Visual Customer Interaction through Emotion Detection and Face Landmarks

Rui P. Duarte*, Carlos A. Cunha*, Valter Borges*, André Ferreira* and David Mota†
*School of Management and Technology
Polytechnic Institute of Viseu, Viseu, Portugal
pduarte@estgv.ipv.pt, cacunha@estgv.ipv.pt, estgv16626@alunos.estgv.ipv.pt, af_af_10@hotmail.com
†Bizdirect Competence Center, Viseu, Portugal
david.mota@bizdirect.pt

***Abstract*—Understanding consumer behavior is a dynamic field, critically important to the success of companies and to consumer satisfaction. It is especially important in scenarios of intense competition, currently characteristic of the retail store industry, where companies fight for every individual customer. A great in-store experience encourages shoppers to become loyal customers, positive word of mouth and referrals. However, the opposite happens if customers' needs are not met, a poor customer experience is provided and further visits of the customer may be at risk. Due to the dimension of several retail stores, a common problem is the location of products and the ability of customers to find them. When this occurs, sales decrease and customer satisfaction is not guaranteed, thus contributing to a poor customer experience. In this paper we present a method that targets user satisfaction, by providing the retail store a tool that detects if a product is not being found. Our approach is twofold: first, we detect if the customer is revealing signs of negative emotions by tracking the facial expressions, and second, the facial position of the customer is tracked to detect if he/she is repeatedly looking at the same place. In each context, a lost factor is updated and when a threshold is passed, the retail store assistant is notified for customer assistance. Results show that this method is well suited for emotion detection and will increase customer satisfaction and retail stores income.**

Keywords–Image recognition; sentiment analysis; activity recognition; face landmarks; user satisfaction; retail environments.

I. INTRODUCTION

Today, the development of technology has a significant impact on society and on the organizations within it. This poses significant challenges for organizations once they are obliged to keep up with developments so fast that they can often suffer if they lack the manageability. New technologies partly determine the way people relate to, and inspire the characterizations of our society. They are the new transmission channels that shape this new world, virtual and technological. Advances in technology allow organizations to be more flexible and open to change, making the most of the opportunities that appear in the market. These opportunities are partially defined by consumers, which are an increasingly visual society. Everything we see as colors, textures, shapes, and images can communicate something to us and the ability to use this type of information is of most importance for companies. In this paper, we focus on the consumer experience by providing an employee assistance to avoid consumer unsatisfactory experiences. It improves the work presented in CENTRIC'2019 [1] by adding a new method based on face landmkarks, which, coupled with the emotion detection method, increases customer satisfaction in a retail store environment.

The concept of shopping has been changing during the years [2]. Today shops are not only the place where customers go to buy products but also the place where they spend part of their time. Thereof, retail stores need to adapt to the needs of customers in order to provide them a positive experience. Two perspectives are present: the customer that wants to find and buy a specific product and the retail store that wants to increase sales. Although in real context scenarios an easy match can be established between perspectives, they have different approaches to achieve a win-win-win solution for the customer–retailer–manufacturer relation. According to Oliver, R.L. [3], it is more challenging to fidelize an existing customer than to attract new ones. However, sometimes, this is not the case: a customer enters the retail store to buy a product, does not find it, and leaves the shop without spending money on that product. This transforms the process into an unsatisfactory experience for all the players involved.

The application of Video Analytics Technology (VAT) in retail dates back more than two decades [4]. More recently, due to advances in computer vision, machine learning, and data analysis, retail video analytics can provide retailers with much more insightful business intelligence [5][6][7]. Thus it promises much higher business value, far beyond the traditional domain of security, authentication, and loss prevention. Examples of this include analysis of store traffic, queue data, customer behavior, and purchase decision making among others. However, it is a complex real-world scenario, and many technical challenges are present for realistic computer vision techniques: changing and uncontrollable lighting conditions, high-level, complex human and crowd activities, cluttered backgrounds, crowded scenes, occlusion, odd viewing angles, low resolution cameras, limited contrast, and low object discriminability [5]. It is well known that VAT mostly focuses on automatic customer detection for the retail store industry. However, customer perspective is of most importance since they acquire products available in stores. One of the potential areas of interest is to determine whether a customer is not finding a specific product. As a consequence, the customer leaves the store without buying it, which does not relate to a win-win-win situation. Thus, it is of most importance to collect more information about customers by using VAT to detect if they are not finding a product and generate triggers to employees informing of the problem. This will increase customer satisfaction and retail store sales.

This paper deals with the role of visual interaction with customers as a strategic resource to promote competitiveness and interactivity between organizations and customers. We target a library assistant, whose main objective is to capture the customer using webcam in real-time and to detect, over a time window, that they need assistance from an employee. For this, a camera has to be in continuous capture of customer

facial information and two algorithms detect if a particular customer is lost and needs assistance. In addition to the main goal described above, other objectives have been identified. The specific contributions of this paper are:

- *Emotion analysis.* To our best knowledge, this is the first scalable attempt to measure negative emotions to determine if a customer is not finding products in a retail stores. These negative emotions are the basis of unsatisfactory behavior of the customer in the context of a purchase.
- *Exploration of face landmarks for customer detection.* Face landmarks feature physical characteristics of customers. From our knowledge this is the first attempt to used them to detect if a customer is lost when looking at products in a shelf. We track the position where a customer is looking at, and determine if the location is repeated.
- *Real-time notification and intervention.* An integrated web platform is developed for the real-time notification of retail stores assistants and intervention with customers when emotions are negative or repeated places where visualy repeated.

The remainder of this paper is organized as follows. Section II briefly reviews works in the field of video analytics technology. Section III details our approach to the underlying problem, and presents a two level method based on negative emotion analysis and face landmarks. Section IV presents a web interface for the retail store assistant, where methods are validated with experimental results carried out in real context scenarios. Section V concludes the paper, providing some hints to future work.

II. Related Work

Automatic detection of human emotions is a complex problem that has been applied to several ordinary problems. Techniques addressing this problem spans several types of data sources. Faces' images are one of the most promising sources for data analytics related to the emotion detection problem and to the physical behaviour of customers.

Our work overlaps with previous research on automatic analysis of human behavior inside retail stores. In this context, several approaches have been studied, like hot zone analysis, automatic activity recognition and sentimental analysis.

