability test, the scenario was made so that the participants were familiar with it prior to performing the tasks, while for the field usability test, the participants had to adapt to the scenario. The usability testing and data-logging method used was thinking aloud. Finally, it was revealed that there are a higher percentage of usability problems in the field than in the laboratory, and although the study helps to understand the limitations in the process of usability evaluation and contributes to an improvement in the design of mobile devices in a social context, it was not possible to draw conclusions about external factors such as noise, privacy when using the device, and mental and physical resources. Ahmad et al. (2018) in his study through game interface design incorporated user observation into usability evaluation. Twenty users (eight males and twelve females) when playing “A Garuda game” were captured every movement of the character they were using for a comparative analysis. The user’s emotions while playing the game were also collected by recording by a laptop/PC camera from the beginning to the end of the game in a way that the user was not aware of. The purpose of this information gathering was to analyze GUI effectiveness, efficiency and satisfaction in order to develop a game toward its usability. A laptop/PC equipped with a camera, the game and a screen recorder software was used in this study. The facial expressions and the screen recordings were made simultaneously using Lotus Screen Cam 8 software, outputting a single video with the two recordings. After the recordings were made, the impact of the emotions was determined by calculating the mean and standard deviation. It concludes that through effective, efficient and satisfactory user experience a better understanding of how to improve game interface design and that a quantitative analysis of user emotions can improve the understanding of usability evaluation.

Through the analysis of Table 3 it can be concluded that there are several authors who use the emotions advocated by Ekman both in a laboratory environment ("L" in column "Setting") and in a wild environment ("W" in column "Setting") and parallel to the use of the same there are several usability evaluation methods such as the Guerilla Usability Testing, Think Aloud Method, System Usability Scale (SUS) and the Cognitive Walktrgouh (CW).

3.2 Facial Emotion Recognition Algorithm

Due to the need to classify small datasets with low resolution images, Salimov and Yoo (2021) needed to develop the small scale deep CNN model using only a small part of memory in terms of the total number of learnable weights showing equally good results than normal CNNs. The architecture of the proposed model was fully created having in general a decrease of window size, this from 2x2 size windows to perform convolution operation for all layers, after the convolutional layers 3 fully connected layers were used.

Mukhopadhyay et al. (2021) created an algorithm with a CNN for statics image classification that makes use of textural images of human faces using Local binary pattern (LBP) to generate their texture. The texture images can be described as fine, coarse, grained, smooth, etc. and have the advantage of giving more low-level details about the images than the normal images.

Focusing also on the creation of a CNN for facial emotion recognition, Modi and Bohara (2021) argued that it could be applied in Augmented reality(AR) technology in order to help Autism Spectrum Disorder(ASD).

Fallahzadeh et al. (2021) in his CNN proposal for facial emotion recognition used a pre-trained model named AlexNet-DCNN. As a conclusion, first, the information contained in the gradients of the original image can increase the accuracy of the system. Next, a large number of epochs can lead to overfitting. Finally, data augmentation and dropout technique can also improve the accuracy and avoid overfitting.

According to Generosi et al. (2020), the ability and need to understand User eXperience (UX) in the wild is important for e-commerce in order to obtain information about the opinion and behaviour of users so that it is possible to customize services, productivity and decision strategies. The possibility of extracting huge amounts of data in the wild is due to the presence of cameras integrated in our smartphones and laptops, not being restricted to the use of a laboratory, thus opening the way to support UX assessment of web-based applications in the wild improving the effectiveness of e-commerce systems. It was created a toolkit with a centralised architecture to be used in web platforms. It consists of a Web Plugin used for the frontend in order to obtain information about the users when they interact with the web application such as clicks, scroll coordination and activation of the webcam located on the computer. It makes use of Deep Learning Platform (DLP) for the backend that uses photos taken from the user to get information about his behaviour and a CNN to classify the users emotions.

Table 3: Summary of covered setting, emotions and usability method evaluation described above.

Author, year **App name**	**Setting**	**Emotions**	**Usability method evaluation**
Aballay et al. (2021)	L	Get Emotion	Google Forms questionnaire
Meyer et al. (2021)	L	Valence and Arousal (Fasel et al., 2002)	System Usability Scale (SUS)
Liu et al. (2021)	L	Enjoyment, boredom, frustration, and anxiety	Questionnaires from Harley et al. (2020) and Poitras et al. (2019) changed
Denecke et al. (2020)	W	Fear, disgust, anger, joy, grief, guilt and shame Berking et al. (2008)	User Experience Questionnaire (UEQ)
Schmidt et al. (2020)	W	Ekman	Guerilla Usability Testing, Think Aloud Method and System Usability Scale (SUS)
Porcu et al. (2019)	L	Valence and arousal	SAM scale
Carlotta Olivetti et al. (2019)	W	Russell's model (Russell, 1980)	User engagement scale (UES)
Johanssen et al. (2019) EmotionKit	L	Ekman and Friesen (1978)	User observation
Galindo et al. (2018)	W	Ekman & (Fasel et al., 2002)	Questionnaire
Ahmad et al. (2018)	L	Plutchik	User observation
Xu et al. (2018)	W	Ekman	CW
Munim et al. (2017)	L	Joy, surprise, fear, engagement and anger	User observation
Kujala and Miron-Shatz (2013)	W	Positive and negative	Questionnaires
Staiano et al. (2012) UX_Mate	W	Frustation and confusion	UX questionnaire and User observation
Champney and Stanney (2007) Emotional Profiling	L	Positive and negative	Emotional questionnaire created, interview and User observation
Yammiyavar et al. (2007)	L	——	Think Aloud Method
Zaman and Shrimpton-Smith (2006)	L	Ekman	QUIS questionnaire
Duh et al. (2006)	L & W	——	CW

In order to test the best model, a confusion matrix plotted as heatmap was used. In conclusion, the study was a success due to the creation of a toolkit able to collect data about user interactions with the web application in order to support UX analysis and using deep-learning to analyse and classify facial expressions "in the wild" through a simple webcam.

