ROMA TRE
UNIVERSITÀ DEGLI STUDI

Dipartimento di Elettronica Applicata
Scuola Dottorale di Ingegneria
Sezione di Ingegneria dell'Elettronica Biomedica, dell'Elettromagnetismo e delle Telecomunicazioni

TESI DI DOTTORATO

EYE DRIVEN HUMAN-MACHINE INTERFACES FOR PEOPLE WITH DISABILITIES

INTERFACCE UOMO-MACCHINA DI TIPO VISUALE PER LO SVILUPPO DI DISPOSITIVI DI AUSILIO ALLA DISABILITÀ

Candidato
Diego Torricelli

Docente guida
Prof. Tommaso D'Alessio

Roma, 1 marzo 2009

Ringraziamenti

Questa tesi è il frutto di tanta passione, lavoro e dedizione, che mi appartengono solo in parte. In queste pagine si concentra il prodotto della sinergia di molte persone, i cui preziosi contributi sono passati, consapevolmente e non, attraverso di me. A loro dedico il mo lavoro.

Il primo grazie va ai miei genitori, che da sempre hanno rispettato e supportato le mie scelte, esprimendo al tempo stesso le loro idee e convinzioni, trasmettendomi fiducia e un prezioso spirito critico.

Grazie ai tesisti, dottorandi e ricercatori del Laboratorio di Ingegneria Biomedica Biolab^{3}, compagni di avventure degli ultimi quattro anni, con cui ho condiviso la passione per la ricerca e imparato l'importanza della comunicazione nel lavoro di gruppo.

Grazie al prof. Tommaso D'Alessio, un riferimento a livello scientifico, ma soprattutto a livello umano. Qualsiasi livello di conoscenza, capacità e ruolo raggiungiamo, siamo e rimaniamo Persone. C'è chi questa cosa non l'ha scordata.

Grazie inoltre a tutti coloro che dedicano i loro sforzi al "fare" e non solo al pensare. Mi riferisco in particolare ai professionisti e volontari nell'ambito del supporto alla disabilità. Mi rendo conto che il "volo pindarico" è sicuramente un buono stimolo, ma inutile se non viene concretizzato.

L'ultimo e il più importante grazie lo dedico ai miei Amici, i vecchi e i nuovi, conosciuti attraverso mille esperienze diverse. Grazie perché con voi sto scoprendo sempre di più quanto sia bello vivere la strada, piuttosto che pensare a raggiungere la vetta. E che sulla strada non sei mai solo.

Sommario in italiano

La presente tesi di dottorato è finalizzata alla progettazione e allo sviluppo di sistemi di ausilio e riabilitazione per portatori di disabilità motorie gravi. La linea di ricerca intende studiare supporti alla soluzione del problema dell'handicap attraverso lo sviluppo di un adeguato insieme di interfacce uomo-macchina, dove con il termine *interfaccia uomo-macchina* si intende un dispositivo per il controllo remoto di una macchina (un computer o un qualsiasi dispositivo elettronico o meccanico) da parte di un soggetto.

Tra i sistemi che ben si prestano ad essere utilizzati da persone con disabilità motorie molto gravi vi sono i cosiddetti puntatori oculari, dispositivi che permettono ad un utente di gestire una macchina, e più specificatamente un computer, attraverso il movimento degli occhi. Il controllo oculare è infatti una di quelle abilità che rimane pressoché inalterata anche nella forme di patologie motorie più gravi. Inoltre, il movimento degli occhi risulta strettamente correlato alla volontà e alla capacità di concentrazione e di attenzione, elementi importanti nel campo della riabilitazione neuromotoria e cognitiva. I puntatori oculari hanno diverse potenzialità applicative. Risultano utilizzati principalmente nel campo della comunicazione, per quelle patologie dove il movimento degli occhi risulta essere l'unico canale comunicativo. In tali contesti il puntatore oculare può restituire alla persona un certo grado di indipendenza comunicativa attraverso la gestione di programmi software per la scrittura e altre applicazioni tipiche, come posta elettronica e navigazione internet. Attualmente, i puntatori oculari disponibili nel mercato presentano diverse problematiche legate principalmente all'usabilità e al costo, che risulta essere nell'ordine della decina di migliaia di euro.

La attività di ricerca svolta propone una soluzione che risulta innovativa rispetto allo stato dell'arte, presentando un puntatore a basso costo ma con un alto rendimento sotto il profilo della usabilità, sia in termini di accuratezza e stabilità della misura sia per la capacità di soddisfare le reali esigenze funzionali dell'utente.

Questa tesi è articolata in 7 capitoli.

Il primo capitolo presenta la motivazione alla base del lavoro e stila le tre ipotesi che il presente lavoro ha il compito di verificare attraverso le sperimentazioni. Tali ipotesi, basate sugli elementi innovativi del sistema sviluppato, affermano:

1. “E’ possibile stimare la direzione dello sguardo con un’accuratezza sufficiente a gestire applicazioni nell’ambito della disabilità senza l’uso di hardware specifico o ad alto costo”.
2. “E’ possibile aumentare la robustezza della stima dello sguardo usando un approccio bio-inspirato basato su reti neurali artificiali”.
3. E’ possibile usare lo sguardo come canale di comunicazioni per fini riabilitativi e non solo assistivi”.

Da un punto di vista generale l’innovazione consiste nel trasferire la complessità dalla parte hardware alla parte software, al fine di dar vita ad un dispositivo che utilizzi hardware facilmente reperibile nel mercato, come una tradizionale webcam. Nessuna tipologia di puntatore oculare in commercio propone un sistema unicamente software e di facile installazione su un tradizionale computer. Un altro elemento di innovazione è rappresentato da un sistema biologicamente ispirato, finalizzato al rendere la misura più robusta rispetto al movimento della testa, che rimane uno dei principali problemi per l’analisi dello sguardo.

Il secondo ed il terzo capitolo presentano il contesto nel quale il lavoro si inserisce. Il secondo capitolo fornisce una panoramica più generale riguardo la disabilità, gli ausili tecnologici e lo sviluppo delle interfacce uomo-macchina con riferimento ai più recenti standard circa l’usabilità e la progettazione centrata sull’utilizzatore. Il terzo capitolo si occupa invece di esplorare il campo della stima dello sguardo sotto un profilo più tecnico, analizzando i metodi e gli approcci presenti nella letterature scientifica. Vengono evidenziati gli aspetti critici e non ancora completamente risolti, come la compensazione del movimento della testa e la robustezza alle variabilità ambientali e intersoggettive.

Nel quarto capitolo il sistema di puntamento oculare sviluppato viene presentato in tutte le sue parti. Il capitolo è suddiviso in tre parti che rappresentano i tre blocchi algoritmici principali della procedura di stima dello sguardo:

i) fase di *inizializzazione*, in cui il sistema, attraverso l’analisi delle immagini acquisite da webcam, riconosce la presenza di un soggetto e si aggancia ad esso, individuando le caratteristiche fisionomiche fondamentali per la stima dello sguardo, come iridi e angoli degli occhi;

ii) inseguimento delle caratteristiche facciali, chiamato in ambito internazionale *tracking*, che si occupa di individuare le suddette caratteristiche in ogni frame del flusso video. Particolari sforzi sono stati dedicati all'incremento virtuale della risoluzione delle immagini.

iii) la stima della direzione dello *sguardo*, parte finale della sequenza algoritmica e cuore del presente lavoro, che riguarda la determinazione di funzioni matematiche non deterministiche per il calcolo delle coordinate del punto osservato sullo schermo del monitor a partire dalle informazioni geometriche sulla fisionomia risultanti dai passi precedenti.

Nell'implementazione di queste fasi, il lavoro ha consistito sia nell'utilizzo di tecniche affermate nell'ambito della Visione Artificiale, sia nella progettazione di nuovi metodi di processamento di immagine e di computo neurale.

Il quinto capitolo propone alcune soluzioni per dimostrare la fattibilità dell'utilizzo del puntatore sviluppato nella gestione di applicazioni in ambito assistivo e riabilitativo. Per quanto riguarda l'ambito assistivo sono stati progettati e realizzati due software, il primo per la scrittura di testi (chiamato eye-typing) ed il secondo per il controllo delle apparecchiature della casa (domotica). L'attenzione è stata rivolta soprattutto all'aspetto grafico e alla struttura logica perché risultassero intuitive, rilassanti ed efficaci. In campo più propriamente riabilitativo, un primo ambito affrontato è quello della riabilitazione neuromotoria post ictus. E' stato proposto il concetto di una piattaforma multimodale basata sull'uso dello sguardo come stimatore dell'intenzionalità, combinato con modulo bio-inspirato per il controllo del movimento del braccio con stimolazione elettrica funzionale (FES) per esercizi di reaching. Il sistema di analisi dello sguardo, predicendo la volontarietà, fornisce indicazioni utili alla macchina che supporta l'arto accompagnandolo verso la direzione desiderata. L'approccio risulta naturale, intuitivo e minimamente invasivo. Un altro campo riabilitativo esplorato riguarda la paralisi cerebrale. In tale contesto lo sguardo è considerato come indicatore quantitativo del grado di capacità di interazione con il mondo esterno. Il protocollo proposto prevede l'utilizzo di filmati di varia tipologia e l'analisi della reazione del bambino durante la loro visione, con il fine di identificare parametri di attenzione, concentrazione, controllo motorio. Il progetto è tutt'ora in corso attraverso una collaborazione con il gruppo di bioingegneria del CSIC (Consejo Superior de Investigaciones Cientificas) di Madrid, Spagna.

I test sperimentali condotti durante i tre anni di lavoro, descritti nel sesto capitolo, hanno verificato la validità del sistema e dei metodi proposti. In particolare, i risultati ottenuti sul sistema di misura dello sguardo hanno mostrato una ottima accuratezza globale rispetto ai sistemi similari presenti in letteratura. Ulteriori test hanno dimostrato l'efficacia del sistema nell'individuare

automaticamente l’utente senza alcun intervento di un operatore esterno. La tecnica, basata sull’analisi del battito di palpebra, si è rivelata essere efficace in termini di accuratezza e costo computazionale, rendendo possibile un funzionamento in tempo reale. Inoltre, tale metodo permette di re-inizializzare automaticamente il sistema in caso di errore nell0inseguimento delle caratteristiche facciali. Il metodo proposto supera la maggior parte dei problemi prodotti dal movimento della testa. Due reti neurali sono state progettate perché apprendessero a calcolare la direzione dello sguardo per posizioni diverse della testa. Lo scopo degli esperimenti è stato quello di testare il sistema per movimenti naturali della testa, che possono avvenire mantenendo una postura confortevole di fronte allo schermo. I risultati confermano la robustezza del sistema, e le funzioni neurali hanno dimostrato essere più performanti delle tradizionali funzioni quadratiche usate in letteratura.

Per dimostrare la fattibilità del sistema per applicazioni per disabilità, i due tipi di interfacce (eye-typing e domotica) sono state progettate, realizzate e testate su soggetti di differenti fasce d’età, che hanno confermato la buona usabilità delle interfacce, in termini di efficacia e soddisfazione.

Gli esperimenti preliminari sulla piattaforma per neuro riabilitazione post-ictus hanno dimostrato una buona capacità del sistema di riconoscere e classificare la direzione dello sguardo diretta su uno di 4 oggetti collocati sopra un piano. I movimenti simulati, guidati da uno stimolatore neurale, hanno provato una buona perfomance in termini di accuratezza della posizione raggiunta, sottolineando l’adeguatezza dell’approccio in un contesto reale.

La tesi si conclude nel capitolo 7, il quale presenta una discussione generale sui risultati ottenuti verificando le tre ipotesi, mette in evidenza i punti di criticità riscontrati e delinea le principali future direzioni di investigazione, focalizzate in particolare sul miglioramento dell’accuratezza dell’analisi di immagine e sull’integrazione con sistemi di movimento della testa in campo tridimensionale. L’approccio bio-inspirato sembra inoltre essere un ottimo strumento computazionale per la realizzazione di sistemi adattivi che apprendano a conoscere il comportamento dello sguardo del soggetto e a incrementare quindi le performance durante l’utilizzo. E’ questa la principale direzione che questa tesi vuole indicare per i lavori futuri che vorranno proseguire la linea di ricerca fin qui sviluppata.

Table of contents

1 Introduction

1.1 Motivation

Ideally, getting a job or education and communicate freely are activities that everybody should be able to do. The last two decades have been characterized by an exponential growth of the use of the technology in the everyday life, that on one side enhanced the potentials of the average citizen. On the other side, since technology is often created without regard to people with disabilities, a further barrier to millions of people arose. In the last years the problem of accessibility has become significant due to factors as a an increasing expectancy of the quality and duration of life and a general growth of sensitivity to human rights. In this sense, public investments are rising in the field of Assistive Technology (AT), that is the discipline that studies and develops technical aids for individuals with disabilities. In the scenario of AT, particular importance resides in the design of new kinds of interfaces, that is the medium through which communication flows between the user and the device. In particular the discipline of human-computer interaction (HCI) studies innovative methods to measure, analyze, interpret the activity of the human in order to provide the machine with appropriate control signals.

In the field of disability the interaction between human and machine plays a crucial role, since each disability needs a proper channel of interaction due to the particular

typology of residual ability. Within the last two decades, new kinds of interface has been developed and commercialized, based on the movement of the eyes, in order to permit people that are unable to move their arms to manage a personal computer and its principal applications. The choice to exploit the movements of the eyes is motivated by the observation that the control of the eye is one of the ability that is not affected in the majority of motor pathologies, at least up to the final stages. The principal problem related with human computer interfaces for disabilities is the trade-off between cost, performance and usability. As some studies confirm [1,2], a large number of technology aids are present in the market but many of them do not match real user needs. Therefore a different approach in the field of research is taking place, i.e. the user-centered design, in which the technology is taken into account as just one of the several important elements in the interface design.

The motivation of this work resides in the will to create innovative eye driven interfaces that could be really *accessible* to a high number of persons and could cover a wide range of disabilities. The efforts have been devoted to realize a system that could be developed respecting the cost-effectivenes, that nowadays represents the most challenging requirement in this field and that prevents these kinds of interfaces from being used in a diffusing way. The thesis introduce another important kind of innovation, represented by the use of new typologies of mathematical functions for the estimation of gaze using artificial neural networks (ANN) in order to account for head movements that represents the other important issue not yet completely solved by the literature.

Finally, in order to establish the rules for design of the interface, diverse standards [3,4,5,6] have been studied. According to them the interface proposed in this thesis has been designed, developed and tested according to the concept of usability, defined as “the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” [4], involving the user in the iterative process of experimentation and development.

1.2 Hypothesis

The thesis aims at designing an innovative, usable and affordable system for gaze estimation to be used as human-computer interface without the use of the hands, specifically addressed to people with high motor disabilities. This general hypothesis can be divided into three particular hypothesis, that will focus on the specific technical goals of the thesis.

Hypothesis n°1: *"It is possible to estimate gaze direction with an accuracy compatible with disability applications without the use of specific or high-cost hardware"*. This hypothesis specifically regards the affordability and the usability of the system. Nowadays all the commercial available systems for remote gaze tracking use specific hardware and the cost is very high as compared to other kinds of computer interfaces. The objective here is to use a simple webcam as the only additional hardware apart from the computer.

Hypothesis n°2: *"It is possible to enhance robustness in gaze estimation using an artificial neural networks (ANN) approach"*. This hypothesis focuses on innovative and bio-inspired methods for the estimation of gaze direction in order to take care of changing scenarios, as the movement of the head, that represent one of the greatest problems for gaze estimation.

Hypothesis n°3: *"It is possible to use gaze as a channel of interaction for rehabilitation purposes"*. So far eye gaze tracking has been used principally as a technological aid for impaired people, with no rehabilitation purposes. This thesis wants to demonstrate that gaze can be used as a valuable channel of interaction in integrated systems for neuro-motor rehabilitation, in particular for stroke rehabilitation.

1.3 Structure of the thesis

This thesis is divided into 7 chapters.

The next chapter presents the context in which the thesis is inserted. A brief description of the terminology used in the field of disability is given. The most relevant eye-driven interfaces and applications present in the market are listed and synthetically analyzed, in order to highlight the technical points of interest. The user-centered design and the concept of usability in the interface design is then introduced with explicit reference to the existing international standards.

In the chapter 3 an extensive description of the State of the Art of the methods for estimating the direction of gaze, i.e. eye-gaze tracking, is given. The crucial problems not yet solved by the literature will be listed in order to point out the positioning of the developed system within the field of eye-gaze tracking technology.

Chapter 4 focuses on the developed eye-gaze tracking system. In this part all the methods are presented without specific relation to the applications. The attention is devoted to the performance of the system in terms of accuracy and robustness.

Chapter 5 describes the applications developed, in the fields of *communication*, *environmental control* and *rehabilitation*. Here the logical thread of the work is represented by the concept of usability, in terms of efficiency and satisfaction of the user.

In chapter 6 the experiments are extensively described, and results on both performance of the methods and usability of the applications are presented.

The thesis ends in the chapter 7 with a general discussion of the work, a revision of the hypotheses and some directions for future work.

2 The context

2.1 Assistive Technology and disabilities

The aim of Assistive Technology is basically eliminating or reducing the handicap consequent to a disability. To clarify the link between disability and handicap it could be useful to go and get some definitions as stated by the WHO [7]:

- an *impairment* is any loss or abnormality of psychological, physiological, or anatomical structure or function;
- a *disability* is any restriction or lack (resulting from an impairment) of ability to perform an activity in the manner or within the range considered normal for a human being;
- an *handicap* is a disadvantage for a given individual, resulting from an impairment or a disability, that limits or prevents the fulfilment of a role that is normal (depending on age, sex, and social and cultural factors) for that individual.

Except for the case of rehabilitation, an assistive device doesn't aim to reduce the disability, but to find alternative ways to make the person overcome the *disadvantage* caused by his/her disability, i.e. the handicap. This is commonly done by exploiting other abilities that haven't be affected by the impairment, namely "residual abilities".

The most common physical impairments relevant to HCI affect the upper body, and include [8]:

- missing or misshapen limbs;
- impairments affecting bone and joint mobility as a result of arthritis, Parkinson's disease, or repetitive stress injuries;
- muscle weakness and/or paralysis as a result of amyotrophic lateral sclerosis (ALS), multiple sclerosis (MS), cerebral palsy, muscular dystrophy, or spinal cord injuries;
- involuntary movements or difficulty controlling voluntary movements as a result of ALS, MS, brain injury, cerebral palsy, Parkinson's disease, or a stroke.

Each pathology is related to one or more residual abilities, depending on the nature and the progress of the pathology. The main problem of assistive technology is to convey an adequate amount of information through a channel that is different from the typical one. As an example, any conversation is normally carried out through spoken words, but the same function can be performed by the language of signs, if the ability to speak is compromised. In such a case the technical aid is represented by the sign language technique: the handicap (impossibility to have a conversation) consequent to a disability (loss of the human speech) is overcome by exploiting a different ability (motor control of hands) that normally is not used for that purpose.

The situations of handicaps consequent to physical impairments are countless, while the residual abilities belong to a relatively smaller ensemble. Among them the eyes have the peculiarity to be not affected by the majority of the motor pathologies. The voluntary control of eye movements thus represent a valuable residual ability. From that observation the idea arises of exploiting this channel of interaction in order to realize interface systems that can be reasonably applied to a vast number of different cases.

2.2 Eye-driven interfaces

All the persons that have good control of the eyes and head could use an eye-gaze tracker (EGT). As a tool for managing a personal computer, eye-driven interfaces are principally addressed to people with complete absence of upper limb voluntary control, since using eye movements for active control is less intuitive than using hands or other parts of the body. Being outside the assistive technology environment, eye-gaze tracking has several applications principally in research fields as Cognitive Science, Psychology, alternative Human-Computer Interaction, Marketing and Medicine (neurological diagnosis).

In the last three decades many efforts have been devoted to develop assistive devices controlled by the movement of the eyes. Different approaches have been proposed [9] and most of them have been proven as accurate, and some of the techniques gave rise to commercially available products [10, 11]. Ideally, these devices could be useful for very wide variety of pathologies. Unfortunately, a quantity of unsolved problems prevent these system to be used diffusely, e.g. the cost and some usability issues.

The requirements that an ideal EGT should satisfy to be applied as an assistive technology device in the interaction with general computer interfaces [12,13] are:

- tracking accuracy and reliability;
- robustness to light conditions and head movements;
- non-intrusiveness (i.e. cause no harm or discomfort);
- real-time implementation;
- minimal burden time for calibration;
- cost-effectiveness.

Common solutions are based on the presence of intrusive devices, such as chin rests to restrict head motion, and some of them need the user to wear equipment, such as ad hoc contact lenses [14], electrodes [15] or head mounted cameras [16], that restrict considerably the natural movements (Figure 2.1). These devices are likely to be rejected by the majority of potential users because of their low comfort.

Figure 2.1 – Examples of more or less intrusive EGTdevices.

Remote eye gaze trackers (REGT) represent an effective non-intrusive solution and are thus becoming prevalent in the context of assistive technology. With these systems, no devices are in physical contact with the user, since the measurement system is represented by a set of one or more cameras placed far-away (i.e. remote) from the user. On the other hand, since they are based on high-technology, they usually have a high cost and require expensive customer support.

Examples of commercially available REGTs are MyTobii [17], ERICA [18], EyeGaze [19], EyeTech [20] (see Figure 2.2). Basically they exploit the capacity of reflection and refraction of the eyeball if illuminated by appropriate kinds of light. The behaviour of the reflected light is analyzed through video inspection and computer vision techniques, and the direction of gaze is then estimated. Some REGT systems analyze the movements of only one eye, while others takes into account both eyes. Detailed information about the technical functioning of REGT systems will be given in the next chapter.

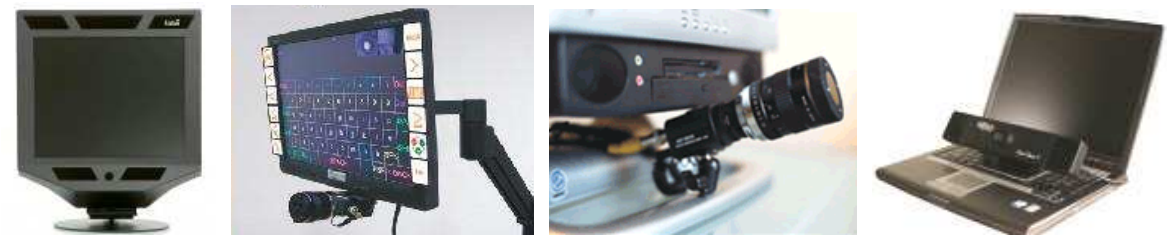

Figure 2.2 – Remote eye-gaze trackers (REGT) as stand-alone solutions, or to be applied to personal computers and laptops. a) MyTobii; b) EyeGaze; c) ERICA; d) EyeTech.

Practically, these systems always work with screen-based application so that estimating gaze direction is transposed to estimating the coordinates of the observed point on the screen. The basic practical purpose of an eye-gaze tracker is to make possible to drive a pointer to the desired location on the screen, i.e. emulating the mouse. For this reason they are also called *eye-pointers*. They are usually supplied with some software for the management of the traditional operative systems (e.g. Windows) and ad-hoc applications and graphical interfaces specifically addressed to the disability. Depending on the user needs and on the typology of disability, this basic eye-pointing functionality can be enhanced and /or considerably changed in order to explore different fields of application.

2.2.1 The problem of communication

The most common impairments to which eye-gaze tracking is addressed are ALS, cerebral-palsy, spinal cord injury and locked-in syndrome. In this cases the eyes are the only way to communicate with the external environment. For these persons, communication is one of the major problems of their life. In order to give aid in this direction, there is the field of Augmentative and Alternative Communication (AAC) defined as “any method of communicating that supplements the ordinary methods of speech and handwriting, where these are impaired” [21]. The idea of *augmentative* communication is to use to the full the residual communication abilities of the impaired person, in order to bypass and/or compensate for areas of impaired function. *Alternative* communication makes use of modalities of communication that are alternative to the speech language, using different codes that could substitute or integrate the alphabetical system, i.e. figures, photos, symbols, etc...

The core of the problem resides in the c*ommunication medium*, that should solve the question of how the meaning of the message can be transmitted. This can be ‘unaided’, for instance by using gesture, facial expression, signing, etc., or it can be ‘aided’, where the persons communicate using some sort of device other than their body, for instance via a communication chart, or an electronic device.

Eye-gaze tracking is used as an high-tech communication medium, and, as all the ‘aided’ mediums, presents some advantage and some disadvantage.

The biggest advantages of aided communication are the flexibility and the richness of communication that can be achieved by creating and/or customizing vocabulary sets and

providing users with special means of accessing them, making them accessible to very young children, non-readers, and individuals with severe intellectual and sensory disabilities. The biggest disadvantage of aided communication is the equipment itself. Having to remember and carry objects around with you, inevitably means something can get forgotten / left behind / lost / broken. Sometimes equipment can be bulky, or heavy, and often it may be very expensive, and there is always the possibility of technical failures. In addition, the computer-based interfaces carry all the problems related to software usability, as understandability, attractiveness, stability, etc..

In the following paragraphs, the principal applications of eye-gaze tracking for people with disabilities will be cited and briefly described. They concern the main activities that can be achieved by means of a personal computer:

- communicate with humans through words;
- control classic computer applications;
- control the household environment.

2.2.2 Eye-typing

When the disability does not affect the cognitive system, the most effective way to communicate is through words. Eye gaze tracking can be of help in this way, through applications that permit to “write with the eyes”. Such systems are called *eye-typing* systems. The two more important factors are the graphical aspect of the interface and the strategy of writing.

The first and more easy solution is the standard QWERTY keyboard (Figure 2.3). Displaying it on a monitor screen, through an eye-gaze tracker the user activates the desired letter, just staring at it.

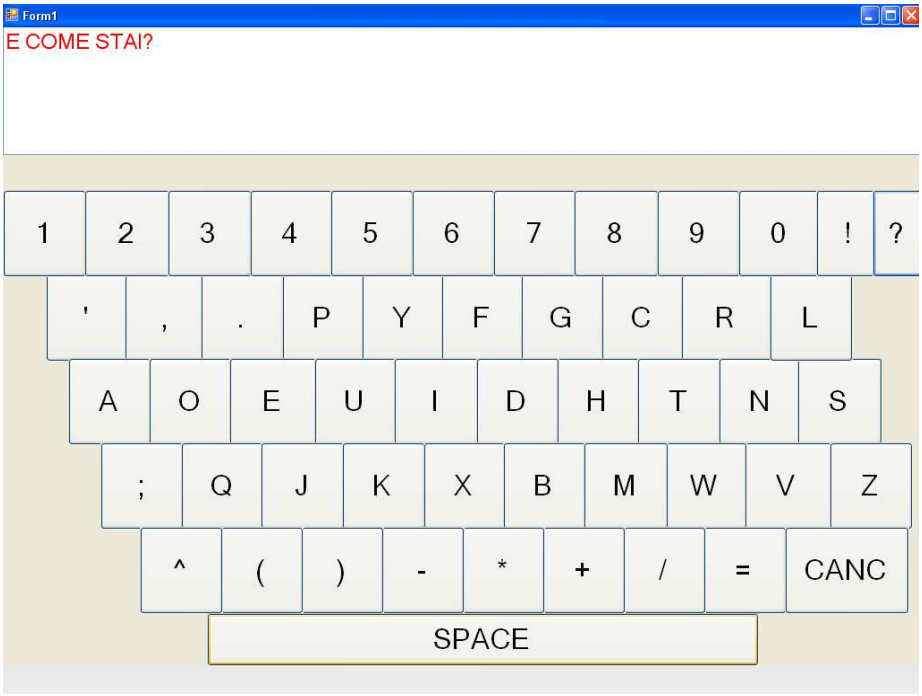

Figure 2.3 – The standard QWERTY keyboard.

This is not at all the most effective solution, since the QWERTY disposition of the letters has been design for writing with both hands.

Other kinds of systems have been proposed in order to improve writing speed and usability in general. Let's briefly analyze three of these solutions, i.e. the "Dasher" [22], the "GazeTalk"[23] and the "UKO-II"[24] systems.

Dasher utilizes a control model that permits to do the basic writing operations. The input is a pointer on the screen monitor, controlled by the mouse or by the eyes. To write with Dasher the user has to focus his/her attention on particular zones on the screen. These zones are dynamic, they change in colour, number and position, depending on the previous selected letters. The algorithm gives priority, i.e. visibility, to the letters that are more probable to be selected. Each time a letter is selected, all the others disappear and a subset of new letters compared, ordered by frequency of use (Figure 2.4).