A. Hot Zone Analysis

Hot zone analysis aims to identify the trajectory of customers within a store. Trajectory analysis unveils spots with more activity and reveal where customers spend their time. Human's head position estimation was explored to create the initial estimates for tracking algorithms. Zhao et al. [8] presented a method for the detection and tracking of several humans in video frames. They propose boundary and shape analysis for human detection. On top of that, a 3D walking model predicts motion templates from the captured frames to track humans. This work was later improved by Zao and Ram [9], through the inclusion of a detection technique for human identification using Markov chain Monte Carlo methods. The method was tested in indoor and outdoor high-density scenes. In the outdoor scenes, false positives appear at far ends and dense edges. In the indoor scenes, the subtraction method gives erroneous foreground blobs. For human segmentation in both scenes, 1000 iterations are necessary to segment human objects. Leykinv and Mihran [10] developed a method where the human head coordinates are extracted from video frames to determine the position of customers in a store. These coordinates are further used to track customers in video sequences captured in crowded environments. The low-level extraction of the customers in a frame and the use of camera calibration to locate customer's head and location in the picture allows them to infer their location in the store.

B. Activity Recognition

The activity recognition is related to the shop behaviour and represents the actions of customers when buying products. Monitoring this behaviour is of most importance to academics and retail stores. Popa et al. [6] analyzed customer behaviour using background subtraction form images. This approach allowed them to detect customers in the entry point and then track them in the system. In [7], Popa et al. improved the method for automatic assessment of customer' appreciation of products. First, they classified customer behaviour by participant observation. Next, they implemented a model for motion detection, trajectory analysis, and face location and tracking for different customers. Sicre and Nicolas [11] resorted to behaviour models for detection of motion, tracking moving objects, and describing local motion. Results have shown that the approach can correctly classify 73% of the frames, for sequences taken in real environments. Later, Frontoni et al. [12] proposed a method to analyze human behavior in shops in order to increase consumer satisfaction and purchases. In their method, they use vertical red, green and blue depth sensors for people counting and shelf interaction analysis. Their results exhibited areas with both positive and negative interactions with products in shelves. They compared their results with ground truth visually recorded, and accuracy varies between 97.2% and 98.5%. Hu et al. [13] investigated the detection of semantic human actions in complex scenes. Their work deals with spatial-temporal ambiguities in frames using bag of instances representing the candidate regions of individual actions. A technique based on the combination of Simulated Annealing and Support Vector Machines has shown better results than standard Support Vector Machines.

C. Sentiment Analysis in Videos

Sentiment analysis is another area of video analytics. This type of problem is related to the problem addressed in this paper, since it acquires the emotional level of the customer. Zadeh et al. [14] addressed this problem using a multimodal dictionary that exploits jointly words and gestures. The approach has shown better results than straightforward visual and verbal analysis. An alternative approach to methods that adopt bag of words representations and average facial expression intensities is presented by Chen et al. [15]. They propose sentiment prediction using a time-dependent recurrent approach that performs fusion of several modalities (e.g., verbal, acoustic and visual) at every time-step. The implementation of the approach using long short-term memory networks has shown significant

improvements over several other approaches. Wang and Li [16] explored sentiment analysis in social media images. The main challenge of the work lies in the semantic gap between visual features and underlying sentiments. Contextual information is proposed to overcome the semantic gap in prediction of image sentiments. The solution was shown effective when evaluated with two large-scale datasets.

III. Approach

The approach presented in this paper is based on a machine learning system that runs in background for the intervention of retail store assistants with costumers. It focuses on the analysis of information obtained from a facial recognition system at two levels: emotion analysis and face landmarks. At the emotional level, when negative emotions are detected, the retail store assistant is notified for customer intervention. At the face landmark level, when a costumer is detected to be looking at the same place several times, the intervention is triggered.

A. Problem Statement

The study of human behavior in retail stores has been carried out in the last years, and it can be interpreted by analyzing the human emotional responses to contexts [17]. Moreover, the tracking of the position of the customer face can also be used to detect patterns of customer behaviour in retail stores.

Figure 1 presents a general view of the specification of the problem at the emotional and physical levels. The example assumes that a costumer is buying a book in a bookstore and is trying to find it in a shelf. A typical behaviour consists on eye motion between books and validation if the book cover is the one that he/she is looking for. This can be represented by emotions that can be positive or negative representing the customer state of mind when looking for a product (represented by the sequential arrows). Typical physical behaviour is also related to the repetition of a position in situations that the product is not being found (the gray circle represents the moment a customer looks more than once to the same place). If these one of these two characteristics are detected, it implies that a costumer is not finding a product and a retail store assistant can go to the customer for assistant.

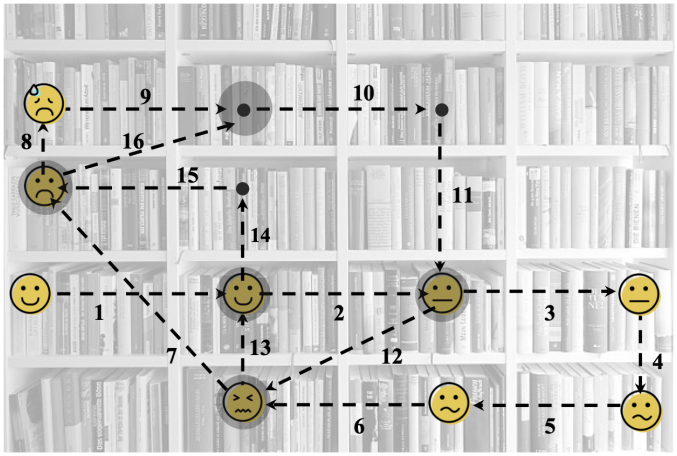


Figure 1. Problem specification for emotion and physical analysis.

At the emotional level, one of the problems that currently exist in customer service is trying to understand their state of mind when inside a store. For that purpose, the detection of emotions from customers will be able to increase the quality of service - the more relevant information about the customer, the better the assistance. The measurement of emotions can be carried out by several applications that are available in the market. These emotions can be either negative or positive. This work aims at the detection of negative emotions in a time window, where sadness is one of the most significant negative emotion to consider. However, manifestation of negative emotions can also be measured using other parameters like anger, disgust, or fear. In this paper we explore the combination of several negative emotions to determine a sadness level, β, used for costumer intervention.

Moreover, at the physical level, this work aims at providing a tool to explore face landmarks that are detected within a repeated context of interaction. Here, when a customer is not finding a product, it is normal that he/she looks around or starts to make random movements, which are indicators of uncertainty. This type of head movement can be captured using facial recognition software and can be used to detect if a costumer is looking at the same products he/she looked before, which may indicate the need for customer intervention, by determining a recurrent level, ρ.