Dubey and Jain (2020) proposed deep learning-based framework through the use of transfer learning for facial expression. In order to improve the accuracy of VGG16, it modified the training model already trained with ImageNet that contains a total of 1000 classes by adding more layers to it. The top layers were removed and added Flatten, Dense, Drop that drops some values in order to avoid overfitting, as well as dense-SoftMax layer for emotion classification.

In order to create a system capable of detecting students' emotions using Haar Cascades detector for their images. Thus, Lasri et al. (2019) created a CNN with four convolutional layers, four max pooling and two fully connected layers able to help teachers to recognize their face emotions during their presentations.

Zhang et al. (2019) created a CNN and in image edge detection. Contrary to the creations of traditional facial expression recognition algorithms that involve CNN, this new method is able to use a CNN with the following innovations: in each layer, the edge of each input image is extracted and then the information contained in the edge is superimposed in each feature image in order to preserve the edge structure information of the texture image able to learn pattern features and reduce the incompleteness caused by artificial design features.

Levi and Hassner (2015) based on the success of Convolutional Neural Networks (CNN) on face recognition problems focused on reducing confounding factors from the input images, with an emphasis on image illumination variations in order to reduce the amount of data required in the model training as well as simplify the problem domain. In order to achieve the proposed goal, they proposed a novel transformations of image intensities to 3D spaces so that it would be invariant to monotonic photometric transformations. The advantage of this model is the ability to look beyond RGB as the input space for CNNs since it gets more information offered by the dataset representation used to train the model.

Focusing on supervised classification trees, Chen et al. (2015b) presented Sparse Coding trees (SC-trees) that aim to solve misclassifications when there are several classes of emotions that map to the same set of features. Several classification tree models were used that are divided into on node-specific dictionaries and classifiers.

Through the analysis of Table 4 it is possible to analyse that Ekman's emotions are strongly used to classify users' facial expressions resorting to several datasets such as CK+, FER2013 and AffectNet. Regarding preprocessing methods, they are used differently by each author, as well as feature extraction methods. Regarding the CNN used for classification, it was verified that VGG16 presents the highest accuracy of 94.84%, followed by DCNN with 93.03%.

4 Proposed Solution

This section introduces a proposal approach to research questions, all participants involved in conducting tests, a recording tool used to record participants, a study support website where participants will interact, the facial emotion recognition algorithm, a study support website for conclusions, quantitative and qualitative analysis, and preliminary work. All the steps are mentioned in the illustration in Figure 2 which are explained in the course of this section.

Table 4: Summary about datasets and Facial Emotion Recognition Algorithm used above.

Author, year	**Datasets**	**# Classes**	**Preprocessing method**	**Feature extraction method**	**Classification**	**Classification Accuracy(%)**
Salimov and Yoo (2021)	FER2013	7 Ekman	Gray scale	—	CNN	A. 85.35%
Mukhopadhyay et al. (2021)	CK+	7 Ekman	Gray scale	Textural Feature Extraction using LBP	CNN	A. 79.56%
Modi and Bohara (2021)	FER2013	7 Ekman	Gray scale, scaling to 48x48	Appearance feature-based	ConvNet	A. 73.58%
Fallahzadeh et al. (2021)	CK+	7 Ekman	Gray scale	Horizontal and vertical gradients	DCNN	A. 93.03%
Generosi et al. (2020)	FER+ AffectNet	7 Ekman	Alignment, centralization of the face, face rotation, images scalted to 64x64 pixels	Facial Landmarks	VGG13 VGG16 VGG19 Inception	A. 75.48% A. 74.48% A. 73.14% A. 75.26%
Dubey and Jain (2020)	CK+	7 Ekman 6	Scaling to 224x224 pixels	—	VGG16	A. 94.84%
Lasri et al. (2019)	FER2013	7 Ekman	Cutting, normalization, scaling to 48x48	Adaboost	CNN	A. 70%
Zhang et al. (2019)	FER2013	7 Ekman	Haar classifier, gray scale, cutting, normalization, histogram equalization	Edge based methods	CNN	A. 85.56%
Levi and Hassner (2015)	AFEW	7 Ekman	Scaling to 224x224 pixels, gray scale and cropping	LBP	GoogleNet	A.54.56%

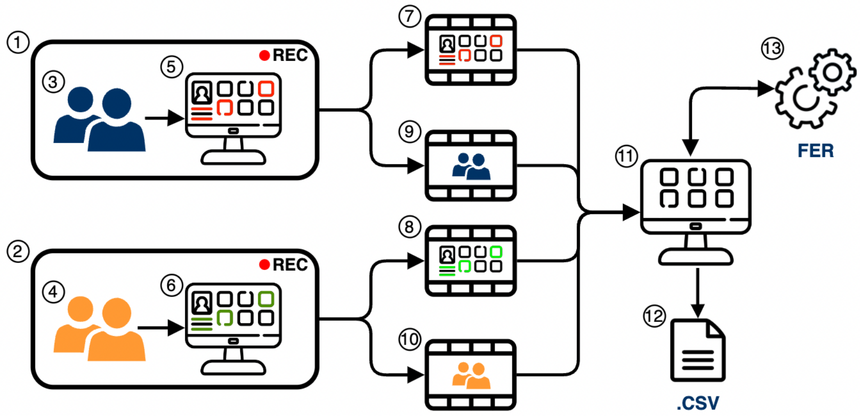


Fig. 2: Illustration of the procedure used in conducting the study.

4.1 Research Questions

Throughout the extensive research related to and described above, some relevant questions will provide additional relevant information for the purpose of drawing final conclusions and fulfill all the objectives mentioned in Section 1.1, including:

RQ1. Does the initial number of usability problems encountered by the expert evaluator increase when facial emotions are considered?

In other words, the use of facial emotions will be able to detect additional usability problems compared to the usability problems that the expert evaluator found without considering the emotional aspects of the users. In order to answer this question, the results will be analysed both for the website with usability problems and without usability problems. For each of them, an average will be made of the total number of additional problems found using emotion across all tests. In this way, we will verify whether, in addition to the usability problems found through Cognitive Walkthrough and Usability Smells, the expert evaluators, by analysing the emotions at all moments of the user's interaction with the system, are able to discover new usability problems that had not been initially detected.

