https://doi.org/10.31436/iiumej.v20i2.1000

UNWEARABLE MULTI-MODAL GESTURES RECOGNITION SYSTEM FOR INTERACTION WITH MOBILE DEVICES IN UNEXPECTED SITUATIONS

HANENE ELLEUCH[1]*, ALI WALI[1], ANIS SAMET[2], ADEL M. ALIMI[1]

[1] *REsearch Groups in Intelligent Machines Lab (REGIM-Lab), University of Sfax, National Engineering School of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia*

[2]*SiFAST: Software and Computing Services Company, Sfax 3003, Tunisia*

**Corresponding author: Hanene.elleuch@ieee.org*

(Received: 26th September 2018; Accepted: 14th August 2019; Published on-line:2nd December 2019)

ABSTRACT: In this paper, a novel real-time system to control mobile devices, in unexpected situations like driving, cooking and practicing sports, based on eyes and hand gestures is proposed. The originality of the proposed system is that it uses a real-time video streaming captured by the front-facing camera of the device. To this end, three principal modules are charged to recognize eyes gestures, hand gestures and the fusion of these motions. Four contributions are presented in this paper. First, the proposition of the fuzzy inference system in the purpose of determination of eyes gestures. Second, a new database has been collected that is used in the classification of open and closed hand gesture. Third, two descriptors have been combined to have boosted classifiers that can detect hands gestures based on Adaboost detector. Fourth, the eyes and hand gestures are erged to command the mobile devices based on the decision tree classifier. Different experiments were assessed to show that the proposed system is efficient and competitive with other existing systems by achieving a recall of 76.53%, 98 % and 99% for eyes gesture recognition, detection of fist gesture, detection of palm gesture respectively and a success rate of 88% for eyes and hands gestures correlation.

ABSTRAK: Kajian ini mencadangkan satu sistem masa nyata bagi mengawal peranti mudah alih, dalam keadaan tak terjangka seperti sedang memandu, memasak dan bersukan, berdasarkan gerakan mata dan tangan. Kelainan sistem yang dicadangkan ini adalah ia menggunakan masa nyata video yang diambil daripada peranti kamera hadapan. Oleh itu, tiga modul utama ini telah ditugaskan bagi mengenal pasti isyarat mata, tangan dan gabungan kedua-dua gerakan. Empat sumbangan telah dibentangkan dalam kajian ini. Anggaran pertama bahawa isyarat gerak mata mempengaruhi sistem secara kabur. Kedua, pangkalan data baru telah dikumpulkan bagi pengelasan isyarat tangan terbuka dan tertutup. Ketiga, dua pemerihal data telah digabungkan bagi merangsangkan pengelasan yang dapat mengesan isyarat tangan berdasarkan pengesan Adaboost. Keempat, gerakan mata dan tangan telah digunakan bagi mengarah peranti mudah alih berdasarkan pengelasan carta keputusan. Eksperimen berbeza telah dijalankan bagi membuktikan bahawa sistem yang dicadang adalah berkesan dan berdaya saing dengan sistem sedia ada. Keputusan menunjukkan 76.53%, 98% dan 99% masing-masing telah dikesan pada pengesanan gerak isyarat mata, genggaman tangan dan tapak tangan, dengan kadar 88% berjaya mengesan gerak isyarat mata dan tangan.

KEYWORDS: *Eye gesture recognition, Hand gesture detection, Multimodal interaction, Fuzzy inference system, Human-computer interaction*

https://doi.org/10.31436/iiumej.v20i2.1000

1. INTRODUCTION

Nowadays, mobile devices are considered like a ubiquitous computer. No matter where the user is, he can access any information and use it to do anything that he/she wants; capturing photos, paying an invoice, accessing emails, etc. Today, the consumption of mobile devices is changing due to the continuous evolution of the hardware as well as the software. However, the question of how providing the most suitable mode of interaction with these devices has not been really solved. All these years, the mobile human-computer interaction is based on one way: the touchscreen. In fact, the touchscreen is simple, clear and intuitive. However, the limits of this mode are clearly visible when using mobile devices in unexpected situations such as moving, driving or using one hand while the other is busy [1], etc. Besides, the minimization of on-screen targets causes occlusions. To cope with these issues, many solutions are proposed to extend mobile human-computer interaction modalities by introducing natural ways of communications like eyes, hands, and voice. Eyes and hands are the most intuitive way of communication. Eyes are our first channel of communication with the surrounding environment. They play an important role in expressing needs and expressional emotions. Also, hands play an important role in non-verbal communication with the purpose of sending messages correctly. For the ubiquitous interaction with mobile devices, it is clear that eyes and hands have many advantages compared to voice because of the background noise that can affect the quality of communication, on the one hand, and on the other hand, gestures are independent of language and do not require any previous knowledge. Several solutions are presented to extend the ways of interaction with mobile devices. Some of these solutions focus on eye tracking by using EEG [2], EOG [3, 4] and Eye Tribe [5]. Other solutions focused on hand gestures recognition are based on gloves [6, 7]. All these solutions require the addition or modification of the device, which is considered as a barrier to adoption and generalization. In contrast, most mobile devices already contain cameras. For this reason, an image-based solution can be more efficient for eye tracking as well as for the hand gesture recognition at one and the same time.

In this paper, a new mobile human-computer interaction system based on eyes and hands' gestures was proposed for a medical application used by doctors to manipulate patients' information during a surgical intervention. The input of this system is a real-time video captured from the front-facing camera of the device. Two modules are responsible for each modality's gestures. The first module aims to recognize eyes gestures from the captured video. The user's face and his eyes are detected first and then the pupils are determined and tracked. A fuzzy inference system is deployed to recognize the pupils' motions. In this work, the main goal is detecting 8 motions in different directions: right, left, up, down, right-up, left-up, right-down, left-down. In addition to these movements, this module can also detect blinks. The second module aims to recognize hand gestures from the input video. Viola and Jones' algorithm is used to detect 2 gestures: fist and palm. To achieve this goal, a new dataset is collected and used for the training phase. A couple of descriptors are combined to improve the accuracy. The recognized gestures coming from different modules are combined and fused in a separate module to execute actions to monitor mobile devices. This system is evaluated by conducting a study on volunteers in real life conditions, the results obtained are promising.

The remainder of this paper is organized as follows: in section 2, the relevant work was detailed in studying the unimodal and multi-modal interaction systems. In section 3, the proposed system is presented and each module is detailed. In section 4, an assessment is conducted for the purpose of evaluating this system. In this section also, a description of the

experiments of each module separately and then the whole system. Finally, the paper is concluded and an example of application is suggested to prove the practical aspect of the proposed system.

2. RELATED WORK

Many research works introduced eyes and hands as alternative modalities to control devices. In this section, the related work was cited that included eyes and hands as the only modality and their combination in the human-computer interaction.