Thus, tracking negative emotions and physical characteristics of faces in the context of a store are open problems, which is of most importance to be solved since they serve the automation of customer-employee contexts, resulting in an increase of the speed of attendance, improve customer satisfaction and increase retail stores sales.

B. Machine Learning Implementation

The performance of machine learning models is deeply dependent on the volume of data available for training models. For that reason, the most accurate models are provided by giants of software that have access to large volumes of data for training models capable of accurate detection of emotions in images. Fortunately, these models are widely available through an Internet accessible API like the IBM Watson [18], Face API [19], Kairos [20], and Amazon Rekognition [21].

In this work, we use Face API [19]. It is a cognitive service developed by Microsoft that provides algorithms to detect, recognize, and analyze human faces in images. Face API features are obtained in two stages: the first is the detection and recognition of face attributes; in the second, a JSON file is returned with the fields that contain face attributes.

Let $\mathcal{C} = \{c_j\}$, $j = 1 \ldots M$, be the number of customers that are detected in the system and $\mathcal{F} = \{f_i\}$, $i = 1 \ldots N$, the number of frames captured in real-time using the Face API for each customer $c_j \in \mathcal{C}$. The detection stage represents the analysis of the existing faces, $\mathcal{F}$, of customers, $\mathcal{C}$, and returns attributes for each $\{f_i\}$. When $\{f_i\}$ is detected, the face rectangle attribute is returned, since it contains the pixels to track $\{f_i\}$ in the image and gets its bounding box.

Within this bounding box, other attributes are returned by the API to the JSON file, namely, face Id, face landmarks, age, emotion, gender, and hair. In this paper all the parameters

TABLE I. USER TESTING IN REAL SCENARIOS: ACTING NORMAL, SIMULATION, FORCE SADNESS, FORCE ANGER AND FORCE HAPPINESS.

Anger (A_p)	Contempt (C_p)	Disgust (D_p)	Fear (F_p)	Happiness (H_p)	Neutral (N_p)	Sadness (S_p)	Surprise (Su_p)	Testing
0	0.001	0	0	0	0.999	0	0	*acting normal*
0.001	0.001	0	0	0	0.985	0.014	0	*simulate scenario*
0	0.002	0	0	0	0.762	0.235	0	*force sadness*
0.004	0.005	0.005	0	0.001	0.962	0.022	0	*sumulate scenario*
0.005	0.002	0.001	0	0.001	0.731	0.261	0	*force sadness*
0	0.002	0	0	0	0.993	0.005	0	*acting normal*
0	0	0	0	1	0	0	0	*force happiness*
0	0.016	0	0	0	0.811	0.172	0	*force sadness*
0.031	0.001	0	0	0	0.967	0.001	0	*simulate scenario*
0.035	0.001	0	0	0	0.966	0.001	0	*force anger*
0	0	0	0	0	0.977	0.023	0	*simulate scenario*
0	0.001	0	0	0	0.905	0.094	0	*simulate scenario*
0	0	0	0	0	0.958	0.041	0	*force sadness*
0	0.089	0.001	0	0	0.58	0.33	0	*force sadness*
0.001	0.027	0	0	0	0.967	0.004	0	*acting normal*
0	0.152	0	0	0.848	0	0	0	*force happiness*
0.172	0.002	0	0	0	0.823	0.003	0	*force sadness*
0.011	0.006	0	0	0	0.962	0.021	0	*simulate scenario*
0.008	0.37	0	0	0	0.621	0.001	0	*force anger*
0.16	0.043	0.001	0	0.001	0.661	0.134	0	*simulate scenario*
0.001	0.025	0	0	0	0.967	0.007	0	*simulate scenario*
0	0.169	0	0	0.009	0.821	0	0	*force sadness*
0.0058	0.011	0	0	0	0.887	0.043	0	*force sadness*
0	0.004	0	0	0.006	0.987	0.004	0	*acting normal*
0	0.001	0	0	0.958	0.04	0.002	0	*force happiness*
0	0	0	0	0	0.857	0.143	0	*force sadness*
0	0	0	0	0	0.84	0.159	0	*simulate scenario*
0.412	0.042	0.09	0.029	0.006	0.57	0.001	0.363	*force anger*
0.001	0.007	0	0	0.001	0.94	0.051	0	*simulate scenario*
0	0.005	0	0	0.038	0.955	0.001	0	*simulate scenario*
0	0.001	0	0	0	0.958	0.041	0	*force sadness*
0	0.001	0	0	0	0.417	0.582	0	*force sadness*
0	0	0	0	0	0.997	0.002	0	*acting normal*
0	0	0	0	1	0	0	0	*force happiness*
0	0	0	0	0	0.965	0.035	0	*force sadness*
0.053	0.004	0	0	0	0.943	0	0	*simulate scenario*
0.127	0.009	0	0	0	0.864	0	0	*force anger*
0	0.0087	0	0	0.036	0.868	0.002	0.007	*simulate scenario*
0	0.001	0	0.001	0.001	0.956	0.033	0.009	*simulate scenario*
0	0.003	0	0	0	0.679	0.318	0	*force sadness*
0	0	0	0	0	0.887	0.113	0	*force sadness*

are considered in three contexts. First, for a general characterization of the costumer, age, gender, and hair attributes are used. These attributes allow the retail store employee to better identify the customer (note that for security policies, the system cannot store the face of the customer). Next, for the emotion analysis (cf., Section III-C), the emotion attribute, containing a set of different emotions, is used to detect negative emotions. Finally, at the facial level (cf., Section III-D), face landmarks are used to track repetition of previously visited positions.

The parameters returned by the Face API are a basis of knowledge for the implementation of the emotion and facial tracking methods presented in the following sections.

C. Emotion Analysis

There are several parameters associated to emotions that are returned by facial recognition systems, namely anger (A_p), contempt (C_p), disgust (D_p), fear (F_p), happiness (H_p), neutral (N_p), sadness (S_p) and surprise (Su_p). In the scope of this work, we only consider negative emotions (A_p, D_p, F_p and S_p) that affect the costumer interaction with the system.

