UNIVERSITY OF SALERNO

FACULTY OF MATHEMATICAL, PHYSICAL AND NATURAL SCIENCES

DEPARTMENT OF PHYSICS "E.R: CAIANIELLO"

DOCTORAL THESIS

IMAGE PROCESSING TECHNIQUES FOR MIXED REALITY AND BIOMETRY

Advisor:

Prof. Andrea F. Abate

Coordinator:

Prof. Roberto Scarpa

Author:

Dott. Stefano Ricciardi

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

in

SCIENCES AND TECHNOLOGIES OF INFORMATION, COMPLEX SYSTEMS AND ENVIRONMENT

Academic Year 2013/2014

"Science is a way of thinking much more than it is a body of knowledge".

Carl Sagan

SUMMARY

Abstract

This thesis work is focused on two applicative fields of image processing research, which, for different reasons, have become particularly active in the last decade: Mixed Reality and Biometry. Though the image processing techniques involved in these two research areas are often different, they share the key objective of recognizing salient features typically captured through imaging devices.

Enabling technologies for augmented/mixed reality have been improved and refined throughout the last years and more recently they seems to have finally passed the demo stage to becoming ready for practical industrial and commercial applications. To this regard, a crucial role will likely be played by the new generation of smartphones and tablets, equipped with an arsenal of sensors connections and enough processing power for becoming the most portable and affordable AR platform ever. Within this context, techniques like gesture recognition by means of simple, light and robust capturing hardware and advanced computer vision techniques may play an important role in providing a natural and robust way to control software applications and to enhance on-the-field operational capabilities. The research described in this thesis is targeted toward advanced visualization and interaction strategies aimed to improve the operative range and robustness of mixed reality applications, particularly for demanding industrial environments.

Biometric recognition refers to the use of distinctive physiological and behavioural characteristics, called biometric identifiers, for automatically recognizing individuals. Being hard to misplace, forge, or share, biometric identifiers are considered more reliable for person recognition than traditional token or knowledge-based methods. Others typical objectives of biometric recognition are user convenience (e.g., service access without a Personal Identification Number), better security (e.g., difficult to forge access). All these reasons make biometrics very suited for Ambient Intelligence applications, and this is especially true for the user's face that is one of the most

common methods of recognition that humans use in their visual interactions. Moreover, face features allow to recognize the user in a non-intrusive way without any physical contact with the sensor. To this regard, the second part of this thesis, presents a face recognition method based on 3D features to verify the identity of subjects accessing the controlled Ambient Intelligence Environment and to customize all the services accordingly. In other words, the purpose is to add a social dimension to man-machine communication thus contributing to make such environments more attractive to the human user.

Chapter 1

Introduction

Today, we live in a image-centric world in which visual communication and comprehension represent crucial ways to interact and learn at any level. The power of images has never been so tangible and valuable in every human activity. Consequently, the interest of scientific research in every aspect of this form of knowledge has never been so strong. The digital revolution has made possible measuring, sampling, processing and even sinthetizing images captured through a new breed of devices almost impossible to imagine a few decades ago, whose tecnological evolution does not know pauses and is itself a stimulus for further development and research in a seamless loop. Computer scientists have the privilege of being at the center of this revolution and research areas such as image processing and image synthesis are constantly pushing the boundary of available technology, with huge effects on the mass-market and on industrial applications as well.

In this context, Mixed/Augmented Reality is emerging from a long period of unfulfilled promises as the related enabling technologies are finally mature enough to unleash the potential of combining computer generated visual contents with the actual enviroment around the user. As the impact of this technology on the everyday life is possibly huge, there is a clear need for addressing the main open challenges that currently still limit its usage, particularly in demanding environments and applications.

On another front of the image related research, the outstanding technological progress that has characterized the development and the pervasive diffusion of high-definition low-noise image sensors, put the basis for ubiquitous biometric applications almost unfeasible in the near past. About twelve years after September 11th 2001, the diffusion of person identification and verification systems has reached a worldwide dimension, as anyone travelling overseas has probably experimented while waiting in a

queue for immigration control procedures. As most biometric systems (face, iris, fingerprint, ear, etc.) rely on image capture and processing to extract and compare user's biometric traits, the worldwide availability of high-performance ubiquitous image capture devices (smartphones, tablet, etc.) is opening the horizon for a new generation of applications.

This thesis, tries to provide a glimpse of these new application fields, focusing on applying techniques of image processing in both the 2D and 3D domains for Mixed Reality and Biometry contexts. The topics covered hereafter, could probably not be considered "mainstream", though they concern aspects which might well have a great impact on the usability of the aforementioned techologies. More in detail, the study presented in the following pages is organized as follows:

- Chapter 2 is dedicated to Mixed Reality topics as detailed below

    - Section 2.1 describes a comprehensive proposal for a mixed reality environment, providing powerful interaction capability with the co-registered virtual/real objects by means of a not-instrumented finger based interface to improving the effectiveness of computer assisted training procedures in mission critical systems,

    - Section 2.2 addresses the topic of "diminished reality". Besides the usual augmenting paradigm common in mixed reality, the proposed approach enables a diminishing visualization strategy allowing the user to see only the fraction of the real object/environment that is visually relevant for the task to be performed.

    - Section 2.3 deals with the occlusion problem related to hand-based interaction in mixed reality. The method described, enables the composition of the virtual objects onto the real background to be performed respecting the distance of each rendered pixel according to the user viewpoint.