RQ2. After successfully completing a task, will the user express positive facial emotions?

Will the user, when interacting with the system and successfully completing a task, express positive emotions? In order to analyze the emotions when a user completes or does not complete a given task, an average will be made of the emotions expressed by users for each of the tasks, both on the website with usability problems and without usability problems.

RQ3. When a user fails to complete a task, will he/she exhibit negative facial emotions?

Similarly, when interacting with the system and failing to complete a task, will the user display negative emotions? These questions will be analyzed in the same manner as in RQ2.

4.2 Procedure

The study will be described objectively, following the order and sequence outlined above (in Figure 2) and described in the following subsections.

4.2.1 Participants

The participants for this study will be selected through convenience sampling and will include more than 20 individuals aged 14-60, of any gender, without physical limitations, and who are frequent users of digital systems. Recruitment will be through direct contact and word of mouth. The participants will be divided into two groups: group A (Figure 2, (3)) and group B (Figure 2, (4)) users.

Before the study starts, all participants must agree to and sign a consent form allowing for the collection and recording of data during the performance of all proposed tasks. This data will be used at all stages of the study.

4.2.2 Recording tool

The test session will start after obtaining user consent for the study. During the test session, the user's facial expressions will be recorded while they interact with the system (Figure 2, (1) and (2)).

The recording will be made using a computer webcam facing the user and simultaneously, the computer screen will be recorded.

We developed a python script that uses the libraries opencv-python 4.7.0.66[1], to record and save the video of the facial expressions, and PyAutoGUI 0.9.53[2] that allows, together with opencv-python 4.7.0.66, to record the screen. The script outputs two videos, facial expressions (Figure 2, (9) and (10)) and screen recorder (Figure 2, (7) and (8)) in .mp4 format. In order to be user-friendly a GUI framework, named tkinter[3], that allows the initialisation and the finalisation of the recording of the user expressions and the monitor screen and allows the expert evaluator to place the title of the test, the group of user (group A or B) and the location of the test (Figure 3).

The names of both videos will consist of the date and time of the recording start, the information written in the GUI (test title, user group (group A or B), and test location), and at the end, "W" if it is the facial expressions video or "S" if it is the monitor screen recording. All information is separated by "+".

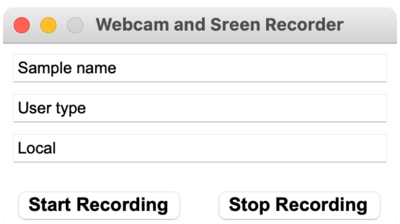


Fig. 3: Webcam and screen recorder application.

[1] https://pypi.org/project/opencv-python/
[2] https://pypi.org/project/PyAutoGUI/0.9.36/
[3] https://docs.python.org/3/library/tkinter.html

4.2.3 Study support website

Websites are the backbone of information exchange used by organizations able to present their products or services to potential customers and are ready for present a variety of methods and tools used to introduce usability problems. In this way, two e-commerce websites named “TechIST” were created with the Django framework, differing in the fact that one presents clear usability problems (Figure 2, (5)) and is used by group A users and the other is not (Figure 2, (6)) and is used by group B users.

The usability problems introduced in a strategic manner capable of negating the basic principles of interface design proposed by Polson and Lewis (1990) on the website with usability problems in a selected set of views will be necessary to use usability evaluation methods capable of supporting the expert evaluators mentioned in Section 2.4.2 such as the Cognitive Walkthrough With Users(CWU) and Usability Smells.

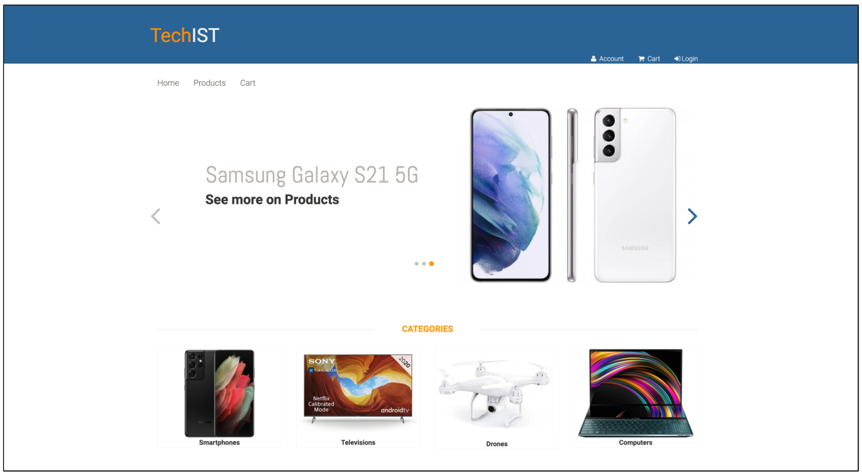


Fig. 4: E-commerce website that will be used by the participants.

Table 5: Tasks used for Cognitive Walkthrough With Users (CWU).

Task id	Tasks
1	Register on the TechIST website
2	Sign in
3	Change your username
4	Buy a samsung A7 tablet and an Iphone 14
5	Add a review
6	Logout

Usability smells, or anomalies in the design of a system that hinder maintenance, understanding, and evolution, will be identified during the study. These “bad smells” can negatively impact the participant’s performance when using the website. A total of 10 usability smells (Table 6) will be considered (Grigera et al., 2017).

After the user of group A or group B finishes the completion of his/her test, the generated logs file will be extracted that contains the exact time

Table 6: Usability smells placed on the website with usability problems.

Event id	Event	Usability Smell
1	Login form submission	No client validation
2	Form submission	Late validation
3	Update user profile	Abandoned form
4	Go to wrong page	Misleading link
5	Long time request	No processing page
6	Sign up form	Unformatted input
7	Form field	Short input
8	Search form	Few search results
9	Click action	Unresponsive element
10	Option selection	Wrong default value

in which each action performed by the user takes place (Figure 2, (3) and (4)).