2.1. Natural user interface based on unimodal interaction systems

Eye-based interaction is deployed as an important input modality. Gaze estimation is used to determine the point-of-regard on the screen of mobile devices [8,9,10,11 ,12]. Wood and Bulling [8] combines cascade classifiers and a shape-based approach for eyes centers detection, limbus ellipse fitting and limbus back projection for gaze estimation. Pino and Kavasidis [9] used Haar classifier and CAMSHIFT to detect and track eyes respectively. The authors estimated the gaze by determining a correlation between the centroid of eyes rectangle and the user's gaze. [10] proposed a dataset for unconstrained gaze estimation for mobile devices. These works proved their robustness by achieving competitive results, however, they suffered from many limits like the execution time and the need of learning phase and calibration preprocessing. Eyes gestures are also used to interact with mobile devices. [13] proposed a new eyes gestures recognition system for mobile devices. The algorithm is based on image processing algorithms to detect and track the user's eyes from video captured from the device. A face detector based on Viola and Jones algorithm is responsible for detecting the face. Another module tracked the detected face. An eye detector detected the eyes from the face which have been tracked using an eye tracker. The eyes gesture recognition is assured by using template-matching. Other methods are also deployed. Miluzzo et *al.* [11] used contour detection to localize the eyes in their system presented EyePhone that combines blinks detection and gaze's estimation to command mobile devices. In [14], the authors proposed an eye gesture input interaction system for smartphones dedicated for people with ALS. This system recognizes gestures by using template-matching but a calibration step is necessary to use this system for the first time. In [15], the authors proposed a system of eyes gestures recognition. This system recognizes 8 gestures to communicate with mobile devices without any additional peripheries.

Hands gestures are widely used in the human-computer interface. Meng et *al.* [16] presented a new system of hand gesture recognition. The hand detection is assured by using skin color detection algorithm and static approach image to separate the hand from the background. The hand gesture recognition is assured by determining the fingers' number and the angle formed between the fingers. This step is based on hand contour detection and fingertips and palm center extraction. The authors proposed the deployment of cloud computing to solve the problem of the slow processing time. Prasuhn et *al.* [17] proposed a HOG-based hand gesture recognition system deployed on the mobile device and applied it to American Sign Language (ASL). A hand detection area based on the skin color algorithm is first applied. The hand shape features are computed by using the HOG descriptor [18]. The recognition step is assured by finding the best matching between the target image and images from the database. Song et *al.* [19] presented a new system of interaction with mobile devices based on hands gestures. The detection step is based on skin color technique and the recognition phase is assured by the random forest algorithm. The experiments showed promising results. In [20], the author proposed a new system of interaction with mobile devices based on static hands gestures. This system is based on skin color for hand

segmentation and SVM for the recognition phase. The hand detection in all these works is based on skin color detection. This technique was proved as a robust method; however, it causes many false positive detections and occlusions with the face region. The response time for [16] and [17] is relatively high. For this reason, [16] deploys a cloud server for the computation part to have more rapid execution time. [16] and [17] are the only systems developed for mobile devices. However, these systems did not treat the case of the face occlusion and the dynamic hand gesture recognition. [16] did not present the recognition results so the robustness of his method cannot be determined. On the other hand, [17] proved that the recognition rate of their system is only 47%. Besides, this system's algorithm is not totally developed on the device, but it is based on a client-server configuration that made the need of the internet indispensable for the operation process. In this paper, a real-time hand gesture recognition based on the camera was proposed. It is running entirely in the device and it can detect hands gesture without any confusion with user's face and any other body part.

2.2. Natural user interface based on multi-modal interaction systems

Many systems are proposed with different combinations of modalities. Gaze and hand gesture were the most popular. Many research works focused on the combination of these modalities in user-computer interaction.

Chatterjee et *al.* [21] used these modalities for the purpose of building an interactive mode with a computer, a series of LEDs and a mobile robot. Each modality is captured by a specific periphery: Eye Tribe Tracker for gaze tracking and Leap motion for recognizing hand gestures. The authors conducted a study to test the efficiency of these modalities applied differently in various scenarios. Hales et *al.* [22] presented a system of object manipulation through gaze and hand gestures. A specific glass assured gaze tracking and a vision-based module assured the recognition of hand gestures. The assessment of this system proved that the combination of these modalities is natural and promising. Pouke et *al.* [23] presented a model of interaction based on gaze and hand gesture for 3D virtual space on tablet devices. The gaze tracking is assured based on corneal reflexion technique and the hand gesture recognition is based on the accelerometer sensor and machine learning algorithms. The author proposed to use the gaze for selecting objects and hand gesture for the manipulation of these objects. Yoo et *al.* [24] evoked the issue of interaction with a large screen cannot be efficient with the traditional tools of interaction like mouse and keyboard cannot be efficient. To this end, the authors presented a new system of human-computer interaction based on gaze and hand gestures.

These works proved the efficacy of multi-modal interaction in comparison with gaze only or hand only gestures. However, in all these works the modalities are treated separately and there is no interaction between them. Besides, these modalities are captured by specific sensors that require adding specific peripheries to the devices.

In this work, eyes and hands gestures modalities are proposed without any additional peripheries. The capture of these modalities is assured by the camera already provided in the mobile device.

3. ARCHITECTURE OF THE PROPOSED SYSTEM: MULTIMOB

The proposed system aims to interact with mobile devices through eyes and hands gestures by providing the possibility to do actions like click, swipe etc. through natural modalities instead of touch way. To this end, the input of this system is a real-time video

https://doi.org/10.31436/iiumej.v20i2.1000

captured from the camera of the devices. There are not any additional peripheries. This system is composed of three main modules. The system overview is presented in Fig.1. The first module is responsible for detecting eyes' regions and then capturing eyes positions for each frame and recognizing gestures. The second module is responsible for the detection and recognition of static hands gestures and uses it to the recognition of dynamic hands gestures. Static hand gestures and eyes gestures resulting from each of these two sub-systems are used in the third module and combined in several ways with the purpose of building a new vocabulary to launch final actions that control the device such as click, rotate, wipe etc. In this system, the possibility to use one or more modality is given, the multimodal gestures fusion system is active only when the user chooses to not use eyes gestures and hands gestures only. In the other case, the user can execute specific actions with the corresponding modality.