- Section 2.4 presents a context adaptive head-up interface, which is projected in the central region of the user's visual field and exploiting gesture based interaction to enable easy, robust and powerful manipulation of the virtual contents which are visualized after being mapped onto the real environment surrounding the user.

- Chapter 3 is dedicated to Biometry and it describes in Section 3.1 a comprehensive face recognition framework based on 3D features to verify the identity of subjects accessing an Ambient Intelligence Environment and to customize all the services accordingly. The face descriptor is based on normal map to enabling fast probe-gallery matching yet it is robust to facial expressions and facial hair by means of specific weighting maps.

Chapter 2

Advanced Interaction and Visualization Methods for Mixed Reality

Over the last decade Augmented/Mixed Reality (AR/MR) technology has become more and more diffused and affordable, dramatically expanding the applicative horizon across fields ranging from aerospace to automotive, from surgery to marketing, proving that the mix between real and virtual has a huge potential for the big enterprise and the mass market as well. The last years in particular, have seen a growing hype about Augmented/Mixed Reality pushed by announcements of new dedicated devices like the Google Glass, just to mention the most known, claiming the ability to augment the vision field with context dependent contents, eventually co-registered to the real world.

It is clear enough that these technologies, apart from technical limitations still concerning tracking accuracy/robustness, or field-of-view wideness, really have a great potential for a broad range of applicative fields and particularly for multimedia training and learning which could finally move from the computer space to the real world.

The growth experimented so far has been stimulated by different factors: a dramatic increase of both general and visual computing power of any kind of computer, a general cost reduction of AR specific devices, like see-trough Head Mounted Displays and motion tracking systems and, last but not the least, the new generations of smartphones, equipped with an arsenal of sensors (hi-res cameras, gyroscopes, accelerometers, GPS, electronic compass, etc.) and enough processing power to became the most portable and affordable AR platform ever. Within this exciting scenario, the research effort should be focused not only on new approaches to the main AR topics (more accurate tracking in outdoor applications, better and lighter hi-res HMD, etc.) but also on open and new challenges as well. The following subsections of this thesis deal with some of these last kind of research topics, and particularly they concern not-instrumented finger-based interface in a MR environment (Section 2.1), diminished reality (Section 2.2), occlusion

handling (Section 2.3) and advanced gesture-based interaction (Section 2.4), also trying to stress the proposed approaches in demanding application contexts.

2.1 Finger Based Interaction for Demanding AR Environments

Mixed Reality technologies often reveal serious limitations when applied to challenging application environments. For these particular context, indeed, specific requirements in terms of tracking accuracy and coverage, augmentation strategies and interaction capabilities determine whether a mixed reality application is useful or not. Mission critical installations such as military and civil radar systems, navigation systems aboard ships and airplanes, high performance communication systems based in airports, ports and oil rigs are just a few examples of high-tech environments featuring an ever growing range of complex hardware and software components. In case of break down of one of these components, it is of paramount importance that the faulty part is repaired as quickly as possible, as the security of a large number of people may be at risk. In this context, the latest advances of Mixed Reality (MR) technologies [1] may prove really useful in assisting on-site operators during servicing and repair activity. Most of on-site interventions in this field depend on trained personnel applying established procedures to complex equipment in relatively static and predictable environments. These procedures are typically organized into well-defined sequences of tasks, concerning specific items in specific locations. A fundamental aspect to be considered is represented by the interaction level available and the related interaction paradigm. The user, indeed, should be able to select what kind of augmenting content to display according to his/her needs by interacting with the MR environment without complicated gear. This section describes a not-instrumented finger based interface to provide effective and reliable visual aid during maintenance operations. This interface has been designed and tested as a part of a comprehensive MR environment aimed to support servicing and repair operations in mission critical systems, but that could be suited to other demanding contexts as well. The proposed architecture is based on a multiple marker-based tracking, and it has been tested in a radar control training facility to assess its benefits and limitations in a real scenario.

Scientific literature presents a number of studies covering the topic of mixed/augmented reality applied to industrial contexts. In 2002, the project ARVIKA [2] fostered the research and the development of AR technologies for production and service in the automotive and aerospace industries, for power/process plants, for machine tools and production gear. Klinker et al. [3] presented an AR system for the inspection of power plants at Framatome ANP, while Shimoda et al. [4] presented an AR solution in order to improve efficiency in nuclear power plants (NPP) maintenance interventions and to reduce the risk of human error. Mendez et al. [5] developed a virtual (AR based) filter to reveal hidden information that is behind an occluding object, to enhance data of interest or to suppress distracting information. Pentenrieder et al. [6] showed how to use AR in automotive design and manufacturing, to analyse interfering edges, plan production lines and workshops, compare variance and verify parts. Still et al. [7] proposed an augmented reality system for aiding field workers of utility companies in outdoor tasks such as maintenance, planning or surveying of underground infrastructure, exploiting geographical information system.

More recently, De Crescenzio et al. [8] described AR based interfaces as valuable tools in preventing manufacture errors in the aviation field. Whatever the context considered, tracking precisely and reliably the user point of view (POV) with respect to six degrees of freedom is of paramount importance for co-registering virtual objects with the surrounding environment. Over the years different technologies have been proposed for this purpose (magnetic, ultrasonic, inertial, computer vision based, hybrid, etc.), each with advantages and disadvantage. However, to this date, none of them can be considered as a general solution, whereas each approach can be suited to a particular domain (indoor/outdoor usage, small/wide/scalable operating volume, presence/absence of ferromagnetic materials or electromagnetic fields, etc.).