4.2.4 Facial Emotion Recognition Algorithm

With the goal of training models in order to classify facial emotions from a video given as input that contains the interaction of each user with the system we will use VGG (Figure 2, (8)), also known as VGGNet. VGG presents an architecture of a classic CNN, standing out as it increases its performance. As presented above in Section 3.2, it has been widely used for the purpose of facial emotion classification and presents performance levels above 70%. It will be trained using the AffectNet and FER2013 datasets since they contain samples in the wild, which is compatible with the environment used. On the one hand, Affectnet contains a large number of facial expressions distributed among the facial emotions presented by Ekman adding one more emotion and, additionally, it contains for each sample the valence and arousal levels that will also be used. On the other hand, FER2013 contains fewer samples than AffectNet and presents all of Ekman's facial emotions. In order to improve the performance of the VGG, the preprocessing and feature extraction steps are taken into account.

4.2.5 Study support website for conclusions

The study support website used to draw conclusions (Figure 2, (11)) aims to receive the videos recorded by the recording tool (webcam and screen recorder) of the interaction of users with the system both of group A and group B so that the user interaction with the system can be analysed by expert evaluators in order to draw conclusions. Due to the particularity that both videos, corresponding to the user interaction with the system, have exactly the same duration it will allow us to:

1. Use the facial emotion recognition algorithm to classify the facial emotions of the corresponding video to be synchronised with the video recording of the computer screen;
2. Process the video of the user's facial expressions and returns the same with the indication of the percentages of each emotion for a given frame as well as a graph of the percentage of each emotion over time.

Due to the fact that there may be more than one expert evaluator to analyse the recorded videos of user interaction with the system, the website created will allow to help them in a way that:

1. Mention the study in question;
2. Allow to write down notes of potential problems and answer the 4 questions mentioned in Section 2.4.2 regarding the cognitive walkthrough for each sub-video of the expert evaluator as well as the user; allow the identification of the expert evaluator who performed and analysed the test;

The website will be built in django using SQLite as dataset in order to store all the information, allow sessions and originate as output a .csv file on all relevant data obtained.

4.3 Preliminary Work

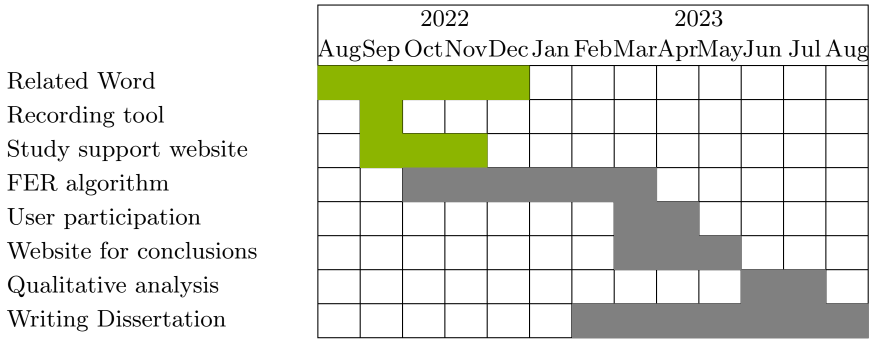
The collection of information on the subject of study has been going on since the beginning of August. Parallel to this, the python script described above that allows recording the user's facial expressions when interacting with the system, the complete e-commerce website and the FER algorithm were developed. Regarding the FER algotirhm, the AffectNet and FER2013 datasets were used in order to train algorithms such as VGG16, ResNet50 and other simple CNNs. The accuracies obtained are around 75% being the highest with the use of VGG16. As mentioned in Section 4.3.4, the algorithm used will be the VGG, in this case the VGG16, so it will be tested in order to improve its performance from the techniques mentioned in the related works (Section 2.2) and with the individual or collective use of the datasets mentioned.

5 Workplan

Table 7 shows the mapping of the work carried out and that will be carried out, dividing each stage into its respective months. Since September, work has begun on the recording tool, the study support website and the choice and training of some emotion classification algorithms as mentioned above, and the VGG training will continue in order to improve its performance until the end of February.

User testing is scheduled to begin in March and continue through April. The data collection website will be available from March to May, with data analysis (both quantitative and qualitative) occurring in June and July. The dissertation will be written from February to August, concurrently with some of these stages. With the green color in the graph it means what has already been done and with the gray color what will be done.

Table 7: Study work schedule.



6 Conclusion

This document presents a research proposal for a study that will analyse the effects of how the use of facial emotions during the process of usability evaluation of a system can influence the results obtained by expert evaluations. The first part of the document is constituted by the background section that presents the main concepts about facial expressions, facial emotion recognition algorithms, facial emotion representation and usability. Afterwards, in order to analyse the works present in the literature, works related to usability evaluation and facial emotion recognition algorithm were explored. The proposed solution includes relevant research questions, the procedure that explains how the study will be carried out, and finally everything that has already been done in the preliminary work. Finally, in the workplan it is referred the division of the stages starting in the month of August 2022 and ending in the month of August 2023 having therefore the duration of 1 complete year the realization of this study.

Bibliography

Aballay, L. N., Aciar, S. V., and Collazos, C. A. (2021). Emotions for virtual learning environments. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 16(3):215–224.

Abduallah, W. M. and Zeebaree, S. R. M. (2017). New data hiding method based on dna and vigenere autokey. *Academic Journal of Nawroz University*, 6(3):83–88.

Abdullah, A. I. (2019). Facial expression identification system using fisher linear discriminant analysis and k-nearest neighbor methods. *ZANCO Journal of Pure and Applied Sciences*, 31(2):9–13.

Abdullah, S. M. S. and Abdulazeez, A. M. (2021). Facial expression recognition based on deep learning convolution neural network: A review. *Journal of Soft Computing and Data Mining*, 2(1):53–65.

Ahmad, I., Abdullasim, N., and Suaib, N. M. (2018). Usability testing on game interface design using video-based behavior analysis. *International Journal of Engineering & Technology*, 7(2.15):142–145.

Allen, I. E. and Seaman, C. A. (2007). Likert scales and data analyses. *Quality progress*, 40(7):64–65.