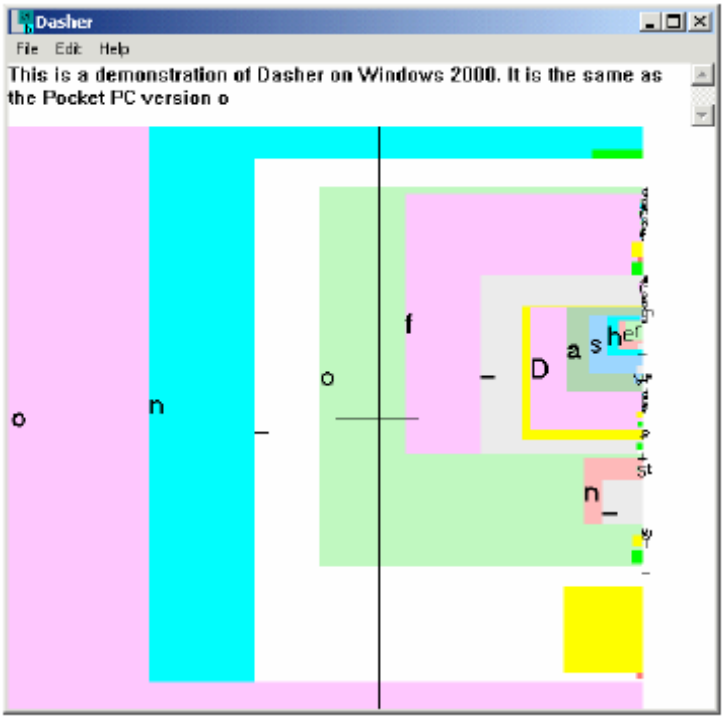

Figure 2.4 – The Dasher eye-typing software.

This algorithm permits to achieve high typing rates, e.g. 15-25 words per minute, but requires good skills of coordination and concentration to maintain in time good performance [25].

Contrary to Dasher, the interface of GazeTalk is very simple (Figure 2.5): the zones are static and the spatial resolution is very low, permitting the use of the interface also with low accuracy eye-gaze trackers. The peculiarity of the system is to have algorithms, based on Markov models, which suggest the user the letters that are most likely to be selected. The typing rate, 6-15 is quite lower than Dasher, but the system in general results really easy to use [25].

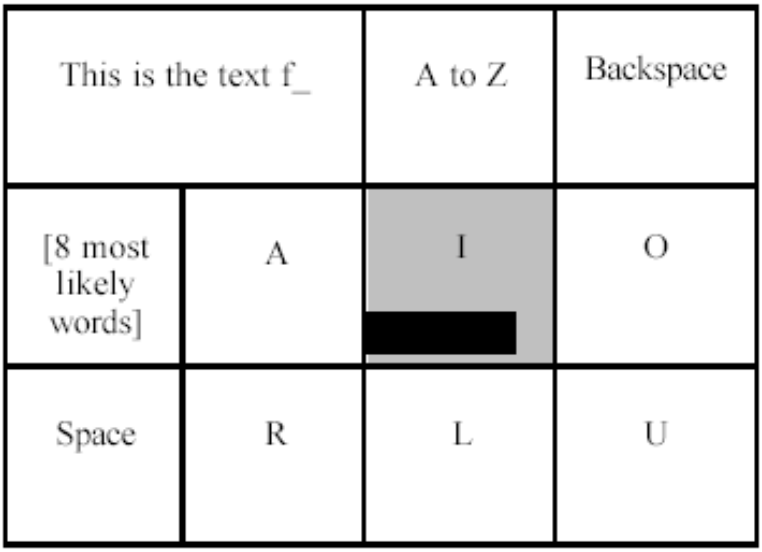

Figure 2.5 – GazeTalk.

The graphical interface of UKO-II is made of a keyboard split into 4 zones. All the letters are visible. A zone is devoted to display the prediction of the word and another one for text representation (Figure 2.6). A genetic approach has been used to minimize the length of the suggested words list. UKO-II is suitable for person with cerebral palsy, because of its simplicity: only four button can be achieve directly. While composing the word, a suggestion of the words, called “ambiguous” is placed on the right. Once the user decides the word, he/she has to voluntarily activate by gaze the desired word among the list of ambiguous ones. UKO-II represents an effective solution between Dasher and GazeTalk in terms of prediction capability and ease of use.

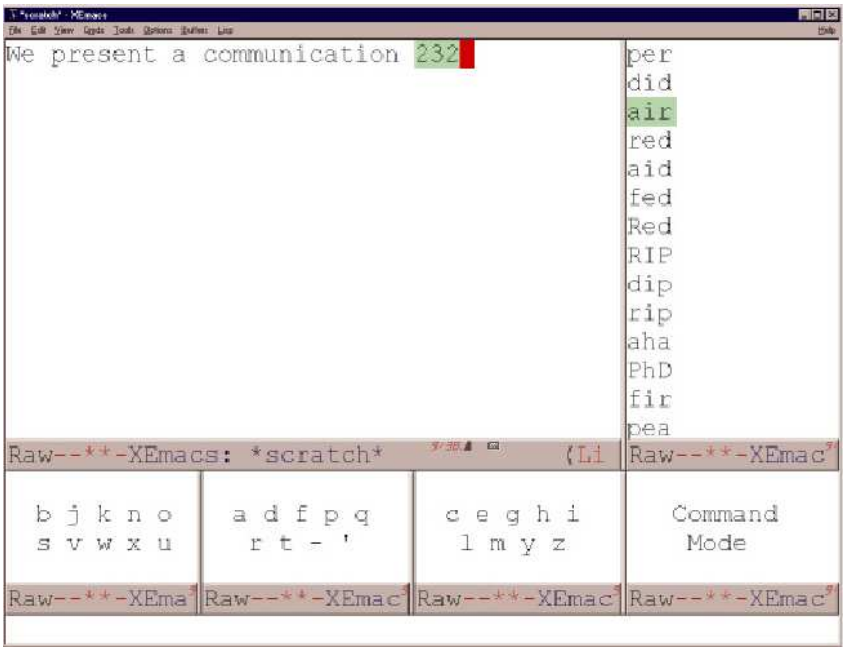

Figure 2.6 – UKO-II.

2.2.3 Mouse emulators

A mouse emulator is any valid alternative to the physical mouse interface. It makes it possible the control of the mouse pointer by using other kind of abilities. It can be hardware or software and it usually works in parallel with the real mouse, in order to permit its contemporary use. An eye-driven mouse emulator is usually embedded in any commercial eye-gaze tracker. It allows managing the traditional applications of a personal computer, by simply drag the pointer with gaze.

The principal problems related to the eye-driven mouse are:

- *accuracy*: usually the accuracy and precision of an eye- gaze tracker are lower than that of the traditional mouse driven by hands. Therefore, normally larger icons are needed. As an alternative, methods for enlarging the selected region are used;
- *trembling*: as a consequence of inaccuracy is the uncertainty of the estimated position, resulting in a tiresome trembling of the pointer. Smoothing filters are applied to make the trajectory of the pointer be more stable and similar to that of a real mouse.

A mouse emulator can represent a simple and easy solution to the problem of accessibility to the personal computer. On the other hand, specific software solutions that do not need mouse emulators are usually preferred, because the users are not very familiar with utilizing their eyes as a continuous pointing device. The motor control strategy of the eye is very different from that of the hands: eyes works much faster and usually anticipates the movement of the hand. “Wait” for the pointer to reach the desired point can result less easy than just stare at an icon waiting for its selection.

2.2.4 Domotic systems

The term “domotics” is the contraction of the word “domus” (Latin for home) and “robotics” automation. Commonly, this neologism is utilized to indicate the automation of some functions in the house by means of electrical devices. The aim is to improve the quality of life of the human beings in household environments, improving domestic organization, security and comfort.

Domotics can contribute considerably to enhance the level of autonomy of the persons affected by motor disabilities. The European Community posed the accent on user-centred approaches by promoting the 6th Framework call on “ageing ambient living societies”, with the goal of providing support to elderly and disabled people in their in-home everyday activities.

Eye-gaze tracking can represent the technology that allow the user to communicate with the appliances to be controlled. According to Bonino et al. [26] the main concern, in this scenario, in the design process of domotic systems based on eye-gaze tracking, efforts should be devoted to:

- usability of the graphical interface;
- reliability of gaze tracking system to changes in the environment (light, movements of the user);
- cost-effectiveness.

In their work they developed a domotic system using a head/eye tracker based on commercial webcams and by adopting a house manager system able to interface different domotic networks. Preliminary results give positive feedback on the feasibility of the

approach and on the capability of the application to improve the interaction between disabled users and household environments.

2.3 Usability

In the design process of an interface, usability is the main key factor to be taken into account. According to ISO 9241 [27], usability is defined as:

*"the extent to which a product can be used by specified users to achieve specified goals with **effectiveness**, **efficiency** and **satisfaction** in a specified **context** of use"*.

Many International Standards have been proposed as an objective form to evaluate interfaces. These concepts are concretized by ISO/IEC 9126-1 [28], and they usually referred to multimedia or software applications. According to Norman [29], one of the major experts in usability, "engineers commonly expect users to understand an application as well as the they themselves do; however, engineers form a complete conceptual map of the product before it is even developed but are often less expert than intended users in the tasks the application will help users perform". The role of the user's environment, defined as persons and institutions, is revealed to be particularly significant. In particular the user must be present in the whole process providing a constant feedback.

2.3.1 Effectiveness, efficiency and satisfaction

Effectiveness is the first key factor of usability. It can be defined as the accuracy and completeness with which specified goals are achieved. Often, during the design of a device, a typical technology-centered approach is followed, and the effectiveness is wrongly considered the most important, if not the only, key factor. ISO/IEC 9126-1 collects some terms which concretized the effectiveness as accuracy, reliability, suitability, stability and security.

Efficiency can be defined as the ratio between the output and the input of any system. It is related with the resources used to get a goal. In the case of human computer interaction, it principally relates to how fast a user accomplish tasks once he or she has learned to use a

system. ISO/IEC 9126-1 describes this term as time behaviour, resource utilization and complexity.

Figure 2.7 – Usability as the trade-off between effectiveness, efficiency and satisfaction.

The last key factor of usability is the user's *satisfaction*, defined as the comfort and acceptability of use, and long term, health and well being. Regarding to satisfaction, technological acceptance appears to remind that sophisticated devices can result difficult to use and generally additional functionality often increases interface complexity, decreasing the accessibility. From this point of view, it is recommendable to keep the simplicity of the interface as much as possible. Cognitive problems play a relevant role because people with cognitive impairments may have difficulty understanding unfamiliar tasks while their previously learned skills may become more interference than assistance [30]. ISO/IEC 9126- provides a list of subterms for satisfaction, including understandability, adaptability, install ability and attractiveness.

A usable interface is a system that represents a good trade-off between effectiveness, efficiency and satisfaction (Figure 2.7).

2.3.2 The user's context

The definition of usability underlines the importance of the context of use. The user's context consists of several groups of influence defined as "users, tasks, equipment and the physical and social environments in which a product is used" [27].

In order to have a complete perspective of the user-centered scenario, the principal agents composing the user's context, as depicted in Figure 2.8, will be briefly described.

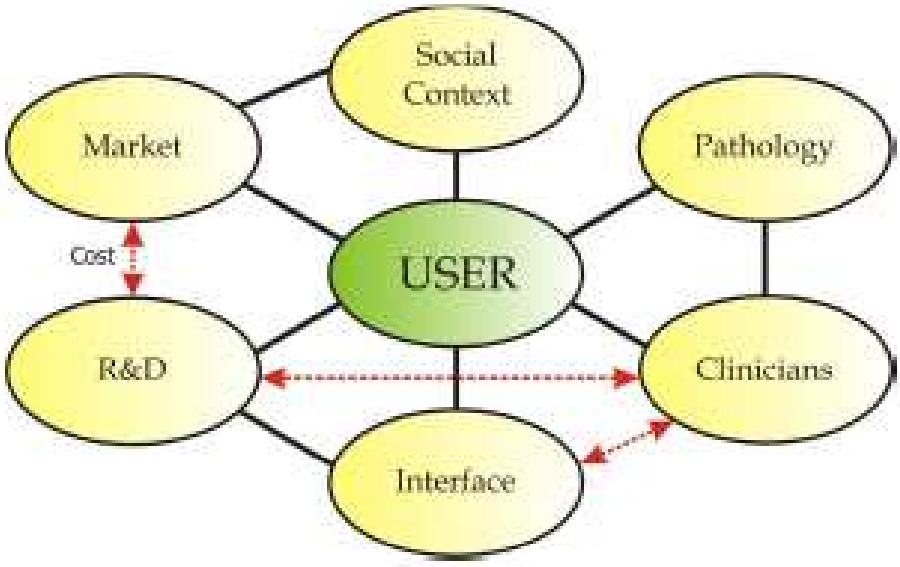

Figure 2.8 – The user's context, with existing connections (in black) and critical issues (red arrows), specifically concerning the design and development processes of the interface.

- **User.** Understanding the user's needs is the essential input to the design process. The participation of users provides relevant data based on their experience and expectations. Users should take part in development and test process providing an interesting feedback to improve details of the prototype. This is an iteration process which should be repeated until a desired outcome is achieved.

- **Pathology.** The study of the nature of the disease allows developing interfaces more adapted to user's needs. Pathology should not be considered only from a physiological and functional point of view, but also under its capacity of influencing user perception and experience, that represent the key factors for satisfaction and usability in general.

- **Social context.** The social context reflects the relationships of the user with the society, intended as family, friends and work community. The quality of this relationship is strongly influenced by the opportunity of freely acting/reacting to the environment. From this point of view, the psychological acceptance of the technological aid plays a crucial role. Often happens that an assistive device that fulfils most of the functionality requirements is doomed to be rejected because of psychological and social involvements.

- **Clinicians.** Professionals who assist elderly may focus the solution on the real needs of the users no matter what is technologically feasible. In this sense, multidisciplinary is essential since the first steps of design. Usability of a technical aid from the clinician/care giver point of view should be also taken into account, particularly when a long term external support is envisaged.

- **Market.** The most limitations for the diffusion of AT solutions are nowadays related to the cost. Whereas technology becomes increasingly sophisticated, investments addressed to commercialize products for people with disabilities do not reach the same level.

- **Research & Development.** The biomedical engineer or researcher should be more and more aware of the multidisciplinary nature of the problem in order to create a useful interface. Neither the efforts should be devoted exclusively to the performance and effectiveness of the device, nor the design should be driven only by technology.

With respect to the Figure 2.8 the red lines represent critical relations that often lose importance during some stages of the process:

- the relation between market and R&D facilitates the post commercialization, taking into account mainly the relation between technology resources and cost;

- clinicians and researchers or engineers should work together (multidisciplinary);

- clinicians and caregivers should be familiarized with the interface because, in many cases, people with strong disabilities will learn using it through them.

3 Eye-gaze tracking technology

3.1 Basic concepts

With reference to human computer interaction, eye-gaze tracking aims at estimating the point of the computer screen the user is looking at. Remote EGT (REGT) is given an additional requirement, that is the impossibility of using any intrusive or wearable device. As a consequence, the existent REGT solutions make use of measurement systems based on video capturing (see Figure 3.1).

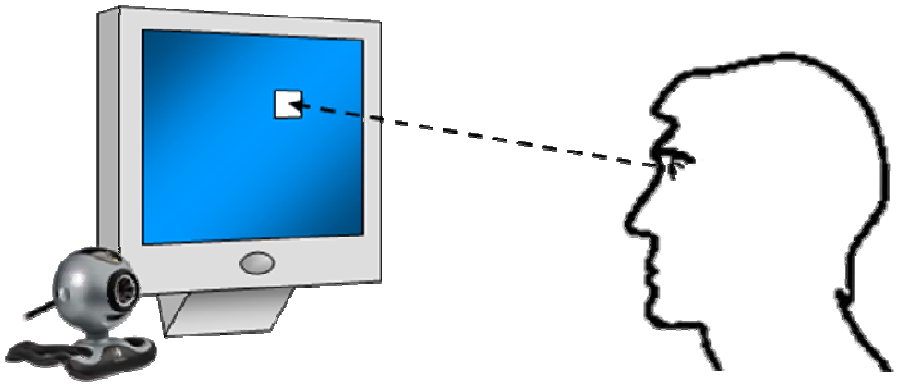

Figure 3.1 – The basic aim of any eye gaze tracker: estimating the point of fixation on a computer screen.

This paragraph presents the basic algorithms that the majority of eye gaze trackers follows, while the next paragraph will deepen the specific technical solutions, highlighting the peculiarity of each technique.

In the majority of view-based REGT systems, the process for the estimation of gaze direction can be broken down into two successive phases: image processing and gaze mapping (see Figure 3.2), that come into action after a preliminary task, i.e. image acquisition, that serves to generate a bi-dimensional digital representation of the 3D reality, making it available for the successive analysis.

The image analysis block gets the multimedia information and extracts the visual features of interest, i.e. position of points of the image that contain relevant information on the geometrical configuration of the eyes. Typical features of interest are the irises, the pupils, the corners of the eyes, the eyelids and reflections of the light on the surface of the cornea. Depending on the eye-gaze tracking method, some kinds of features rather than others are privileged.

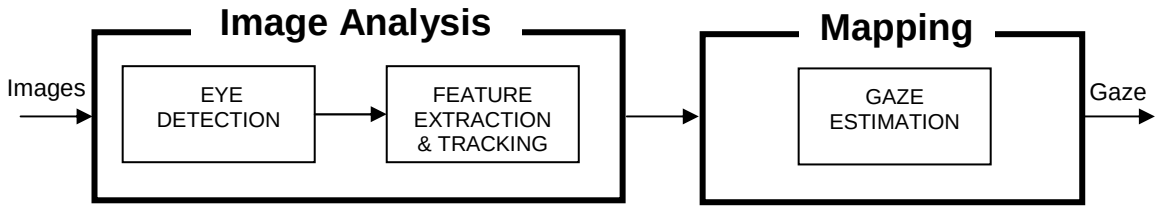

Figure 3.2 – A general eye-gaze tracking algorithm. The image analysis block gets the approximate location of the eyes and then extracts and track the features of interesting, while the mapping block calculates the coordinates of the observed point on the screen.

The task of gaze mapping is to process the geometrical relations between the features in order to calculate the direction of gaze, that is the coordinates of the observed point in the screen reference frame. To realize an accurate mapping between the space of the image representation and the space of the gaze representation, it is necessary to deal with two issues, that is the typology of the mapping function and the procedure of calibration. The choice of the typology of the mapping function can fall upon either deterministic solutions, where the physics of the problem is modeled, or non-deterministic methods, where the function is built and optimized through training procedures based on experimental data. The calibration procedure is used to set the parameters of the mapping function in order to assure that the results of measurement will meet specified criteria, e.g. concerning accuracy and

robustness. Calibration is usually performed by asking the user to look at a point moving on the screen on a specified path and then applying iterative processes to set the coefficients of the function (see Figure 3.3).

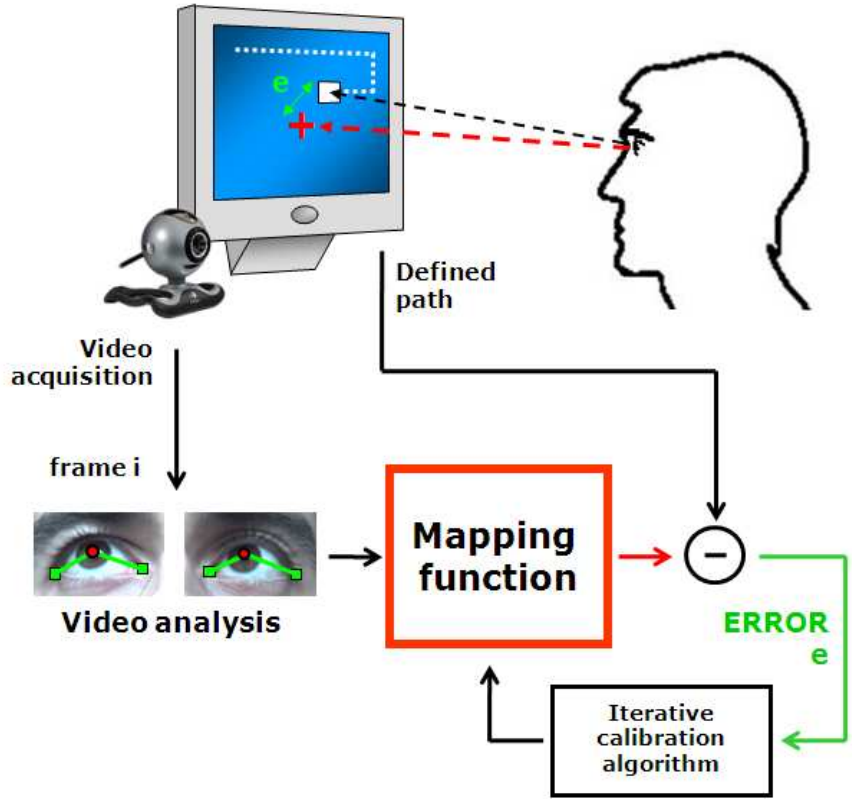

Figure 3.3 – The calibration procedure. While the cursor is moving on the screen on a predefined path, the mapping function calculates the gaze (red arrows). Through an iterative process, the error (depicted in green) between the exact and the estimated position is used to adjust the internal coefficients of the mapping function.

3.2 REGT technology

Among the REGT systems, two classes of approaches exist: one widespread and effective solution is based on active infrared (IR) illumination [31, 32, 33, 34]; the second relies on the analysis of videos capturing eye movements under natural light conditions, and is commonly referred to as view-based or appearance-based [35, 36, 37]. Both approaches are described in the following.

3.2.1 Infrared-based REGT

The IR-based techniques utilize active illumination from infrared light sources to enhance the contrast between the pupil and the iris. The light beam produces two effects on the eye. The first one, called bright-eye effect, is similar to the red-eye effect in photography: the camera records a bright circle, which is the light passing through the pupil and reflected by the retina. The second effect is the reflection of the light on the corneal surface (the external surface of the eye), seen by the camera as a small bright point called "glint" or "corneal reflection" (Figure 3.4). Assuming that the eye is a sphere that rotates around its centre, the glint does not move with the eye rotation, whereas the pupil does. For this reason the glint can be considered as a reference point.

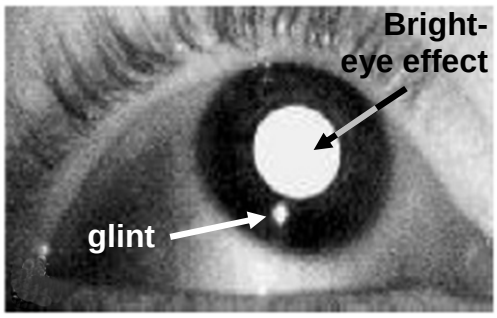

Figure 3.4 – The infrared-based scenario. The bright-eye effect and the glint can be easily detected and their relative position is used as the input of the mapping function.

After grabbing the eye image, pupil and glint can be detected by appropriate image processing algorithms (e.g. [38]). Gaze detection is then computed by applying a mapping function that, starting from the 2D pupil-glint vector, calculates the point observed on the screen. To this latter aim, different mapping functions have been proposed [39, 40, 41]. Among them, the most commonly used is based on a second order polynomial function defined as:

$$\begin{cases} s_x = a_0 + a_1 x + a_2 y + a_3 xy + a_4 x^2 + a_5 y^2 \\ s_y = b_0 + b_1 x + b_2 y + b_3 xy + b_4 x^2 + b_5 y^2 \end{cases}$$

where (sx, sy) are the screen coordinates, and (x, y) is the pupil-glint vector. A calibration procedure is thus needed to estimate the 12 unknown variables {a0, …, a5, b0, …, b5}. To perform the calibration, the user is requested to look at N points on the screen. Since each calibration point defines 2 equations, the system is likely to be over constrained with 12 unknowns and Nx2 equations, and can be solved e.g. through least squares analysis.

The accuracy of the aforementioned IR-based methods corresponds to a standard deviation of about 1-1.5 degrees, which represents a good figure for the requirements of the vast majority of the interactive applications. Moreover, the calibration is relatively fast and easy, and the required computational time allows real-time interaction. Thus, IR-based REGT systems are prevalent in both research and commercial environment. Nevertheless, several important issues have yet to be solved, and at present hinder IR-based REGT systems usability:

- changes in light conditions, especially in presence of sun light, which interferes with the infrared illumination, generating other kinds of reflective phenomena [9, 13, 42]. Some solutions have been proposed for this issue [38];
- bright eye effect variability in different subjects. This effect has been verified as preventing IR-based techniques from providing consistent and reliable results in gaze direction estimation [10].
- High cost of the technology. Most of the commercially available REGTs cost several thousand euros. There's agreement that EGT systems may find use as future computer input devices only if they become convenient and inexpensive [43].

In order to overcome these drawbacks, different classes of REGT systems has become frequent in research literature, i.e. view-based REGT. These will be described in the following.

3.2.2 View-based REGT

In the classical view-based approaches, intensity images of the eyes are grabbed by traditional image acquisition devices and then processed to extract information on the eye configuration. No ad-hoc hardware is usually needed. On the other hand, more challenging

efforts are required in terms of image processing, as compared to the IR-based REGT systems, in order to detect face and eye features.

In the work by Baluja and Pomerleau [35] no explicit features are required. Each pixel of the image is considered as an input parameter to the mapping function. A 15x40 pixels image of the eye is processed, corresponding to a 600-dimension input space vector. An Artificial Neural Network (ANN), trained to model the relationship between the pixel values and the 2D coordinates of the observed point on the screen, is used as the mapping function. In the calibration procedure, the user is requested to look at a cursor moving on the screen on a known path made of 2000 positions. An accuracy of 1.5 degrees is reported.

Similar to the previous work, Xu et al. [37] present a neural technique as well. The image of the eye is segmented to precisely locate the eye and then processed in terms of histogram normalization to enhance the contrast between the eye features. As in Baluja and Pomerleau's method, the ANN receives 600 image pixel values and returns a vector of possible (x,y) coordinates of the estimated gaze point with an accuracy of around 1.5 degrees. 3000 examples are used for the training procedure.

Tan et al. [36] proposed a method that considers the image as a point in a multi-dimensional space, by using an appearance-manifold technique: an image of 20 x 20 pixels can be considered as a 400-component vector in a 400-dimensional space. In this work, 252 images from three different subjects were taken, using each one as the test and the others as the manifold. The calibration is done by looking at markers on a screen, while taking the picture of the eye for each position. The reported accuracy, as calculated by the leave-one-out approach, is very good (0.38 degrees).

Zhu and Yang [41] propose a method for gaze estimation based on feature extraction from an intensity image. Both the irises and the inner corners of the eye are extracted and tracked with sub-pixel accuracy. Then, through a linear mapping function, they calculate the gaze angle, using only two points for the calibration. The reported accuracy is 1.4 degrees.

3.3 The problem of head motion

Compensating for head motion still represents the greatest limitation for most of the remote eye gaze trackers [44]. In principle, as the head moves from its original position, a new calibration is needed. Thus, a strong requirement for the system to work properly is to maintain the head perfectly still. At the same time, this requirement is by far very restrictive

for most of the applications. Several researchers over the last years have dedicated their efforts to the solution of this issue, with the goal of allowing the user to move the head freely, yet maintaining the accuracy in an acceptable range.

To compensate for head movements, most of the related methods are based on a 3D model of the eye [31, 34, 45, 46, 47]. In such methods, two or more cameras are used to estimate the 3D position of the eye. Through the use of mathematical transformations, the line of gaze is then computed. In some of these systems [32, 45], pan and tilt cameras with zoom lenses are required to follow the eye during head movements. An original approach proposed by Yoo and Chung [34] uses a multiple-light source that gives rise to a calibration-free method. Accurate results are also obtained by Park [48], which uses a three-camera system to compensate for head motion.

Even if these approaches seem to solve the problem of head motion, the complexity of the proposed systems, driven by the need of additional hardware, prevents them from being routinely used in a home environment.

To accommodate this need, Ji and Zhu proposed a fairly simple method to compensate for head motion, based on Artificial Neural Networks (ANN), and a single IR-based camera [3]. They criticize the linear/quadratic mapping functions because they cannot take into account the perspective projection and the orientation of the head. Hence, they proposed a different set of eye features: beside the pupil-glint vector, other four factors are chosen for the gaze calibration to get the mapping function. A Generalized Regression Neural Network (GRNN) is then used to accomplish the task, by using a normalized input vector comprising the six factors. As shown in the Figure 3.5 the GRNN can classify one of 8 screen regions with a success of 95%.