Computer vision, in both the marker-based [9] [4] and marker-less [10] [11] variants, is generally recognized as the only tracking methodology that has the potential to yield non-invasive, accurate and low cost co-registration between virtual and real [12]. As the MR interface described in the following pages should be able to working on equipment rich of small components and operating in the vicinity of strong electromagnetic fields, the choice of a multi-marker tracking method seems adequate to

deliver high accuracy and robustness. This last scenario also highlights the relevance of a proper interaction capability within the augmented environment, which could not be properly addressed by conventional input device like mouse and keyboard, as the user usually stands upright, eventually moving. Wei et al. [13] introduced a MR framework featuring voice commands, but a hand-based interface would rather be more suited to the scope. As a hardware solution (instrumented gloves plus wrists tracking) would provide accurate hands capturing but would also reduce system's acceptability, a more feasible option is to exploit image-based techniques to track hands in real time [14]. The simple and robust approach described hereafter is based on the recent work by Mistry and Maes [15] and rely on colored caps worn on index and thumb fingers to track their position and gestures, providing effective and natural interaction within the MR environment. A final aspect to be considered is the computing/visualization hardware required by the system to operate. The growing diffusion of new-generation smartphones/tablets promise to deliver cheaper and more usable [16] AR platforms [17]. This is probably true if the interaction is always mediated by the touch-screen, but when the interaction also implies a contact with physical environment, the user is forced to hold the device with one hand while operating with the other hand behind the device's screen. If this is the case, a prolonged working session is likely to become a stressful experience. For this main reason a video see-through HMD and a backpack enclosed notebook has been preferred over a tablet computer.

2.1.1. Overall system architecture

The system proposed is schematically depicted in Figure 1. It is composed by three main components. The Mixed Reality Engine (MRE) is in charge of user's head tracking, scene augmentation/rendering and servicing procedures management. The User-System Interface captures fingers position/gestures enabling the human-computer interaction, while the Maintenance Database contains the working environment setup, the virtual contents and the maintenance procedures required for the system to work. To start the assisted servicing procedure, the user has to wear a video-based see-through HMD, a backpack enclosed notebook and a few fingertips caps required for contactless interaction. This architecture is further detailed in the following subsections.

2.1.2. Mixed reality engine

As mentioned in the previous sections, the tracking system developed exploits optical markers for estimating user's perspective. It is based on the well-known ARToolkit open source AR library [18] by implementing new functions and designing a marker configuration optimized for the application context considered. Typical marker based tracking systems operating under controlled conditions (i.e. avoiding or at least reducing strong reflections, extreme shadows and excessive camera noise) are able to track the user's point of view provided that the head mounted camera entirely frames a single marker.