The basic idea of our method is presented in Figure 1 (which includes both representations of emotions and physical motion). When a customer arrives at a shelf, Face API captures his emotions, and a sadness level β is set to zero. This factor updates in the presence of negative emotions, and once a threshold is passed ($\beta > 50\%$), the assistant is asked to go to the customer. Negative emotions manifest in several ways, and one of the most critical parameters is the sadness parameter, $S_p \in [0..1]$ (values near 1 correspond to the total manifestation of sadness). Therefore, once a frame captures a customer with a high value of sadness, it may be an indicator of a potential product not being found by a customer. Other parameters like A_p, D_p or F_p are also present in negative emotions, and their contribution is analyzed in this paper.

To determine the weights to consider in each of the negative emotions, an empiric study (presented in Table I) was carried out with users that were asked to express several emotions: S_p, N_p, D_p, H_p and simulate the action of looking for a product and not finding it, referred to as *Simulated*. In the emotion tests considering H_p and N_p, these parameters have high values, representative of the tested emotion. In the tests for forced sadness and simulation, S_p has low values in most cases, which is justified by the fact that the sadness emotion can result in false positives. However, in this case, the presence of other

negative emotions is visible, with small values of A_p, D_p and F_p. Analyzing the impact of these parameters in the emotion is an essential factor to determine how to infer sadness when S_p should be naturally present and is not.

In this context, two types of tests were carried out: first, the evaluation of the impact of each negative emotion and, second, the presence of all negative emotions. In the first test, results obtained ($A_p = 47\%$, $D_p = 16\%$, $F_p = 6\%$ and $S_p = 91\%$), show that negative emotion is present in the tests. However, excluding S_p, the other negative emotions are not feasible to be used individually to complement the sadness test, since they are present in a small number of tests, which are not representative of the sample. In the second test, was considered the cumulative presence of all negative emotion parameters ($A_p + D_p + F_p + S_p > tol$) for the same scenario (forced sadness and simulation), as shown in Table II.

TABLE II. TOLERANCE TESTS FOR $A_p + D_p + F_p + S_p > tol$

	Tolerance (tol)				
	0.0	0.01	0.02	0.03	0.04
Cumulative negative emotions (%)	97.22	83.73	80.16	74.32	68.26

Results show that when $tol = 0.0$, 97.22% of the tests reveal the presence of cumulative negative emotions, which is very representative of the tested scenario. The rate decreases for $tol \geq 0.01$. Therefore, when S_p is not representative in a sadness test, the alternative of considering cumulative negative emotions has success rate of 97.22%. Recall that these criteria are used only to improve the success rate of retail store assistants interventions and are used in two contexts: in the evident presence of sadness (high values of S_p) and in the presence of signs of sadness ($A_p + D_p + F_p + S_p > tol$, for low values of S_p). The resulting method is presented in Algorithm 1.

Algorithm 1: Emotion-based intervention method

```
Data: C                          ◁ detected customers C = {c_j}
Data: F                          ◁ API frames F = {f_i}
Result: β, I    ◁ β = sadness level, I = Intervention
begin
    foreach c_j ∈ C do
        β_j ← 0.0                ◁ set sadness level to zero
        I ← false                ◁ no intervention required
        foreach f_i ∈ F do
            A_{p_i} ← A_p ∈ f_i      ◁ get anger from f_i
            F_{p_i} ← F_p ∈ f_i      ◁ get fear from f_i
            S_{p_i} ← S_p ∈ f_i      ◁ get sadness from f_i
            D_{p_i} ← D_p ∈ f_i      ◁ get disgust from f_i
            if (S_{p_i} > 0.5) then
                β_j ← β + 0.1        ◁ update sadness level
            else if (A_{p_i} + F_{p_i} + S_{p_i} + D_{p_i} > 0) then
                β_j ← β + 0.05       ◁ update sadness level
            if (β_j > 0.5) then
                I ← true             ◁ intervention required
        end
    end
end
```

The algorithm starts by scanning if a customer is detected by the Face API and its faceId is generated. The sadness level of each customer, β_j, is set to zero, and frames are captured while the customer is detected in the system. For every captured frame, the Face API returns negative emotion values that are stored for processing. Every time the algorithm captures evidence of sadness ($S_{p_i} > 0.5$ or signs of sadness ($A_{p_i} + F_{p_i} + S_{p_i} + D_{p_i} > 0$), the value of β_j is updated in a factor of 0.1 or 0.05, respectively. When the sadness level passes a threshold of 0.5, the assistant is informed that a customer needs intervention.

An important consideration is that our system does not retain personal information of a customer. After detection by a camera, only a faceId is generated to uniquely identify the characteristics of that customer. If he/she leaves the system, the method still continues to try to track the faceId of the customer for five minutes. After that period, the information of the faceId is removed from the database, but the face attributes are kept. With this, personal information of users is not stored, therefore, it does not allow the system to track a specific customer. If the customer is again detected in the system, he/she will be assigned a new faceId.

D. Physical Motion Analysis

At the physical level, a human face is composed of sets of points that can be well identified. These points, called face landmarks go from pupils to the tip of the nose. Face landmark detection is a computer vision technique developed automatically detect some particular landmarks in human faces using machine learning algorithms. The accurate identification of facial landmarks is a process by which a number of complicated image analysis problems are solved. This identification has been extended outside the domain of image research and into other applications, such as the medical field [22][23], animation [24][25], face reconstruction [26] and security [27]. In [28] and [29], a complete review of facial landmark identification techniques is presented.

Face API features 27 predefined landmark points in a face describing physical characteristics of a face: eyebrows, eyes, nose and mouth. In this paper we explore the landmarks associated to the nose, more concretely, the nose tip. In the Face API, the nose tip is captured in (x, y) coordinates, which is important to detect motion in a captured frame. In this context, the method presented in Algorithm 2 detects if a customer is looking at a (x^*, y^*) point in the neighborhood of a previous point (x, y) captured in a previous frame (see Figure 1). If that occurs, the likelihood of a product not being found increases.

As in Section III-C, once a costumer is detected, a faceId is generated, and the recurrent level, ρ, is set to zero. As a customer looks for products, new frames are captured and nose tip coordinates are determined. In this context, let f_j be the present frame and (x_j^*, y_j^*) the the nose tip corresponding coordinates. For each capture frame, if it is in the neighbourhood of a previous face ($|x_a^* - x_p| < tol$ and $|y_a^* - y_p| < tol$), then it is assumed that the customer is looking at the same place. This increases ρ in 0.01 until $\rho > 0.5$, and the retail store assistant receives notification for intervention.