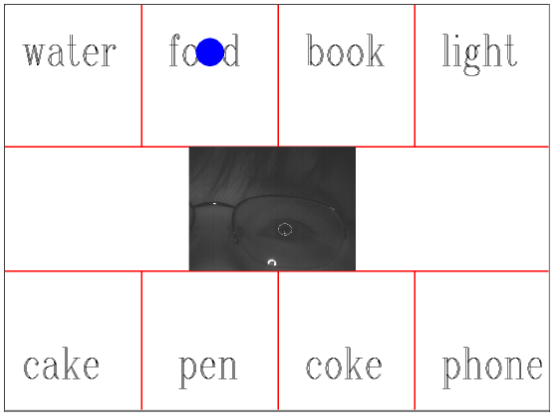

Figure 3.5 – The 8-region interface designed by Ji and Zhu for their neural-based eye gaze tracker.

Solving the issue related to head motion is thus common to both IR-based and view-based REGT systems: if it is likely that this problem is close to the solution for the first ones, work still needs to be done in view-based systems.

3.4 Feature tracking in view-based REGT

Tracking can be defined as the problem of estimating the trajectory of an object in the image plane as it moves around a scene [49].

In general, object tracking can be complex due to:

- loss of information caused by projection of the 3D world on a 2D image;
- noise in images;
- complex object motion;
- non-rigid or articulated nature of objects;
- partial and full object occlusions;
- complex object shapes;
- scene illumination changes;
- real-time processing requirements.

In eye-gaze tracking, the “object” is represented by the eye on the whole and comprehends, in particular, some more specific features whose changes in relative position can describe the movement of the eyeball. In the realm of eye-gaze tracking, accuracy is crucial: small eye movements generate big changes in gaze direction. This problem is even more significant in the case of low resolution images, as happens in webcam-based systems.

The techniques used to track the features should realize a good trade-off between:

- accuracy;
- robustness to changes in shape, pose, size and brightness;
- real-time implementation.

With respect to the IR-based approaches, where the two kinds features, i.e. corneal reflection and glints, represent circular/elliptical shapes, in the view-based approach the scenario is more complex: the image is taken in the visible spectrum, so that filtering processes to isolate the features can't be applied; moreover the features suffer from considerable changes in shape because of the motion of irises and eyelids.

REGT systems accomplish tracking of the eyes by using shape tracking [50, 51], looking for contours in intensity or colour images [52], or exploiting image gradients [53], projections [54] and templates [55, 56]. Among all, the most accurate, but sometimes unstable and computationally expensive, is the technique of template matching.

Template matching aims at finding small parts of an image which match a template image. There are different approaches to accomplish template matching. Some are faster than others, and some find better matches. To enhance the robustness, some constraints can be made, concerning rigid body assumptions and/or restriction on the relative velocities between the features.

3.5 Blink detection

A critical point of both initialization and feature tracking regards the ability to detect the eye position in the image reference system. The use of blink detection as an estimator of eye localization is prevailing in computer vision. By detecting and locating the eyelid movements, blink-based methods are effectively able to initiate and update the eye tracking process (see Figure 3.6). Regardless of the way the gaze tracking systems are put into play, the number of research activities on blink detection are in line with the hypothesis that revealing blink is also a robust way to monitor user intention and interaction.

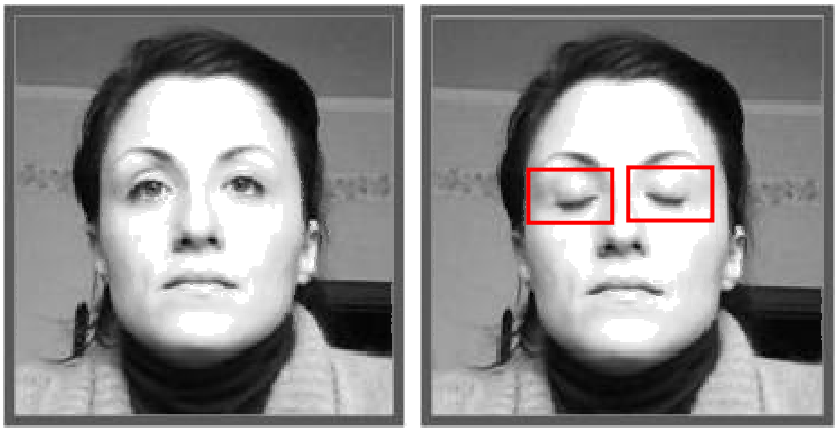

Figure 3.6 – Blink detection can easily serve to detect the location of the eyes by individuating the regions of high movement.

In these regards, Black et al. [57] proposed an optical flow based method and reported a percentage of success in blink recognition of 65%. By combining the optical flow with frame differencing, Bhaskar et al. [58] distinguished between blinking eyelid movements and other eye movements; the reported accuracy is 97% with an image resolution slightly higher (768x576 pixels) than the one of a commercial webcam. Gorodnichy [59] compensates for head motion by introducing the so-called second-order change detection system, based on the detection of blinking: the image difference over three consecutive frames is combined to filter out the linear changes of intensity due to the head movement. The system works in real time with 160x120 pixel images.

The technique called “Blink Link” has been introduced by Grauman et al. [60], where the Mahalanobis distance of the frame difference is computed to locate the regions that more likely represent the eye locations: after extracting an “open-eye” template, blinks are then detected via normalized cross-correlation functions. This technique has been used to determine the voluntary/involuntary nature of blink and a rate success of 95,6% has been achieved. The main drawback of this method derives from the necessity of an off-line training phase for different distances of the user from the camera. A similar approach is presented in the work by Chau and Betke [61], where the open-eye template window is determined through a filtering process based on six geometrical parameters. If the user distance from the screen changes significantly or rapid head movements occur, the system is automatically reinitialized. The system works with low-cost USB cameras in real time at 30 fps. Experiments with tests on eight subjects yielded an overall detection accuracy of 95.3%. Morris et al. [62] propose a real time blink-based method of eye features detection for

automatic initialization of eye tracking applications. Results show a percentage of successful eye-blink detection of 95% with a good robustness to environment changes with images of 320x240 pixels. Nevertheless, the method presents a limitation on anthropometric constraints hypothesis, i.e. the existence of a straight line connecting the four corners of the eyes, that would limit the accuracy of the method. Moreover the presence of fixed parameters would restrict the use to conditions where the distance to the camera remains almost unchanged.

In fact, issues that still need to be tackled mainly refer to the degree of robustness to head movement and to the use of blink detection for accurate and reliable feature extraction. The utilization of the blink detection technique does not necessarily limit to the initialization process (i.e. to determine where the eye is in the first frame), but can help in a number of different situations:

1) calibrate the system. Calibration in this context refers to defining the user-specific features to track. This could be done either manually or automatically and should be performed rarely since it usually needs a certain amount of time;
2) initialize the tracking. The initialization procedure is the step that follows the calibration. It has the aim of creating the actual templates of the features to track, and it is normally performed every time the user starts the session. It should be automatic and fast;
3) reset or update the system: whenever the tracking algorithm “loses track”, the blinking detection algorithm can help reallocate the regions of interest on the eyes, making the tracking restart successfully;
4) help user system interaction, by providing the user with the possibility to control system actions and functions, through voluntary blinking. This could be useful for interactive applications, even if other channels of interaction are usually preferred, since involuntary blinks often occur.

4 The developed tracker

4.1 Introduction

The proposed eye-gaze tracker has been developed to best match with the following requirements:

- non-intrusiveness;
- cost-effectiveness;
- tracking accuracy and reliability;
- robustness to light conditions and head movements;
- real-time implementation;
- minimal burden time for calibration;

Within the REGT scenario, the view-based approach has been considered the most suitable solution for a good trade-off between the over-mentioned requirements. The great advantage of this approach is that no special hardware is needed (just a webcam), making the system cost effective and easy to install. On the other hand, particular efforts are needed in the processing stage, since usually the images grabbed by a traditional camera are

sensitive to changing in light conditions, and the appearance of the face features (like the eyes) are very complex and user-dependent. Here, the accuracy and reliability of the tracking plays an important role, since this is the stage that most affects the global performance of the system.

The approach used to calculate the direction of gaze represents the core of the novelty introduced by this thesis. Since the system aims at a free-head condition, traditional mapping functions have been substituted by new kinds of mapping functions, based on neural approaches. The motivation of that lies in the belief the typical ability of the neural nets to generalize could serve to account for the different sources of variability present in this context.

The proposed eye-gaze tracker is constituted of three blocks (see Figure 4.1).

The first block is about the initialization phase, wherein the subject is detected by the machine and the features of interested are accurately individuated. This phase is based on blink detection, a widely used technique to estimate the blink of a person and here used in a novel way to detect eye location and to individuate the corners of the eyes. Blink detection is also used to re-initialize the system in case of tracking failure and as a channel of command by the user. The blink detection module is an “always present” agent of this system.

The second block is about the feature tracking, i.e. the estimation of the position of the features of interests within each frame of the video stream. The method makes use of traditional approaches, like template matching based on normalized cross-correlation. Slight modifications to the algorithms and some integrations with other image processing methods have been made to increase to robustness.

The third block is about gaze estimation, i.e. the estimation of the point of the real world the user is staring at. This point can belong to a computer screen or be represented by an object in the space. The traditional mapping functions, principally developed for the IR-based scenario, have lead to scanty results as applied to the present view-based approach, in particular with the presence of head movements. Therefore this part of the thesis has been devoted to the design and implementation of other kinds of mapping functions based on neural networks, to principally account for changes in head pose. Within this phase particular attention has been paid to the design of a fast and easy calibration.

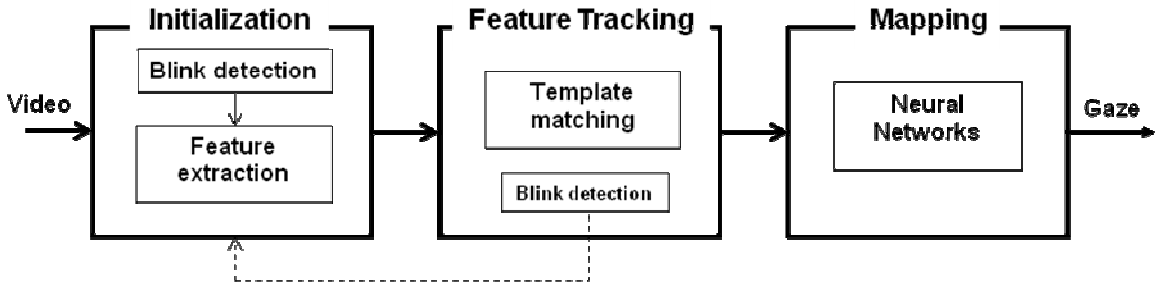

Figure 4.1 - Proposed eye-gaze tracking procedure.

4.2 Initialization

The initialization phase is intended as the preliminary detection of the features of interest within the image, such as head, face, eyes and other more specific features. In the view-based scenario, where no infrared light sources have been used, the point of interests are represented by the irises and one or more reference points on the face, represented, in this case, by the corners of the eyes (Figure 4.2).

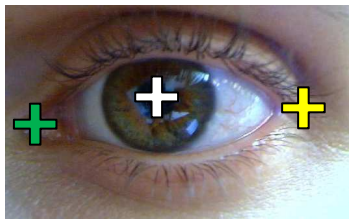

Figure 4.2 - Features of interest.

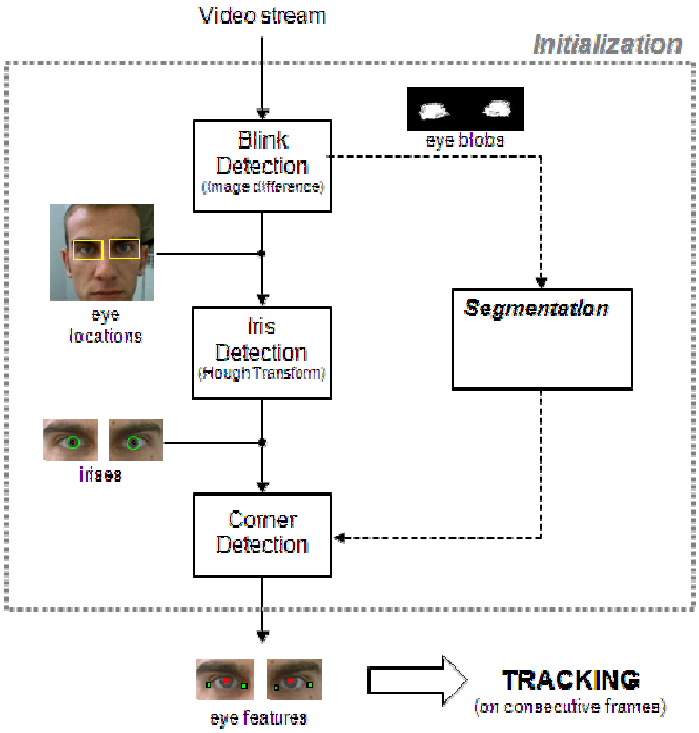

Figure 4.3 - The initialization procedure. Three consecutive steps have been adopted and integrated in the initialization procedure: blink detection, iris detection and corner detection. This procedure allows the system to extract the feature templates to be used by the subsequent tracking procedure.

The proposed initialization algorithm is a method to find eyes, irises and eye-corners in the very first frames of the video stream. The method does not require a specific position/orientation of the head in the space. The only requirement is to maintain the head still for a very short lapse of time (a couple of seconds), during which the user is requested to blink in front of the camera.

The algorithm is composed of three main blocks (see Figure 4.3): i) a blink detection module, that individuates the eyes and extracts information on the shape of the eye sockets, ii) an iris detection module, that finds the irises through a circle recognition method and iii) a corner detection algorithm, that merge information from irises and shape of the eyes, estimating the location of the internal and external corners of the eyes, by means of image segmentation methods.

4.2.1 Blink detection

The proposed blink detection method is able to detect blink with no specific requirements on head pose and distance of the subject from the camera. The only constraint imposed to the user is to maintain the head still during blinking in front of the camera.

The algorithm is based on a simple operation of frame differencing among couples of consecutive frames. Considering ***I****(x,y,t)* the image grabbed by the webcam at time (frame) t, the difference image ***D****(x,y,t)*, representing the regions where changes of intensity level occur, is obtained as follows:

$$\boldsymbol{D}(x,y,t) = |\boldsymbol{I}(x,y,t) - \boldsymbol{I}(x,y,t-1)|$$

$\boldsymbol{D}(x,y,t)$ represents the starting point of a sequence of checks that are performed in order to determine if it contains a blink event or not.

STEP 1) Thresholding.

The first step aims to eliminate low levels of changes, caused either by noise or negligible movements. An automatic thresholding, based on mean value and standard deviation of the non-zero pixels, is performed: any pixel whose value is less than the threshold is set to zero, while the pixel values greater the threshold are set to 1, returning the binary image ***M****(x,y,t)*:

$$\mathrm{M}(x,y,t) = \begin{cases} 1 & if\ \mathbf{D}(x,y,t) \geq \mu(\mathbf{D}) + 8 \cdot \sigma(\mathbf{D}) \\ 0 & \text{otherwise} \end{cases}$$

STEP 2) Filtering.

The image ***M****(x,y,t)* is then undergone a kind of filtering process that eliminates small areas resulting from noise. A median filter of size 7x7 pixels is placed only over the non-zero pixels. The pixel where the filter is centered is given a value corresponding to the number of pixels in the neighborhood, divided by the size (7^2) of the filter. This value, is then binarized using a threshold, defined empirically, obtaining the binary image ***F****(x,y,t)*:

$$\mathbf{F}(x,y,t) = \begin{cases} 1 & tf \ \frac{\sum_{i=x-3}^{x+3} \sum_{j=y-3}^{y+3} \mathbf{M}(i,j,t)}{7^2} \geq \varphi \ and \ \mathbf{M}(x,y,t) = 1 \\ 0 & otherwise \end{cases}$$

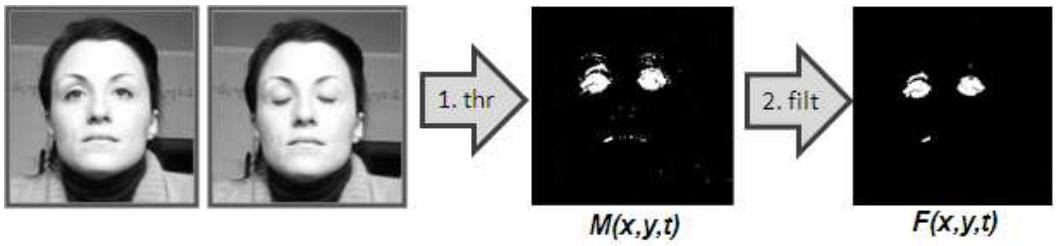

Figure 4.4 - link detection. The image difference between two consecutive frames is followed by a gray-to-black and white conversion and a filtering process to detect the presence of a blink.

STEP 3) Clustering.

The image $\mathbf{F}(x,y,t)$ contains a certain number of white regions, namely blobs. Due to not uniform changes in pixel values during the eyelid movements, each blob could appear fragmented in smaller zones (namely sub-blobs), very close to each other. The aim of this step is to label correctly each sub-blob, deciding to which cluster it belongs to. This is done through projecting all the sub-blobs over the x-axis and subsequently over the y-axis and giving the same label to those blobs that are closer to each other than a predefined length, used as a threshold. The value of this threshold is based on the total length of the projection of the blobs over the x-axis; no erosion-dilation are used since original difference image would not change.

The projection over the x-axis is represented by the vector $\boldsymbol{H}(x,t)^{\mathbf{H}(x,t)}$, obtained by the following operation (see Figure 4.5a):

$$\boldsymbol{H}(x,t) = \sum_{j=1}^{Y} \boldsymbol{F}(x,j,t)$$

where Y represents the height of the image $\boldsymbol{F}(x,y,t)$.

The clustering threshold $\tau(t)$ is set to:

$$\tau(t) = \frac{\max\left(\mathbf{sgn}\left(\mathbf{H}(x,t)\right) \cdot \mathbf{x}\right) - \min\left(\mathbf{sgn}\left(\mathbf{H}(x,t)\right) \cdot \mathbf{x}\right)}{T_c}$$

where the numerator represents the width of the global white zone (considering all the blobs) and " · " denotes the vector multiplication. If the distance between two separate sub-blobs is lower than $\tau(t)$, then they are considered belonging to the same cluster. T_c represents the ratio between the threshold value of the distance, i.e. $\tau(t)$ and the maximum width. The greater T_c, the smaller $\tau(t)$.

The vector $\boldsymbol{H}(x.t)$ is then labeled. The horizontal label vector is defined as follows:

$$\boldsymbol{L_H}(x,t) = \begin{cases} 0 & if \ \boldsymbol{H}(x,t) = 0 \\ l_H & otherwise \end{cases}$$

where the natural number l_H starts from 1 and increases of 1 each time the distance between the actual value and the previous (with respect to x) non-zero value of $\boldsymbol{H}(x,t)$ is greater than $\tau(t)$.

Thus, for each label l_H a binary image $\boldsymbol{S_{lH}}(x,y,t)$ is created (see Figure 4.5b). $\boldsymbol{S_{lH}}(x,y,t)$ has been called partial clone of the original image $\boldsymbol{F}(x,y,t)$: it's an image that contains only the blobs of $\boldsymbol{F}(x,y,t)$ that are labeled l_H:

$$\boldsymbol{S_{lH}}(x,y,t) = \begin{cases} F(x,y,t) & if \ \boldsymbol{L_H}(x) = l_H \\ 0 & otherwise \end{cases}$$

Each $\boldsymbol{S_{lH}}(x,y,t)$ is then labeled among y direction exactly as done for the x-axis, and a number of partial clone images, namely $\boldsymbol{S_{lH,lV}}(x,y,t)$ $\boldsymbol{S_{lH,lV}}(x,y,t)$, are then generates, where l_V indicates the vertical projection label of $\boldsymbol{S_{lH}}(x,y,t)$.

To sum up, the described clustering process generates N binary images, each one containing only the blobs of $F(x,y,t)$ belonging to the same cluster (see Figure 4.5c).

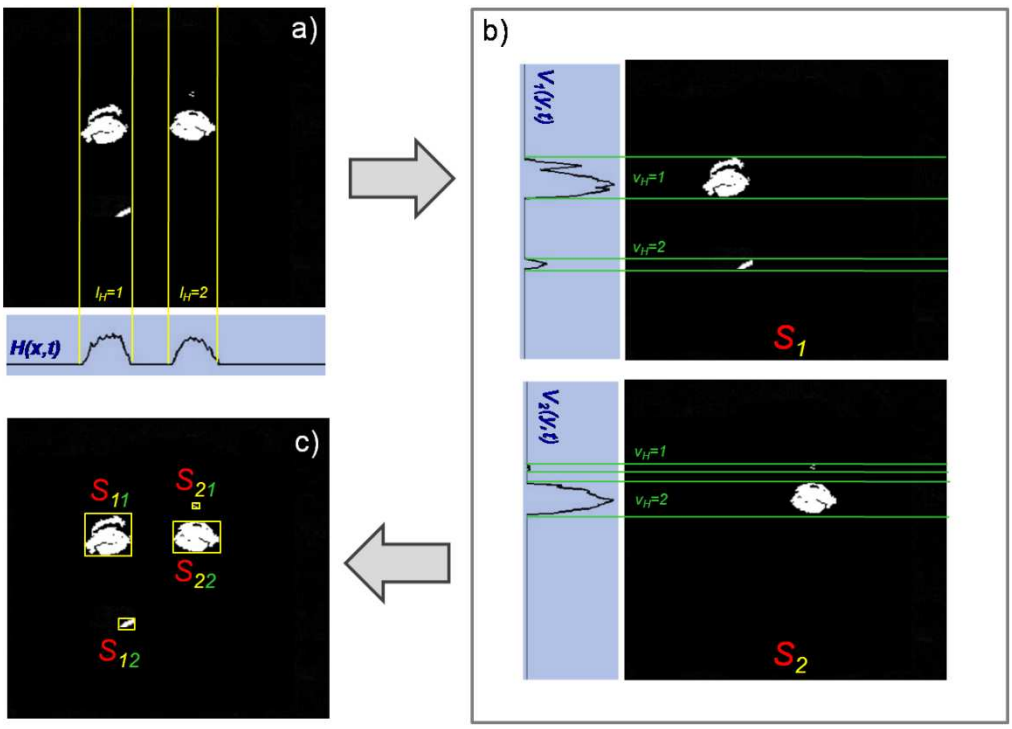

Figure 4.5 – Clustering process. a) The image is clustered through projecting blobs on x-axis. b) The same labeling process, is applied on y direction to each sub-images resulting from the previous step. c) Step a) and b) are merged to finalize the clustering of the blobs.

STEP 4) Single eye morphology check.

This step deals with the analysis of the single blob morphology, i.e. $S_{IH,IV}(x,y,t)$, in order to classify it as "eye" or "non-eye". The classification is based on the ratio between the two dimensions of the clusters. The hypothesis here is that the width of an eye cannot be smaller than the height, nor bigger than three times the height.. Defining w and h as the width and height of the smaller rectangle containing all the non-zero values of $S_{IH,IV}(x,y,t)$, the decision process results:

$$\text{classification}_{\text{result}} = \begin{cases} \text{EYE} & \text{if } 1 \leq \frac{w}{h} \leq 3 \\ \text{NON} - \text{EYE} & \text{otherwise} \end{cases}$$

"Non-eye" clusters, and the corresponding $S_{IH,IV}(x,y,t)$ images, are discarded. The remaining $S_{IH,IV}(x,y,t)$ images will pass through the following, and last, step.

STEP 5) Eye-pair morphology checks.

From this moment the blobs are not analyzed separately. All the possible combination of blobs-pairs is checked to be a possible eye-pair. The checks regards symmetry, human proportions and head pose constraints. The methods have been designed to be independent from the camera-user distance.

The symmetry check is based on the fact that blink is a symmetric movement. Unfortunately, at this point of analysis there's no information about the axis of this symmetry (we don't know the position of the head). Anyway, it can be observed that the size S_1 and S_2 of the two clusters should be not too different (size here refers to the number of pixels belonging to the cluster). A threshold value on the ratio of the two sizes is then used:

$$symmetry_{check} = \begin{cases} TRUE & if\ 0.5 \leq \frac{S_1}{S_2} \leq 2 \\ FALSE & otherwise \end{cases}$$

The check on proportion is designed to account for the anatomy of the face, involving the relation between the distance between the eyes, d, and the mean size of the eyes, S_{MEAN}.

$$proportion_{check} = \begin{cases} TRUE & if\ 1 \leq \frac{d}{S_{MEAN}} \leq 3 \\ FALSE & otherwise \end{cases}$$

S_{MEAN} is defined as the mean value between the two diagonals of the rectangles containing the two clusters, giving a raw estimation of the size of the clusters; d is the distance between the two centroids of the clusters.

The head pose check accounts for the rotation of the head in the plane parallel to the camera image plane. A maximum inclination of 45 degrees is admitted. Mathematically, the inclination is defined as the angle between the line passing through the centroids of the two clusters and the horizontal line.

$$headpose_{check} = \begin{cases} TRUE & if\ \arctan\left|\frac{yB_1 - yB_2}{xB_1 - xB_2}\right| \leq 45° \\ FALSE & otherwise \end{cases}$$

where xB_i and yB_i are the x and y coordinates of the centroid of the cluster i.

The pairs of clusters that satisfy the above morphological checks are considered “eye blinks”. In the case of positive detection of two or more “eye blinks”, the pair with the lowest inclination will be elected.

4.2.2 Iris detection

The aim of this part of the algorithm is to find the initial positions of the irises. This information is essential to give the system the possibility to track the irises in the following frames of the video stream.

By considering the centroids of the two clusters as a first guess estimation for the irises position, it is possible to determine two rectangular regions of interest (ROI), one for each eye (Figure 4.6) , whose dimensions are based on the distance between the eyes. These regions of interest are then used for the successive phases of the tracking.

The gray intensity images of the eyes are then extracted from the last frame of the initialization video and processed by an algorithm based on two techniques, namely edge detection and Hough transform, as following explained.

Edge detection finds the pixels of the image where high contrast occurs, turning out to be appropriate to detect irises, since iris and sclera (the white part of the eye) present high contrast in brightness. Among the different existing kinds of edge detectors, i.e. Sobel, Prewitt, Roberts and Canny [63,64], the Sobel operator accomplishes the task with higher specificity in terms of border detection: even if Canny operator results to be in general more robust to light changes, a very high number of edges is detected within the eye image, making the discrimination of the correct iris edge very difficult. With the Sobel operator a lower number of edges is detected. Nevertheless, thank to its high contrast, the iris edge always belongs to the detected borders, making it possible to automatically choose the correct iris border with Hough transform technique.

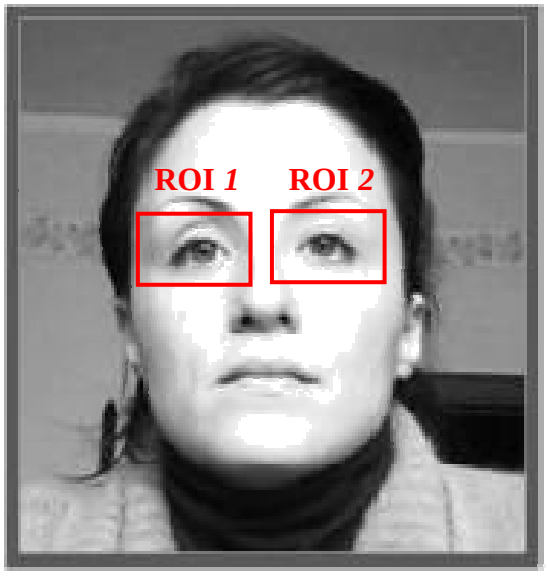

Figure 4.6 – Regions of interest (ROI) obtained by blink detection.

The Hough transform is a feature extraction technique aimed at finding a certain class of shapes, such as lines, circles and ellipses, within a binary image resulting from edge detection [29]. In the case of circumferences, the value of the radius is needed to make the algorithm works. Each circumference detected within the image is given a vote, representing the number of pixels of the image that belongs to that circumference.

In the case of iris detection, since the exact value of the radius is unknown, the algorithm has been applied iteratively for different radius values. The range of radius values is based on a first guess value obtained automatically by dividing the inter-eyes distance by a constant factor, set to 13:

$$\frac{dist_{EYES}}{13} - n \leq R \leq \frac{dist_{EYES}}{13} + n$$

where 2n+1 is the size of the range.