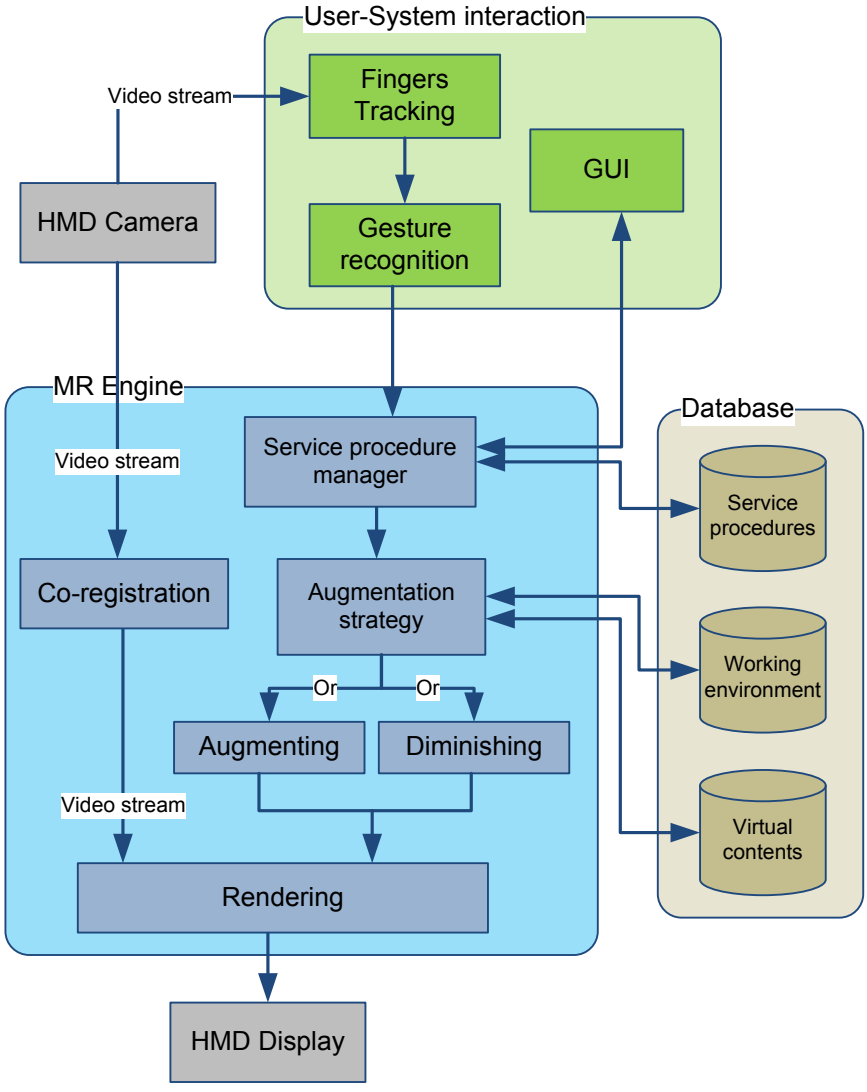

Figure 1: *The overall schematic view of the system.*

This simple solution, often adopted for desktop based AR applications, forces the user to continuously aim at the marker, holding it in the center of the visual field to reduce the risk of detection miss. Moreover, this configuration often involves the need to use large markers (10x10 cm. or more is fairly common), because the accuracy of user's tracking directly depends on precise estimate of marker's apparent position/orientation that, in turn, is affected by the amount of error in measuring marker geometrical features, which are proportionally easier to detect on a larger pixel surface. On the other side, arranging a large marker in the middle of operational environment could simply be unfeasible for many application contexts characterized by uneven surfaces or it could even interfere with the operations.

In this study, many of these issues are addressed by exploiting multiple markers, thus delivering an inherently more robust and more accurate tracking even using small markers. Factors like the average distance between user and augmented object, required tracking volume, camera's focal length and resolution, have to be carefully considered when designing the marker configuration as many of them depend on the particular operating environment. In our test-bed, a set of six 4x4 cm sized markers (see Figure 2) provides an optimal tracking coverage of approximately 60x60x60 cm with an equivalent co-registration error within 2 mm, which is below the size of most small parts. As the relative position of each marker with respect to the absolute reference system is known, when more than one marker is recognized each approximated estimate of camera's position/orientation (relative to a particular marker) contributes in reducing the overall co-registration error through a weighted average strategy based on the quality and number of visual features recognized (see Figure 3). To this regard it has to be remarked that the rotational component of camera tracking has a greater impact on augmentation accuracy compared to the positional component. In fact, even a degree of error may produce a visible misalignment between virtual-to-real as the distance from the tracked point increases. To minimize this effect the origin of absolute reference system (respect to which is expressed the position of any virtual content) is deliberately located in the geometric center of the marker-set to further reduce the rotation co-registration error of all objects falling within it.

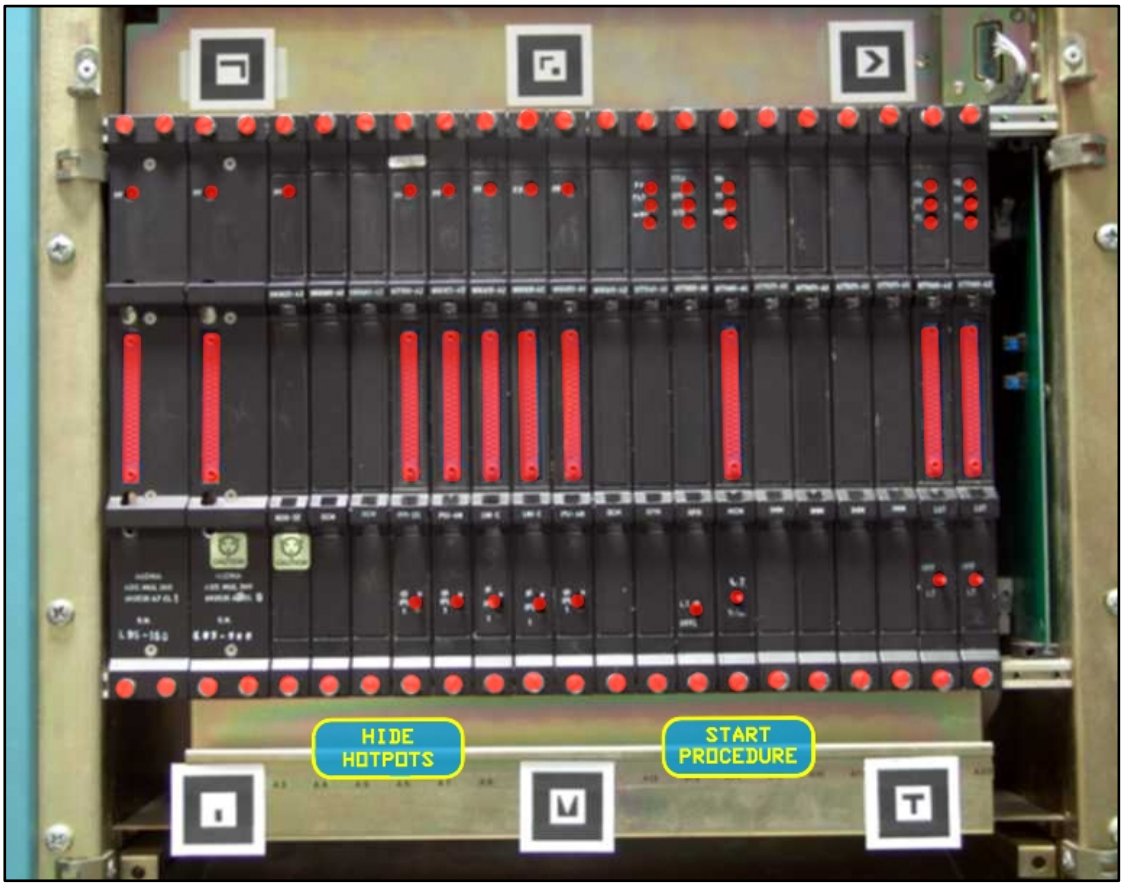

Figure 2: *The full marker set and all the available hotspots highlighted after successful co-registration calibration.*

Additionally, to reduce unwanted camera shaking tracking data was smoothed out by means of a damping function. The marker set is easily scalable. For instance, by adding other six markers (arranged in two strips of three, placed 60 cm. above and below the basic set) an optimal tracking volume of 180x60x60 cm (adequate for a full-size industrial rack) is seamlessly achieved. Besides an embedded calibration function aimed to measure and correct camera's lens distortion, a manual procedure allows the user to fine-tune co-registration between the real camera and its virtual counterpart in charge of rendering the required graphics.