Almeida, D., Campos, J. C., Saraiva, J., and Silva, J. C. (2015). Towards a catalog of usability smells. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pages 175–181.

Álvarez, V. M., Velázquez, R., Gutiérrez, S., and Enriquez-Zarate, J. (2018). A method for facial emotion recognition based on interest points. In *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)*, pages 1–4. IEEE.

Asarbakhsh, M. and Sandars, J. (2013). E-learning: the essential usability perspective. *The clinical teacher*, 10(1):47–50.

Bani, M., Russo, S., Ardenghi, S., Rampoldi, G., Wickline, V., Nowicki, S., and Strepparava, M. G. (2021). Behind the mask: Emotion recognition in healthcare students. *Medical science educator*, 31(4):1273–1277.

Bashyal, S. and Venayagamoorthy, G. K. (2008). Recognition of facial expressions using gabor wavelets and learning vector quantization. *Engineering Applications of Artificial Intelligence*, 21(7):1056–1064.

Berking, M., Wupperman, P., Reichardt, A., Pejic, T., Dippel, A., and Znoj, H. (2008). Emotion-regulation skills as a treatment target in psychotherapy. *Behaviour research and therapy*, 46(11):1230–1237.

Biswas, S. and Sil, J. (2015). An efficient expression recognition method using contourlet transform. In *Proceedings of the 2nd International Conference on Perception and Machine Intelligence*, pages 167–174.

Brooke, J. et al. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

Carlotta Olivetti, E., Violante, M. G., Vezzetti, E., Marcolin, F., and Eynard, B. (2019). Engagement evaluation in a virtual learning environment via facial expression recognition and self-reports: A preliminary approach. *Applied Sciences*, 10(1):314.

Caroppo, A., Leone, A., and Siciliano, P. (2020). Comparison between deep learning models and traditional machine learning approaches for facial expression recognition in ageing adults. *Journal of Computer Science and Technology*, 35(5):1127–1146.

Champney, R. K. and Stanney, K. M. (2007). Using emotions in usability. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 51, pages 1044–1049. SAGE Publications Sage CA: Los Angeles, CA.

Chen, C.-H., Lee, I.-J., and Lin, L.-Y. (2015a). Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders. *Research in developmental disabilities*, 36:396–403.

Chen, K., Comiter, M. Z., Kung, H., and McDanel, B. (2015b). Sparse coding trees with application to emotion classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 77–86.

Chin, J. P., Diehl, V. A., and Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 213–218.

Chowdary, M. K., Nguyen, T. N., and Hemanth, D. J. (2021). Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications*, pages 1–18.

Cossetin, M. J., Nievola, J. C., and Koerich, A. L. (2016). Facial expression recognition using a pairwise feature selection and classification approach. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 5149–5155. IEEE.

Dagher, I., Dahdah, E., and Al Shakik, M. (2019). Facial expression recognition using three-stage support vector machines. *Visual Computing for Industry, Biomedicine, and Art*, 2(1):1–9.

Dahmane, M. and Meunier, J. (2014). Prototype-based modeling for facial expression analysis. *IEEE Transactions on Multimedia*, 16(6):1574–1584.

Darwin, C. and Prodger, P. (1998). *The expression of the emotions in man and animals*. Oxford University Press, USA.

de Souza Santos, F., Vinícius Treviso, M., Gama, S. P., and de Mattos Fortes, R. P. (2022). A framework to semi-automated usability evaluations processing considering users' emotional aspects. In *International Conference on Human-Computer Interaction*, pages 419–438. Springer.

Denecke, K., Vaaheesan, S., and Arulnathan, A. (2020). A mental health chatbot for regulating emotions (sermo)-concept and usability test. *IEEE Transactions on Emerging Topics in Computing*, 9(3):1170–1182.

Desmet, P. (2003). Measuring emotion: Development and application of an instrument to measure emotional responses to products. In *Funology*, pages 111–123. Springer.

Dix, A., Finlay, J., Abowd, G. D., and Beale, R. (2003). *Human-computer interaction*. Pearson Education.

Dourado, M. A. D. (2018). Usability heuristics for mobile applications a systematic review.

Dubey, A. K. and Jain, V. (2020). Automatic facial recognition using vgg16 based transfer learning model. *Journal of Information and Optimization Sciences*, 41(7):1589–1596.

Duh, H. B.-L., Tan, G. C., and Chen, V. H.-h. (2006). Usability evaluation for mobile device: a comparison of laboratory and field tests. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, pages 181–186.

Ekman, P. and Friesen, W. (1978). *Facial action coding system (facs)*. Manual. Book.

Fabian Benitez-Quiroz, C., Srinivasan, R., and Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570.

Fallahzadeh, M., Farokhi, F., Harimi, A., and Sabbaghi-Nadooshan, R. (2021). Facial expression recognition based on image gradient and deep convolutional neural network. *Journal of AI and Data Mining*, 9(2):259–268.

Farkhod, A., Abdusalomov, A. B., Mukhiddinov, M., and Cho, Y.-I. (2022). Development of real-time landmark-based emotion recognition cnn for masked faces. *Sensors*, 22(22):8704.

Fasel, I. R., Bartlett, M. S., and Movellan, J. R. (2002). A comparison of gabor filter methods for automatic detection of facial landmarks. In *Proceedings of Fifth IEEE international conference on automatic face gesture recognition*, pages 242–246. IEEE.

Fernandez, A., Insfran, E., and Abrahão, S. (2011). Usability evaluation methods for the web: A systematic mapping study. *Information and software Technology*, 53(8):789–817.

for Standardization, I. O. (2010). *Ergonomics of Human-system Interaction: Part 210: Human-centred Design for Interactive Systems*. ISO.

Galindo, J. A., Dupuy-Chessa, S., Mandran, N., and Céret, E. (2018). Using user emotions to trigger ui adaptation. In *2018 12th International Conference on Research Challenges in Information Science (RCIS)*, pages 1–11. IEEE.

Gao, Y., Leung, M. K., Hui, S. C., and Tananda, M. W. (2003). Facial expression recognition from line-based caricatures. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 33(3):407–412.