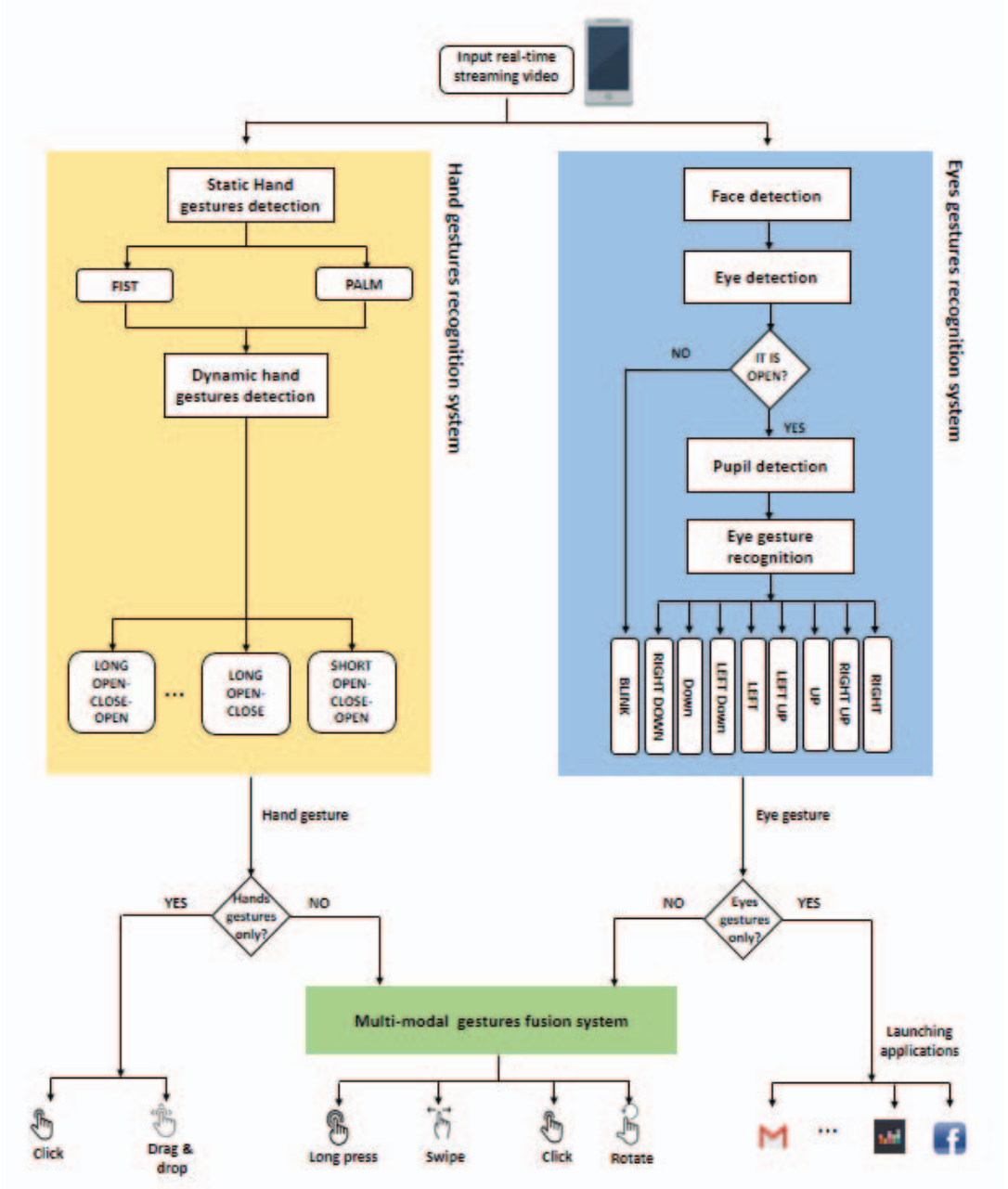

Fig 1 System overview

https://doi.org/10.31436/iiumej.v20i2.1000

3.1. Eyes gesture recognition for mobile devices

This subsystem is composed of successive steps.

- Face and eyes detectors based on Viola and Jones algorithm [25] are launched for each given frame.
- The eye detector indicates if the eyes are open or closed. If the eyes are detected as open, the next sub-module is launched.
- The pupil's detection is based on the reduction of the region of interest (ROI) in the eye and the determination of the darkest zone in the eyes [26].
- Once the pupils are detected, the eye gestures recognition sub-module will determine and extract the positions of pupils. These positions in the next step will be provided as inputs for a fuzzy inference system to recognize the corresponding motion.

3.1.1. Face and eyes detection

In the beginning, this module detects the user's face. This detection is assured by the Viola and Jones algorithm. This classifier, known by its ability to be running in real-time conditions, is provided by the OpenCV library [27] that achieves a true detection rate of 98.5% [28]. The face region is extracted and used as a location for eyes detection. The classification is based on boosted cascade Haar-like features. In fact, the determination of eyes regions includes also a classification about the state of the eyes to determine if they are open or closed.

Eyes blink is an involuntary reflex. This motion is intuitive and very rapid. The duration of the blink is on average between 100 ms and 400 ms [4]. For this reason, to differentiate between the spontaneous and the intentional blinks, a threshold is fixed, fixed at 500 milliseconds, of eye's closing duration. Any blink that lasts for a time longer than this threshold is considered as an intentional blink.

3.1.2. Eyes' gestures recognition

Once the eyes' region is detected, a reduction of this zone is required. In fact, the result eyes zone contains several parts of the eyes like the eyelid. This part is not needed to detect the pupils; for this reason, the first step in this sub-module is keeping only the small part of the iris and its surrounding. In this ROI, a pupil detector is charged to localize the pupils. In fact, the pupils are characterized by their common features. The pupils are the darkest zone in the eyes. So, to find the pupil's zone, a search of the darkest pixel is applied, once the pupil is localized a tracking task is launched by the template matching [26, 15].

Eye gesture recognition is defined in this work as the motion of pupils in a specific direction. In this module, the target gestures to be detected are the movements of the eyes to the lowing directions: right, right up, up, up left, left, left down, down and down right direction. These gestures have been chosen because they are simple and do not tire the eyes. However, the determination of each gesture cannot be precise. There is an overlap between 2 consecutive gestures for instance between left and left down like Fig.2 shows. To solve this issue, a fuzzy-based approach for the classification of eyes' motions based on pupils tracking was proposed.

https://doi.org/10.31436/iiumej.v20i2.1000

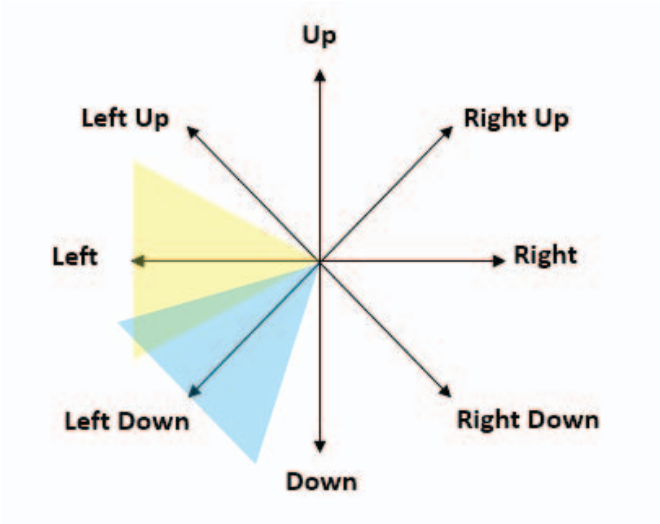

Fig 2 Eyes gestures

The recognition of the eye's gestures used basically the pupils' position extracted from each frame. The variation of these positions in successive frames is exploited in the motion's classification. This variation is computed as shown in the following equations:

$$dx = P2(x) - P1(x) \quad (1)$$

$$dy = P2(y) - P1(y) \quad (2)$$

Where d_x and d_y are the variations between 2 pupils' positions P_1 and P_2 at the x and y-axis respectively computed in pixels. In fact, the eyes are naturally in continuous variation. So, any movement that is lower than a specific threshold $Val_{Threshold}$ cannot be taken in consideration of. The sign of the value of d_x and d_y can be precise in the next step presented by the pseudo-code in Fig. 3

A fuzzy inference system (FIS) based on a set of rules to assure the phase of eyes gestures recognition as presented in Fig.4. These rules are defined in a natural way according to the directions of d_x and d_y, for example, if there is a variation on the x-axis without any variation on the y-axis, the final direction is right.