Algorithm 2: Physical-based intervention method

Data: $\mathcal{C}$ ◁ detected customers $\mathcal{C} = \{c_j\}$
Data: $\mathcal{F}$ ◁ API frames $\mathcal{F} = \{f_i\}$
Result: ρ, $\mathcal{I}$ ◁ ρ = recurrent level, $\mathcal{I}$ = Intervention

begin
  foreach $c_j \in \mathcal{C}$ **do**
    $\rho_j \leftarrow 0.0$ ◁ set recurrent level to zero
    $\mathcal{I} \leftarrow false$ ◁ no intervention required
    $f_j \leftarrow$ Get Current Frame
    $N_{x^*} \leftarrow N_x \in f_j$ ◁ get nose tip x coordinate from f_j
    $N_{y^*} \leftarrow N_y \in f_j$ ◁ get nose tip y coordinate from f_j
    foreach $f_i \in \mathcal{F}$ **do**
      $N_x \leftarrow N_y \in f_i$ ◁ get nose tip x coordinate from f_i
      $N_y \leftarrow N_y \in f_i$ ◁ get nose tip y coordinate from f_i
      if $(|N_{x^*} - N_x| < tol$ *and*
      $|N_{y^*} - N_y| < tol)$ **then**
        $\rho_j \leftarrow \rho_j + 0.1$ ◁ update recurrent level
      if $(\rho_j > 0.5)$ **then**
        $\mathcal{I} \leftarrow true$ ◁ intervention required
    end
  end
end

IV. Experimental Design and Results

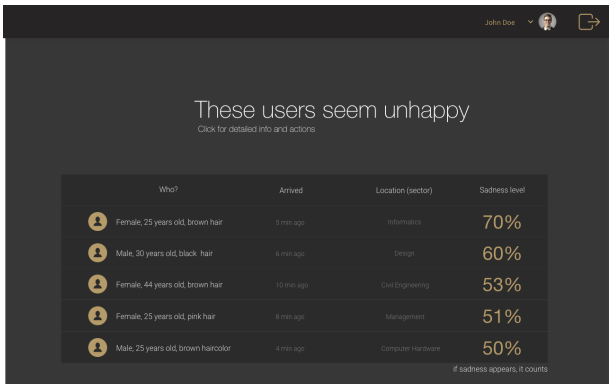
The algorithm presented in the previous section runs in background and processes information that can be visualized by the retail store assistant in an web application (see Figure 2).

A. Web Interface

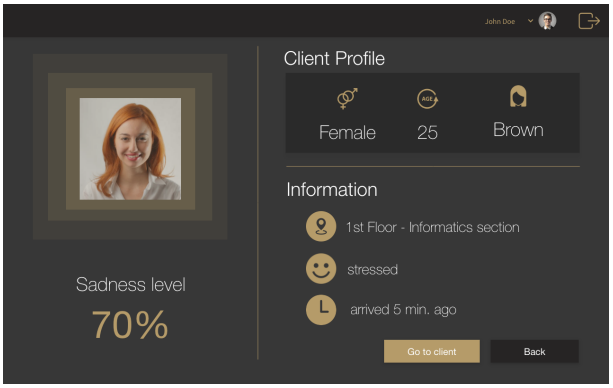
The design and implementation of a web interface for the retail store assistant was of must importance to carry out a pilot study. The assistant has access to the notifications management page, presented in Figure 2a). This page is updated in real time and contains a list of customers that require intervention. Here, some general information of the customers is provided for better identification.

When the assistant selects a customer for intervention, Algorithm 1 and Algorithm 2 (running in background) stop increasing β and ρ for that customer, respectively, and these values are stored. Otherwise, they would reach the threshold value for all customers in the time elapsed between the interaction and the time to go to the customers. When the assistant selects a customer, general information is provided (such as hair color, age, gender, location in store, the emotion revealed by the customer and how long the customer is in the system). In addition, the assistant has the possibility to attend the customer or to cancel and return to the call management page as shown in Figure 2b) and Figure 2c).

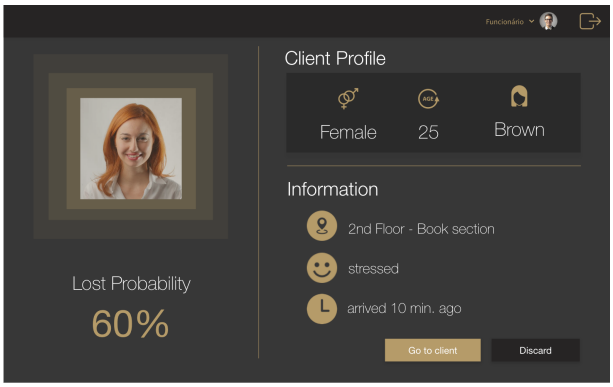
The intervention level starts when the assistant clicks in the "go to client" button and the page changes so that feedback data can be provided by the assistant, which possesses relevant information regarding the intervention with the customer, as shown in Figure 2d). It is important to note that while the



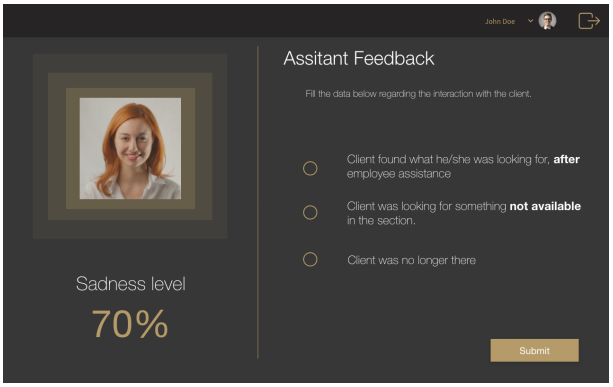
(a) List of customers for assistant intervention.



(b) User info and emotions.



(c) User info and face landmarks.



(d) Assistant feedback.

Figure 2. Web interface for retail store assistant intervention.

assistant is attending the customer, no further changes in the customer emotions are captured. It is intended to capture the emotions that have caused the customer to exceed the emotional threshold and not to register emotion changes while being under intervention.