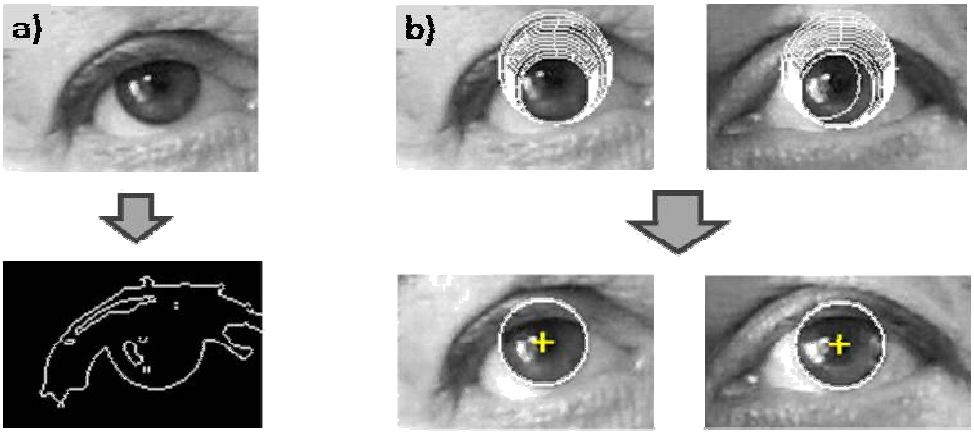

Figure 4.7 – Iris detection. The Sobel operator is applied to detect the edges in the image (a) and a modified Hough transform algorithm is then used to choose the two most voted circumferences among the candidates (b).

For each value of radius, a set of circumference candidates is detected, each one with its vote. To estimate the correct pair from the set of candidates, a modification of the original Hough algorithm has been applied: each group of similar circumferences, in terms of radius and center position, is merged into a new circumference and a new vote is assigned to it, coming from a weighted sum of the single ones. The most voted circumferences of each eye are then compared with the candidates of the other eye to choose the pair of circumference with the biggest total vote (see Figure 4.7). This process has shown a good behaviour over different light conditions even in asymmetric illumination of the face.

4.2.3 Corners detection

To determine the eye corners, a further processing of the results of the procedure of blink detection is needed. As mentioned, the frame-to-frame difference images highlight the zones in which a movement occurs, based on the fact that during a movement the corresponding pixel of the image changes its value. We are interested in getting the shape of the eye as accurately as possible by separating the zones with large inter-frame motion (eyelids) from those with small one. To individuate the appropriate threshold value in an automatic way, a number of consecutive difference images, belonging to the same blink, are summed.

The resulting image is filtered to determine the eye corner position, as explained in the following.

For the inner corner, a search window is generated, not including the iris. Within the window, the mean value and the standard deviation of the values of the image are used to define the threshold for the image binary process. As shown in Figure 4.8 the most lateral pixel of the binary image is considered as the estimated inner corner.

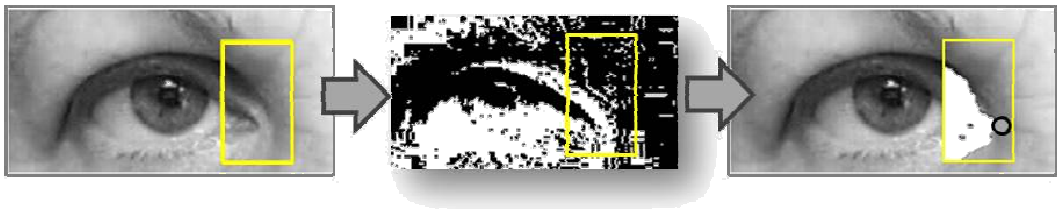

Figure 4.8 – Inner corner detection. From the knowledge of the position of the irises, a search window is created. Then the blobs coming from the blink detection block are filtered to detect the inner extreme of the eyelid movements, i.e. the inner corner of the eye.

For the external corner a search window is created over the external area of the eye, starting from the estimated iris position. 10-level quantization is then applied to the intensity image. By eliminating the brighter levels, the line of the upper eyelid can be easily identified. The external extremity of this shape will be considered as the external corner (see Figure 4.9). In the experimental testing section, the results of this technique, over different light conditions and distance of the user from the camera, will be shown.

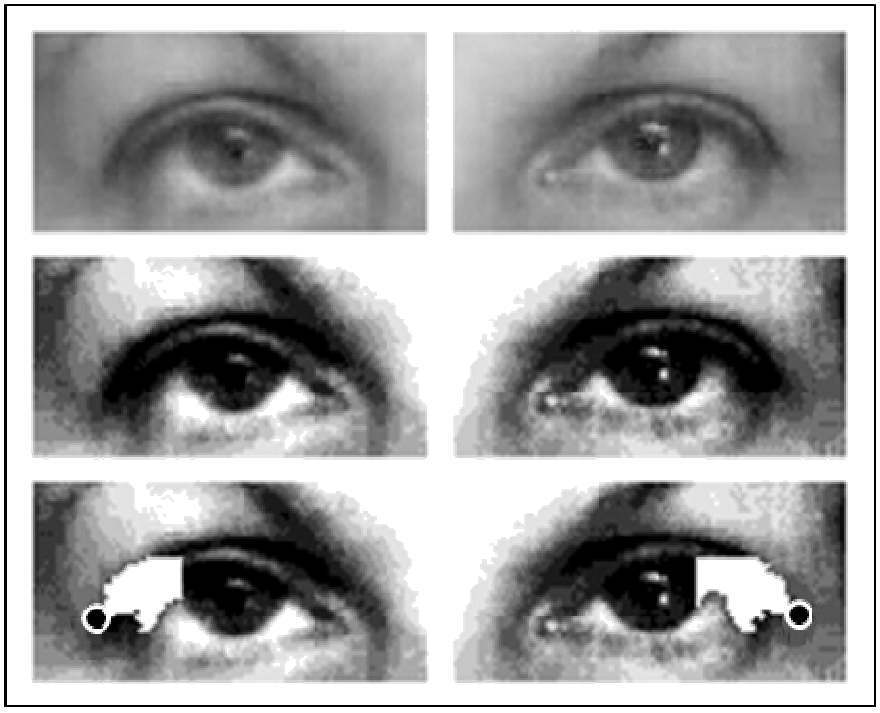

Figure 4.9 - External corner detection. The gray level image of the eyes is enhanced using histogram equalization. A binary image (the white areas) is created by eliminating the pixels with higher luminance. The external extremes are then taken as the external corners of the eyes.

4.3 Feature tracking

As mentioned in the previous chapter, tracking the features of eye is a particularly complex task, since irises and eyelids moves with respect to the head and the head moves with respect to the camera, making the features change in term of translation, rotation, size and shape. Moreover the brightness of the image can change, modifying the appearance of the feature, even without presence of movements.

The tracking methods developed here are based on template matching, that is the technique for finding parts of an image which match a template image. Among the different template matching methods, the normalized cross-correlation [65] (NCC) is particularly appropriate for those applications in which the brightness can vary, due to lighting and exposure conditions.

Normalized cross-correlation calculates for each pixel the similarity between the template and the region of the image under the template, as the following equation expresses:

$$\gamma(u,v) = \frac{\sum_{x,y}\left[f(x,y) - \overline{f}_{u,v}\right]\left[t(x-u,y-v) - \overline{t}\right]}{\sqrt{\sum_{x,y}\left[f(x,y) - \overline{f}_{u,v}\right]^2 \sum_{x,y}\left[t(x-u,y-v) - \overline{t}\right]^2}}$$

where

$f(x,y)$ is the image,

$\overline{t}$ is the mean value of the template,

$\overline{f}_{u,v}$ is the mean of the image $f(x,y)$ in the region under the template.

The crucial point of template matching is the choice of the template. A wrong choice can bring the system to be unstable or not accurate.

4.3.1 Iris and corners tracking

Hough transform, the technique used for initialization, showed low stability as applied for iris tracking. It fails when the iris is positioned sideways, because the boundary between iris and sclera is visible only on one side of the iris, making the searching for circumferences be unstable.

Instead, template matching through NCC showed good stability and accuracy.

A particular kind of template has been chosen, called virtual template. As shown in Figure 4.10I the virtual template is an image with a black circle (representing the virtual iris) over a white background. The size of the template is initially set imposing the size of the black circle equal to the iris detected by the initialization module.

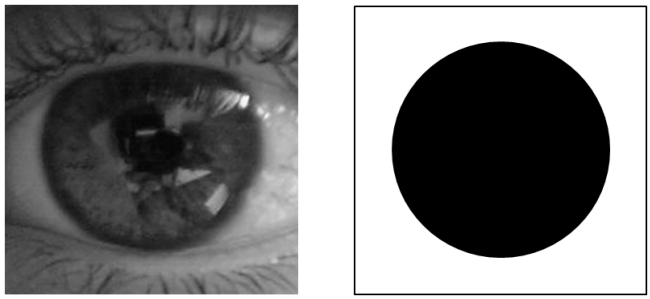

Figure 4.10 – Real iris and virtual template.

To improve robustness to movements toward/away from the camera, the size is dynamically changed, frame by frame, by maintaining the new size of the template proportional to the distance between the eyes.

The cross-correlation does not directly provide the position of the iris, that is instead calculated by analyzing the local maxima of the output function *γ(u,v)*. As a matter of fact, the correlation values does not assume high values, since the virtual template does not represents a real iris and therefore it never matches exactly with the image. As a consequence the highest local maxima have values very close to each other. Thus, sometimes the iris position does not belong to the absolute maximum. For instance, it frequently happens that the absolute maximum falls upon the eyebrow.

Empirical observation ensure that the iris always takes part of the first three greater maxima. The solve the problem of estimating the right maximum, a check on the circular shape of image areas is then accomplished, in order to discard the local maxima that fall upon zones that are not similar to a circle. The word “shape” is not meaningful in case of intensity images as is. Instead, the correlation function $\gamma(u, v)$ contains some useful information. As depicted in Figure 4.11 $\gamma(u, v)$ can be seen as a surface in 3D space. This surface results to be continuous around the maxima values of correlation. If we cut the surface with an horizontal plane in the proximity of the peaks, it happens that the shape of the cross section is circular if the maximum refers to the iris, while a different shape is obtained in the other cases, as for the eyebrow. The shape check is then performed to decide which one of the three greater local maxima of $\gamma(u, v)$ corresponds to the iris.

Differently from the iris, the corner template is directly extracted from the image during the initialization process (Figure 4.12). The resizing process follows the same rules of iris tracking, i.e. the ratio between the inter-eyes distance and the size of template is maintained constant.

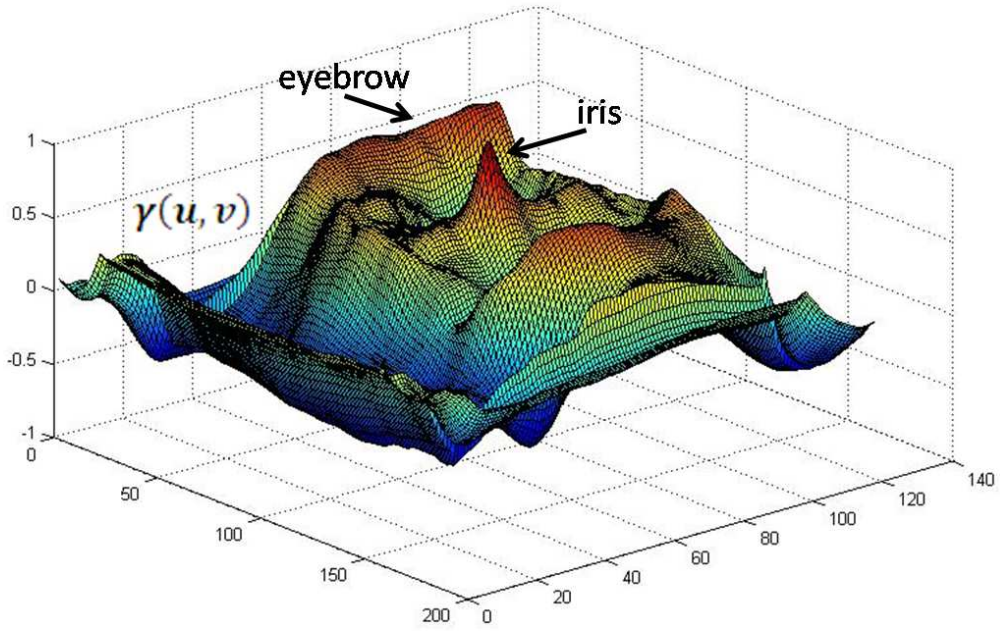

Figure 4.11 – The correlation image. The 3D shape of the surface around the local maximum of the iris is different respect to the other local maxima.

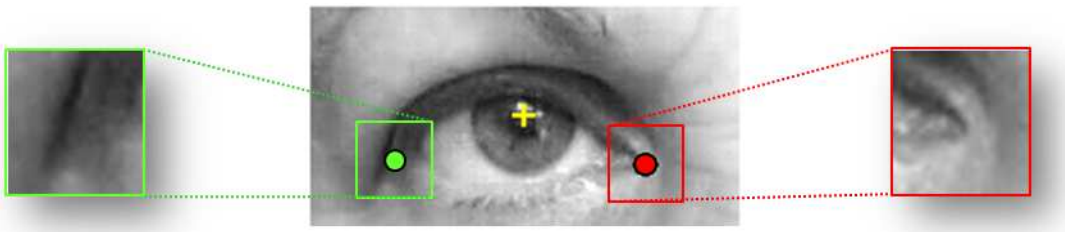

Figure 4.12 – Internal and external corner template extraction.

4.3.2 Sub-pixel optimization

The maximum spatial resolution of a single frame during video capturing by a webcam is 640x480 pixels. Considering that the size of the head fits the image frame, the dimension of one eye is approximately 100x60 pixels. The analysis of the eye movements showed that the range of variation is very small, especially for the vertical displacement,

which is 7/8 pixels. The consequence the low resolution is that even low noise in image acquisition or small errors in image analysis can strongly affect the calibration procedure, i.e. the calculation of the mapping function coefficients. To overcome this problem a sub-pixel technique has been adopted to increase the spatial resolution.

The sub-pixel technique is a mathematical operation applied on the image. A new image (namely the mask in the following) is created, with the difference that the size of the pixels of the new image is not integer, but decimal. In other words, each pixel of the original image, is divided in several pixels of the mask. The mask is then overlapped to the image and, in the area around the estimated feature, the correlation function $\gamma(u,v)$ is interpolated and the new regional maximum is found (see Figure 4.13), making the coordinate of the irises become decimal.

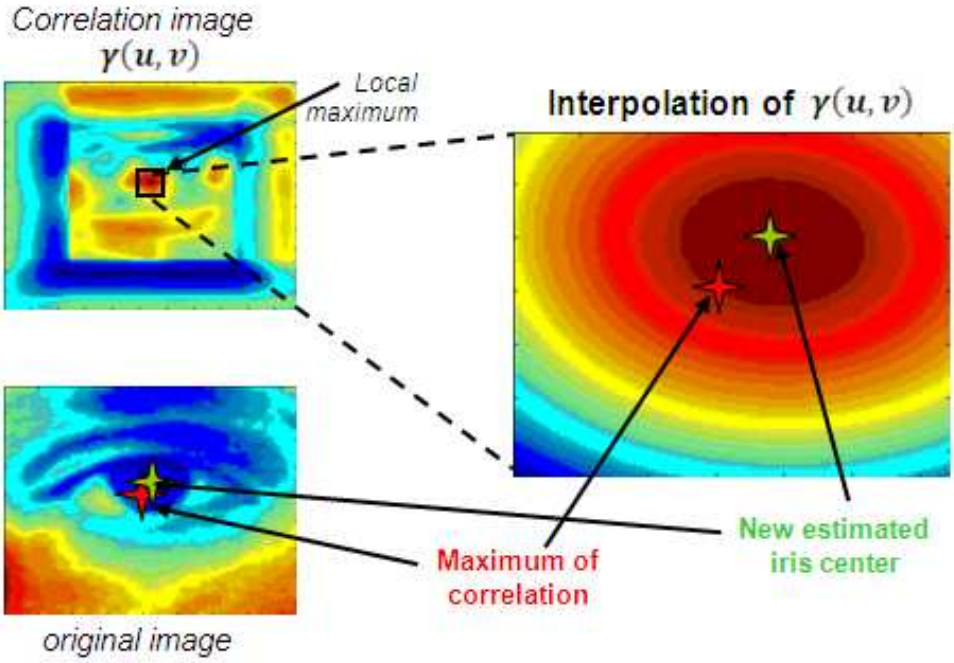

Figure 4.13 – Sub-pixel optimization. The correlation image is interpolated using a mask with a higher spatial resolution. The new local maximum of the interpolated function gives the new position of the iris.

4.4 Gaze estimation

As displayed in Figure 4.14 the procedure for the estimation of gaze is composed of two blocks: the first one, described in the paragraphs above, makes use of image processing algorithms to extract and track the features of the eyes; the second block accomplishes the task of mapping the geometric visual information with the gaze direction, by using artificial neural networks.

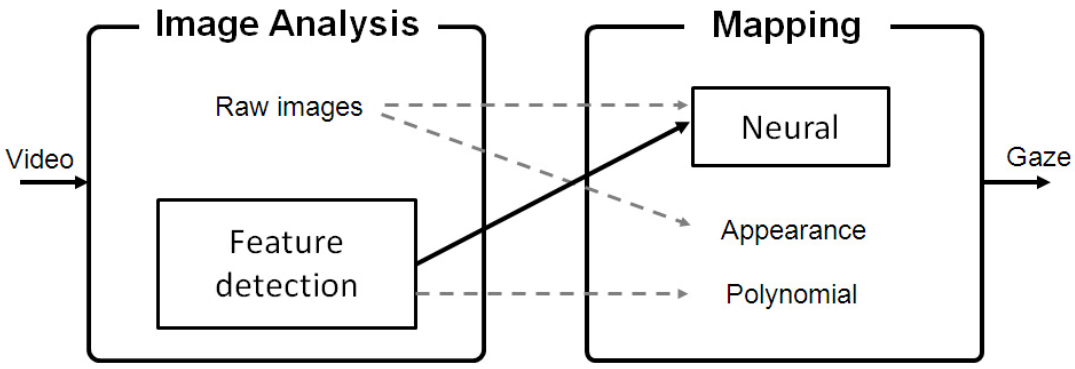

Figure 4.14 – General scheme of a procedure for a view-based REGT system. The dashed arrows indicate the classical view-based techniques. The solid arrow shows the proposed approach, where feature detection is combinedwith neural mapping to make the system robust to light changes and head movements.

The neural approach has been chosen for the ability of the nets to learn by examples, to smoothly adapt to changing contexts and to generalize for unseen configurations. The proposed approach arises from the belief that, once provided with a proper set of input space vectors and training examples, the net will be able to compensate for different head positions. The aim is to leave the user free from cumbersome equipment to maintain the head fixed.

Due to the underlying complexity, the neural architecture requires an accurate design that regards the choice of:

- the typology of the input;
- the internal structure of the net, i.e. the number of hidden layers and neurons;
- the kind of training procedure.

In the next paragraphs the chosen input set, the different nets and the training procedures will be detailed.

4.4.1 Input space selection

The choice of the input set has the role of giving the net the appropriate information about the relative position of the eye within the eye-socket like on the pose of the head with respect to the screen. This information will be combined by the net to build a non linear regression function that takes into account those head movements that can occur during the visual task. Basically two features are needed: one refers to the pupil, and the other one, called reference point, to appropriately describe the movement of the head. The latter has been identified as the corners of the eye.

As opposed to what happens with infrared-based techniques, the image of the eye in visible light spectrum does not permit to identify a reference point on the surface of the eye. Whereas the "glint" of IR-based systems is pretty insensitive to the rotation of the eyes around its optic axis, the vector connecting the eye corner with the pupil turns out to be extremely sensitive to small head rotations. For this reason, and for the unavoidable uncertainty of feature tracking with low-resolution images, a redundancy has been introduced by considering, for each eye, two vectors connecting respectively the pupil with the inner and the external corner. The resulting vector of inputs thus consists of twelve parameters: eight of them come from the magnitudes and angles of the distance vectors previously mentioned; the remaining four consist of the x and y coordinates of the two external corners. All of them are depicted in Figure 4.15.

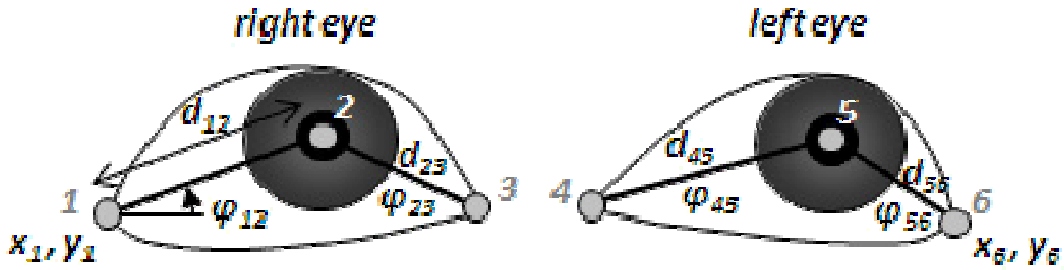

Figure 4.15 – Geometric features of the eyes (numbered from 1 to 6). Magnitudes and angles of the distance vectors represent 8 of the 12 inputs. The remaining 4 inputs are the x and y coordinates of the external corners of the eyelids. For the sake of clarity, only one of the angle features is represented.

4.4.2 The neural structures

Two different architectures of neural networks have been used to explore and approximate the mapping properties of the gaze function. The first net, a multilayer perceptron (MLP) [66], is considered to have a strong ability to provide compact and robust representation of mapping in real-world problems. The other net is represented by a general regression neural network (GRNN) [67], which is recognized to have better performance in function approximation tasks with respect to traditional neural networks. A GRNN has also been used in previous work [14], based on infrared-based techniques.

Both networks have a 12-neuron input layer receiving the vector of the eye parameters. Each architecture consists of two distinct nets that separately calculate the s_x and s_y coordinates of the screen point.

The MLP design aims at finding the optimum configuration with the following variables: number of hidden layers, number of neurons, number and typology of training trials and number of training epochs. To evaluate the best configuration, accuracy is taken as the performance index. The structures of the chosen MLP are displayed in Figure 4.16.

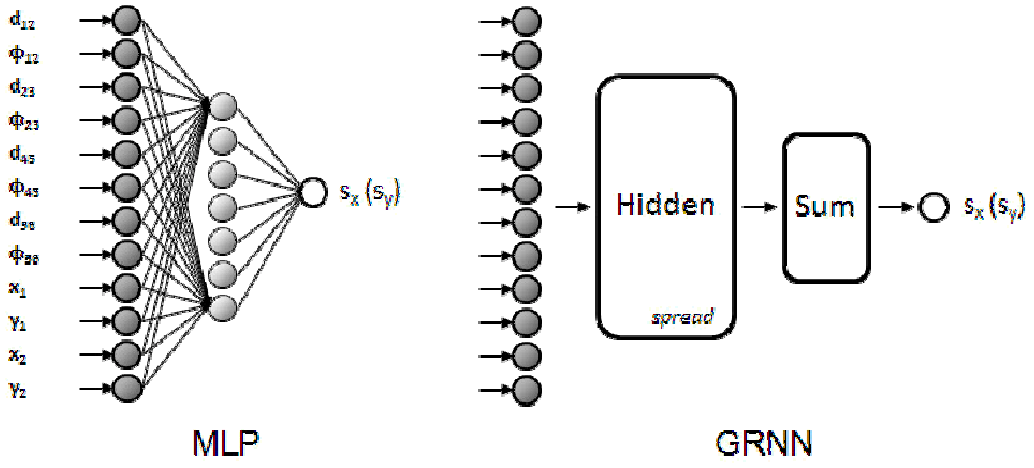

Figure 4.16 – Network architectures. The multilayer perceptron (MLP, left) and the general regression network (GRNN, right), both with a 12 input layer and a single output neuron for the calculation of either the x or the y coordinate of the observed point on the screen.

The GRNN, based on the radial basis architecture, consists of an input layer, a hidden layer, a summation layer and an output layer. The hidden layer has as many neurons as there

are input/target vectors, while the number of nodes of the summation layer equals the number of output neurons plus one. The hidden layer contains a parameter called spread that determines the width of an area in the input space to which each neuron responds: the higher the spread, the smoother the function. The choice of the best configuration lies basically in the optimal estimation of the spread, which constitutes the only user-defined internal parameter.

In order to detect the most appropriate structure, multiple configurations of MLP and GRNN have been tested. The optimization of the performance of each net has been focused on the ability to classify the gaze direction over the zones of the screen.

More than 50 configurations of MLP have been tested, with one, two and three hidden layers and a total number of neurons ranging from 5 to 420. For the experimental results see the chapter 6 .

4.4.3 The training procedure

In the neural realm, training is a fundamental process for making the net "learn" the environment in which it will work. Mathematically, during training the internal weights of each neuron are modified in order to give a confident output.

Every training procedure requires a corresponding input-output set, i.e. for each input, the desired output is to be known. This could be done exactly as for the calibration of a traditional mapping function, asking the subject to look at a cursor moving on the screen over different known positions. To make the net account for head movement, the input set is composed of examples from different trials with changes in head position. During each trial the subject is free to maintain a comfortable position, i.e. natural movements are allowed. Each frame grabbed by the webcam is processed to extract the eye features. For each observed position, a certain amount of grabbed frames are taken to build up the training set, while the remaining set is used to validate the training procedure. In the Figure 4.17 an example of training procedure is proposed, with a cursor moving on a 15 positions path, and 7 grabbed frames for each position. N trials cover a wide area of movements, as displayed in Figure 4.17.

For the MLP a Resilient Back Propagation training algorithm (RBP) has been chosen. At each epoch of training two kinds of errors have been evaluated and compared each other.

The first one is internally calculated by the RBP algorithm at each iteration, and the second one, the validation error, is obtained by using the validation set (2 frame per position) as input. The end of the training is established by comparing these two values with predefined thresholds, ranging between 10-4 and 10-3.

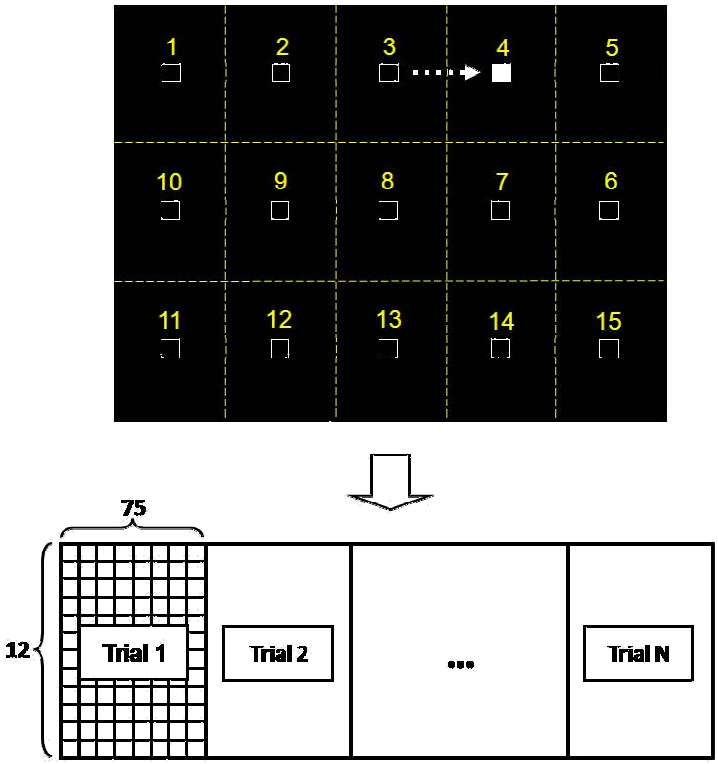

Figure 4.17 – The training procedure. Example with 15 points and 7 frames grabbed per position. The upper panel represents the screen with the 15 positions that the user is asked to look at during one trial. For each position, 5 of the 7 video frames are processed and analyzed to build up a 75 columns (5x15) matrix, each column representing a 12 parameters vector. The remaining 2 frames per position will be used to validate the algorithm.

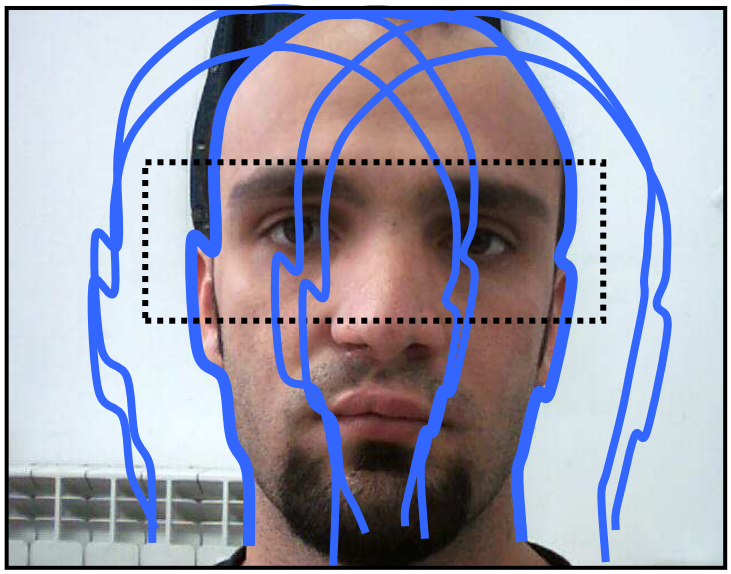

Figure 4.18 - Profiles of head shift as obtained during the training procedure. The dotted rectangle delimits the range of eye movements.