Each of the six degree of freedom plus the camera's focal length and markers' thresholding can be precisely adjusted. This task is performed only once unless physical or environmental changes occur in equipment's configuration.

The hierarchical representation adopted provides an increasingly detailed description while proceeding from the highest to the lowest level. An *environment description* file, indeed, contains the precise positioning of any relevant equipment (e.g. a system rack) including all the associated items (e.g. the boards located within the rack) by means of specific tags. In a similar way, for any item, an *object description* file contains a tag list of all the hotspots associated to it (e.g. switches, screws, warning lights, connectors, etc.). The MR engine, according to the aforementioned descriptors, builds up a virtual scene by means of a DOM XML parser, while another XML based language, Xpath, is used to query the application database to retrieve the required data. The MR engine also performs another crucial task: the maintenance procedure management. Each generic maintenance procedure can be represented as a deterministic finite automaton (DFA). According to this approach, a particular state represents a maintenance step and its links define the execution order.

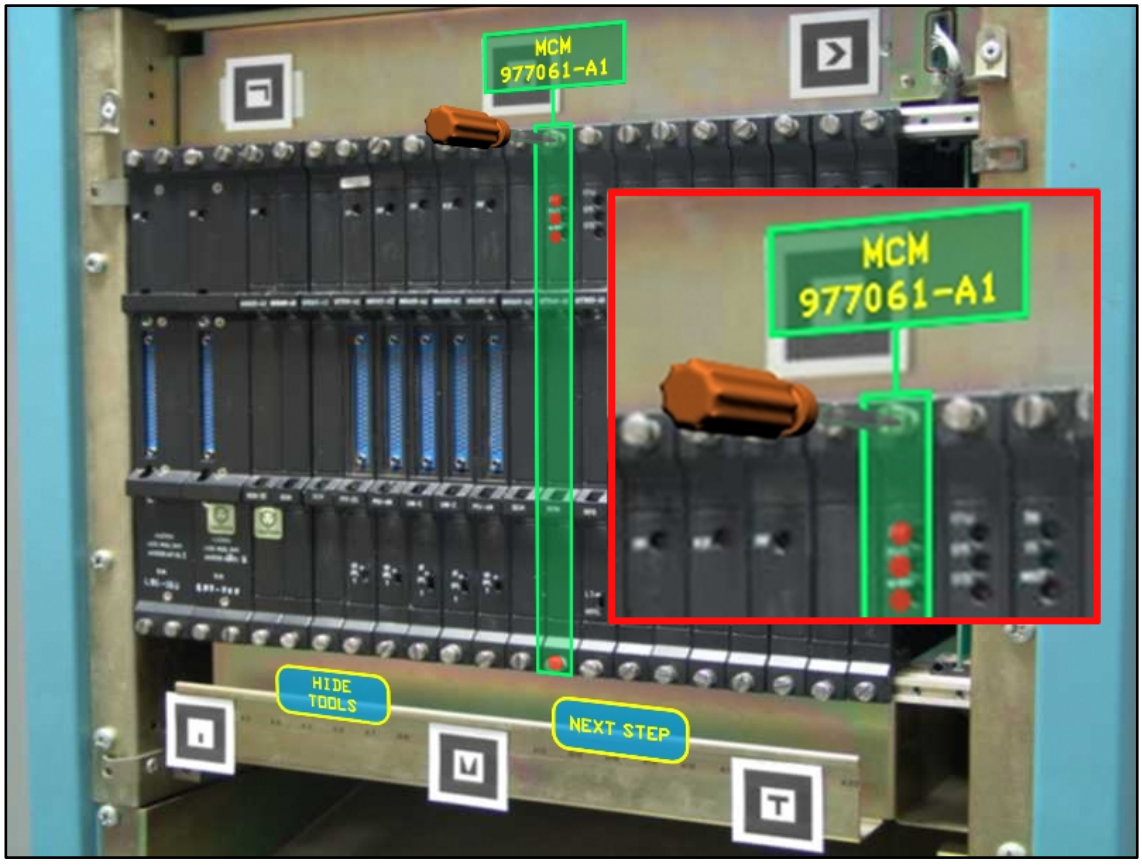

Figure 3: *Panel augmented by virtual labels, tools and GUI (inset) a magnified view showing the small amount of co-registration error*

DFA result particularly suited to model both simple and complex maintenance procedure in an easy, verifiable and legible way. The DFA representation of a particular procedure is converted in a XML file where a <*step*> tag defines a state. Any possible path through the automaton defines a procedure's file.By this approach a single XML procedure file defines a specific execution order in a maintenance procedure. At runtime, this file is progressively parsed, and in every moment the user can switch to the next or previous task by means of the contact-less interface. A fragment of a generic procedure step is shown below.

```
<step equipmentRef=”Server.xml”>
<label rgbText="default" rgbBackground="default">
Unscrew the two fixing screws
</label>
<hotspot>screw</hotspot>
<tool>screwdriver</tool>
</step>
```

This particular step refers to the device *Server.xml.* A label informs user that there are two screws which have to be unscrewed. By parsing Server.xml the engine locates the screws and highlights them by means of a blinking spot. Because a screwdriver is required to perform the step, the <tool> tag allows the MR engine to locate and load the proper 3D model from the virtual content repository to render it onto the corresponding screw showing how to perform the task.