B. Results

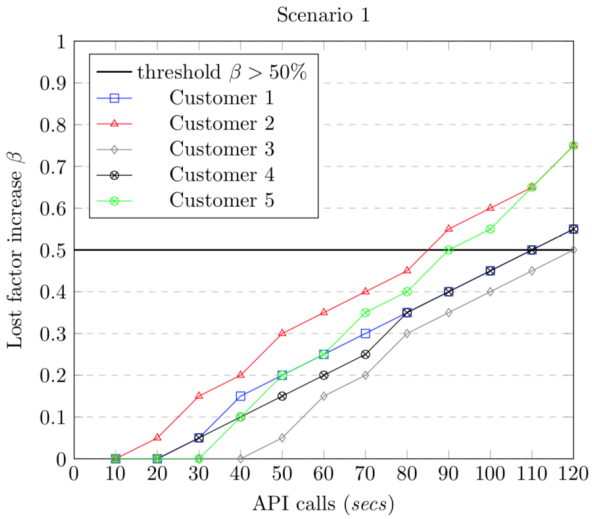
We have tested the approach described in Section III by carrying out a pilot study at both the emotional and physical levels. Books were placed in shelves with a camera placed to capture emotions and face landmarks. Five customers where asked to find a book, from twenty available books, in three scenarios:

- Scenario 1: The book is not available in the products placed in the shelves.
- Scenario 2: The book is in the shelves, but very similar to other books, making it difficult to be found.
- Scenario 3: The book is available in the shelves and easy to be identified.

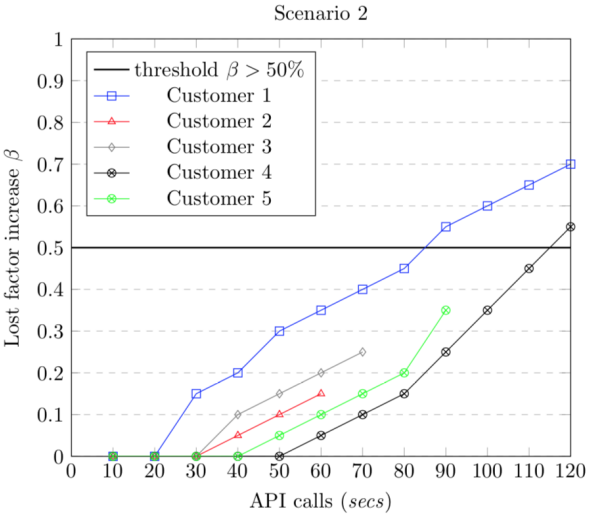
Results obtained are presented in Figure 3 and Figure 4, which refer to the emotion detection and face landmark detection, respectively. To provide flexibility to the system, the assistant can decide the moment of the intervention. As previously referred, when the sadness level threshold is passed, the assistant web page is updated with the customer information. However, if the assistant considers that the sadness level is not increasing with time, he/she can decide not to go to the customer. However, if the customer continues to reveal cumulative negative emotions or head motion is present, the assistant then makes the decision to assist him. Moreover, if all assistants are occupied, the system continues to increase the lost levels of a customer, until an assistant is available.

At the emotional level, for scenario 1 (Figure 3a)), costumers reveal signs of cumulative unhappiness, $(A_{p_i} + F_{p_i} + S_{p_i} + D_{p_i} > 0)$, or sadness $(S_p > 50\%)$ as they realize that they are not finding the product. The sadness level threshold is passed for all customers after a few iterations of API calls. The variation of the sadness level cumulative response is due to the fact that, in the API calls, the customer can reveal one of both negative emotions tested. This implies that there can be an increase of 0.05 or 0.1, depending on the most prevalent negative emotion in each detection. In this context, the web interface for the assistant is updated with the data related to the new customer that requires intervention (see Figure 3a)). For all the customers of the tested scenarios, the assistant reported option two in the feedback page (see Figure 2d)). In scenario 2, three customers found the product, after some iterations and left the system. The other two reached the sadness level threshold. For them, the assistant reported option one in the feedback page. Finally, in scenario 3, all the customers found the product after a few iterations of API calls, never reaching the sadness level threshold, thus, not requiring assistant intervention.

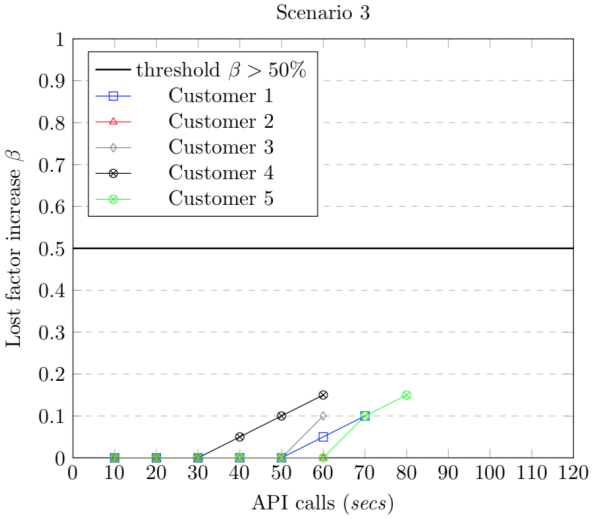
At the physical level, the same scenarios were considered and different customers were asked to carry out the study. Figure 4 presents the results obtained by applying Algorithm 2, and similar results were obtained, when compared to the emotional tests. However, more API calls were required in



(a) Book is not available in shelves.

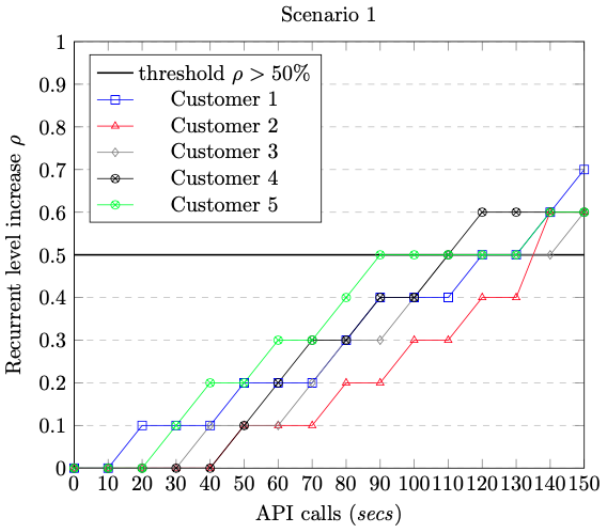


(b) Book is similar to other products.