The FIS used in this approach is a the Mamdani type because there are many outputs so, it is the most intuitive and suitable in this case.

The fuzzification of the inputs and the outputs variables is presented in Fig.5, Fig.6, and Fig.7.

```
if (d_x > Val_Thresold) then
|  d_x ← POSITIVE
else
|  if (d_x < −Val_Thresold) then
|  |  d_x ← NEGATIVE
|  else
|  |  d_x ← ZERO
|  end
end
```

Fig 3 Pseudo-code of defining the sign of d_x and d_y

https://doi.org/10.31436/iiumej.v20i2.1000

- Rule 1: if d_x is POSITIVE and d_y is ZERO then Gesture is RIGHT
- Rule 2: if d_x is POSITIVE and d_y is POSITIVE then Gesture is RIGHT UP
- Rule 3: if d_x is ZERO and d_y is POSITIVE then Gesture is UP
- Rule 4: if d_x is NEGATIVE and d_y is POSITIVE then Gesture is LEFT UP
- Rule 5: if d_x is NEGATIVE and d_y is ZERO then Gesture is LEFT
- Rule 6: if d_x is NEGATIVE and d_y is NEGATIVE then Gesture is LEFT DOWN
- Rule 7: if d_x is ZERO and d_y is NEGATIVE then Gesture is DOWN
- Rule 8: if d_x is POSITIVE and d_y is NEGATIVE then Gesture is RIGHT DOWN
- Rule 9: if d_x is ZERO and d_y is ZERO then Gesture is NOGESTURE

Fig 4 Fuzzy rules of eyes gesture recognition

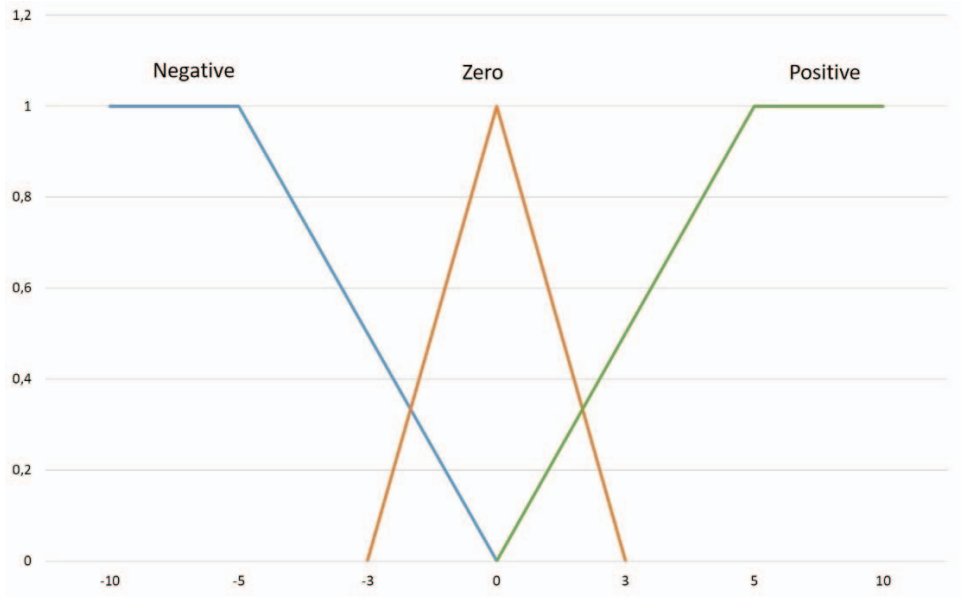

Fig 5 Fuzzification of the input variable d_x

2.1. Hands gesture recognition for mobile devices

The second system aims to recognize hands gestures. A system of hand gesture recognition is developed. It is based on the 2 types of gestures: the static and the dynamic ones. The definition of the static hand gesture is the stationary posture of a hand formed by fingers and the dynamic hand gesture as the succession of consecutive hand positions.