4.5 Final structure of the tracker

In conclusion, the developed interface is based on a view-based eye-gaze tracking, that works with a traditional low-cost video acquisition device (a webcam) without any additional equipment, besides a personal computer with a screen.

The algorithms of feature detection and tracking have been designed to work real-time and to automatically adapt to the user, with no intervention from an operator.

The system automatically re-initialize in the case the tracking fail, e.g. for sudden occlusions, since it is always present a blink detection module that estimates the positions of the eyes of the subject anytime he/she blinks.

The algorithms for the estimation of gaze are bio-inspired. They use neural networks in order to account for slight changes in head position that could occur when the user is managing a computer without devices that keep the head still.

As a consequence of the experimental tests for the accuracy and robustness of the eye-gaze system (see chapter 6), the graphical interface has been set to fit a grid of 15 zones. This kind of spatial resolution permits the system to estimate gaze with very good accuracy. Moreover HCI interfaces used for people with disability usually contain a reduced number

of icons of relatively large dimensions, which can roughly be approximated by dividing the screen in 8÷15 areas.

5 Developed applications

This chapter is devoted to present four typologies of applications that have been designed to be used in the context of eye-gaze tracking. The applications have been designed taking into account the specification of the eye gaze tracker developed, in order to demonstrate that it is possible and worthy using low cost eye gaze tracking for many kinds of applications, since most of them does not necessarily require very high accuracy.

Two of the solutions proposed regard assistive technology for the standard communication needs of disable people, in particular eye-typing and domotics. The third application concerns the field of rehabilitation. It is about a multi-modal platform for upper arm rehabilitation in stroke, in which gaze is taken as an estimator of the intentions of the user in reaching tasks. The aim of this third solution is to demonstrate that eye gaze tracking can be used actively not only as an assistive device, but also for motor rehabilitation. The fourth application is about AAC (augmentative alternative communication) for children affected by cerebral palsy. It is the idea of a system that merges eye tracking with other kinds of sensors, as IMU (inertial measurements units) to permit, through a unique device, to monitor the functional residual abilities of the child and individuate the specific channels of communication to be used in rehabilitation and communication tasks. These fourth idea goes in the direction of using eye gaze tracking in multi-modal systems to account for the variety and the progression of the disability.

5.1 Assistive technology solutions

The first step in the design of the interfaces here developed consisted in setting the spatial resolution of the graphical interface. The spatial resolution depends on the accuracy of the eye-gaze tracker. As depicted in Figure 5.1 the experimental tests of the tracker (see chapter 6) showed good performance with a 15-position interface, i.e. 5 columns by 3 rows. Other experimental tests have been performed by varying the number of zones. The results showed that the 15-position represents a good trade-off between number of zones and reliability. For this reason the interfaces presented in this paragraph have been developed with a 5x3 spatial resolution.

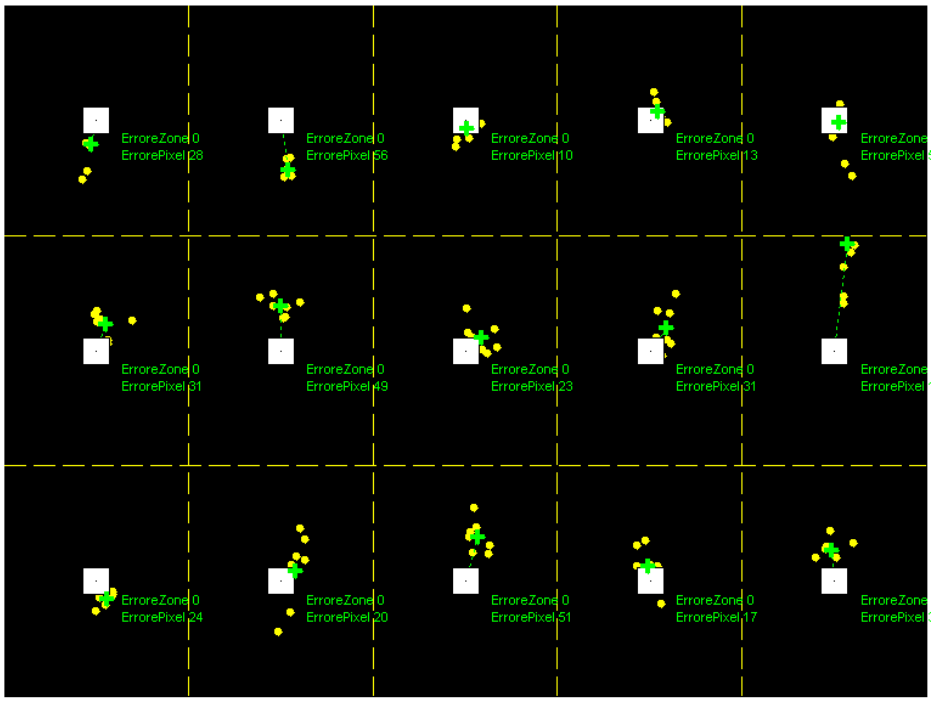

Figure 5.1 – Performance of the gaze tracker on a 15-position interface. The white squares represent the observed points while the dots represent the estimated positions.

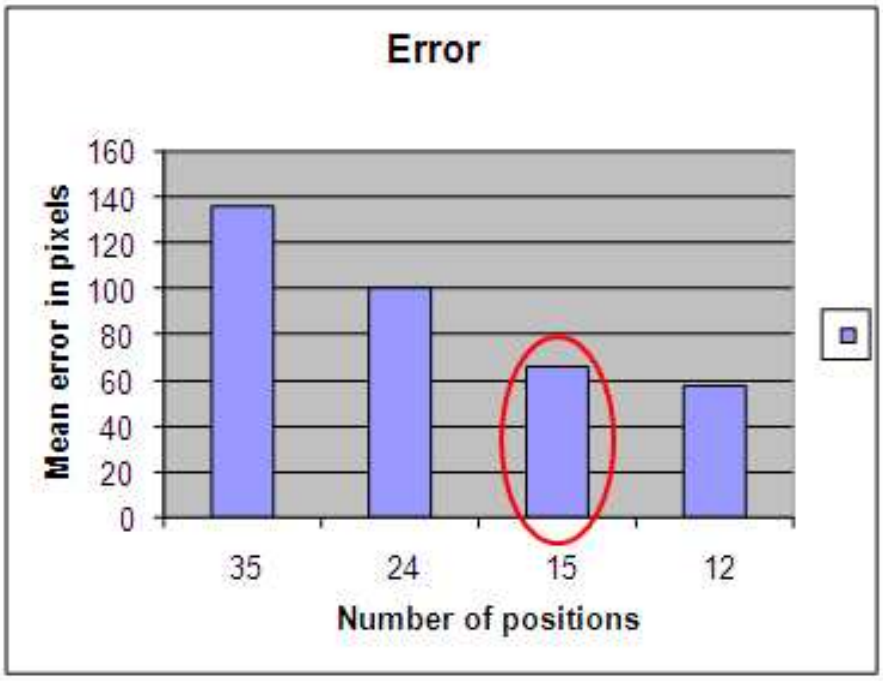

Figure 5.2 – Analysis of the error for different spatial resolutions of the interface.

5.1.1 The Midas Touch Effect

Interfaces based on eye gaze tracking suffer from the so-called Midas Touch Effect [68]. It occurs when an eye tracker is used not only for pointing, but also for clicking, that is the case of all the practical applications. The Midas Touch effect causes users to inadvertently select or activate any target they fixate upon. By issuing a click only when the user has fixated on a target for a certain amount of time (dwell time click), the Midas Touch effect can be controlled, but not entirely removed. The Midas Touch effect can be completely avoided by issuing clicks via an alternate input modality, such as a manual button, voice command or voluntary movement of the eyelids.

5.1.2 Eye Typing

Eye typing systems enable to write a text with the only movement of the eyes.

In the design of such kind of interface, usability is the main guideline to follow. To best design the interface, various users have been involved in the work, by asking them to try the preliminary solutions and feed their opinion/suggestion back. After several attempts

the final version of the graphical interface results as depicted in Figure 5.3. The main important issues concerning the layout and the functioning are listed and briefly explained below:

- *Letter positioning*. Because of the lack of a sufficient number of icons, a grouping of letters was necessary. The letters have been placed as in a telephone keyboard. These choice has been adopted for different reasons. Firstly, for intuitiveness: the telephone device is very widespread, and since 10 years the telephone keyboard has been used to write the short messages (SMS). The technique is well accepted, tested and does not present critical problems. Moreover, often people with disabilities come from a past with normal abilities. Looking at a familiar layout and functioning can enhance acceptability. The letters are all visible in the first two rows. Numbers and punctuation marks can be selected by selecting the first icon.

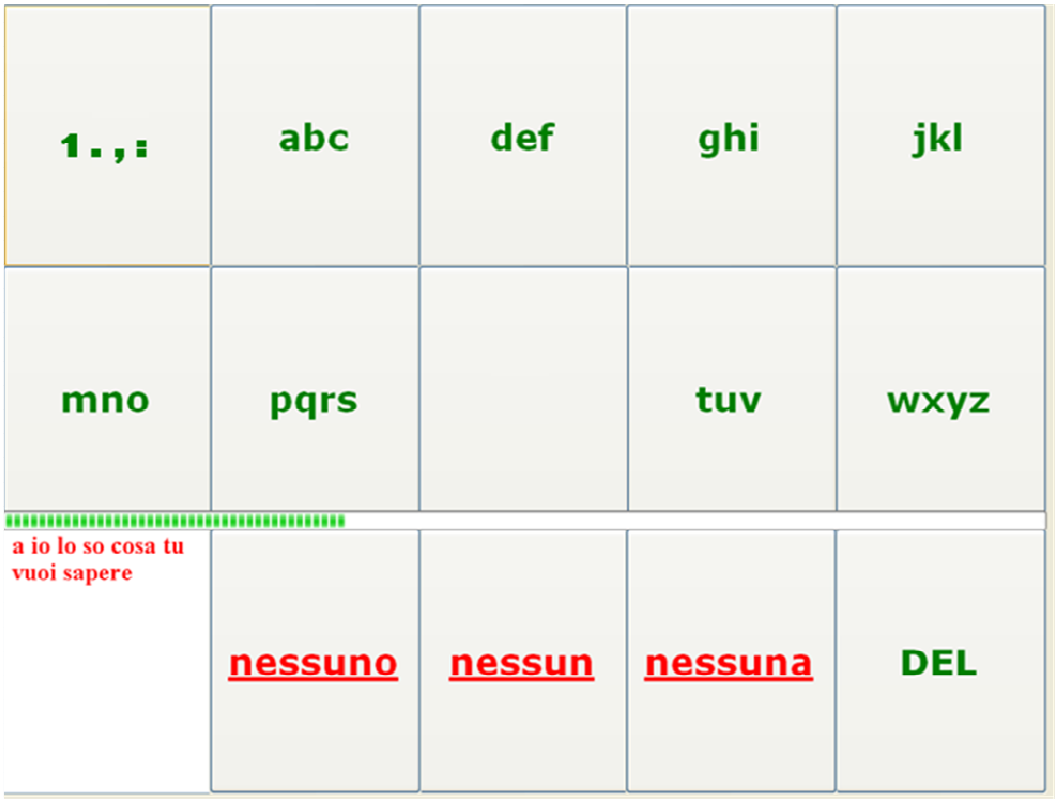

Figure 5.3 – Layout of the graphical interface.

- *Inactive-zones*. A non-active icon is placed at the center of the screen, to permit the user to rest while looking at the screen avoiding the Midas Touch Effect. By a process of voluntary action it is possible anyway to activate this icon to calibrate, re-initialize, and enter higher level commands such quitting the program. Another zone is used to display the text typed. That zone it is non-active so that the user can read the text anytime he/she needs.
- *Predictive algorithm*. The system contains a software for the prediction of the words, in order to increase the writing speed. The algorithm follows the same logic of the predictive algorithms present in mobile phones. Basically it is based on a database of words. Each word of the database is associated to a sequence of numbers (namely code in the following) that represents the sequence of the icons to select to compose the word. For instance, to write the word “Hello” it is necessary to select the code 4-3-5-5-6. While typing the word, the system individuates the words of the database whose code matches with the typed code. In the case two or more words match with the typed code, the system puts the candidate words in the three windows at the bottom of the interface. These three words are chosen by sorting all the candidates by a parameter called *use*. The *use* parameter is a natural number that increases by 1 every time that word is typed in the text. The *use* parameters gives information of the words that are more frequently used by the user, so that these words will appear very soon in the windows of the interface, so that the user can select the word before it is completely typed.

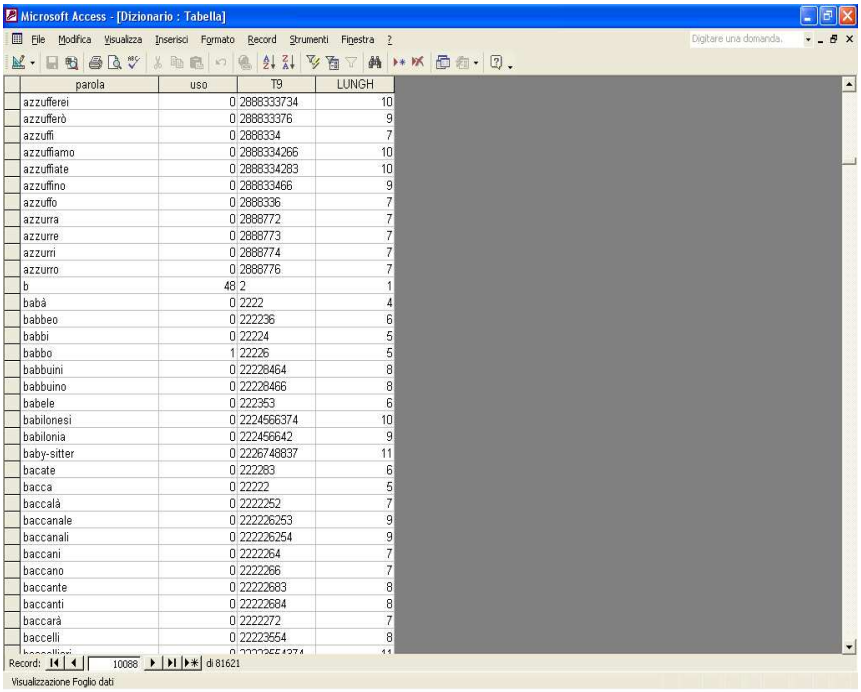

Figure 5.4 – The word database used by the predictive algorithm.

- *Icon selection.* The system adopts a dwell time approach, whose duration can be varied depending to the ability and experience of use of the user. Blinking is used only for higher level actions, e.g. to quit the program.

Figure 5.5 – The predictive algorithm. The code associated to the word "Hello"

5.1.3 Domotics

The objective of this part of the thesis is the design of a graphical interface to be used in the context of a domotic system (Figure 5.6).

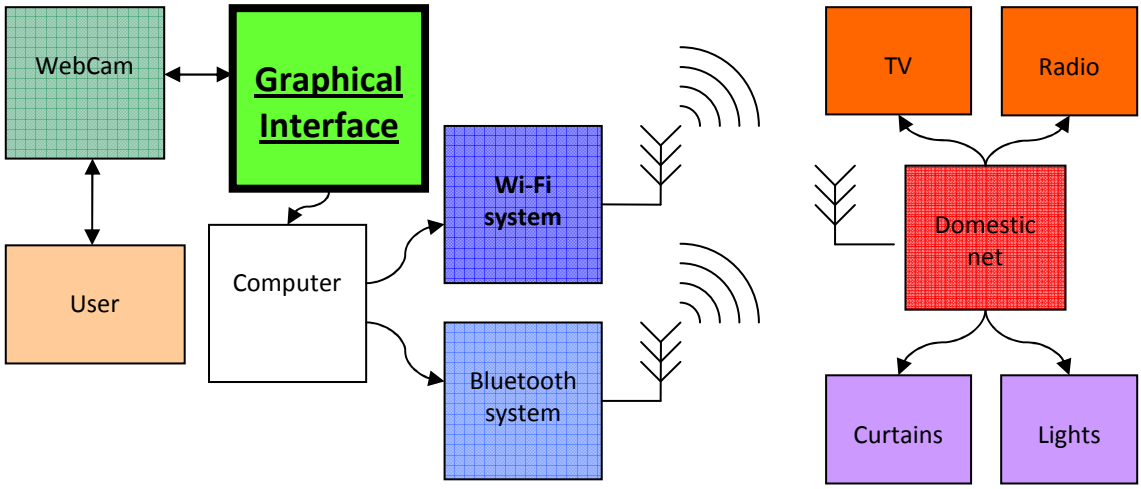

Figure 5.6 – Scheme of a domotic system.

The interface aims at managing the principal electrical and electromechanical devices present in the household environment with a unique simple interface. Compared to the eye-typing system, it constitutes an higher level algorithm: the eye-typing system could be considered as a sub-system of this interface, if we consider the computer and its applications as an domestic-device. The algorithm is organized in menus and sub-menus, following a three-level approach (Figure 5.7). The first level contains the main areas by which the devices are classified. They are:

- *Home*. Here are placed the devices that are strictly related to the basic living necessity, as lighting, heating, air-conditioning, and security.
- *Multimedia*. This part is dedicated to fun activities, as listening to the music, watching movies and television, reading.
- *Communication*. The communication menu is devoted to the principal systems for communicating with other people, as telephone, e-mail, sms and chat.
- *Call caregiver*. Even if the system is conceived to be completely managed by the user, he /she will never stay alone at home. The possibility to call

immediately the caregiver is an important requirement. For this reason, the menu “call caregiver” is completely dedicated to this function.

- *Emergency*. In the case the caregiver cannot help for a general emergency, four kinds of emergency calls have been provided: one addressed to police, one for fire, one for general technical assistance and the last one for medical assistance.

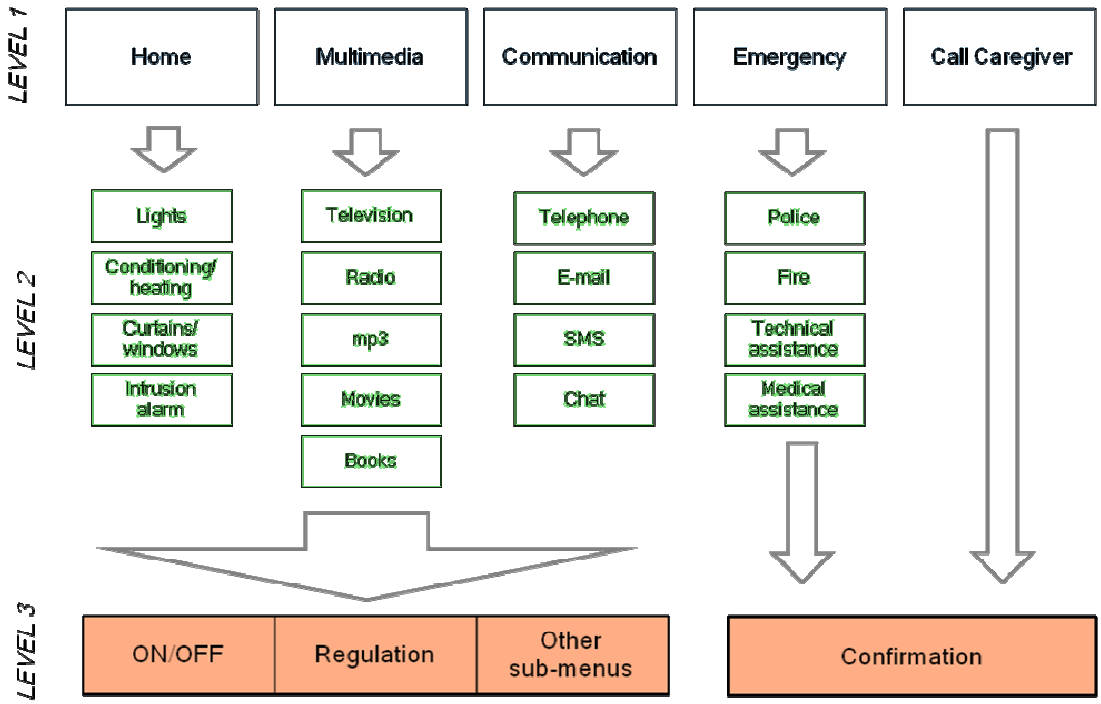

Figure 5.7 – Scheme of the three-level domotic algorithm.

The second menu level concerns the choice of the single device. The third (and eventually the subsequent) level regards the regulation of the single functionalities, as turning on/off the device, regulating power intensity or the frequency, and opening sub-functionalities, (e.g. brightness/contrast in the television sub-menu). For the emergency calls, a confirmation of the selection is present, to avoid involuntary calls.

Graphically, the interface is constituted of buttons placed in the screen within a 5x3 grid.

Each button opens a sub-menu and each level of menu appears in a different part of the screen (Figure 5.8). The first level, i.e. the main menu, is displayed at the bottom of the

screen. The second level appears at the top part of the screen any time one of the icon of the main menu is selected. The same logic is followed by the third level, that is placed in the central row. The potential fourth level, when present, does not appear in a different part of the screen but substitutes the third level.

This multi-level mechanism permits the user to easily access to each sub-function without losing contact with the path followed and with the system in the whole.

As displayed Figure 5.9 the location and the size of the icons does not exactly match the zones defined by the 5x3 grid. This choice has been driven by the necessity to leave some static portions of the screen between the three rows, in order to highlight with more clarity each different level and to leave space to insert titles and written communications.

Figure 5.8 – Levels of the algorithm.

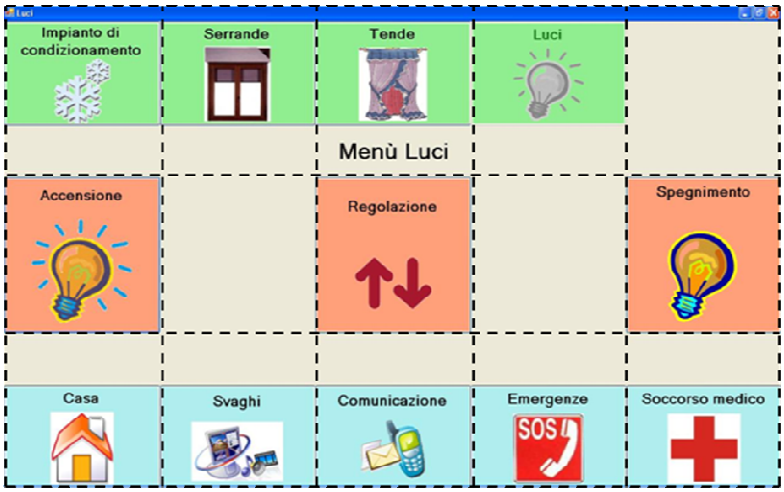

Figure 5.9 – Location of the icons within the 5x3 grid.

Each level of the menu corresponds to a different color to facilitate the individuation of the functions and to avoid losing the sense of direction. The colors of the icons and background have been chosen to make the interface be relaxing and not stressful even after an prolonged use. The use of titles has been reduced to the minimum, giving way to big and simple images to communicate the meanings of the buttons. A title between the upper and central rows, always tell the user where he/she currently is.

The functioning of the interface is based on a dwell time approach. The two principal events are:

1) *the user looks at a button*. Instantly the button is highlighted, but not yet selected. This visual feedback aims at giving to the user information about the current state of the system. This is particularly important in these kinds of systems where the mouse pointer is not visible on the screen. In fact, the functioning here is not based on mouse-emulating approach;
2) *the user stares at the button* previously highlighted for an amount of time greater than the dwell time. This events generates the selection of the function represented by the button. The button color switch to red and, in the case a sub-menu exists, this is opened. The selected button remains highlighted of red to remind the user which command was selected last.

5.2 A Multimodal interface for neuro-motor rehabilitation

The present paragraph proposes the proof of concept of a multimodal platform for neuro-rehabilitation, based on the use of gaze as an estimator of intentionality, combined with a bio-inspired arm control module, based on Functional Electrical Stimulation (FES) to assist the patient in planar reaching tasks.

5.2.1 Gaze and FES in stroke rehabilitation

It is widely recognized that stroke is an age related disease and the World Health Organization [69] claimed that 15 million people suffer stroke worldwide each year. Only about 700.000 of these subjects will regain useful arm movement which is regarded as one of the most important factors for regaining functional independence [70, 71]. Upper limb function is clearly a major problem and therefore new rehabilitation systems for its recovering are needed. Functional Electrical Stimulation (FES) assisted rehabilitation programs have been extensively recognized as a possible therapy for subjects affected by neurological pathologies [72, 73, 74]. The majority of the systems which implement FES therapy drives the electrical stimulator by means of external devices [75, 76, 77, 78] whose intrusive presence, in addition to the FES, is not desirable for patients with neurological injuries. To this aim, a non-invasive FES-assisted rehabilitation system for the upper limb had been proposed by Goffredo et al. [79], where the electrical stimulator is driven by a biologically inspired neural controller and a markerless arm tracking system. Following this rationale and considering that recent studies have shown that when stimulation is associated with a voluntary attempt to move the limb, improvement is enhanced [80, 81], a system which includes the user's intention to move is particularly attractive.

The use of gaze analysis for controlling a FES-based rehabilitation exercise is a novelty in the research area and is based on some considerations coming from neurological and visuo-motor studies:

i) eyes' movement does not result damaged in the majority of stroke cases [82];

ii) during reaching tasks, gaze anticipates arm's movements, thus giving usueful information about movement intentionality [83, 84, 85, 86];

iii) in well structured task protocols, when particular concentration and attention are required, the user usually maintains his gaze fixed on the target during the execution of the movement [87, 88].

Therefore we believe that gaze could represent a valuable channel of interaction between the patient and the rehabilitation machine.

The aim of this work is to present the proof of concept of a new 2D multimodal platform which integrates FES and a specifically designed system for user's intention prediction, based on subject's gaze interpretation.

5.2.2 The multimodal platform

As depicted in Figure 5.10, the proposed platform is composed of a gaze analysis system based on a webcam, a bio-inspired arm controller and a FES stimulator for making an impaired arm reach the aimed position with a 2D movement. The key-point is to give priority to the contact between the user and the real environment (real objects placed on a desk) in which he/she moves, instead of using monitor screen to map the movements of the arm with movements of a virtual object.

In particular, this paper aims at describing and testing two of the systems shown in Figure 5.10:

1) Intention prediction module
2) Arm control module

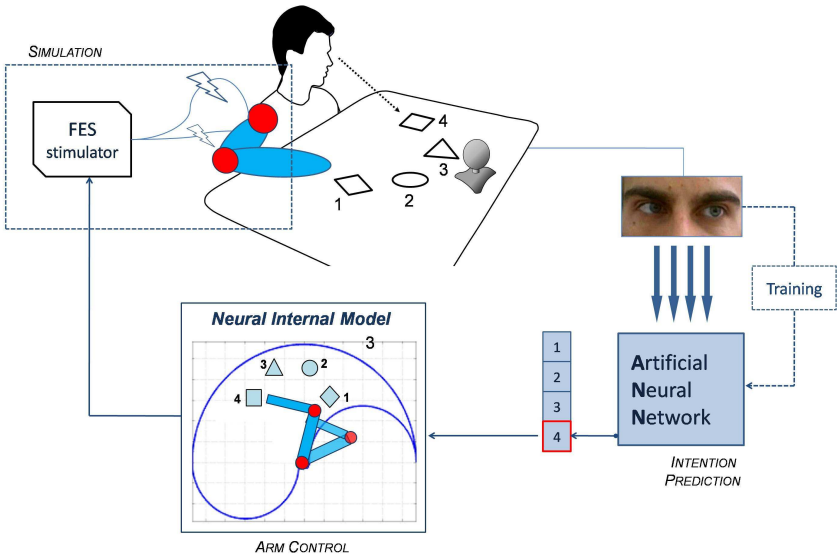

Figure 5.10 - Scheme of the rehabilitation platform.

5.2.2.1 The intention prediction module

The intention prediction module accomplishes the task of estimating the gaze direction. The system is an optimized version of [89] and, after a training procedure, it classifies the gaze extracted from images captured with a commercial webcam and selects the intended object to be reached on a table.