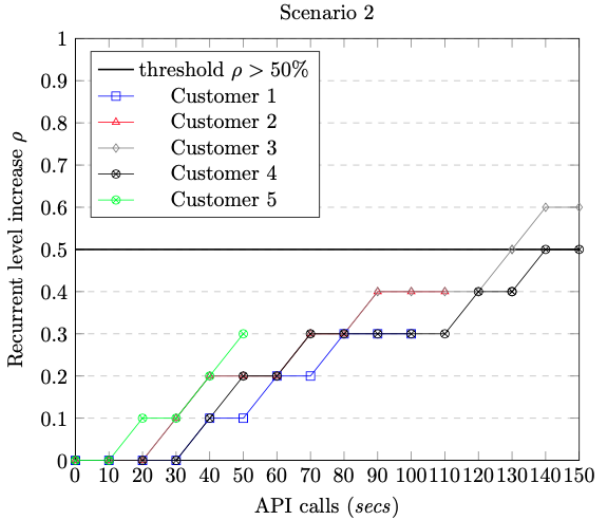


(c) Book is well identified in shelves.

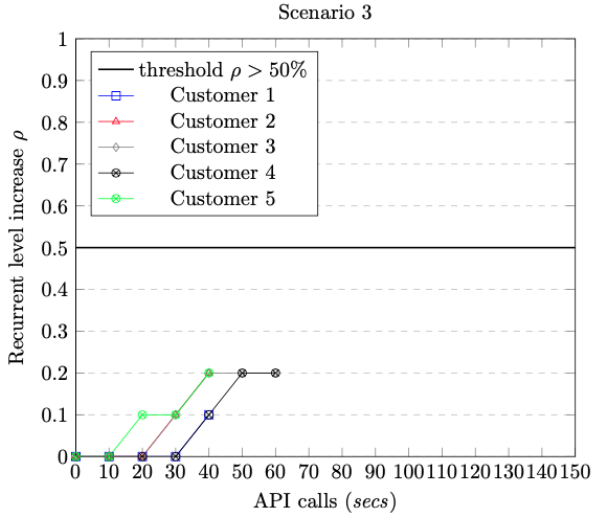
Figure 3. Results obtained for emotion tests with five customers in the three scenarios.



(a) Book is not available in shelves.

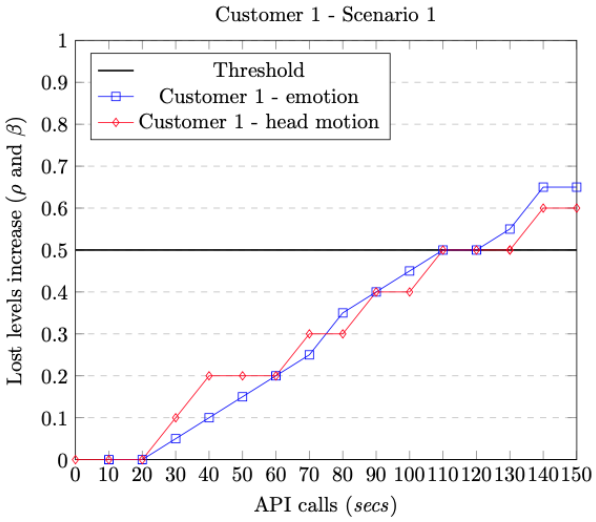


(b) Book is similar to other products.

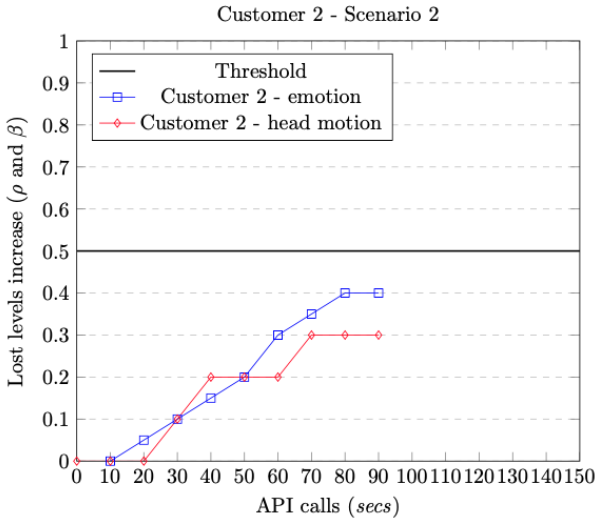


(c) Book is well identified in shelves.

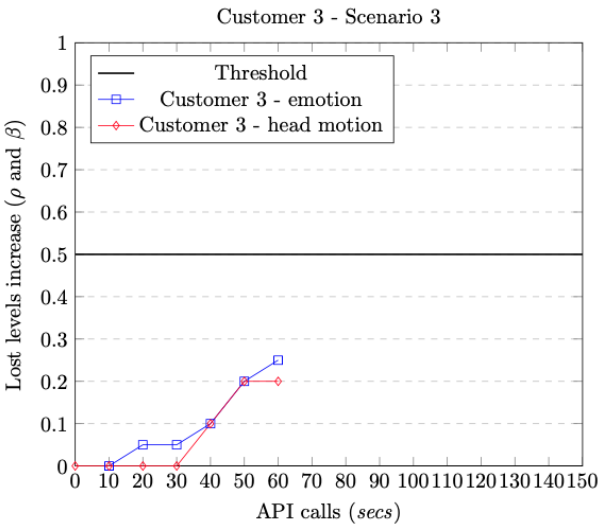
Figure 4. Results obtained for the detection of face position for five customers in the three scenarios.



(a) Book is not available in shelves.



(b) Book is similar to other products.



(c) Book is well identified in shelves.

Figure 5. Emotion and face analysis for customers in the three scenarios.

some tests to achieve customer intervention. In the context of Scenario 1, depicted in Figure 4a), after some iterations, customers started to look at previously visited places, which increased the recurrent level, ρ, and forced it to pass the threshold, which implied in customer intervention. In the other scenarios (Figures 4b) and 4c)), results show that most customers found the book. Results show that the method detects if a customer is not finding a product.

A final test was carried out with three customers to compare the accuracy of each method. Each customer was asked to find a book in the context of the defined scenarios. Results are presented in Figure 5 and reveal a correlation between emotion analysis and the tracking of previously visited places, for all the tested scenarios.

V. Conclusion and Future Work

This paper presented two novel scalable methods based on visual recognition of customer emotions and face landmarks when buying products, using Face API. The method uses a camera to capture the manifestation of negative emotions at two levels: the effective manifestation of sadness and evidence of sadness, in a set of frames. Concurrently, the method detects if the customer is looking at previously visited places, by extracting face landmarks. The evaluation methodology shows that both methods present good results in real scenarios. Additionally, the implementation of an intuitive web interface allows retail shops assistants to carry out interventions with customers, if the emotional and recurrent thresholds are passed. This interface will greatly assist retail stores to have an understanding of which customers require intervention and provide the necessary help in real-time. The natural implications are an increase in sales and customer satisfaction.