Generosi, A., Ceccacci, S., Faggiano, S., Giraldi, L., and Mengoni, M. (2020). A toolkit for the automatic analysis of human behavior in hci applications in the wild. *Advances in Science, Technology and Engineering Systems*, 5(6):185–192.

Georgsson, M., Staggers, N., Årsand, E., and Kushniruk, A. (2019). Employing a user-centered cognitive walkthrough to evaluate a mhealth diabetes self-management application: A case study and beginning method validation. *Journal of biomedical informatics*, 91:103110.

Giannopoulos, P., Perikos, I., and Hatzilygeroudis, I. (2018). Deep learning approaches for facial emotion recognition: A case study on fer-2013. In *Advances in hybridization of intelligent methods*, pages 1–16. Springer.

Gonçalves, D., Fonseca, M., and Campos, P. (2017). Introdução ao design de interfaces. *Lisboa: FCA-Editora de Informática, Ltda*.

Granollers, T. and Lorés, J. (2005). Cognitive walkthrough with users: an alternative dimension for usability methods. In *Proc. HCI International, Las Vegas*.

Granollers, T. and Lorés, J. (2006). Incorporation of users in the evaluation of usability by cognitive walkthrough. In *HCI related papers of Interacción 2004*, pages 243–255. Springer.

Grigera, J., Garrido, A., Rivero, J., and Rossi, G. (2016). Automatic detection of usability smells in web applications. *International Journal of Human-Computer Studies*, 97.

Grigera, J., Garrido, A., Rivero, J. M., and Rossi, G. (2017). Automatic detection of usability smells in web applications. *International Journal of Human-Computer Studies*, 97:129–148.

Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). Multi-pie. *Image and vision computing*, 28(5):807–813.

Handrich, S., Dinges, L., Al-Hamadi, A., Werner, P., and Al Aghbari, Z. (2020). Simultaneous prediction of valence/arousal and emotions on affectnet, aff-wild and afew-va. *Procedia Computer Science*, 170:634–641.

Happy, S. and Routray, A. (2014). Automatic facial expression recognition using features of salient facial patches. *IEEE transactions on Affective Computing*, 6(1):1–12.

Harley, J. M., Lajoie, S. P., Tressel, T., and Jarrell, A. (2020). Fostering positive emotions and history knowledge with location-based augmented reality and tour-guide prompts. *Learning and Instruction*, 70:101163.

Harms, P. and Grabowski, J. (2014). Usage-based automatic detection of usability smells. In *International Conference on Human-Centred Software Engineering*, pages 217–234. Springer.

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA.

He, J., Yu, X., Yu, L., and Sun, B. (2019). Facial emotion and action unit recognition based on bayesian network. In *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, pages 363–368.

Hernandez-Matamoros, A., Bonarini, A., Escamilla-Hernandez, E., Nakano-Miyatake, M., and Perez-Meana, H. (2015). A facial expression recognition with automatic segmentation of face regions. In *International Conference on Intelligent Software Methodologies, Tools, and Techniques*, pages 529–540. Springer.

Hess, U. and Thibault, P. (2009). Darwin and emotion expression. *American Psychologist*, 64(2):120.

Hickson, S., Dufour, N., Sud, A., Kwatra, V., and Essa, I. (2019). Eyemotion: Classifying facial expressions in vr using eye-tracking cameras. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1626–1635. IEEE.

Hollender, N., Hofmann, C., Deneke, M., and Schmitz, B. (2010). Integrating cognitive load theory and concepts of human–computer interaction. *Computers in human behavior*, 26(6):1278–1288.

Huisman, G., Van Hout, M., Van Dijk, E., Van Der Geest, T., and Heylen, D. (2013). Lemtool: measuring emotions in visual interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 351–360.

Ionescu, M. and Ralescu, A. (2005). Fuzzy hamming distance based banknote validator. In *The 14th IEEE International Conference on Fuzzy Systems, 2005. FUZZ'05.*, pages 300–305. IEEE.

Jain, D. K., Shamsolmoali, P., and Sehdev, P. (2019). Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters*, 120:69–74.

Ji, Y. and Idrissi, K. (2012). Automatic facial expression recognition based on spatiotemporal descriptors. *Pattern Recognition Letters*, 33(10):1373–1380.

Jindal, T. (2016). Importance of testing in sdlc. *International Journal of Engineering and Applied Computer Science (IJEACS)*, 1(02):54–56.

Johanssen, J. O., Bernius, J. P., and Bruegge, B. (2019). Toward usability problem identification based on user emotions derived from facial expressions. In *2019 IEEE/ACM 4th International Workshop on Emotion Awareness in Software Engineering (SEmotion)*, pages 1–7. IEEE.

John, B. (1995). Why goms? *interactions*, 2(4):80–89.

Josephs, L. (2005). Emotions revealed: Recognizing faces and feelings to improve communication and emotional life, by paul ekman. henry holt and com. *American Journal of Psychoanalysis*, 65(4):409–411.

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., and Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science*, 306(5702):1776–1780.

Kalwad, P. D., Kanakaraddi, S. G., Chikaraddi, A. K., Preeti, T., and Gull, K. C. (2022). Xception: Facial expression detection using deep learning techniques. In *Proceedings of Third International Conference on Intelligent Computing, Information and Control Systems*, pages 339–353. Springer.

Kanade, T., Cohn, J. F., and Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Proceedings fourth IEEE international* *conference on automatic face and gesture recognition (cat. No. PR00580)*, pages 46–53. IEEE.

Kefal, A., Oterkus, E., Tessler, A., and Spangler, J. L. (2016). A quadrilateral inverse-shell element with drilling degrees of freedom for shape sensing and structural health monitoring. *Engineering science and technology, an international journal*, 19(3):1299–1313.

Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401.

Kollias, D. and Zafeiriou, S. (2021). Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*.

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., and Pantic, M. (2017). Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36.

Kujala, S. and Miron-Shatz, T. (2013). Emotions, experiences and usability in real-life mobile phone use. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1061–1070.