The algorithm follows two main steps. Firstly, a video-based eye tracking system accomplishes the tracking of specifically selected eye features, which are then sent to a hierarchical classification module.

The gaze tracker is based on a commercially available webcam and works under visible light spectrum and head-free conditions, so that the user does not strictly require to maintain the head position fixed. Through image processing algorithms, based on segmentation and template matching, three features of each eye are tracked, i.e. irises and the corners of the eyes (see Figure 5.11).

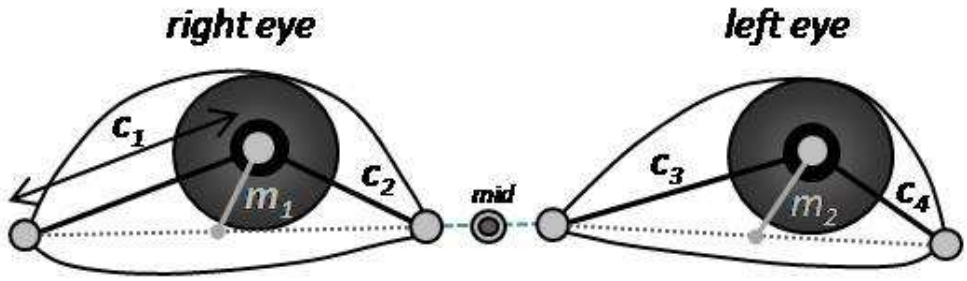

Figure 5.11 - Geometric eye parameters.

Geometric parameters are then extracted and sent to the classifiers. For each eye, three vectors have been selected, i.e. two corner vectors between the corners and the iris and a medium vector connecting the midpoint of the two corner's segments with the iris. In order to include information regarding the global head movements, an additional virtual feature has been included: the midpoint between the two internal corners (named mid in the following).

The classifier has been designed in order to solve two main issues about the gaze detection which have been widely described in literature [90]. There is, in fact, the need of a "rest zone", i.e. a range of gaze directions that do not produce any activation. In a practical way, it is a zone to look at when the user doesn't want to interact with the machine. Moreover, there is the "Midas Touch" problem, that deals with the fact that not all the objects that are observed are object of voluntary attention.

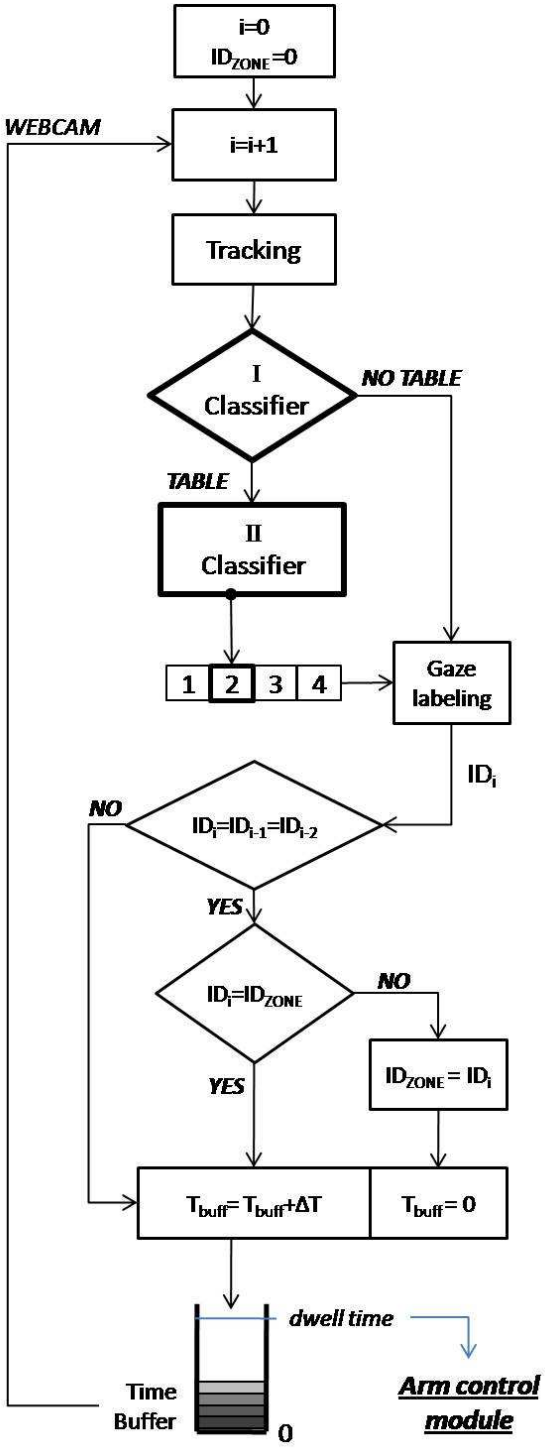

Figure 5.12 - The algorithm of intention prediction through gaze analysis.

Therefore, in the proposed gaze interpretation system, two classifiers and a decision strategy module have been designed, as shown in the scheme in the Figure 5.12.

The classification procedure assigns an identification number (IDi) to the current frame in order to indicate the observed zone. Subsequently, the estimate is refined by means of the decision strategy module, based on a dwell time approach, as explained hereinafter.

The first classifier reports if the user is looking at the table, where the objects are placed, or not. The practical purpose of this classifier is to allow the user to voluntarily interrupt the rehabilitation exercise by simply looking away from the table, and therefore to start again the task by looking again to an object on the table.

The main discriminating direction for this classification results to be the up-down movement. Therefore, the vertical components of the two medium vectors (m1 and m2) have been considered. Additionally, a third parameter has been conceived, i.e. the vector connecting the mid point at the frame i with the mid at the first frame, in order to account for head movements. As depicted in Figure 5.13 (where only two out of three components are shown for clarity purposes), the two classes can be easily and linearly separable. For this reason a k-nearest method with k=15 [91] has been adopted to solve the problem.

The second classifier aims at identifying the observed object on the table. The objects are spread on the table along a right-left direction (see Figure 5.10), with a small range of variation along the close-distant axis. The eye movements are then principally distributed along the x direction, with smaller components along the y direction. Because of the multidimensionality and non linearity problems introduced by head movements, a neural approach is proposed.

A general regression neural network (GRNN) has been adopted, due to its good ability in function approximation tasks [92]. The GRNN, based on the radial basis architecture, consists of an input layer, a hidden layer, a summation layer and an output layer (see Figure 5.14). Actually, the net does not perform classification, but regression. The classification task is performed afterwards by comparing the estimated output with the closest class value assigned to each object in the training phase.

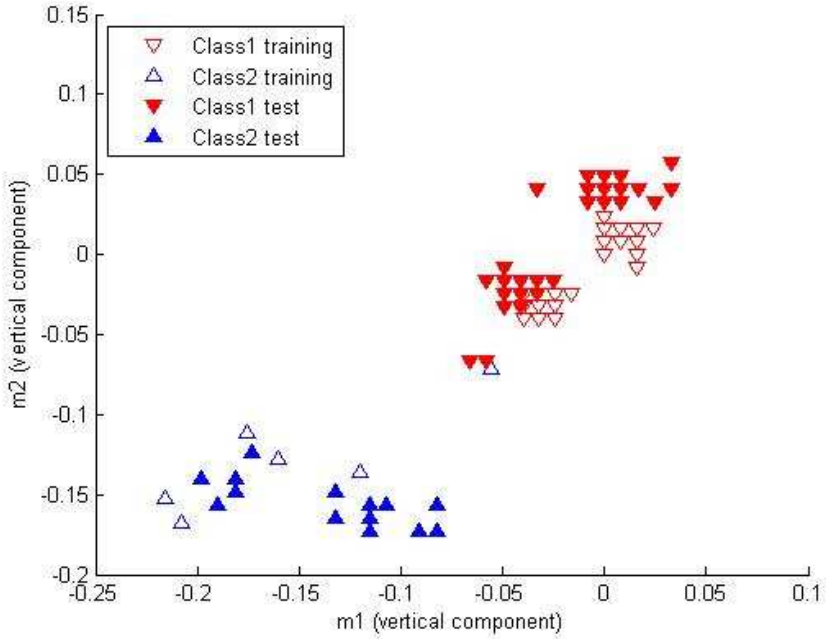

Figure 5.13 - First gaze classification: distribution of the parameters. The upward-pointing triangles represent gazes away from the table, while the downward-pointing triangles indicate that the user is looking to the objects.

The geometrical parameters used by the net are 9, whereof 4 are the internal corner vectors components (c2 and c3), 4 are the medium vectors components (m1 and m2) and 1 takes into account the head rotation around the vertical axis and is represented by the difference between the two eyes' width.

All the parameters used in the algorithms are normalized on the mean eyes' size in order to reduce the error introduced by the relative position between the user and the camera.

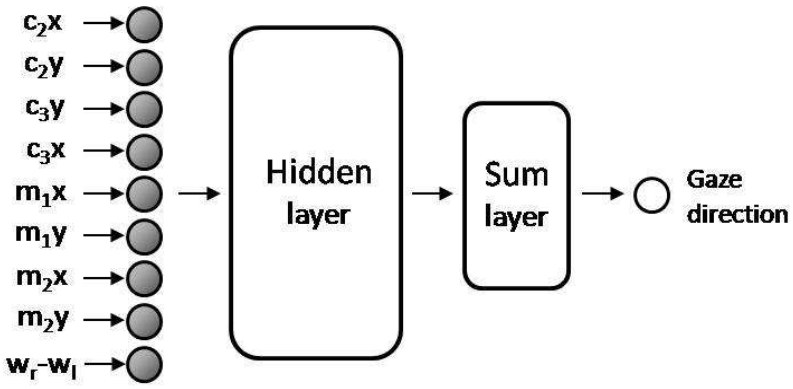

Figure 5.14 - The GRNN used to identify the observed object.

The dwell time approach, frequently used in gaze based human-computer interaction [93], measures the amount of time the user stares the object: every gaze that rests upon an object for a shorter period of time does not produce any outcome.

Therefore, in the present method, the result of every frame classification is compared with the previous ones and if the observed zone is not changing, a time buffer is incremented. Once the time buffer reaches the dwell time (500 ms), a trigger is sent to the arm control module to drive the arm toward the desired zone.

Since our classification approach is based on image processing and single tracking errors could be present, the control strategy module includes also a low-pass frequency filter for fast changes in gaze classifications.

5.2.2.2 The arm control module

The information about active/rest status of the subject gaze is given to the subsequent module of arm control.

The arm control module is based on a bio-inspired neural internal model of the arm [94], which gives the muscular stimulation patterns that will drive the impaired arm assisted through FES towards the desired location on the table (Figure 5.15).

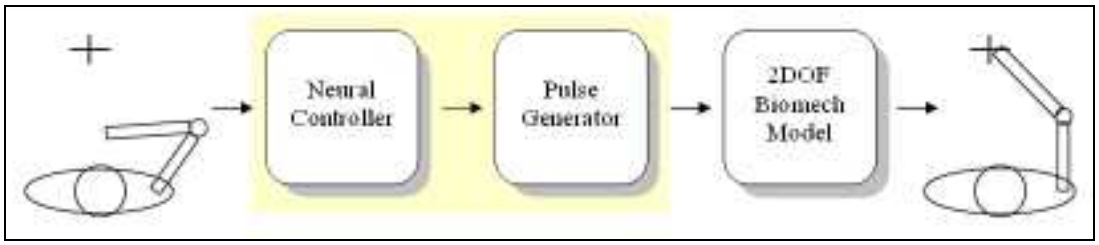

Figure 5.15 - The structure of the bio-inspired arm controller.

To accomplish this objective, a specifically trained multilayer perceptron uses the estimated coordinates of the target position to give the FES stimulator the stimulation patterns to activate the shoulder and elbow joint muscles for planar movements. The system uses a 2D biomechanical model of the human arm to calculate the arm trajectory, and the

Hill's muscle model to solve the problem of mapping between the electrical stimulation and the internal joint torques generation. For further details the reader is addressed to [27].

Within this study a simulation has been performed to test the accuracy of the whole system to drive the arm toward the observed target.

5.3 A Multimodal system for cerebral palsy

The work here presented aims at studying and defining new techniques of interaction between persons with neuro-motor impairments and the computer through facial expressions and gestures. The study is specifically addressed to children affected by cerebral palsy. Two kinds of movements have been taken into account, i.e. the 3D movement of the head and the movement of the eyes. The main objectives of the proposed method are:

- individuating parameters to quantify the level of functional ability (physical, communicative, interactive) of the child;
- individuating alternative channels of communication between the child and the external world, in order to design other kinds of communication devices, specifically addressed to this type of pathology;

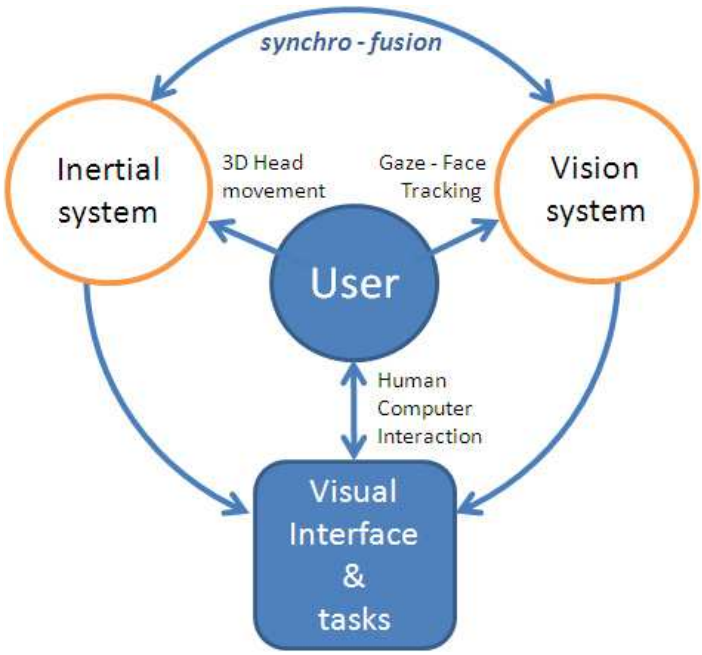

Figure 5.16 – Scheme of the multimodal platform.

The system is composed of three modules, as depicted in Figure 5.16:

1) an inertial analysis unit (IMU) to estimate the 3D pose of the head;
2) a video analysis module to extract information from gaze and face movements of a subject looking at a computer screen;
3) Video tasks to evoke the visuomotor reaction of the child. The video tasks are constituted by a cartoon character moving on a white background. Several movement configuration have been designed:
   i) continuous movements on rectilinear trajectories,
   ii) continuous movements on curved trajectories,
   iii) movements on discrete positions;
   iv) fixed targets at the center of the screen.

These configurations have been designed to evaluate the coordination of the head and gaze, reactivity to unexpected stimulus, attention and concentration.

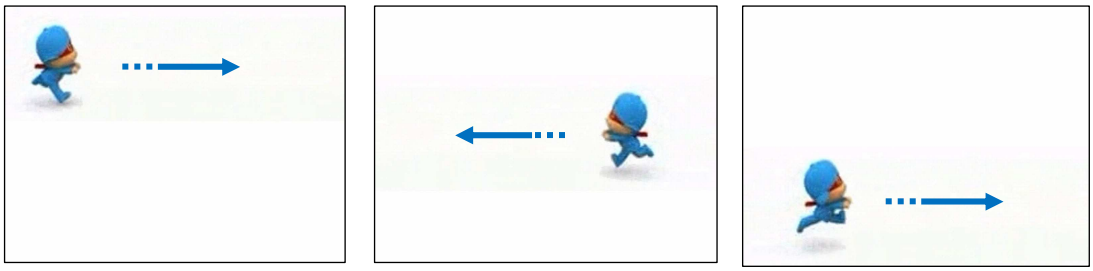

Figure 5.17 – Example of a video task, where a cartoon character moves on predefined trajectories.

6 Experimental testing

This chapter is devoted to evaluate the methods presented in the previous two chapters, concerning eye-gaze tracking and its applications. Regarding the eye-gaze tracker, the experiments have been principally focused on the performance of the techniques of feature detection/tracking and gaze estimation. The applications concerning eye-typing and domotics have been evaluated in terms of usability. The third part of the experimental session is dedicated to the two multi-modal systems for stroke rehabilitation.

6.1 Eye-gaze tracking

The experimental tests on the developed eye-gaze tracker have been carried out to evaluate the performance of the device, in terms of accuracy and robustness under conditions that mimic a realistic context of use.

This section has been organized as follows:

1) testing of the algorithms of initialization, i.e. iris and corner detection, over different subjects, light conditions and distances to the camera;

2) testing of the different neural gaze mapping functions compared with other methods applicable to the view-based context.

All the experimental trials have been performed with a system composed of a commercial webcam (QuickCam® Communicate STX™, Logitech®, 30 fps, 640x480 pixels), a Personal Computer with a Pentium-4 3.0 GHz processor, and a 17'' monitor.

6.1.1 Initialization

6.1.1.1 Experimental procedure

Five subjects with age ranging from 22 to 57 years have been recruited. For each subject 5 videos have been analyzed (see Figure 6.1), one very close to the camera (25 cm), one pretty far (80 cm) and three videos at a normal distance (40 cm) with different conditions of illumination, high (900 lux), low (250 lux) and asymmetrical, thus resulting in a total number of 25 trials. During the trial the subject is requested to blink three times while looking to the camera.

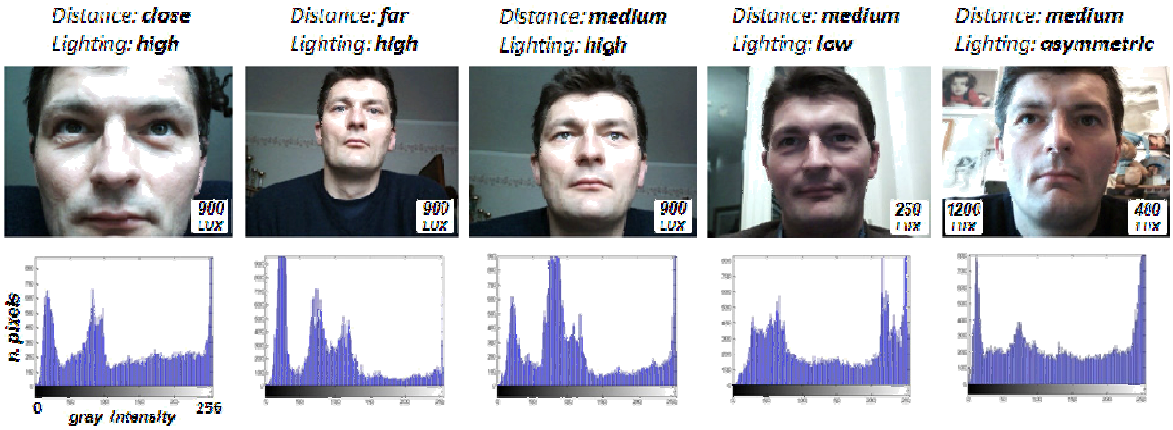

Figure 6.1 – Illumination and distance modifications. Five different combinations of illumination and distance from the camera have been tested for the algorithm. Below each picture, the histogram representation of the luminance distribution is depicted.

The accuracy has been evaluated in terms of percentage of correct estimation: for each video, a value ranging from 0 to 4 was assigned to each estimated position, based on the distance between the estimated position and the one manually determined by one independent researcher. As an example, considering a 80x40 eye image, the relation between performance values and distance ranges is depicted in Table 6.1.

Distance between manual and automatic detection	Performance value	Correctness value
0 – 1 (pixels)	4	100 %
2 - 3	3	90%
4 - 6	2	50 %
7 - 9	1	20 %
> 9	0	0 %

Table 6.1 - Feature detection: assigned performance values corresponding to the distance between the automatically and manually detected feature (for a 80x40 pixels eye image).

6.1.1.2 Results

The method shows a good performance in non-extreme conditions, even for high and low illumination (Table 6.2). Only a strong lateral illumination makes the technique fail in the corner identification.

	Percentage of successful detections			
	Blink	Iris	Inner corner	External corner
Close (25 cm)	100%	100%	90%	100%
Far (80 cm)	100%	80%	100%	80%
Medium distance (40 cm) (high lighting)	100%	90%	100%	80%
Medium distance (40 cm) (low lighting)	100%	90%	100%	100%
Medium distance (cm) (asymmetric lighting)	100%	90%	30%	60%

Table 6.2 - Performance of the feature detection algorithm.

6.1.2 Gaze estimation

In order to detect the most appropriate structure of mapping function, multiple configurations of MLP and GRNN have been considered. More than 50 configurations of MLP have been tested, with one, two and three hidden layers and a total number of neurons ranging from 5 to 420.

6.1.2.1 Experimental procedure

The experimental protocol has been designed as follows: for each trial session the user is asked 1) to seat in front of the screen, 2) to execute a repetition of three blinks to permit to the system to automatically detect the eye features, and then 3) to perform the visual task, looking at the cursor moving on the 15-position path. At the end of the trial the subject is asked to stand up and move away from the seat, and then go back again in front of the screen to start a new trial. During the session the subject is free to move the head in a natural way, still avoiding sudden and/or large movements, at a distance from the screen of approximately 46 cm (corresponding to a distance from the camera of 43 cm). The test set includes 8 trials with the head approximately centred with respect to the trained volume, and

slight movements allowed (Figure 6.2). In particular the head translation lies within an area of 3x3 cm in x and y directions, and the variation along z direction is of ± 1 cm.

Figure 6.2 – Head movements: dashed line: range of head movement in the training phase; dotted line: head motion during the preliminary test trials. Sample of head movement during the execution of the visual task: approximately 3x3x2 cm movements have been observed.

The performance of each net has been evaluated in terms of the ability to classify the gaze direction over the 15 zones on the screen, expressed by the percentage of correctly estimated zones.

6.1.2.2 Results

MLP neural net has been firstly examined in order to find the best configuration to be consequently tested. As shown in Figure 6.3, the mean performance does not vary significantly with the number of neurons. A slight improvement (2-3%) occurs in the case of two and three hidden layers. Among the overall set, some nets reach higher recognition percentages. In particular, for the calculation of the x coordinates, some configurations achieve a performance of 96% to 98%, while for the y coordinate the best score (96.7%) is reached with a 3-layer configuration.

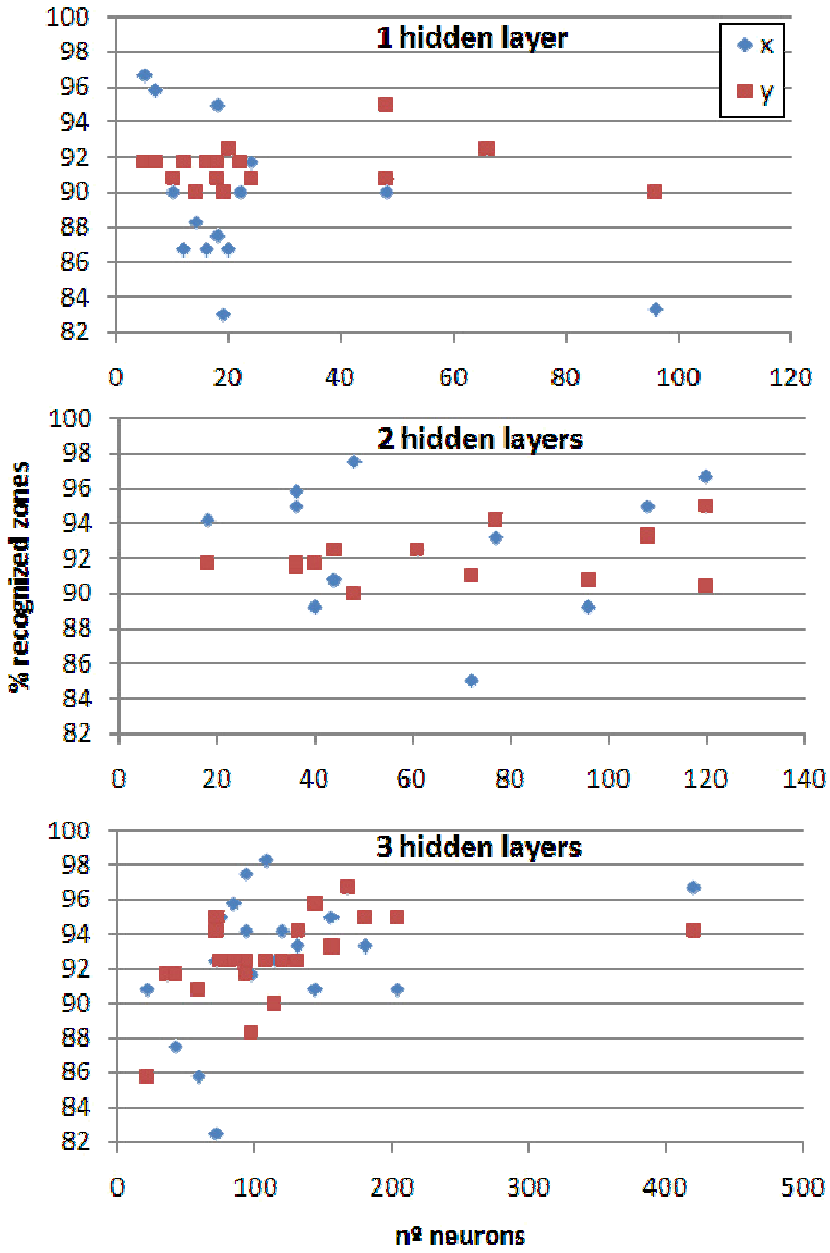

Figure 6.3 - MLP preliminary results. Classification performance as a function of the number of neurons of the MLP net across the different number of hidden layers.

Also the GRNN has been tested in order to find the best configuration parameters, i.e. in this case is represented by only one parameter, the spread. As the figure 13 shows, the optimum value of the spread has been set at 0.3 for the x and 0.33 for the y, leading to a performance of respectively 96.7% and 93.3% of zone recognition performance.

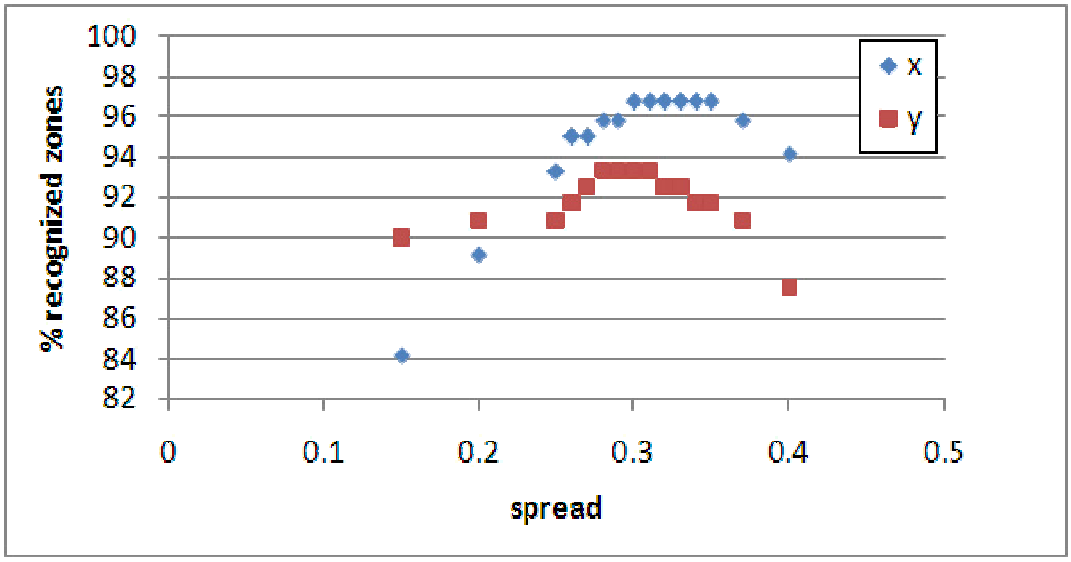

Figure 6.4 - GRNN preliminary results. Classification performance as a function of the spread parameter.

The nets above selected have been tested again, loosening some of the constraints to the head motion, in particular concerning the distance to the camera. The test set is in this second situation composed of two head positions, one far, at 53 cm from the screen and the other closer, at 41 cm from the camera. The different MLP configurations have shown a low capability of calculating the gaze correctly for these two distances, while the GRNN has shown a very good performance for the higher distance. In the case of close distance, some errors occur especially for the y coordinate estimation (Table 6.3).

	Percentage of zone recognition zones			
	far (53 cm)		near (41 cm)	
	X	Y	X	Y
MLP	45%	40%	39%	55%
GRNN	100%	100%	80%	60%

Table 6.3- Performance of the neural networks for different distances from the camera.