2.1.3. Finger based contact-less interface

The contact-less interface developed frees the user from the usage of any tangible I/O device to communicate with the system. The user indeed, has only to wear small rubber caps of different colors over his thumb and index fingertips, eventually of both hands.
The image-based tracking exploits the same video stream used for the camera tracking to achieve fingertips detection and tracking, thus yielding a reduction of computational cost compared to a solution based on a dedicated camera and a simpler hardware

configuration. Each finger is associated to a different color according to a simple enrollment procedure repeated for each finger and generally performed only the first time the system starts-up. The fingertips samples captured are analyzed in the HSL color space to extract the dominant hue and saturation ranges, while the lightness component is used to filter out eventual highlights. At runtime, the image grabbed from the camera is subsampled (by a factor of 4 to 8 times) to both reduce the effect of camera noise and to optimize system's performance. For each pixel in the subsampled image whose HLS levels fall within the ranges defined during enrollment, a recursive search for similar (color wise) neighbors is performed until a region of 20x20 pixels is explored. If at least one half of the pixels inside this region matches with the original pixel, then the engine recognizes that region as one of the colored caps to track (see Figure 4). This approach resulted both reliable and responsive, granting a sustained frame rate always well above 30 frame per second for an image resolution of 640x480 pixels (typical for most HMD cameras). Finger tracking enables a rich interaction paradigm that can be exploited in many different ways. For instance the user may query the working environment to learn more about it by simply moving the index finger over any hotspot (i.e., a screw, a button, a handle, a led indicator and so on) according to his point of view to obtain visual info about a particular component. Moreover, finger tracking enables operating the system by using a graphical user interface (see Figure 5). The main challenge with an intangible GUI is related to the interaction paradigm, which has to manage the lack of physical contact with the interface elements (buttons, slider, toggles etc.). Indeed, when using a conventional (tangible) interface, the kinesthetic feedback provides an important confirm of the operations performed. To address this issue, a time based interaction paradigm was exploited, requiring the user to hold the finger in position for a defined (around one second) amount of time to trigger the associated function. A visual feedback, in the form of a small progress bar drawn over the GUI element selected, inform about the selection state (i.e. hold the finger until the progress is over). The same paradigm is used during a servicing procedure to move from a step to the next one or previous one as well as to play/pause/rewind an animated virtual tool showing how to perform a specific task.