https://doi.org/10.31436/iiumej.v20i2.1000

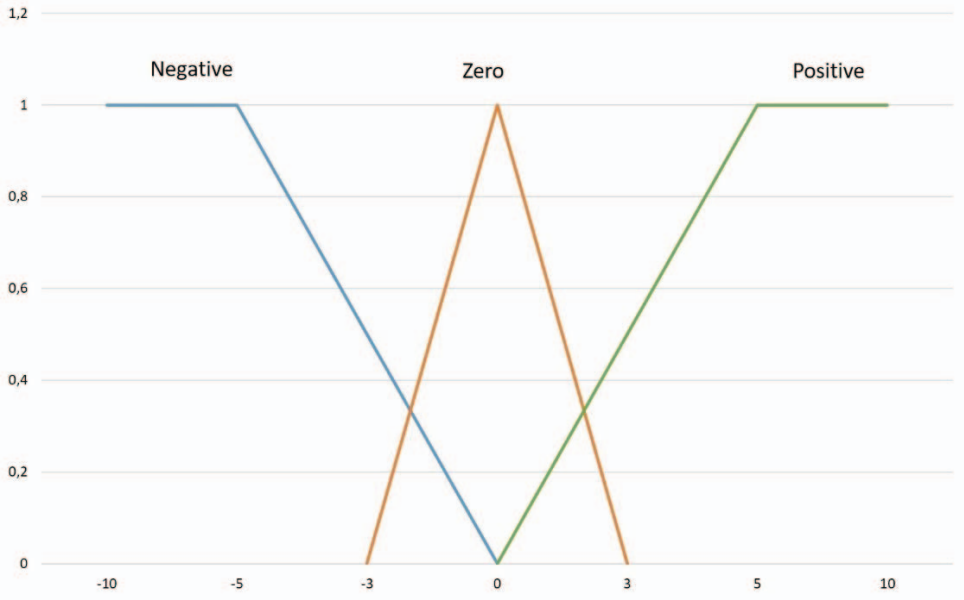

Fig 6 Fuzzification of the input variable d_y

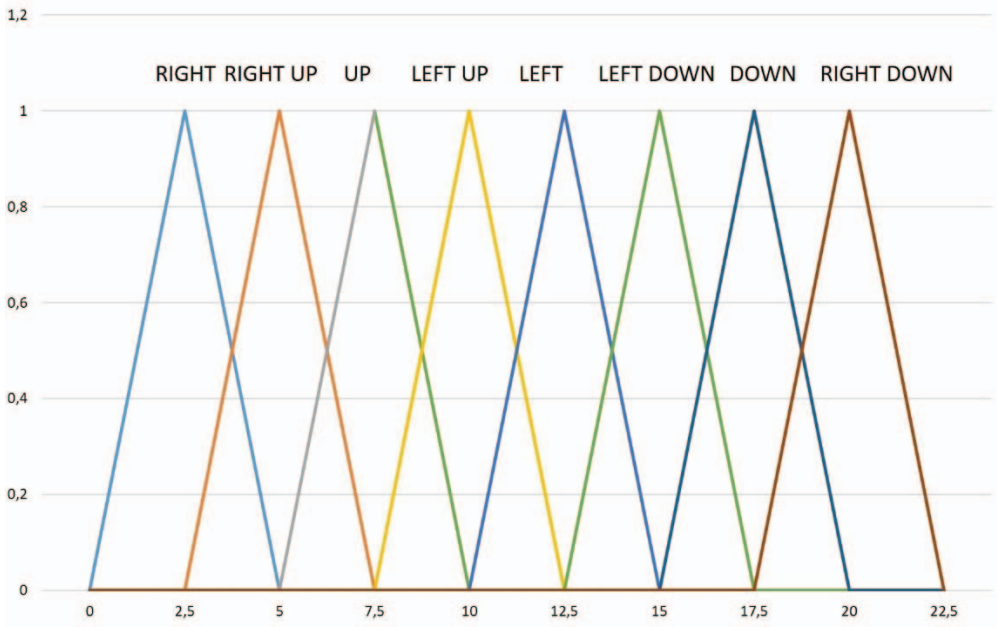

Fig 7 Fuzzification of the output variable Gesture

2.1.1. Static hand gesture recognition

To recognize hand gestures, independent classifiers were created for each gesture based on Viola and Jones algorithm [25]. These detectors are based on images converted in the grayscale. A cascade of strong classifiers is obtained by threshold on scalar features [28] using the Adaboost algorithm. For this system, two hand gestures were chosen: fist (closed hand) and palm (opened hand). To ensure having the highest success recognition rate, 3 classifiers were created for each gesture based on Haar-like, LBP, and HOG descriptors.
Dataset: For the training, 7716 and 7611 positive images for the fist and the palm, respectively and 12933 negative images collected from our private dataset [29] and other

https://doi.org/10.31436/iiumej.v20i2.1000

different public databases [30,31] were used. For the collection of this dataset, 21 volunteers have participated. The images collected are in indoor conditions with different backgrounds and hands positions. A low resolution of 72 dpi is used. The parameters of each window are described in the following table:

Table 1 Parameters of hand gestures detectors

	Windows	Stages
Fist detector	[32,33]	5
Palm detector	[42,32]	5

To have the highest success rate, a test of our classifiers is conducted according to three descriptors on 200 images for the closed hand and 200 images for the opened hand. The recall, the precision and the false positive per image (FPPI) were computed. A reminder of the metrics 'equations related to the recall, the precision, and the FPPI is presented in the following equations:

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$FPPI = \frac{FP}{TP+FN} \tag{5}$$

Where TP, FP, and FN are respectively the value of the true positive, false positive, and false negative of the detection rate. To determine these values, a bounding-box-based representation of the ground truth and the detected hand gestures are created. The PASCAL criterion [32] is applied. It is based on the computation of the bounding bow overlap (BO_i) of the detected bounding box (D_i)and the ground truth (G_i) as in this equation:

$$BO_i = \frac{|D_i \cup G_i|}{|D_i \cap G_i|} \tag{6}$$

A TN detection is obtained if the value of (BO_i) exceeds the value of the threshold. An FP detection is determined when a bounding box is detected and the value of (BO_i) is less than the threshold. An FN is obtained when there is no bounding box detected although the hand gesture is visible. The test results for the fist and the palm detection are shown in tables 2 and 3, respectively.

To improve our obtained results, the used descriptors are combined with the purpose of increasing the true detection rate and decreasing the false detection.
So, the common region of each bounding box of the two descriptors is computed by using Eq. 6. If the result is more than 0.5, the number of TP is increased and the result is the union of the two bounding boxes.

https://doi.org/10.31436/iiumej.v20i2.1000

Table 2 Recall, precision and FPPI values of the fist detection

	Recall (%)	Precision (%)	FPPI
HOG	98	62.61	0.585
HAAR	100	26.49	2.775
LBP	100	38.24	1.615

Table 3 Recall, precision and FPPI values of the palm detector

	Recall (%)	Precision (%)	FPPI
HOG	90	50.04	0.885
HAAR	98.5	18.97	4.2
LBP	100	35.14	1.845

To have the best result, all possibilities of combinations are tested and the results are presented in tables 4 and 5 for the fist and the palm gesture, respectively.

Table 4 Recall, precision and FPPI of the fist detector for each combination of descriptors

	Recall (%)	Precision (%)	FPPI
HOG+HAAR	98.5	98.5	0.015
HOG+LBP	97.5	97.5	0.025
HAAR+LBP	94	94	0.09

Table 5 Recall, precision and FPPI of the palm detector for each combination of descriptors

	Recall (%)	Precision (%)	FPPI
HOG+HAAR	91	93.84	0.06
HOG+LBP	99	96.11	0.04
HAAR+LBP	79.5	85.94	0.145

https://doi.org/10.31436/iiumej.v20i2.1000

(a) Fist detection results on some examples from our dataset

(b) Fist detection results on some examples from NUS dataset

Fig. 8 Fist detection results

These tables show that the high result is obtained by the combination of HOG and HAAR for the fist detection and HOG and LBP for the palm detection. To check the robustness, these classifiers are tested on a public database NUS [33]. 200 images from this database have been tested for each open and closed hand. The recall, precision and FPPI found are respectively: 95%, 93.59% and 0.065 for the fist and 96%, 97.46% and 0.025 for the palm. These results are promising mainly in comparison with the work of [34] that achieves 94,36% of success rate. Fig. 9 and Fig. 8 presented the detection result for each palm and

https://doi.org/10.31436/iiumej.v20i2.1000

fist gesture for both our database and NUS database. In fact, the detection is successful in several cases as Fig. 8 and Fig. 9 show; there are simple and complex backgrounds and there are pictures captured in outdoor and indoor conditions. This static hand gesture recognition is the same regardless of the color skin, the gender and the size of the hand.

(a) Palm detection results on some examples from our dataset

(b) Palm detection results on some examples from NUS dataset

Fig. 9 Palm detection results

https://doi.org/10.31436/iiumej.v20i2.1000

2.1.2. Dynamic hand gesture recognition

The second module developed in this system is the dynamic hand gesture recognition. In fact, a dynamic hand gesture is a sequence of static hand gestures.
As the dynamic hand gesture is a natural way of communication, there is a need to differentiate between the intentional one and the gesture used to execute a command. For this reason, vocabulary is built to define each gesture with the commands. The dynamic hand gesture is composed by a sequence of static hand postures that determine the beginning and the end of one gesture.
A tree of vocabulary is presented in Fig.10 to recognize three gestures that each one is responsible for executing a command: a click, a drag, and a drop. All motions begin with the initial gesture which is the opened hand and are finished with the same gesture.