According to the obtained results we can state that the MLP structure gives more accurate results for a given distance to the camera, while the GRNN is more robust than MLP even for changes in the z direction, still maintaining the percentage of correct recognition at a high level. For this reason we consider the GRNN structure more suitable for free head conditions.

In the following, an analysis of the global accuracy and robustness of the GRNN will be presented. The accuracy has been calculated in terms of mean and standard deviation of the gaze error, i.e. the error between the real observed position and the estimated values, expressed in terms of pixel (e_{pxl}) and angular degrees ($e_{\deg ree}$), according to the following equation:

$$e_{\deg ree} = \arctan \frac{e_{pxl}}{d_{pxl}}$$

with d_{pxl} representing the distance between the subject and the screen plane expressed in pixels. As in the previous paragraph, the percentage of zone recognition will be used as a measure of the accuracy of the system.

Two kinds of results are reported, first considering each trial separately to determine the robustness to head movements (Table 6.4), then by reporting the accuracy on each zone over the different trials (Table 6.5 and Figure 6.5) to highlight the distribution of the accuracy over the screen.

The estimated mean accuracy is approximately 1.6 degrees on x direction and 2.6 degrees on y direction with a standard deviation respectively 1.4 and 1.9 degrees, leading to a rate of successful recognition of 94.7%. Exceptions have occurred in three zones, where the performance decreases to values under 90%.

Trial	Distance (cm)	Mean Value ± Standard Deviation		Percentage of zone recognition	
		X	Y	X	Y
1	48	**1.6 ± 1.2 (degrees)** 36 ± 26 (pixels)	**3.3 ± 2.1** 78 ± 46	100 %	80 %
2	47	**1.4 ± 0.9** 32 ± 21	**1.2 ± 0.8** 31 ± 20	100 %	100 %
3	46	**2.2 ± 1.3** 47 ± 27	**4.1 ± 3.0** 91 ± 63	93.3 %	66.7 %
4	46,5	**1.7 ± 1.2** 38 ± 27	**1.3 ± 1.1** 32 ± 26	100 %	100 %
5	45	**1.3 ± 1.1** 29 ± 24	**1.5 ± 0.8** 37 ± 19	100 %	100 %
6	45	**2.0 ± 1.8** 45 ± 40	**3.5 ± 2.1** 80 ± 45	93.3 %	93.3 %
7	45	**1.5 ± 1.0** 34 ± 25	**1.8 ± 1.8** 43 ± 42	100 %	100 %
8	45	**1.9 ± 1.3** 61 ± 28	**2.4 ± 1.5** 59 ± 33	86.6 %	100 %
9	53	**1.2 ± 1.3** 29 ± 30	**2.8 ± 1.2** 66 ± 26	100 %	100 %
MEAN		**1.7 ± 1.2** (deg)	**2.4 ± 1.6** (deg)	**97 %**	**93.3 %**

Table 6.4- Accuracy of the GRNN for 9 different test trials.

Zone	Mean Value ± Standard Deviation (degrees)		Zone recognition
	X	Y	
1	1.8 ± 1.8	2.1 ± 1.3	93.3 %
2	0.7 ± 3.1	3.4 ± 2.2	93.3 %
3	1.0 ± 1.5	2.5 ± 1.0	100 %
4	1.4 ± 1.7	1.8 ± 1.3	100 %
5	1.7 ± 1.0	1.2 ± 1.6	100 %
6	2.8 ± 0.7	3.5 ± 3.1	93.3 %
7	0.2 ± 2.0	2.6 ± 3.1	93.3 %
8	0.4 ± 1.1	2.8 ± 2.4	93.3 %
9	0.3 ± 2.0	2.5 ± 2.7	93.3 %
10	4.2 ± 1.5	2.3 ± 2.8	80 %
11	3.9 ± 1.4	3.5 ± 1.8	100 %
12	0.3 ± 2.5	2.7 ± 1.8	100 %
13	0.3 ± 1.9	2.8 ± 2.3	100 %
14	0.1 ± 1.9	4.0 ± 2.5	93.3 %
15	2.7 ± 1.6	5.9 ± 3.1	86.7 %
Mean	**1.4 ± 1.7**	**2.9 ± 2.2**	**94.7 %**

Table 6.5 - Accuracy of the GRNN on the 15 zones.

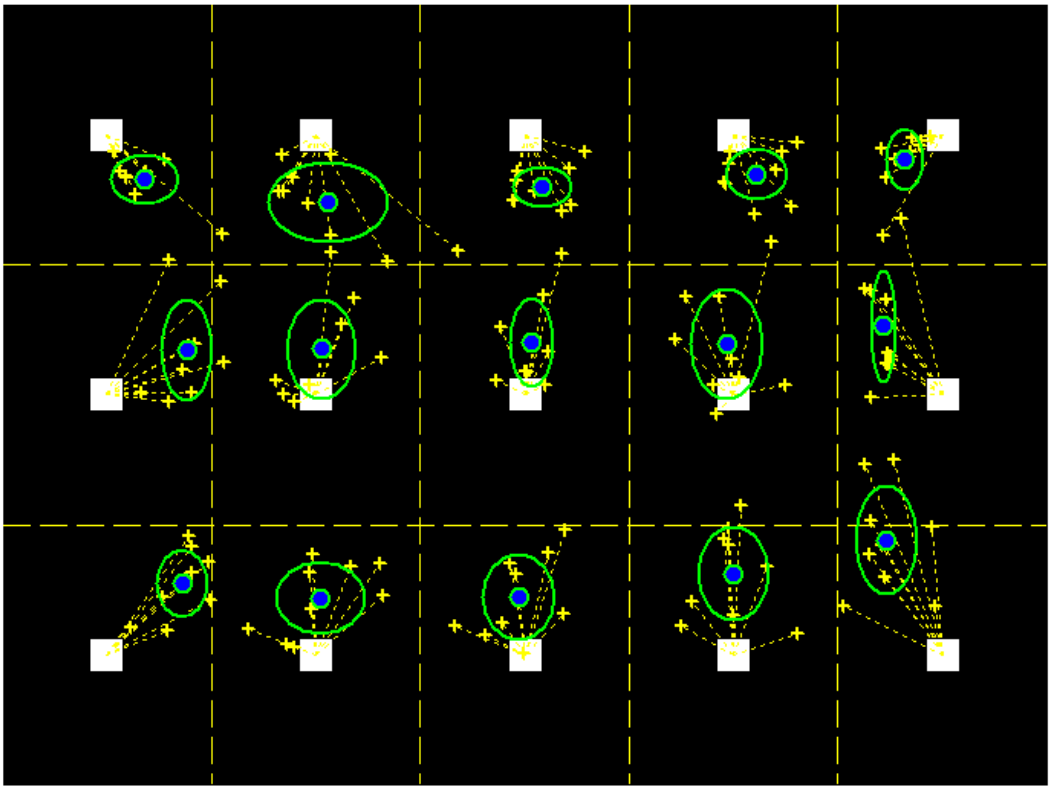

Figure 6.5 - Accuracy evaluation. White squares represent the observed positions. Crosses correspond to the estimated values (1 for each test trial). Small circles represent mean values among the trials, while the ellipses stand for the standard deviations in x and y direction.

Moreover, the neural approach has been compared with a polynomial mapping function that typically used in the context of gaze tracking. In particular, the second order equation system mostly used in the infrared-based systems has been implemented.:

$$\begin{cases} s_x = a_0 + a_1 x + a_2 y + a_3 xy + a_4 x^2 + a_5 y^2 \\ s_y = b_0 + b_1 x + b_2 y + b_3 xy + b_4 x^2 + b_5 y^2 \end{cases}$$

The pupil-glint vector used in the infrared based techniques is not reproducible in the view-based approaches, and thus a new vector has been considered as a proxy of the pupil-glint: it is obtained by connecting the pupil with the midpoint of the segment connecting the two corners.

The quadratic function takes as inputs the two components of the vector, returning the coordinates of the screen point. A 15-point calibration and a least square solution have been

used to solve the over-constrained problem, as proposed by Morimoto [32]. The results depicted in Figure 6.6 demonstrate the better performance of the neural approach with respect to the polynomial interpolation method in terms of mean values and standard deviations, showing an accuracy of 1.7º ± 1.2º for the x direction and 2.4º ± 1.6º for the y direction, with respect to the accuracy of the quadratic function, respectively of 2.9º ± 1.9º and 3.6º ± 1.9º for x and y directions.

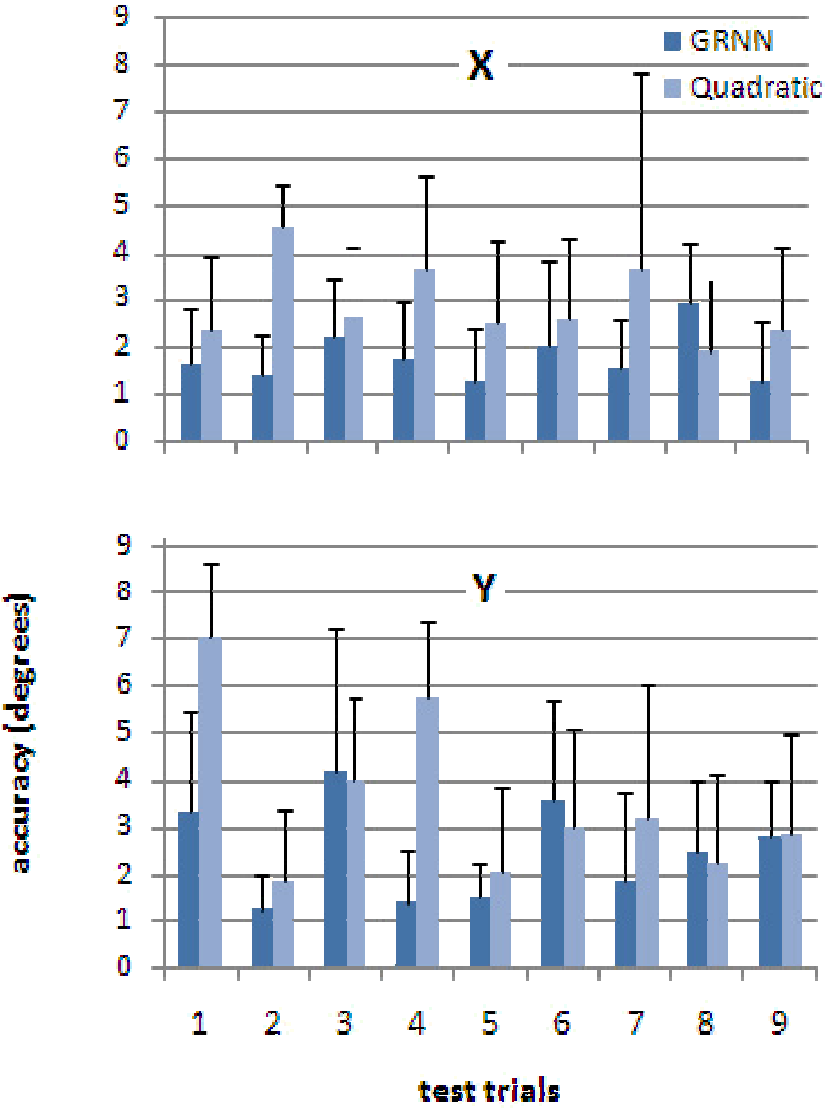

Figure 6.6 - Comparison to the quadratic mapping. Mean values (histograms) and standard deviations of the error in degrees for the different trials.

6.1.3 Discussion

The proposed REGT system shows reliable and accurate global results. The uncertainty of gaze estimation has been proven to come from two main factors. The first one refers to the eye features tracking: in some extreme cases, if the gaze is directed towards the very lowest part of the screen and off the centre, the iris tracking algorithm does not achieve a high accuracy, due to occlusions from the eyelids and significant changes in the iris shape, so that the template matching is not as accurate as for the intermediate positions. This is what happens for the zones 14 and 15 (low right in the screen, see figure 14) where the estimation accuracy increases to higher, yet acceptable, values of error.

The second source of inaccuracy is due to the nature of the input set, i.e. magnitudes and orientation of the corner-iris vectors, which might describe a non-predictable distribution over the 12-dimensional space in presence of head movements. The neural mapping seems to represent a proper solution for this problem, overcoming the non-linearity introduced by head and eyes movements. The GRNN structure has been proven to be more effective than MLP in compensating head motion along the z direction, while the MLP network seems to achieve higher values of accuracy for distances very similar to the ones present in the training set. Since the system is aimed at compensating for movement in a 3D space, the GRNN has been considered the most appropriate solution. The neural mapping outperforms the quadratic mapping function, as one would expect by considering the possibility of having the user look at the same point on the screen by having different head positions.

The system has also been proven to be robust to light changes. By using an initialization procedure based on the movements of the eyelids during blinking, the feature extraction and tracking is almost independent from the presence of light changes, both in intensity and direction. Moreover this preliminary procedure attains to automatically initialize the procedure, avoiding an external user to select the features to track.

The estimated accuracy of the system is pretty good, yielding values comparable with most of all view-based REGT ones, except for the work of Xu et al. [37], where the accuracy is, however, calculated through the leave-one-out procedure, whose use is controversial, since it usually underestimates the true prediction error [95].

As for the classification performance, results are in favour of a very high classification rate under natural head motion with a 15-zone user interface environment, which represents a valuable trade-off between usability and timing cost.

The 5-trial calibration procedure, necessary to let the system learn and correctly estimate the gaze direction, is user specific and requires about 8 minutes (5 minutes of visual task plus one minute of rest between two consecutive trials): the burden time is acceptable for user comfort.

6.2 General applications

6.2.1 Evaluating usability of interfaces

Evaluating the usability of an interface during its development is an important action that permits to involve user point of view within an iterative development process (Figure 6.7).

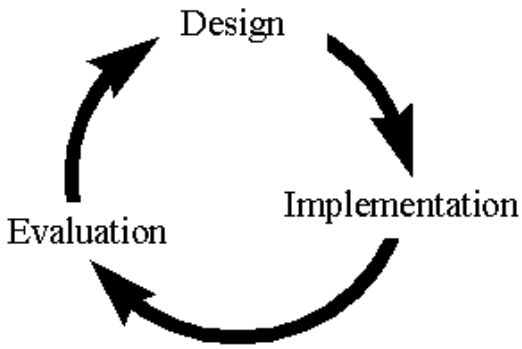

Figure 6.7 – Iterative development process

Several approaches to usability inspection were proposed. They can be divided into *empirical* and *non-empirical*.

6.2.1.1 Empirical evaluation

Empirical evaluation means that information is derived from actual users of the system or people who resemble users. There are various techniques to collect information from empirical evaluation. The most important are:

- **Questionnaires and interviews**. Questionnaires and interviews can be created with numerical and yes/no questions (closed-ended questions) or with open-ended questions, where uses formulate their own responses.
- **Performance measurement.** It is what the experts usually call "usability testing". Normally, the user is given tasks to complete, and the evaluator measures relevant parameters such as percentage of tasks or subtasks successfully completed, time required for each task or subtask, frequency and type of errors, and duration of pauses, indications of user frustration, and the ways in which the user seeks assistance.
- **Thinking-aloud protocols**. It is a kind of usability test in which the user is asked to continuously explain what he or she is thinking. The benefit is that the developer can more readily understand the user's mental processes and, especially, what is causing a problem.

Apart from choosing the most appropriate method, another crucial task is to devise an appropriate test methodology. Some of the central issues in devising a test are:

- choosing the subjects;
- deciding how many subjects you need. Experts of usability (Nielsen et al. [96]) argue convincingly that 5 is enough;
- deciding the nature of the tasks/questions;
- setting the time limits.

6.2.1.2 Non-empirical evaluation

Non-empirical evaluation consists of advice and other information that does not come from users or potential users. Non-empirical information normally derives, directly or indirectly, from experts. The information can certainly be valuable, and it's almost always easier and cheaper to obtain than empirical data. The main disadvantage of non-empirical evaluation is that the evaluators must be experts, as suggested by Nielsen. A second disadvantage is that several evaluation experts are usually needed.

6.2.2 Eye-typing

6.2.2.1 Experimental procedure

An empirical evaluation approach has been adopted to evaluate the interface. Eight subjects have been recruited with age ranging from 24 to 28 years. Each subject was asked to use the interface to write a text dictated by voice along 5 minutes. At the end of this task, a questionnaire has been submitted to the user, composed of nine questions, each one concerning a principle of usability:

1) Is the interface simple?
2) Is it intuitive?
3) Is it difficult to learn?
4) Do you understand what you do?
5) Is the debugging easy?
6) Is the interface familiar?
7) Is it entertaining?
8) Is it relaxing?
9) Is it fast to use?

6.2.2.2 Results and discussion

The results shown in Figure 6.8 go in two directions. On one hand the interface results familiar, understandable and, after a little practice, intuitive and entertaining. On the other hand some user had difficulty to learn the functioning of the “T9-like” typing mode, and to manage the debugging process. The answers about velocity did not produce particularly significant results.

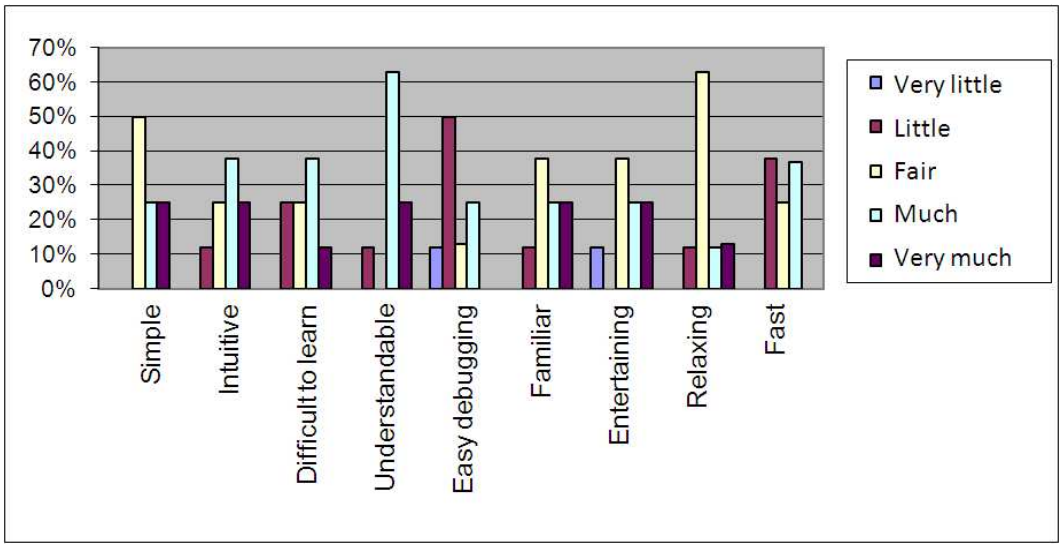

Figure 6.8 – Results of the questionnaire.

The main drawback of the interface is represented by the debugging process. In fact, in this version, to correct a mistake the user has to select a button for each letter that is not correct. In the case a predicted word is selected erroneously, the user has to select the clear button as many time as the number of letters that compose the word, making the functioning very slow and tiring. In the next version, a different debugging strategy should be taken into account.

6.2.3 Domotics

6.2.3.1 Experimental procedure

The usability tests for this interface have been proposed to 10 subjects, five of them with an age ranging from 25 and 30 years, and the other five ranging from 60 to 65 years. The choice of the age of the subjects has been driven by the necessity to focus on the different attitudes and needs of young and elderly persons.

The test is represented by a procedure of tasks and a final questionnaire. Moreover, during the execution of the tasks, all the spontaneous observations from the user have been taken note.

The task procedure is constituted of six steps:

1) a brief explanation about the interface functioning is given to the user;
2) in order to make the user understand the basic functions, he/her is asked to utilize the program during a couple of minutes;
3) the user is given a more detailed explanation about the interface, its objectives and other technical aspects;
4) the user is asked to execute a simple task, as for instance switch the lights on.
5) The user is asked to execute a more complex task, as regulating the temperature of the air-conditioning system.
6) The user is brought through a wrong path and then is asked to go back to the right path;

Once the procedure is concluded, the user is requested to respond to a questionnaire. To each closed-ended question the user can choose among the following answers:

a) Very much
b) Much
c) Fair
d) Little

The questionnaire is composed of 11 questions:

- Is the graphical aspect of the interface nice?
- Is it easy to use?
- Is the meaning of each button clear?
- Is the hierarchy of the program clear?
- Does it consider all the main household functions?
- Is it intuitive?
- Is it difficult to learn?
- Is it easy to debug?
- Is it entertaining?
- Is it relaxing with regards to the colors?
- Is it fast to use?

6.2.3.2 Results and discussion

The testing procedure has been performed without particular difficulties by the users. Only two persons have had some hesitations in the execution of the complex task (point number 5 of the procedure); despite that they solved the problem by themselves without any external suggestion.

As shown in Figure 6.9, the interface received a good judgment on the whole even for the elderly people, whereas 80% never had experience with a computer.

In particular the interface has been considered very easy to use, and with meaningful buttons.

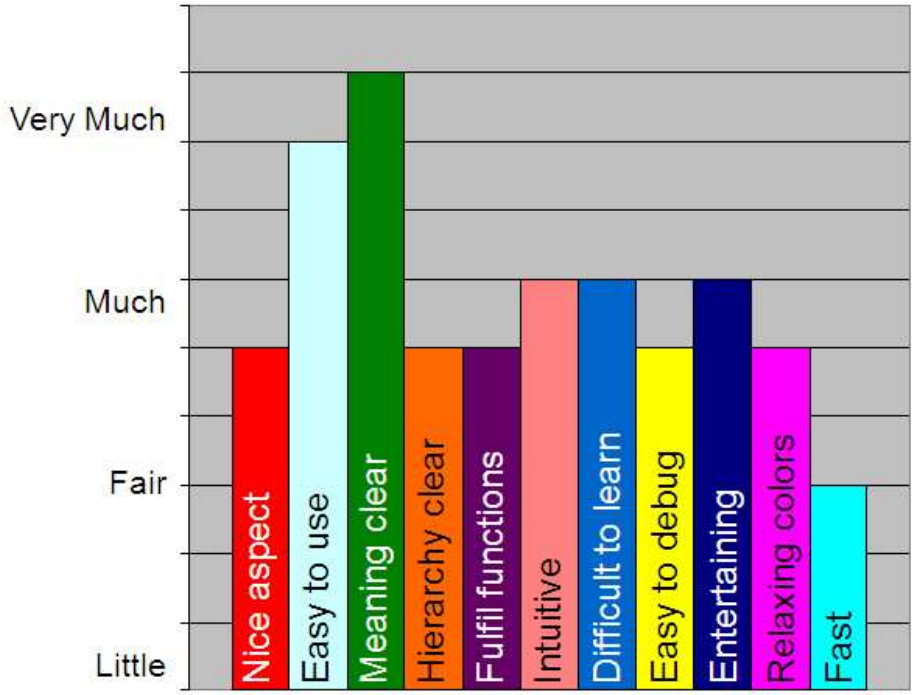

Figure 6.9 – Global results on usability testing.

Figure 6.10 shows the results classified by age bracket. To almost all the questions the young users answered more positively with respect to the older ones, with exception for the velocity of selection. Older people where more satisfied than younger regarding the velocity, the choice of colors and the fulfilling of the possible household functions.

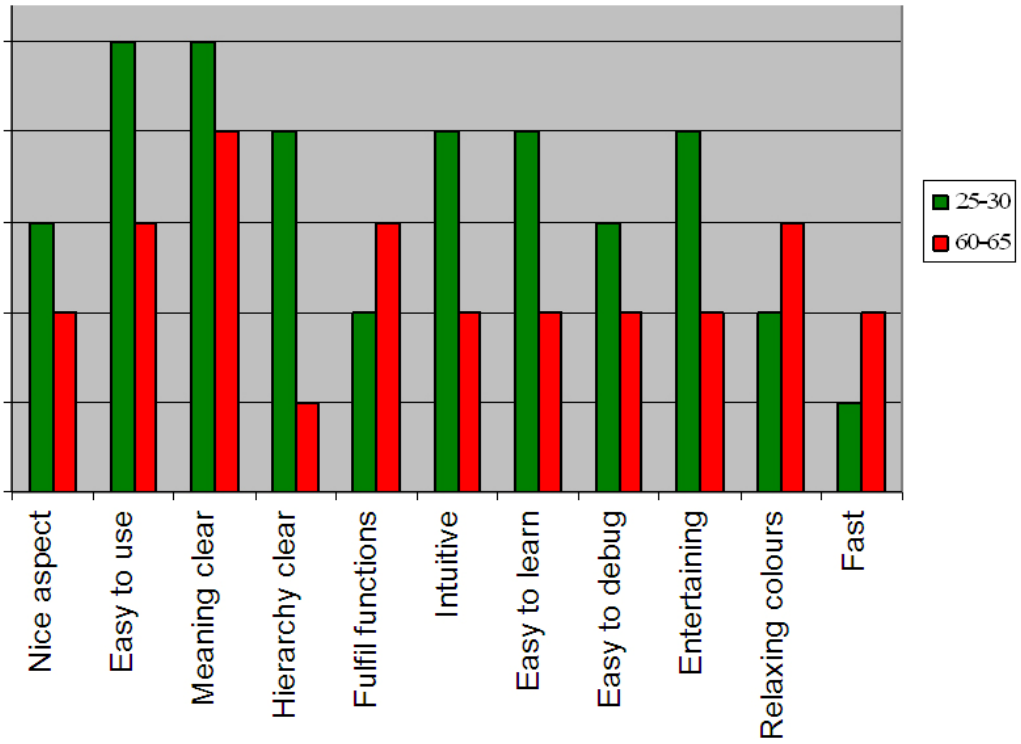

Figure 6.10 – Results classified by age bracket

Younger people have considered the velocity of selecting not sufficient, while older users did not feel uncomfortable with velocity. This can be due to the fact that people that are used to modern kinds of performing computer interfaces are generally not well-disposed to a dwell-time approach, where the user had to wait with the pointer on the button for almost 1 second to put it into action. Completely different is the case of elderly people, whereas most of them are not used to spend all the day on computers. Another difference that has been noticed is that the younger people did not pay much attention about the appreciation of the colours and the completeness of the household functions, while older users have demonstrated to be more sensible to the topic, judging positively the choices performed.

It is important to point out that in the current version of the interface, not all the functions have been implemented. With reference to Figure 5.7 the most complex functions as chats or composing an SMS have not been realized since the design of this interface has been performed in the mean time of the eye-typing interface, necessary for the above mentioned applications.

6.2.4 Eye driven platform for stroke rehabilitation

6.2.4.1 Experimental procedure

The experimental setup includes a Logitech Pro9000 webcam with spatial resolution of 640x480 pixels and time resolution of 30 fps; an Acer Extensa 5200 laptop, with a Intel Celeron 1.9 MHz processor; a rehabilitation board (Figure 6.11), used for fixing the objects and the webcam.

Three healthy subjects, two males and one female, with age between 25 and 30, have been recruited for the tests. The experimental protocol consists of 5 trials for each subject: 1 trial is used for the classifier training and the remaining 4 to test the system.

During the tasks the users are asked to stare sequentially at the objects for at least one second. Subsequently the users gaze randomly far away from the board for at least three seconds. No restrictions to head movement have been imposed.

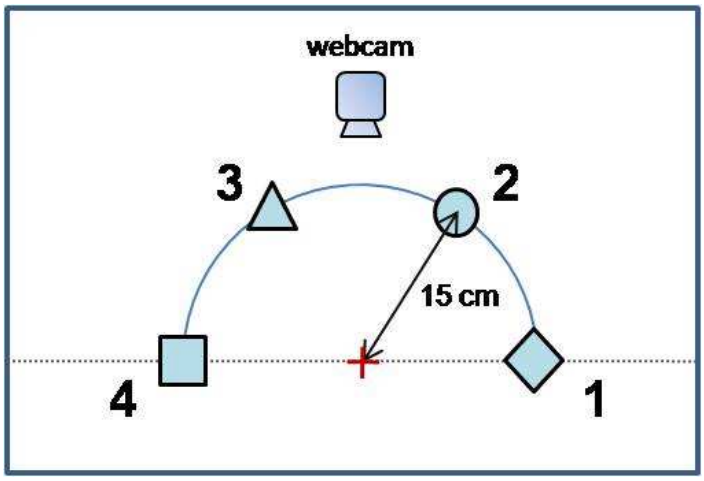

Figure 6.11 - The proposed rehabilitation board.

The accuracy of the gaze system has been evaluated in terms of percentage of Correct Classification Rate (%CCR).