Lasri, I., Solh, A. R., and El Belkacemi, M. (2019). Facial emotion recognition of students using convolutional neural network. In *2019 third international conference on intelligent computing in data sciences (ICDS)*, pages 1–6. IEEE.

Laugwitz, B., Held, T., and Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and usability engineering group*, pages 63–76. Springer.

Laurans, G. and Desmet, P. (2012). Introducing premo2: New directions for the non-verbal measurement of emotion in design. In *Out of Control: Proceedings of the 8th International Conference on Design and Emotion, London, UK, 11-14 September 2012*.

Levi, G. and Hassner, T. (2015). Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 503–510.

Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the asq. *ACM Sigchi Bulletin*, 23(1):78–81.

Lewis, J. R. (1995). Computer system usability questionnaire. *International Journal of Human-Computer Interaction*.

Li, B. and Lima, D. (2021). Facial expression recognition via resnet-50. *International Journal of Cognitive Computing in Engineering*, 2:57–64.

Li, F. and Jagadish, H. (2012). Usability, databases, and hci. *IEEE Data Eng. Bull.*, 35(3):37–45.

Li, H., Xiao, X., Liu, X., Guo, J., Wen, G., and Liang, P. (2022). Heuristic objective for facial expression recognition. *The Visual Computer*, pages 1–12.

Li, J., Jin, K., Zhou, D., Kubota, N., and Ju, Z. (2020). Attention mechanism-based cnn for facial expression recognition. *Neurocomputing*, 411:340–350.

Liu, Y., Yuan, X., Gong, X., Xie, Z., Fang, F., and Luo, Z. (2018). Conditional convolution neural network enhanced random forest for facial expression recognition. *Pattern Recognition*, 84:251–261.

Liu, Y. S., Hankey, J., Lou, N. M., Chokka, P., and Harley, J. M. (2021). Usability and emotions of mental health assessment tools: Comparing mobile app and paper-and-pencil modalities. *Journal of Technology in Human Services*, 39(2):193–211.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer* *society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE.

Magallanes, Y., Sánchez, J. A., Molina-Rueda, A., and Méndez, Y. A. (2012). Towards an emotional validation of heuristic approaches for usability evaluation. *Acta Universitaria*, 22:119–125.

Mahatody, T., Sagar, M., and Kolski, C. (2010). State of the art on the cognitive walkthrough method, its variants and evolutions. *Intl. Journal of Human–Computer Interaction*, 26(8):741–785.

Mahersia, H. and Hamrouni, K. (2015). Using multiple steerable filters and bayesian regularization for facial expression recognition. *Engineering Applications of Artificial Intelligence*, 38:190–202.

Matos, A., Filipe, V., and Couto, P. (2016). Human-computer interaction based on facial expression recognition: A case study in degenerative neuromuscular disease. In *Proceedings of the 7th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*, pages 8–12.

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160.

McDuff, D., Kaliouby, R., Senechal, T., Amr, M., Cohn, J., and Picard, R. (2013). Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 881–888.

Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (ferc). *SN Applied Sciences*, 2(3):1–8.

Mellouk, W. and Handouzi, W. (2020). Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science*, 175:689–694.

Meng, Z., Liu, P., Cai, J., Han, S., and Tong, Y. (2017). Identity-aware convolutional neural network for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 558–565. IEEE.

Meyer, M., Siemon, D., and Robra-Bissantz, S. (2021). Emotion-based is support for customer-salesperson interaction.

Modi, S. and Bohara, M. H. (2021). Facial emotion recognition using convolution neural network. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1339–1344. IEEE.

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

Mollahosseini, A., Hasani, B., Salvador, M. J., Abdollahi, H., Chan, D., and Mahoor, M. H. (2016). Facial expression recognition from world wild web. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 58–65.

Moradian, S., Krzyzanowska, M. K., Maguire, R., Morita, P. P., Kukreti, V., Avery, J., Liu, G., Cafazzo, J., Howell, D., et al. (2018). Usability evaluation of a mobile phone–based system for remote monitoring and management of chemotherapy-related side effects in cancer patients: mixed-methods study. *JMIR cancer*, 4(2):e10932.

Mukhopadhyay, M., Dey, A., Shaw, R. N., and Ghosh, A. (2021). Facial emotion recognition based on textural pattern and convolutional neural network. In *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, pages 1–6. IEEE.

Munim, K. M., Islam, I., Khatun, M., Karim, M. M., and Islam, M. N. (2017). Towards developing a tool for ux evaluation using facial expression. In *2017 3rd international conference on Electrical Information and Communication Technology (EICT)*, pages 1–6. IEEE.

Nan, Y., Ju, J., Hua, Q., Zhang, H., and Wang, B. (2022). A-mobilenet: An approach of facial expression recognition. *Alexandria Engineering Journal*, 61(6):4435–4444.

Nayak, S., Nagesh, B., Routray, A., and Sarma, M. (2021). A human–computer interaction framework for emotion recognition through time-series thermal video sequences. *Computers & Electrical Engineering*, 93:107280.

Nielsen, J. (1994a). Guerrilla hci: Using discount usability engineering to penetrate the intimidation barrier. *Cost-justifying usability*, pages 245–272.

Nielsen, J. (1994b). *Usability engineering*. Morgan Kaufmann.

Nielsen, J. and Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 206–213.

Nielsen, J. and Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 249–256.

Nilsson, E. G. and Følstad, A. (2012). Effectiveness and efficiency as conflicting requirements in designing emergency mission reporting. In *I-UxSED*, pages 20–25. Citeseer.

O'Brien, H. L. and Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69.

Oliveira, E. and Jaques, P. (2013). Classificação de emoções básicas através de imagens capturadas em vídeos de baixa resolução. *Revista Brasileira de Computação Aplicada ISSN 2176-6649*, 5:40–54.

Ounachad, K., Oualla, M., and Sadiq, A. (2020). Geometric feature based facial emotion recognition. *International Journal*, 9(3).

Owusu, E., Zhan, Y., and Mao, Q. R. (2014). A neural-adaboost based facial expression recognition system. *Expert Systems with Applications*, 41(7):3383–3390.