- Click: the user must open his hand, close it and open it again in a duration that does not exceed the minimal time t_{min}.
- Drag: the user should open and close his hand for a duration more than the minimal time t_{min} then a tracking module is responsible for following the hand's position.
- Drop: is executed once the user opens his hand after the drag action.

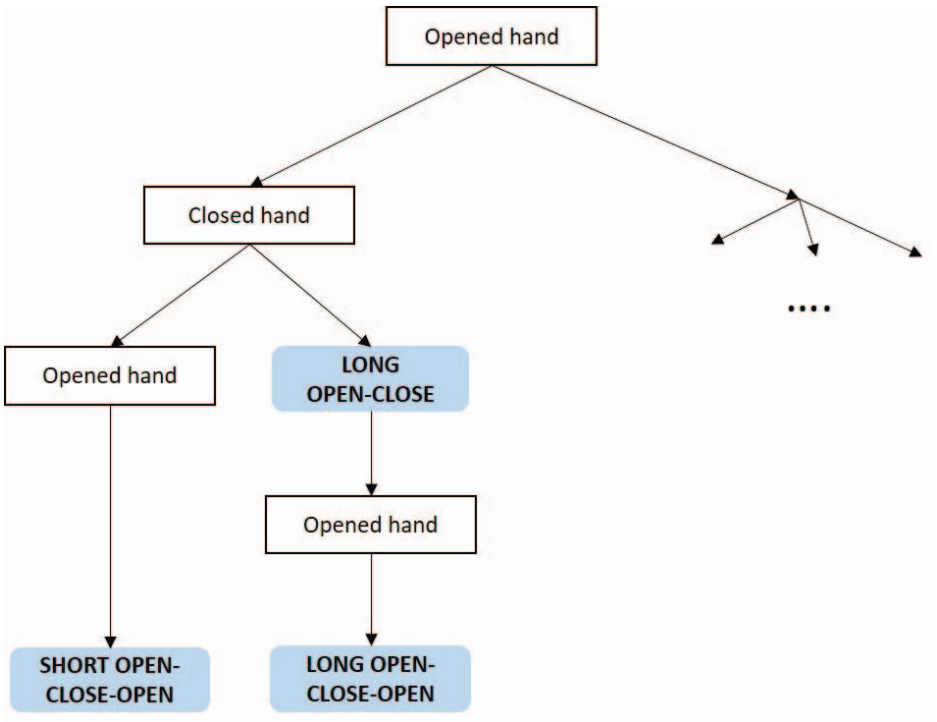

Fig. 10 Dynamic hand gestures vocabulary

2.2. Multi-modal fusion based gestures for interaction with mobile devices

The hands and eyes are in related movements. These couple modalities have an important role in the enhancement of the efficiency of controlling mobile devices. The coupled modalities provide for us additional information that can serve in the separation between the intentional and natural motions [35].

https://doi.org/10.31436/iiumej.v20i2.1000

To this end, the fusion of gestures recognized by the previous modules specific for eyes and hands gestures is proposed.

In fact, the fusion domain is divided into three levels: data, feature and decision levels [36].

- Data-level fusion is applied in the phase of receiving information captured from all available inputs. It aims to conserve all this information. However, in this type of fusion, the pre-processing step is absent and so the generated noise cannot be eliminated [36].
- Feature-level fusion combines the features generated from each input modality. Basically, this fusion's category requires that the generated features' vectors have the same representation. This condition cannot be available in many cases particularly when the input modalities have different natures.
- Decision-level fusion is the most used thanks to its capability to operate with different types of modalities. The final decision is taken based on the local decision generated from each modality. In comparison with other fusion levels, decision-level fusion has not any constraint of invariance input or vectors' representation because each module analyses its related modality separately.

The comparison of these fusion-levels leads us to use the decision-level fusion. This is the most convenient in our case. In fact, the eyes and hands gesture recognition modules are captured from the same camera. However, each module analyzed the captured video and generated a decision regarding each modality. The variety of the decisions classified by each module, the ability to fuse them and the respect of the real-time conditions are the important constraints that specify the tool used for the decision fusion step. The decision tree is one of the most known techniques in the supervised classification. It is characterized by its simplicity in analyzing and interpreting thanks to its flexibility in the determination of the input variables and the rules used for the classification task.

The fusion module takes 3 attributes as inputs, one from eyes gestures and two from hand gestures with the purpose of executing actions to command mobile devices. It is noteworthy that the hands gesture recognition module has 2 classes: palm and fist and the eyes gesture recognition module has 9 classes*: right, right up, up, left up, left, left down, down, right down* and *blink*. For the fusion module, a choice is made to consider that *the right*, *right up* and *right down* as *right* direction and *left*, *left up* and *left down* as *left* direction. As final decisions, 6 classes are defined: *long press, rotate*, *click*, *zoom*, *swipe in the right* and *swipe in the left*. Table 6 details how these decisions are organized.

Table 6 Training table of eye and hand gestures fusion module

Eye gesture	Hand gesture 1	Hand gesture 2	Decision
Blink	Close	Close	Long press
Right	Close	Close	Rotate
Left	Close	Close	Rotate
Up	Close	Close	Rotate
Down	Close	Close	Rotate
Blink	Open	Open	Zoom
Right	Open	Open	Swipe in Right
Left	Open	Open	Swipe in Right
Up	Open	Open	Swipe in Left
Down	Open	Open	Swipe in Left
Right	Open	Close	Click
Left	Open	Close	Click

https://doi.org/10.31436/iiumej.v20i2.1000

Up	Open	Close	Click
Down	Open	Close	Click
Right	Close	Open	Click
Left	Close	Open	Click
Up	Close	Open	Click
Down	Close	Open	Click

3. Experimental results and discussion

In this part, the twofold system is evaluated with the purpose of validating its robustness. To this end, we used an Android-based tablet for the deployment of our system. This device has an android version of 4.2.1 running on NVIDIA Tegra 3 Quad-Core and a 2 Go of RAM. The front-facing camera has a 2 Megapixels resolution.

3.1. Eyes gesture recognition for mobile devices

In this module, these libraries are used: OpenCV [27] image processing and Fuzzy Lite libraries [37] for the classification. 8 participants aged between 25 and 31 years, who are unfamiliar with eyes gestures recognition systems, volunteered to test our system. First, each participant is asked to sit in front of the tablet and save a distance between 25 and 40 cm to keep the whole face captured by the camera. Then, they are asked to close their eyes and move it in the 8 directions. To validate this module, different metrics are computed: the sensitivity, the false positive rate (FPR) defined by these equations:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

$$FPR = \frac{FP}{FP+TN} \quad (3)$$

Where the value of TP, FN, FP, TN are computed as in this way:

- There is a TP when a correct eyes gesture is recognized.
- There is a FN when none or a false eyes gesture is recognized.
- There is a FP when an eyes movement is detected but there is no gesture.
- There is a TN when there is no gesture is detected and the user keeps his eyes in a stationary position.

The result of this assessment and a comparison with the most related work are presented in table 7.