The simulation tasks reflect the path carried out by the eyes' movements, so that, starting from a rest position (red cross in the Figure 6.11), a sequential reaching of the

objects 1 to 4 is driven complying with the following procedure. Any time the outputs of the gaze classification fill up the time buffer, they are transformed into target coordinates to be given to the neural controller. The take into account the inter-subject variability, the muscolo-skeletal model is updated, resting on the weight and height of each subject. A simulation of a complete reaching movement is then performed and a measure of the accuracy is accomplished. The accuracy has been evaluated in terms of mean value of the absolute distance error between the target position and the position of hand after the movement.

6.2.4.2 Results and discussion

Table 6.6 shows the ability of the system to recognize if the user is voluntarily interrupting the rehabilitation exercise (I classifier) or is looking to a object on the rehabilitation board (II classifier). The first classifier's outputs present a %CCR higher than 96% and are influenced by intra-subject variability. Furthermore, the obtained results on the object's gaze, with a mean value of 96.79%, are particularly encouraging and the sensitivity of the proposed system to the target's position is very low.

	I classifier	II classifier			
		Object 1	Object 2	Object 3	Object 4
Subject 1	97.2 %	98.6 %	93.0 %	93.3 %	99.3 %
Subject 2	96.0 %	97.4 %	100 %	99.0 %	93.0 %
Subject 3	97.9 %	92.2 %	96.3 %	98.6 %	100 %

Table 6.6 – Gaze Correct Classification Rate (%CCR)

The results of the simulation of the FES-assisted movements (Table 6.7) show that the position of the hand lies less than 2 cm apart from the target, with a duration of movement of less than 1 second.

		Gaze + simulation			
	Movement	Start to 1	1 to 2	2 to 3	3 to 4
Subject 1	Reaching error (cm)	2.2	0.3	1.1	1.4
	Movement duration (ms)	610	615	617	554
Subject 2	Reaching error	2.3	2.7	1.2	1.5
	Movement duration	616	555	650	542
Subject 3	Reaching error	2.6	0.8	1.3	1.9
	Movement duration	640	641	668	554

Table 6.7 - Reaching accuracy in the FES-assisted movement simulation.

The preliminary experiments here presented have shown a good ability of the system to recognize an active gaze from a passive one, and to classify the direction of gaze among 4 objects placed on a rehabilitation board. The simulation of movements driven by the neural stimulator have proved good performance in terms of end position and duration of movement, remarking the adequacy of the approach to the real reaching context. The relevance of the results goes towards two directions:

1) it has been proven that the developed eye-gaze tracker can predict the direction of a goal-oriented movement without using any specific hardware. These can be particularly interesting in the field of telerehabilitation, whereas the patients executes the therapy from home with a simple and affordable setup;

2) it has been realized a system that gives priority to the contact between the user and the real environment in which he/she moves. A minimally invasive setup, assured by a FES approach and the absence of monitor screens for the execution of the task go in the direction of creating a natural environment in which the user could concentrate directly on the task, on the target to reach, and on proprioception.

Future developments will regard experiments on healthy and pathological elderly people to test performance and usability of the system. Further developments will be include the use of a FES controller that will substitute the simulation part of the experiment.

7 Conclusions

The work presented in this thesis aimed at developing an innovative assistive device for people with motor disabilities, based on eye-gaze tracking.

With respect to the state of the art, the innovation is represented by two main factors. From a technological point of view, the proposed system has been designed to work with low-cost and completely off-the-shelf hardware. As a matter of fact the cost is the main factor that prevents eye-gaze tracking from being used in a diffusing way. The second cause of innovation is represented by the use of new typologies of mathematical functions for the estimation of gaze. In particular, a bio-inspired approach has been followed, using artificial neural networks (ANN) in order to account for the high non-linearity introduced by changes in head movements. Accounting for head movements represents the other important issue not yet completely solved by the literature.

Under an applicative point of view, the thesis demonstrated the feasibility of the developed eye-gaze tracker in the field of disability. For this purpose two kinds of computer-based applications have been realized and tested, concerning written communication (i.e. eye-typing) and household environment control (i.e. domotics). An additional area of study has been explored, that is the use of eye-gaze tracking for rehabilitation purposes, in particular addressed to stroke and cerebral-palsy.

The following paragraphs will deepen the over mentioned conclusions by verifying the stated hypothesis, highlighting the limitations of the methods and suggesting interesting further work.

7.1 Hypothesis verification

Specifically, the experimental tests aimed at verifying the following hypothesis (see Figure 7.1 and paragraph 1.2):

- Hypothesis n°1: *"It is possible to estimate gaze direction with an accuracy compatible with disability applications without the use of specific or high-cost hardware*". The obtained results on eye-gaze tracking shown a global average accuracy of around 2 degrees and a classification rate of 95% in zone recognition by dividing the monitor screen in 15 zones. Further experimental tests proved the effectiveness of the system to automatically detect the user without any intervention by an external operator. The technique, based on blink detection, resulted to be efficient in terms of accuracy and computational cost, allowing a real-time functioning. Moreover, the blink detection module together with an automatic failure detection estimator permit to automatically re-initialize the system in the case the tracking fails due to occlusions or wide movements.

In order to verify the feasibility of such accuracy for computer-based applications, two kinds of interfaces, i.e. eye-typing and domotic control, have been designed to fit the 15-zone layout. Several tests on subjects of different ages confirmed the good usability of the interfaces, in terms of efficacy and satisfaction.

- Hypothesis n°2: *"It is possible to enhance robustness in gaze estimation using artificial neural networks (ANN)"*. The core of the proposed approach overcomes most of the issues induced by head movement, that is one of the most serious problems in eye-gaze tracking. Two neural network structures have been designed to learn the gaze behavior of the user under natural head motion: head shift was performed during the training sessions and a test session with head motion in 3-D has been performed. The aim of the experiments was to test the system under natural head movements during a visual task, maintaining a comfortable posture. The obtained results confirm the reliability and robustness of the proposed system, in which the neural mapping has been proven to outperform traditional quadratic approaches.

- Hypothesis n°3: *"It is possible to use gaze as a channel of interaction for rehabilitation purposes"*. This thesis proposed the proof of concept of a multimodal platform for neuro-rehabilitation, based on the use of gaze as an estimator of intentionality, combined with a bio-inspired arm control module to assist the patient in planar reaching tasks. A minimally invasive setup, assured by a FES approach and the absence of monitor screens for the execution of the task, moves towards the creation of a natural environment

in which the user could concentrate directly on the target to reach. Preliminary experiments here presented have shown a good capability of the system to recognize and to classify the direction of gaze when choosing among 4 objects placed on a rehabilitation board. The simulation of movements, driven by the neural stimulator, have proved good performance in terms of reaching position, remarking the adequacy of the approach to the real reaching context. Moreover the work here presented started a study on new techniques of interaction between persons with neuro-motor impairments and the computer through facial expressions and gestures, in particular for children affected by cerebral palsy, in order to contribute to the definition of new protocols to evaluate the level of functional ability (physical, communicative, interactive) of the child and to design other kinds of communication devices, specifically addressed to this type of pathology.

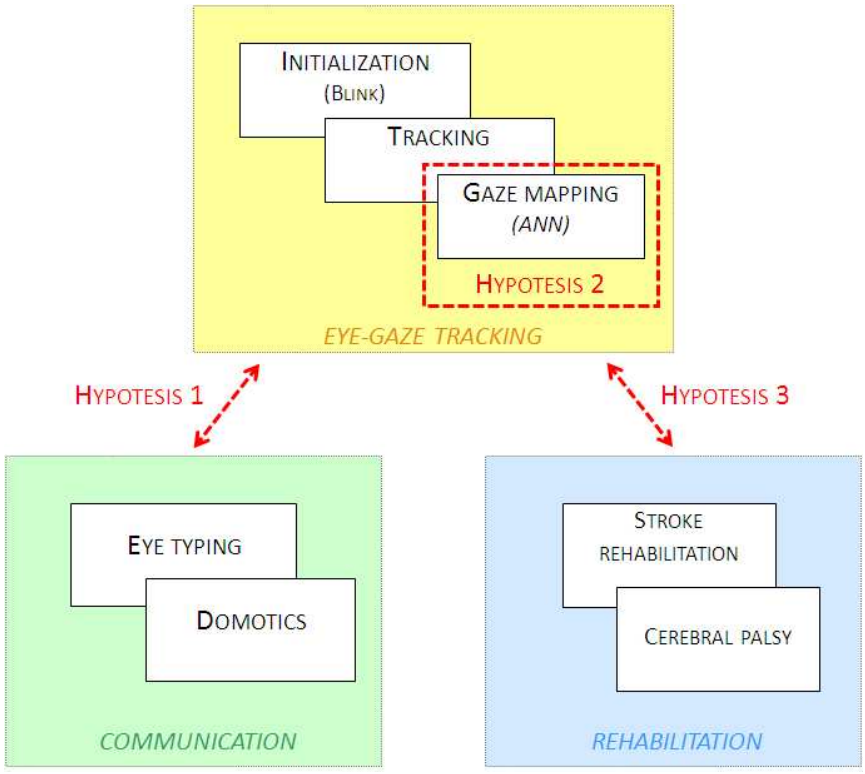

Figure 7.1 – Hypothesis of the work.

7.2 Limitations

With respect to the first and second hypothesis, the present thesis principally focused on the trade-off between accuracy and robustness of eye-gaze tracking, keeping fixed the assumption of low-cost. The developed system showed a good performance if compared with the view-based approaches present in the literature. Nevertheless, compared to the commercial trackers based on infrared technology, the proposed system is less accurate and less robust, preventing it from being used as a stand-alone input device. The principal source of inaccuracy concerns the spatial resolution of the images. The resolution of the eyes within the image frame is quite low, since commercial webcams nowadays work at a maximum resolution art 640x480 pixels during real-time grabbing. As a consequence, the iris movements suffer of quantization errors, that are transferred to errors in gaze estimation. Resolution could be enhanced in three ways:

i) virtually, by sub-pixel techniques. This technique, applied in the present thesis, does not always furnish accurate results, since it is based on mathematical models that in some cases could not match with the reality;

ii) applying an hardware zoom to the camera, or placing the camera really close to the face. This solution causes the drawback of restricting head movements to prevent the eyes from going out of the image frame. Therefore it can be suitable only in specific cases whereas the head does not move, such as in ALS pathology;

iii) using higher resolution cameras. The feasibility of this solution depends on the availability high resolution webcam, not present to this day due to global market policies.

The second limitation of the proposed eye-gaze tracker concern the robustness of big head movements, i.e. the movements that exceed the value of 3 cm. This problem has been overcome in literature only in the context of IR-based tracker. In the view-based context there is still not a concrete and valid solution. With regard to this, the present thesis and some works in literature foresee that neural computing represents a good way to go, not yet extensively explored.

7.3 Future goals

With respect to the limitations highlighted in the previous paragraph, this thesis threw light on possible paths to follow in order to improve the work done so far.

Regarding the accuracy and robustness of the feature tracking, other kinds of template matching techniques should be designed, since cross-correlation requires considerable computational time. Kalman filtering and face detection can be helpful to restrict the areas of analysis and to make the technique be more robust to changes in light conditions.

Future efforts should be also devoted to develop new strategies to address the issue of larger head movements Accounting for large head motion can be performed by combining different approaches. For example, inserting 3D information on the pose of the head can be extremely of help, since usually the user rotates his/her face in the direction of the observed object. Such information, together with a tree-based algorithm, can be used to perform a course to fine approach for gaze detection. The 3D analysis could be done either by using a single webcam together with 3D models of the head or by using additional webcams and sterophotogrammetry methods, paying particular attention to the fact that computational cost grows with the increase of the number of cameras. Inertial sensors (IMU) placed on the head can be used to calibrate and validate the computer vision methods of tridimensional head pose estimation.

Another issue related to the enhancement of gaze tracking is the use of hybrid gaze mapping. Even if the geometrical problem related to gazing is known, some factors as the movement of the head, the projection of the 3D world into a 2D representation and the morphological variability of the subjects introduce high, and not always predictable, non-linearities. This problem can be overcome by the use of deterministic gaze mapping function in combination with neural networks: the deterministic approach will take into account the geometrical problem of gazing, while the neural approach would make the generic model adapt to the specific subject/condition.

References

1 European Commission. “Access to Assistive Technology in the European Union” Web access: http://ec.europa.eu/employment_social/publications/2004/cev5030 03_en.html.
2 G.Eizmendi and J.M. Azkoitia. Technologies and active strategies for an ageing population. Assistive technology. Issue nº2. April 2007.
3 N. Bevan. International Standards for HCI and Usability. International Journal of Human Computer Studies, 55(4), 533-552
4 ISO 9241. Ergonomics of Human System Interaction by ISO
5 ISO/IEC FDIS 9126-1. Software engineering. Product quality. Part1.
6 ISO DTS 16071.Guidance on accessibility for human-computer interfaces (2002)
7 World Health Organization. International classification of impairments, disabilities and handicaps. Geneva: WHO, 1980.
8 Andrew Sears and Mark Young, “Physical Disabilities and Computing Technologies: An Analysis of Impairments”, in The Human-Computer Interaction Handbook, 488
9 D.W. Hansen, and A.E.C. Pece, Eye tracking in the wild, Computer Vision and Image Understanding 98 (2005) 155–181.
10 T. Hutchinson, K.J. White, K. Reichert, L. Frey, Human-computer interaction using eye-gaze input, IEEE Transactions on Systems, Man, and Cybernetics 19 (1989) 1527–1533.
11 SMI, SensoMotoric Instruments GmbH, Teltow, GERMANY. http://www.smi.de
12 M. Adjouadi, A. Sesin, M. Ayala, and M. Cabrerizo, in: K. Miesenberger, J. Klaus, W. Zagler, and D. Burger (Eds.), Computers Helping People with Special Needs, Springer Berlin / Heidelberg, Germany, 2004, pp. 761/769.
13 C.H. Morimoto, M.R.M. Mimica, Eye gaze tracking techniques for interactive applications, Computer Vision and Image Understanding 98 (2005) 4–24.
14 D.A. Robinson, A method of measuring eye movement using a scleral search coil in a magnetic field, IEEE Transactions on Biomedical Engineering 10 (1963) 137-145.
15 M.J. Coughlin, T.R. Cutmore, T.J. Hine, Automated eye tracking system calibration using artificial neural networks, Computer Methods and Programs in Biomedicine 76 (2004) 207-220.
16 L.E. Hong, M.T. Avila, I. Wonodi, R.P. McMahon, G.K. Thaker, Reliability of a portable head-mounted eye tracking instrument for schizophrenia research, Behavioural Research Methods 37 (2005) 133-138.
17 Tobii Technology AB, Stockholm, SWEDEN - http://www.tobii.se
18 Eye Response Technology, Charlottesville, USA - http://www.eyeresponse.com
19 LC Technologies, Inc., Eye gaze systems, Virginia, USA - http://www.eyegaze.com/
20 EyeTech Digital Systems - http://www.eyetechds.com/
21 Millar-Scott, Call Centre 1998.
22 Ward, D. J. & MacKay, D. J. C. Fast hands-free writing by gaze direction. Nature, 418, 838, (2002)
23 Hansen, J.P., Hansen, D.W., Johansen, A.S., Bringing Gaze-based Interaction Back to Basics in Universal Access In HCI, C. Stephanidis (ed.), Lawrence Erlbaum Associates. 2001, p.325-328
24 Harbusch, K. and Kühn, M. 2003. Towards an adaptive communication aid with text input from ambiguous keyboards. In Proceedings of the Tenth Conference on European Chapter of the Association For Computational Linguistics (EACL 2003)
25 COGAIN, Sixth Framework programme priority 2. Information Society Technologies IST. Charter 6 Joint programme of activities, 2004
26 D. Bonino, A. Garbo, An Accessible Control Application for Domotic Environments, First International Conference on Ambient Intelligence Developments, September 2006, Sophia-Antipolis, pp.11-27. Ed. Springer-Verlag, ISBN-10: 2-287-47469-2
27 ISO 9241. Ergonomics of Human System Interaction by ISO.
28 ISO/IEC FDIS 9126-1. Software engineering. Product quality. Part1.
29 D. A. Norman, The Design of Everyday Things, vol. 1, MIT Press, Boston, Massachusetts, 1st edition, 1998.
30 Hanson, V.L. Web access for elderly citizens. WUAUC, Alcacer do sal, Portugal, 2001.

31 S. Shih, J. Liu, A novel approach to 3d gaze tracking using stereo cameras. IEEE Transactions on Systems, Man, and Cybernetics 3 (2003) 1–12.
32 C. Morimoto, D. Koons, A. Amir, M. Flickner, Pupil detection and tracking using multiple light sources, Image and Vision Computing 18 (2000) 331–336.
33 Q. Ji, and Z. Zhu, Non-intrusive Eye and Gaze Tracking for Natural Human Computer Interaction, MMI-Interaktiv 6 (2003), ISSN 1439-7854.
34 D.H. Yoo, M.J. Jin Chung, A novel non-intrusive eye gaze estimation using cross-ratio under large head motion, Computer Vision and Image Understanding 98 (2005) 25–51.
35 S. Baluja, D. Pomerleau, Non-intrusive gaze tracking using artificial neural networks, in: J.D. Cowan, G. Tesauro, and J. Alspector (Eds.), Advances in Neural Information Processing Systems (NIPS) 6, Morgan Kaufmann Publishers, San Francisco, CA, 1994, pp. 753-760.
36 K.H. Tan, D.J. J Kriegman, N. Ahuja, Appearance-based Eye Gaze Estimation, in: Proceedings of the IEEE Workshop on Applications of Computer Vision—WACV02, 2002, pp. 191–195.
37 L.Q. Xu, D. Machin, P. Sheppard, A Novel Approach to Real-time Non-intrusive Gaze Finding, in: Proceedings of the Ninth British Computer Vision Conference-BMVC, 1998, pp. 428-437.
38 Y. Ebisawa, Improved Video-Based Eye-Gaze Detection Method, IEEE Transactions On Instrumentation and Measurement, 47 (1998) 948-955.
39 C.H. Morimoto, M.R.M. Mimica, Eye gaze tracking techniques for interactive applications, Computer Vision and Image Understanding 98 (2005) 4–24.
40 Z. Zhu, Q. Ji, Eye Gaze Tracking Under Natural Head Movements, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 918-923.
41 J. Zhu, J. Yang, Subpixel eye gaze tracking, Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition, 2002, pp. 124–129.
42 A.T. Duchowski, A breadth-first survey of eye tracking applications, Behavior Research Methods, Instruments, and Computers 34 (2002) 455–470.
43 R.J.K Jacob, and K.S. Karn, Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises, in: R. Radach, J. Hyona, and H. Deubel (Eds.), The mind's eye: cognitive and applied aspects of eye movement research, North/Holland/Elsevier, Boston, MA, 2003, pp.573-605.
44 Z. Zhu, Q. Ji, Novel eye gaze tracking techniques under natural head movement IEEE Transactions on Biomedical Engineering 99 (2007) 1-1.
45 R. Newman, Y. Matsumoto, S. Rougeaux, A. Zelinsky, Real time stereo tracking for head pose and gaze estimation, Proceedings of the fourth IEEE international conference on automatic face and gesture recognition, 2000, pp. 122-128.
46 D. Beymer, M. Flickner, Eye gaze tracking using an active stereo head, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, vol. II, Madison, WI, 2003, pp. 451–458.
47 T. Ohno, N. Mukawa, A Free-head, Simple Calibration, Gaze Tracking System That Enables Gaze-Based Interaction, ETRA 2004: eye tracking research and applications symposium, pp.115-122, 2004.
48 K.R. Park, A Real-Time Gaze Position Estimation Method Based on a 3-D Eye Model, IEEE Transactions on Systems, Man and Cybernetics, 37 (2007) 199 – 212.
49 Alper Yilmaz, Omar Javed, Mubarak Shah: Object tracking: A survey. ACM Comput. Surv. 38(4): (2006).
50 Khosravi, M. H., Safabakhsh, R., 2005. Human Eye Inner Boundary Detection Using A Modified Time Adaptive Self-Organizing Map. In: Proc. IEEE Intl. Conf. Imag. Proc. 2, pp. 802-805.
51 Lam, K.M., Yan, H., 1996. An Improved Method for Locating and Extracting the Eye in Human Face Images. In: Proc. 13th Intl. Conf. Pattern Recog. 3, pp. 411-412.
52 Schwerdt, K., Crowley, J.L., 2000. Robust face tracking using color. In: Proc. 4th IEEE Intl. Conf. Automatic Face and Gesture Recognition, pp. 90–95.
53 Kothari, R., Mitchell, J.L., 1996. Detection of eye locations in unconstrained visual images. In: Proc.IEEE Intl. Conf. Imag. Proc. 3, pp. 519–522.
54 Zhou, Z.H., Geng, X., 2004. Projection functions for eye detection. Pattern Recogn. 37(5), 1049–1056.

55 D'Orazio, T., Leo, M., Cicirelli, G., Distante, A., 2004. An algorithm for real time eye detection in face images. In: Proc. 17th Intl. Conf. on Patt. Rec. (ICPR'04) 3, pp. 278-281.
56 Kawaguchi, T., Hidaka, D., Rizon, M., 2000. Detection of eyes from human faces by hough transform and separability filter. In: Proc. IEEE Intl. Conf. Imag. Proc. 1, pp. 49–52.
57 Black, M.J., Fleet, D.J., Yacoob, Y., 1998. A framework for modeling appearance change in image sequences. In: Proc. Sixth Intl. Conf. Comp. Vis. (ICCV'98), 660-667.
58 Bhaskar, T.N., Keat, F.T., Ranganath, S., Venkatesh, Y.V., 2003. Blink detection and eye tracking for eye localization. In: Proc. Conf. Convergent Technologies for Asia-Pacific Region, pp. 821–824.
59 Gorodnichy, D.O., 2003. Second order change detection, and its application to blink-controlled perceptual interfaces. In: Proc. IASTED Conf. on Visualization, Imaging and Image Processing, pp. 140-145.
60 Grauman, K., Betke, M., Gips, J., Bradski, G., 2001. Communication via Eye Blinks - Detection and Duration Analysis in Real Time. In: Proc. IEEE Comp. Soc. Conf. Comp. Vis. Patt. Recogn, pp. 1010-1017.
61 Chau, M., Betke, M., 2005. Real Time Eye Tracking and Blink Detection with USB Cameras. Tech. Rep. 2005-12 Boston University Computer Science.
62 Morris, T., Blenkhorn, P., Zaidi, F., 2002. Blink detection for real-time eye tracking. J. Netw. Comput. Appl. 25, 129-143.

63 Parker, James R., Algorithms for Image Processing and Computer Vision, New York, John Wiley & Sons, Inc., 1997, pp. 23-29.
64 Canny, John, "A Computational Approach to Edge Detection," IEEE Transactions on Pattern Analysis and Machine Intelligence,Vol. PAMI-8, No. 6, 1986, pp. 679-698.
65 Lewis, J. P., "Fast Normalized Cross-Correlation," Industrial Light & Magic.
66 S. Haykin. Neural Networks: A Comprehensive Foundation. 1998, Prentice Hall.
67 Specht, D.F. (1991). A general regression neural network. IEEE Transactions on Neural Networks, 2:568-576, 1991.
68 Jacob, R.J.K. The Use of Eye Movements in Human- Computer Interaction Techniques. ACM Transactions on Information Systems 9 (3), 1991, pp. 152-169.
69 World Health Report - 2007, from the World Health Organization
70 Royal College of Physicians. Stroke: towards better management. Royal College of Physicians, London, 1989.
71 D. T. Wade, R. Langton-Hawer, V. A. Wood, C. E. Skilbeck, and H. M. Ismail. The hemiplegic arm after stroke: measurement and recovery. Journal of Neurology, Neurosurgery, and Psychiatry, 46(4):521–524, 1983.
72 Popovic MB, Popovic DB, Sinkjaer T, Stefanovic A, Schwirtlich L: Restitution of reaching and grasping promoted by functional electrical therapy. Artif Organs 2002, 26:271-275.
73 Galen SS, Granat MH, Study of the Effect of Functional Electrical Stimulation (FES) on walking in children undergoing Botulinum Toxin A therapy. In Proceedings of the First FESnet Conference: 2-3 September 2002; Glasgow. Edited by Hunt KJ and Granat MH. Glasgow: University of Strathclyde; 2002: 31-32.
74 Gritsenko V, Prochazka A: A functional electric stimulation-assisted exercise therapy system for hemiplegic hand function. Archives of Physical Medicine and Rehabilitation 2004, 85:881-885.
75 Freeman, C. T., Hughes, A. M., Burridge, J. H., Chappell, P. H., Lewin, P. L. and Rogers, E. (2006) Iterative Learning Control as an Intervention Aid to Stroke Rehabilitation. In: UKACC Control 2006 Mini Symposium: InstMC Control Concepts in Healthcare and Rehabilitation, 31 August 2006, Glasgow, Scotland, UK.
76 Giuffrida JP, Crago PE: Reciprocal EMG control of elbow extension by FES. Ieee Transactions on Neural Systems and Rehabilitation Engineering 2001, 9:338-345.
77 Hendricks HT, MJ IJ, de Kroon JR, in 't Groen FA, Zilvold G: Functional electrical stimulation by means of the 'Ness Handmaster Orthosis' in chronic stroke patients: an exploratory study. Clin Rehabil 2001, 15:217-220.
78 Popovic D, Stojanovic A, Pjanovic A, Radosavljevic S, Popovic M, Jovic S, Vulovic D: Clinical evaluation of the Bionic Glove. Archives of Physical Medicine and Rehabilitation 1999, 80:299-304.

79 Goffredo M., Bernabucci I., Schmid M., Conforto S., A neural tracking and motor control approach to improve rehabilitation of upper limb movements, Journal of NeuroEngineering and Rehabilitation 2008, 5:5 (5 February 2008).
80 J. H. Burridge and M. Ladouceur. Clinical and therapeutic applications of neuromuscular stimulation: A review of current use and speculation into future developments. Neuromodulation, 4(4):147–154, 2001.
81 T. Sinkjaer and D. Popovic. Peripheral nerve stimulation in neurological rehabilitation. In 3rd world congress in Neurological Rehabilitation, Venice, Italy, April 2002.
82 Daniela Zambarbieri – Movimenti oculari – Pàtron Editore (book).
83 K. Takakiy, D. Aritay, S. Yonemoto and R.Taniguchi, Using gaze for 3-D direct manipulation interface, FCV2005 The 11th Japan-Korea Joint Workshop on Frontiers of Computer Vision
84 Jeff Bennett Pelz, Visual Representations in a Natural Visuo-motor Task, PhD thesis, Department of Brain and Cognitive Sciences, The College Arts and Sciences, University of Rochester, Rochester, New York, 1995
85 Land, M. F., & Hayhoe, M. (2001). In What Ways Do Eye Movements Contribute to Everyday Activities. Vision Research, 41(25-26), 3559-3565.
86 Pelz, J. B., Canosa, R., & Babcock, J. (2000). Extended Tasks Elicit Complex Eye Movement Patterns. In Eye Tracking Research & Applications (ETRA) Symposium (p. 37-43). Palm Beach Gardens, FL.
87 Hoffmann H., Schenck W., Möller R., Learning visuomotor transformations for gaze-control and grasping, Biological Cybernetics, Volume 93, Issue 2, Pages 119-130, Aug 2005.
88 L. E. Sibert, R. J. K. Jacob, Evaluation of Eye Gaze Interaction, Proc. of the CHI 2000, ACM, New York, pp. 281-288
89 D. Torricelli, M. Goffredo, S. Conforto, M. Schmid, and T. D'Alessio, A Novel Neural Eye Gaze Tracker, Proceedings of the Second International Workshop on Biosignal Processing and Classification - (BPC 2006), 2006, pp. 86-95.
90 Jacob, R. J. K. (1993). Eye-movement-based human-computer interaction techniques: Toward non-command interfaces, in H. R. Hartson & D. Hix, eds, `Advances in Human-Computer Interaction', Vol. 4, Ablex Publishing Corporation, Norwood, New Jersey, chapter 6, pp. 151-190.
91 Mitchell T (1997), Machine Learning, McGraw-Hill.
92 Specht, D.F. (1991). A general regression neural network. IEEE Transactions on Neural Networks, 2:568-576, 1991.
93 Jacob, R. J. K. (1991). `The use of eye movements in human-computer interaction techniques: What you look at is what you get', ACM Transactions on Information Systems 9(3), 152-169.
94 I. Bernabucci, S. Conforto, M. Capozza, N. Accornero, M. Schmid, and T. D'Alessio, "A biologically inspired neural network controller for ballistic arm movements," J Neuroeng Rehabil, vol. 4, pp. 33, 2007.
95 H.A. Martens, P. Dardenne, Validation and verification of regression in small data sets, Chemometrics and Intelligent Laboratory Systems 44 (1998) 99-121
96 Nielsen J., Mack R. L. (eds.) (1994), 'Usability Inspection Methods', John Wiley & Sons.