Paternò, F., Schiavone, A. G., and Conti, A. (2017). Customizable automatic detection of bad usability smells in mobile accessed web applications. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–11.

Poitras, E. G., Harley, J. M., and Liu, Y. S. (2019). Achievement emotions with location-based mobile augmented reality: An examination of discourse processes in simulated guided walking tours. *British Journal of Educational Technology*, 50(6):3345–3360.

Polson, P. G., Lewis, C., Rieman, J., and Wharton, C. (1992). Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of man-machine studies*, 36(5):741–773.

Polson, P. G. and Lewis, C. H. (1990). Theory-based design for easily learned interfaces. *Human–Computer Interaction*, 5(2-3):191–220.

Porcu, S., Uhrig, S., Voigt-Antons, J.-N., Möller, S., and Atzori, L. (2019). Emotional impact of video quality: self-assessment and facial expression recognition. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE.

PS, N. and Aithal, P. (2022). Real-time customer satisfaction analysis using facial expressions and head pose estimation. *International Journal of Applied Engineering and Management Letters (IJAEML)*, 6(1):301–312.

Rahul, M., Kohli, N., Agarwal, R., and Mishra, S. (2019). Facial expression recognition using geometric features and modified hidden markov model. *International Journal of Grid and Utility Computing*, 10(5):488–496.

Revina, I. M. and Emmanuel, W. S. (2021). A survey on human face expression recognition techniques. *Journal of King Saud University-Computer and Information Sciences*, 33(6):619–628.

Ribeiro, R. F., de Meneses Campanhã Souza, M., de Oliveira, P. A. M., and de Alcântara dos Santos Neto, P. (2019). Usability problems discovery based on the automatic detection of usability smells. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 2328–2335.

Rosyidah, U., Haryanto, H., and Kardianawati, A. (2019). Usability evaluation using goms model for education game "play and learn english". In *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pages 1–5. IEEE.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Salimov, S. and Yoo, J. H. (2021). A design of small scale deep cnn model for facial expression recognition using the low resolution image datasets. *The Journal of the Korea institute of electronic communication sciences*, 16(1):75–80.

Sauro, J. (2015). Supr-q: A comprehensive measure of the quality of the website user experience. *Journal of usability studies*, 10(2).

Schmidt, T., Schlindwein, M., Lichtner, K., and Wolff, C. (2020). Investigating the relationship between emotion recognition software and usability metrics. *i-com*, 19(2):139–151.

Sekaran, S. A. R., Lee, C. P., and Lim, K. M. (2021). Facial emotion recognition using transfer learning of alexnet. In *2021 9th International Conference on Information and Communication Technology (ICoICT)*, pages 170–174. IEEE.

Shackel, B. (1981). 1984. the concept of usability. In *Proceedings of the IBM Software and Information Usability Symposium. IBM Corporation, Poughkeepsie, NY*, pages 1–30.

Siddiqi, M. H., Ali, R., Khan, A. M., Park, Y.-T., and Lee, S. (2015). Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields. *IEEE Transactions on Image Processing*, 24(4):1386–1398.

Siddiqi, M. H., Ali, R., Sattar, A., Khan, A. M., and Lee, S. (2014). Depth camera-based facial expression recognition system using multilayer scheme. *IETE Technical Review*, 31(4):277–286.

Silva, P. (2015). Davis' technology acceptance model (tam)(1989). *Information seeking behavior and technology adoption: Theories and trends*, pages 205–219.

Souza, O. T., Souza, A. D. d., Vasconcelos, L. G., Baldochi, L. A., et al. (2021). Usability smells: A systematic review. In *ITNG 2021 18th International Conference on Information Technology-New Generations*, pages 281–288. Springer.

Staiano, J., Menéndez, M., Battocchi, A., De Angeli, A., and Sebe, N. (2012). Ux_mate: from facial expressions to ux evaluation. In *Proceedings of the Designing Interactive Systems Conference*, pages 741–750.

Uçar, A., Demir, Y., and Güzeliş, C. (2016). A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering. *Neural Computing and Applications*, 27(1):131–142.

Vieira, E. A. O., SILVEIRA, A. C. d., and Martins, R. X. (2019). Heuristic evaluation on usability of educational games: A systematic review. *Informatics in Education*, 18(2):427–442.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee.

Wang, E. and Caldwell, B. (2002). An empirical study of usability testing: heuristic evaluation vs. user testing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 46, pages 774–778. SAGE Publications Sage CA: Los Angeles, CA.

Webb, N., Ruiz-Garcia, A., Elshaw, M., and Palade, V. (2020). Emotion recognition from face images in an unconstrained environment for usage on social robots. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Wharton, C., Rieman, J., Lewis, C., and Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In *Usability inspection methods*, pages 105–140.

Wilhelm, T. (2019). Towards facial expression analysis in a driver assistance system. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–4. IEEE.

Xu, N., Guo, G., Lai, H., and Chen, H. (2018). Usability study of two in-vehicle information systems using finger tracking and facial expression recognition technology. *International Journal of Human–Computer Interaction*, 34(11):1032–1044.

Yaacoub, A., Assaghir, Z., Makki, S., and Almokdad, R. (2019). Diagnosing clinical manifestation of apathy using machine learning and micro-facial expressions detection. In *Proceedings of the 2019 3rd International Symposium on Computer Science and Intelligent Control*, pages 1–6.

Yammiyavar, P., Clemmensen, T., and Kumar, J. (2007). Analyzing nonverbal cues in usability evaluation tests. In *International Conference on Usability and Internationalization*, pages 462–471. Springer.

Zaman, B. and Shrimpton-Smith, T. (2006). The facereader: Measuring instant fun of use. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*, pages 457–460.

Zebari, D. A., Haron, H., Zeebaree, S. R., and Zeebaree, D. Q. (2019). Enhance the mammogram images for both segmentation and feature extraction using wavelet transform. In *2019 International Conference on Advanced Science and Engineering (ICOASE)*, pages 100–105. IEEE.

Zhang, H., Jolfaei, A., and Alazab, M. (2019). A face emotion recognition method using convolutional neural network and image edge computing. *IEEE Access*, 7:159081–159089.