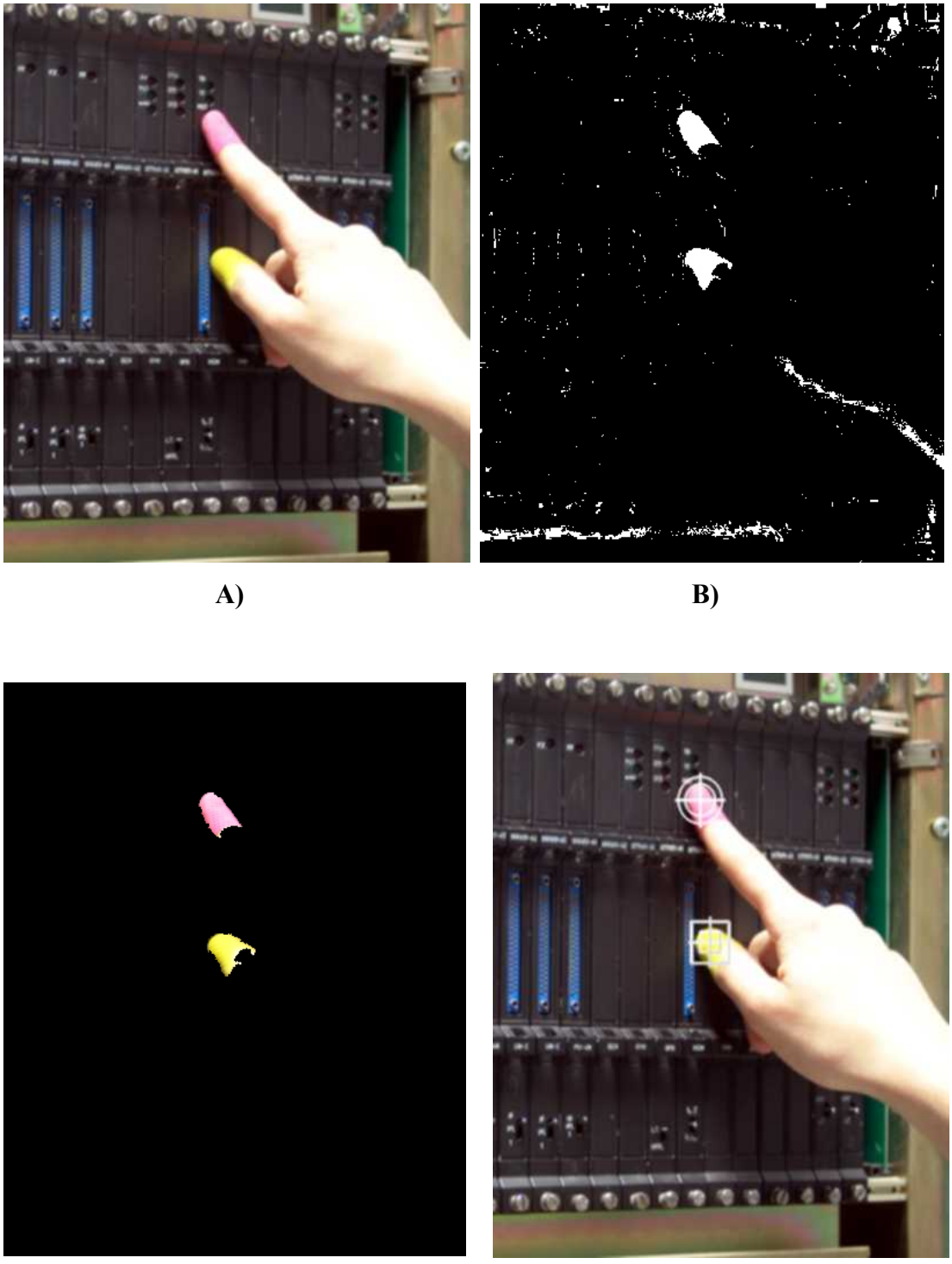

Figure 4 – *Fingertips tracking in four steps: (A) original frame grabbed from the camera; (B) candidate pixels highlighted (C); matching regions found; (D) resulting tracking*

Additionally, as the system is designed to track up to four colored caps, multi-finger gestures can be used to provide more powerful interaction modalities, like object picking, zoom or rotation.

2.1.4. Experiments and comments to the results

Two kinds of experiments have been conducted on the system described above, to assess both the performance of the tracking approach and the overall usability of the MR environment applied to a training facility. The hardware used for the experiments includes a notebook, featuring Intel I5 processor and Nvidia GeForce 9 series graphics board and an ARVision-3D video see-through HMD from Trivisio, equipped with two 800x600 LCD display and two 640x480 cameras capturing the surrounding environment at 30 FPS (see Figure 6.).

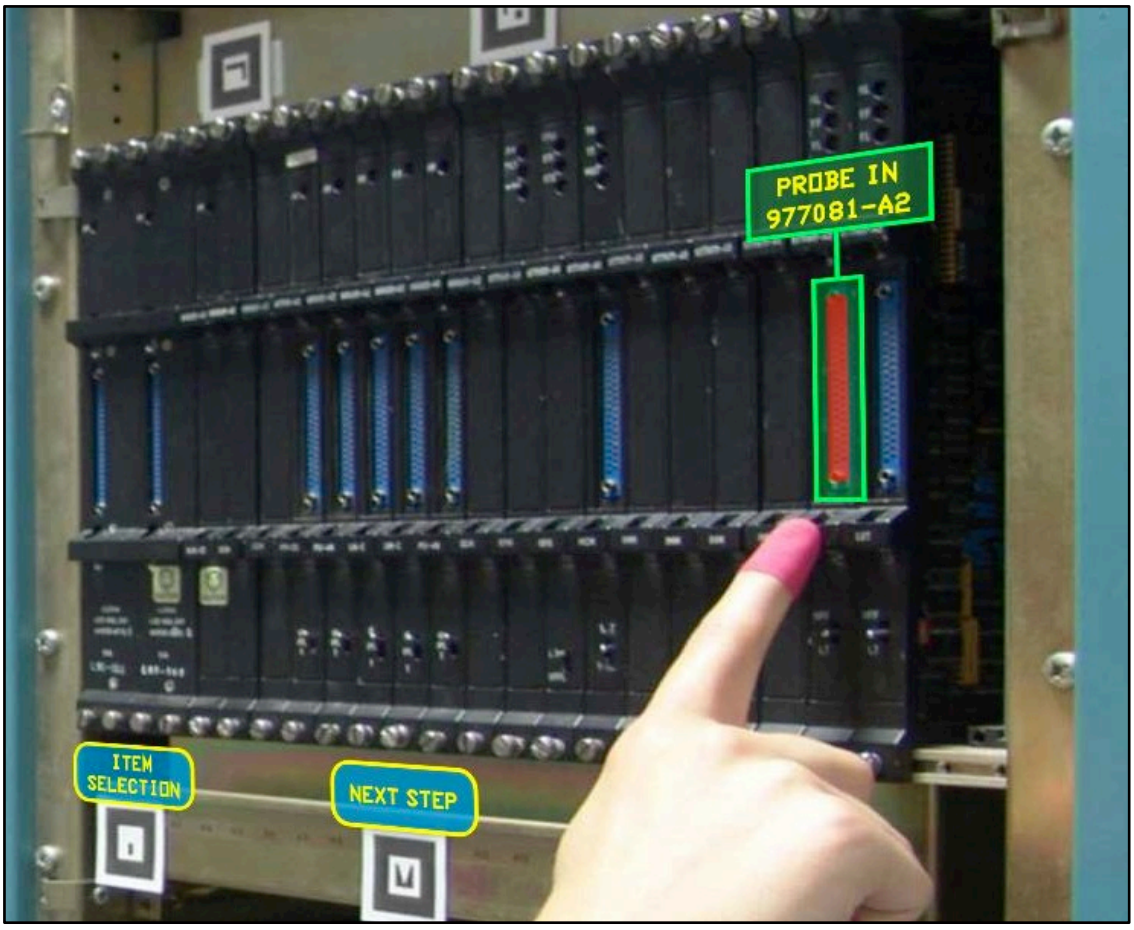

Figure 5: *User interacting with the servicing assistance environment through a contactless finger based interface while performing a simulated maintenance procedure.*