Table 7 The result of eyes gesture recognition module

	Accuracy [%]	Sensitivity [%]	FPR [%]
Our system	76.53	85	27
Vaitukaitis et *al.* [13]	60.0	28.3	17.6

This assessment showed promising results achieved by our system. In comparison with related work, the proposed system attained 76.53% of accuracy while [13] achieved 60%.

https://doi.org/10.31436/iiumej.v20i2.1000

Another test is conducted to investigate the material consumption of this module. The measurement of the physical resources consumption of RAM and CPU is computed. The rapidity of our developed algorithm can be induced in the number of frames per second (FPS). The results are presented in table 8.

Table 8 Computation of RAM and CPU consumption

	RAM	CPU	FPS
Our system	90	40	24
Miluzzo et *al.*[11]	56.51	65.4	-

As shown in table 8, our subsystem consumes 90% and 40% of RAM and CPU, respectively. These numbers showed the limit of this module in terms of RAM consumption. It is noteworthy that this module consumes RAM much more than the system proposed by [11], in the contrast to CPU. Besides, this module proved the fact that it can be running in real-time condition by achieving 24 FPS. It is so far advantageous when compared with similar systems [13] and [38] which had lower than 15 FPS.

3.2. Hands gesture recognition for mobile devices

For this module, a study is conducted on 10 volunteers that they never use any touch-less application based on eyes and hands gestures. This test is repeated in two scenarios depending on the external environment and the user's state whether he is sitting or walking:

- Scenario 1: the user is in laboratory conditions and takes the tablet when he is sitting.
- Scenario 2: the user is in outdoor conditions and takes the tablet when he is walking.

In this evaluation, each participant is asked to move one hand with the purpose of executing actions corresponding to the vocabulary presented in the subsection 3.2.2. The initial hand position is palm gesture. The results of this experiment are shown in table 9.

Table 9 Experiment result for dynamic hand gesture recognition

	Scenario 1	Scenario 2
Click	100%	90%
Drag+Drop	100%	100%

The experiment results prove that the recognition of the dynamic gesture reaches 100% in scenario 1 and 90% in scenario 2 for both motions. In fact, these results depend on the performance to the recognition of the static and dynamic gestures. The error coming from the static gesture causes automatically an error in the dynamic gesture.

https://doi.org/10.31436/iiumej.v20i2.1000

3.4. Multi-modal fusion based gestures for interaction with mobile devices

To validate our approach in this module, a study is conducted on 5 volunteers aged between 24 and 30 years. Each participant is asked to operate the tablet to execute the proposed actions. Fig.11 shows that our module can achieve promising results. The success rate attains 100% for Long press and Zoom actions and 80% for others. Indeed, the results of fusion's gestures are due mainly to the result of eyes and hands gestures modules. Note that every error of the eye's gestures recognition or in the hand gestures recognition modules causes a decision error.

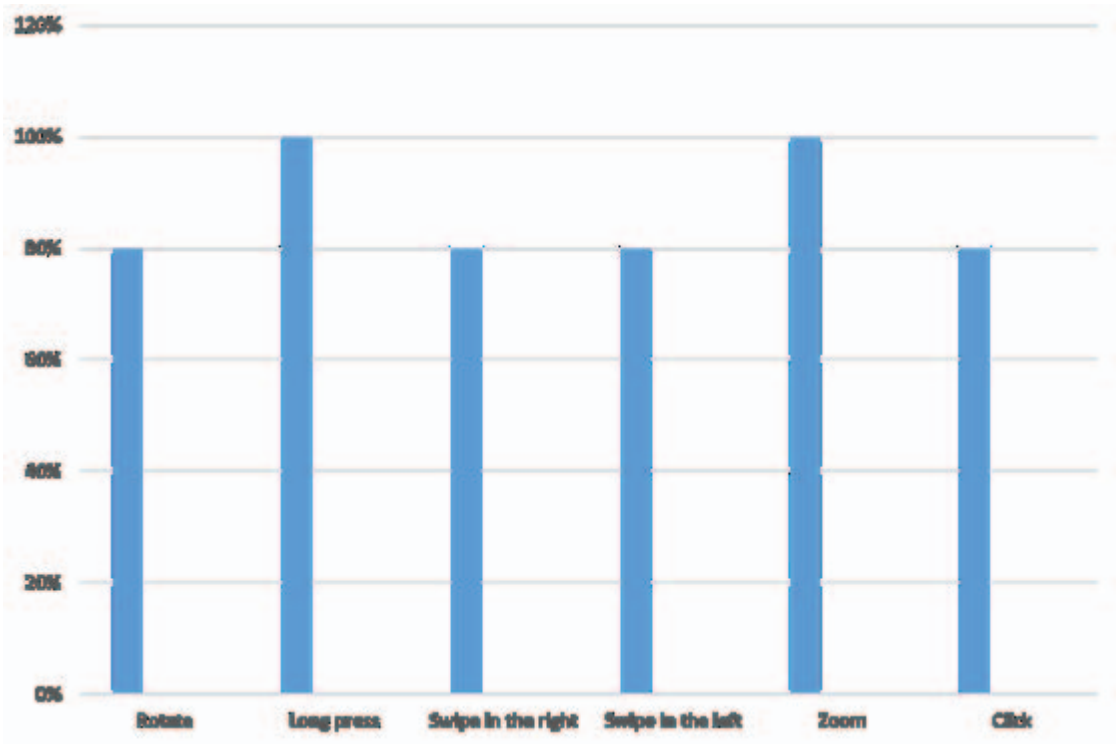

Fig. 11 Experiment results of multi-modal fusion based gestures subsystem

4. CONCLUSION

In this paper, a new system for mobile HCI based on eyes and hands gesture recognition is presented. This system uses the eyes' gestures to execute commands based on detection of the eyes and gaze direction. For the hand gesture recognition, two independent classifiers are created and dedicated to fist and palm gesture. A combination of different descriptors is used to have the highest success rate and the minimum false detection. From the static gestures, dynamic gestures are created. They are based on precise vocabulary to execute commands.

To evaluate our system, a study is conducted on 10 participants in two different scenarios for each eye and hands gesture. The experimental assessment showed that this system can reach promising results.

To study the feasibility of the proposed system, an application for medical uses is developed. This application provides the doctor with the possibility to consult and manipulate a patient's medical images and information that he can need during the surgical intervention.

Actually, the surgeon cannot touch anything that is not sterilized during the surgery; also his hands are busy all the time. As a consequence, he is obliged to use touchless ways to interact with his device.
First, the doctor chooses the modality that he/she wants to use to control his/her devices.
Then, he/she can swipe the medical images and the patient's information by using their eyes with a blink and an eye movement.
The hand gestures can serve to move a cursor to provide the manipulative actions in the application like clicking on buttons etc.

ACKNOWLEDGMENT

The research leading to these results has received funding from the Ministry of Higher Education and Scientific Research of Tunisia under the grant agreement number LR11ES48. The research and innovation are performed in the framework of a thesis MOBIDOC financed by the EU under the program PASRI.