Future work will follow two directions, mostly focused on Artificial Intelligence (AI). A first approach will use to anticipate the needs of customers based on the previous emotional analysis. This will allow retail stores to determine which products are not being found and reorganize stores in order to better allow the correct identification of products. Moreover, the lost levels (sadness and recurrent) were obtained empirically. It will be essential to use AI as a mean to adjust these parameters.

Acknowledgments

This work is funded by National Funds through the FCT - Foundation for Science and Technology, I.P., within the scope of the project Ref. UIDB/05583/2020. Research Centre in Digital Services (CISeD) and the Polytechnic of Viseu.

References

[1] V. Borges, R. P. Duarte, C. A. Cunha, and D. Mota, "Are you lost? using facial recognition to detect customer emotions in retail stores," *CENTRIC 2019 : The Twelfth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services*, p. 49–54, Nov. 2019. [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid=centric_2019_3_30_30031

[2] M. H. Moss, *Shopping as an entertainment experience*. Lexington Books, 2007.

[3] R. L. Oliver, "Whence consumer loyalty?" *Journal of Marketing*, vol. 63, no. 4, pp. 33–44, 1999, ISSN: 00222429.

[4] R. M. Bolle, J. H. Connell, N. Haas, R. Mohan, and G. Taubin, "Veggievision: A produce recognition system," in *Proceedings Third IEEE Workshop on Applications of Computer Vision (WACV'96)*. IEEE, Dec. 1996, pp. 244–251, ISBN: 0-8186-7620-5.

[5] J. Connell, Q. Fan, P. Gabbur, N. Haas, S. Pankanti, and H. Trinh, "Retail video analytics: an overview and survey," in *Video Surveillance and Transportation Imaging Applications*, vol. 8663. International Society for Optics and Photonics, 2013, p. 86630X.

[6] M. Popa, L. Rothkrantz, Z. Yang, P. Wiggers, R. Braspenning, and C. Shan, "Analysis of shopping behavior based on surveillance system," in *2010 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, Oct. 2010, pp. 2512–2519, ISBN: 978-1-4244-6588-0.

[7] M. C. Popa, L. Rothkrantz, C. Shan, T. Gritti, and P. Wiggers, "Semantic assessment of shopping behavior using trajectories, shopping related actions, and context information," *Pattern Recognition Letters*, vol. 34, no. 7, pp. 809–819, May 2013, ISSN: 0167-8655.

[8] T. Zhao, R. Nevatia, and F. Lv, "Segmentation and tracking of multiple humans in complex situations," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, vol. 2. IEEE, Dec. 2001, pp. II–II, ISBN: 0-7695-1272-0.

[9] T. Zhao and R. Nevatia, "Stochastic human segmentation from a static camera," in *Workshop on Motion and Video Computing. Proceedings.* IEEE, Dec. 2002, pp. 9–14, ISBN: 0-7695-1860-5.

[10] A. Leykin and M. Tuceryan, "A vision system for automated customer tracking for marketing analysis: Low level feature extraction," in *Human Activity Recognition and Modelling Workshop*, vol. 3. Citeseer, 2005, pp. 6–13.

[11] R. Sicre and H. Nicolas, "Human behaviour analysis and event recognition at a point of sale," in *2010 Fourth Pacific-Rim Symposium on Image and Video Technology*. IEEE, Nov. 2010, pp. 127–132, ISBN: 978-1-4244-8890-2.

[12] E. Frontoni, P. Raspa, A. Mancini, P. Zingaretti, and V. Placidi, "Customers' activity recognition in intelligent retail environments," in *New Trends in Image Analysis and Processing (ICIAP 2013)*. Springer Berlin Heidelberg, 2013, pp. 509–516, ISBN: 978-3-642-41190-8.

[13] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang, "Action detection in complex scenes with spatial and temporal ambiguities," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 128–135, ISBN: 978-1-4244-4420-5.

[14] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, Nov. 2016, ISSN: 1941-1294.

[15] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI '17)*. New York, NY, USA: ACM, 2017, pp. 163–171, ISBN: 978-1-4503-5543-8.

[16] Y. Wang and B. Li, "Sentiment analysis for social media images," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2015, pp. 1584–1591, ISBN: 978-1-4673-8493-3.

[17] R. Ekman, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[18] "Ibm watson - visual recognition," URL: https://www.ibm.com/watson/services/visual-recognition/ [accessed: 2020-11-10].

[19] "Microsoft cognitive services: Face api," 2019, URL: https://azure.microsoft.com/en-us/services/cognitive-services/face/ [accessed: 2020-11-10].

[20] "Kairos apis and sdks," 2019, URL: https://www.kairos.com/ [accessed: 2020-11-10].

[21] "Amazon rekognition - video and image," URL: https://aws.amazon.com/rekognition [accessed: 2020-11-10].

[22] D. L. Guarin, J. Dusseldorp, T. A. Hadlock, and N. Jowett, "A machine learning approach for automated facial measurements in facial palsy," *JAMA facial plastic surgery*, vol. 20, no. 4, pp. 335–337, 2018.

[23] A. T. Balaei, K. Sutherland, P. A. Cistulli, and P. de Chazal, "Automatic detection of obstructive sleep apnea using facial images," in *2017 IEEE*

14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, 2017, pp. 215–218.

[24] K. Liu, A. Weissenfeld, J. Ostermann, and X. Luo, "Robust aam building for morphing in an image-based facial animation system," in *2008 IEEE International Conference on Multimedia and Expo*. IEEE, 2008, pp. 933–936.

[25] S. Ioannou, G. Caridakis, K. Karpouzis, and S. Kollias, "Robust feature detection for facial expression recognition," *Journal on image and video processing*, vol. 2007, no. 2, pp. 5–5, 2007.

[26] U. Park and A. K. Jain, "3d face reconstruction from stereo video," in *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*. IEEE, 2006, pp. 41–41.

[27] C. Pradhan, D. Banerjee, N. Nandy, and U. Biswas, "Generating digital signature using facial landmlark detection," in *2019 International Conference on Communication and Signal Processing (ICCSP)*, Apr. 2019, pp. 0180–0184.

[28] B. Johnston and P. de Chazal, "A review of image-based automatic facial landmark identification techniques," *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, p. 86, 2018.

[29] O. Celiktutan, S. Ulukaya, and B. Sankur, "A comparative study of face landmarking techniques," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 13, 2013.