REFERENCES

[1] Nagamatsu T, Yamamoto M, Sato H. (2010). Mobigaze: Development of a gaze interface for handheld mobile devices. In Proceedings of Extended Abstracts on Human Factors in Computing Systems, pp: 3349-3354.
[2] Samadi M R H, Cooke N. (2014). EEG signal processing for eye tracking. In Proceedings 22nd European Signal Processing Conference. pp. 2030-2034
[3] Dhuliawala M, Lee J, Shimizu J, Bulling A, Kunze K, Starner T, Woo W. (2016). Smooth eye movement interaction using EOG glasses. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, 307-311. doi: 10.1145/2993148.2993181.
[4] Bulling A, Ward JA, Gellersen H, Troster G. (2011). Eye movement analysis for activity recognition using electrooculography, IEEE transactions on pattern analysis and machine intelligence, 33(4):741-753
[5] TheEyeTribe [http://www.theeyetribe.com/]
[6] Chen P, Wang P, Wang J, Yao Y. (2017). Design and motion tracking of a strip glove based on machine vision. Neurocomputing. doi: https://doi.org/10.1016/j.neucom.2017.03.098.
[7] Dipietro L, Sabatini A M, Dario P. (2008). A survey of glove-based systems and their applications. IEEE Transactions on Systems, Man, and Cybernetics, Part C (App/lications and Reviews) 38(4): 461-482. doi: 10.1109/TSMCC.2008.923862
[8] Wood E, Bulling A. (2014) Eyetab: Model-based gaze estimation on unmodified tablet computers. In Proceedings of the Symposium on Eye Tracking Research and Applications, pp: 207-210
[9] Pino C, Kavasidis I. (2012). Improving mobile device interaction by eye tracking analysis. In Proceeding of Federated Conference on Computer Science and Information Systems. pp:1199-1202.
[10] Huang Q, Veeraraghavan A, Sabharwal A. (2017). Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. In Machine Vision and Applications. 28(5): 445-461. doi: 10.1007/s00138-017-0852-4
[11] Miluzzo E, Wang T, Campbell A T. (2010). Eyephone: activating mobile phones with your eyes. In Proceedings of the second workshop on Networking, systems, and applications on mobile handhelds, pp: 15-20.
[12] Iqbal N, Lee H, Lee S Y. (2013). Smart user interface for mobile consumer devices using model-based eye-gaze estimation. In IEEE Transactions on Consumer Electronics, 59(1): 161-166
[13] Vaitukaitis V, Bulling A. (2012). Eye gesture recognition on portable devices. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp: 711-714

[14] Zhang X, Kulkarni, H, Morris M R. (2017). Smartphone-based gaze gesture communication for people with motor disabilities. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp: 2878-2889.

[15] Elleuch, H., Wali, A., Samet, A., Alimi, A.M. (2016). A real-time eye gesture recognition system based on fuzzy inference system for mobile devices monitoring. In Proceedings of International Conference on Advanced Concepts for Intelligent Vision Systems, pp. 172-180.

[16] Meng X, Cheung C M, Ho K L, Lui K S, Lam E Y, Tam V. (2012). Building smart cameras on mobile tablets for hand gesture recognition. In Proceeding of Sixth International Conference on Distributed Smart Cameras, pp: 1-5

[17] Prasuhn L, Oyamada Y, Mochizuki Y, Ishikawa H. (2014). A hog-based hand gesture recognition system on a mobile device. In Proceedings of IEEE International Conference on Image Processing, pp. 3973-3977

[18] Dalal N, Triggs B. (2005). Histograms of oriented gradients for human detection, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1:886-893

[19] Song J, Soros G, Pece F, Fanello S R, Izadi S, Keskin C, Hilliges O. (2014). In-air gestures around unmodified mobile devices. In Proceedings of the 27th annual ACM symposium on User interface software and technology, pp: 319-329

[20] Elleuch, H., Wali, A., Samet, A., Alimi, A.M. (2015). A static hand gesture recognition system for real time mobile device monitoring. In Proceedings of the 15th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 195-200.

[21] Chatterjee I, Xiao R , Harrison C. (2015). Gaze+ gesture: Expressive, precise and targeted free-space interactions, Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 131-138

[22] Hales J, Rozado D, Mardanbegi D. (2013). Interacting with objects in the environment by gaze and hand gestures. In Proceedings of the 3rd international workshop on pervasive eye tracking and mobile eye-based interaction; pp: 1-9

[23] Pouke M, Karhu A, Hickey S, Arhippainen L. (2012). Gaze tracking and non-touch gesture based interaction method for mobile 3d virtual spaces. In: Proceedings of the 24th Australian Computer-Human Interaction Conference, pp. 505-512

[24] Yoo B, Han J J, Choi C, Yi K, Suh S, Park D, Kim C. (2010). 3d user interface combining gaze and hand gestures for large-scale display. In Proceedings of Extended Abstracts on Human Factors in Computing Systems, pp: 3709-3714

[25] Viola P, Jones M J. (2004). Robust real-time face detection. International journal of computer vision. 57(2): 137-154

[26] Elleuch H, Wali A, Alimi A M. (2014). Smart tablet monitoring by a real-time head movement and eye gestures recognition system. In Proceedings of International Conference on Future Internet of Things and Cloud (FiCloud); pp. 393-398.

[27] OpenCV [http://www.opencv.org/]

[28] Gonzalez-Ortega D, Diaz-Pernas F, Martinez-Zarzuela M, Anton-Rodriguez M, Diez-Higuera J, Boto-Giralda D. (2010). Real-time hands, face and facial features detection and tracking: Application to cognitive rehabilitation tests monitoring. Journal of Network and Computer Applications. 33(4): 447-466

[29] REHG [http://www.regim.org/publications/databases/rehg/]

[30] MIT [http://web.mit.edu/torralba/www/indoor.html]

[31] Memo A, Zanuttigh P. (2018). Head-mounted gesture controlled interface for human-computer interaction. Multimedia Tools and Applications. 77(1): 27-53. doi: 10.1007/s11042-016-4223-3.

[32] Everingham M, Van Gool L, Williams C K, Winn J, Zisserman A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision. 88(2): 303-338

[33] NUS [https://www.ece.nus.edu.sg/stfpage/elepv/NUS-HandSet/]

[34] Pisharady P K, Vadakkepat P, Loh A P. Attention based detection and recognition of hand postures against complex backgrounds. International Journal of Computer Vision.101(3) :403-419

[35] Elleuch H, Wali A, Alimi M A, Samet A. (2017). Interacting with mobile devices by fusion eye and hand gestures recognition systems based on decision tree approach. In Ninth International Conference on Machine Vision, pp: 103-410.

[36] Dumas B, Lalanne D, Oviatt S. (2009). Multimodal interfaces: A survey of principles, models and frameworks. Human machine interaction. 5440:3-26. doi: https://doi.org/10.1007/978-3-642-00437-7_1

[37] FuzzyLite [http://www.fuzzylite.com/]

[38] Skodras E, Fakotakis N. (2015). Precise localization of eye centers in low resolution color images. Image and Vision Computing. 36: 51-60