Only the left camera has been used for scene capture. During operations the notebook was contained inside a small backpack. Both the accuracy and the robustness to camera motion of the multi-marker tracking have been tested. The measurements take into account the global error amount due to the combined effect of position and rotation errors. The results are summarized in Figure 7, and they overall confirm that accuracy delivered is more than adequate for the target application. the At a distance of 30 cm an augmented point results offset respect the real position only for 1.64mm if only one marker is recognized. This error amount falls under 1.09 mm when two markers are detected. Due to rotation error if the same point is seen under an angle of 45° the error increases to 2.19 mm for a single marker. The same evaluations have been done at distance of 60cm and 90cm. Not surprisingly, as the distance increases, the number of markers recognized increases too. This ensures that the error remains small even at greater distance. In fact at 90 cm with six markers recognized the error is of just 2 mm meaning that for the user the perceived error is nearly negligible. Another important aspect affecting the tracking accuracy is represented by the camera angular velocity.

Figure 6 *User wearing HMD and colored fingertip caps.*

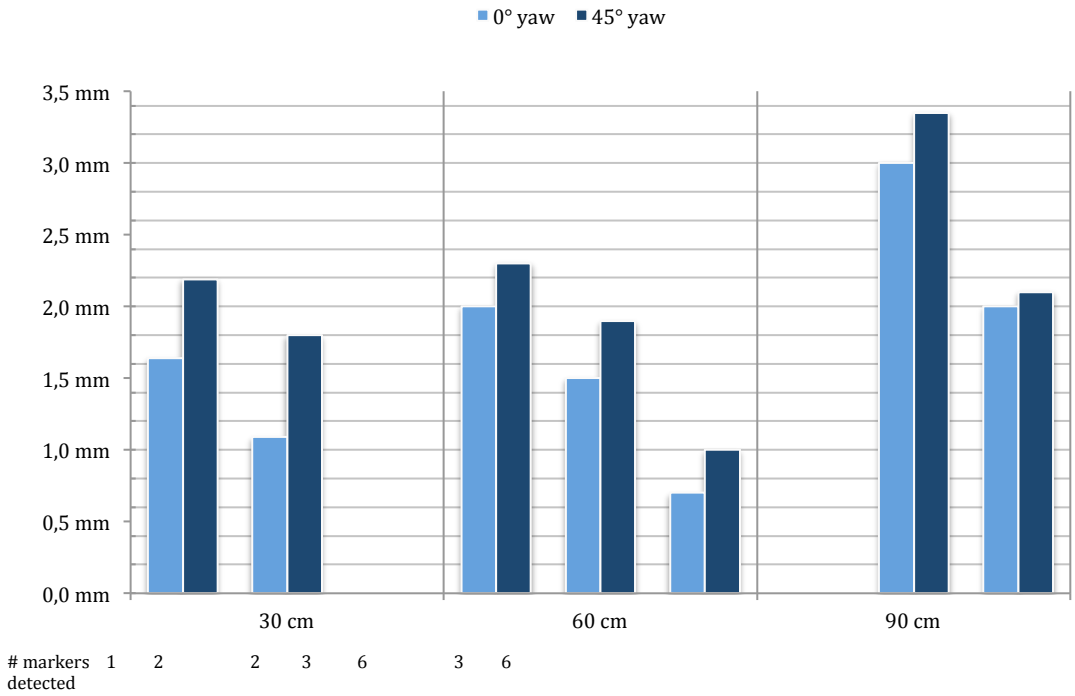

Figure 7: *Co-registration error measured according to number of markers detected, user-markers distance and off-plane angle*

When the user rotates his head rapidly (this happens mostly with respect to the vertical axis), the video stream may results in blurred frames negatively affecting markers detection and recognition due to insufficient image contrast. This issue is directly related to the camera's capturing speed, so the higher the frame rate the lower the blur produced and consequently the higher the angular velocity allowed. In the case considered, (the HMD's camera operating at a common 30 FPS) the system was able to track the user reliably until the angular velocity is below 2 rad/sec. Over this limit a tracking failure is very likely, however as soon as the speed slow down the system recover form the error condition almost instantaneously. In any system evaluation, user testing is of great relevance in confirming the validity and the effectiveness of solutions adopted. To this aim, a user questionnaire has been prepared to assessing the perceived quality of the interaction after performing a number of tasks significant to the operating context considered. The evaluation sessions involved ten users, selected among

specialized technicians with no previous experience of either MR systems or contactless interfaces. The following is a list of the tasks performed by the testers:

- *Load a new servicing procedure.*
- *Select a particular hotspot.*
- *Select a function from the GUI.*
- *Toggle between two functions.*
- *Perform a servicing procedure.*

In the final questionnaire, the questions were presented using a five-point Likert scale, where respondents specify their level of agreement to a statement. In order to avoid any bias, some statements were in positive form and others in negative one. This was taken into account in the final assessment of results. The following is the list of the proposed statements:

1. *Available finger based functions are easy to perform*
2. *Functions are too many to remember them*
3. *Interacting by fingers is not intuitive*
4. *It is easy to select objects*
5. *Visual aids are clear and useful*
6. *It is easy to operate the contact-less GUI*
7. *The type and number of available functions to interact with objects is not sufficient*
8. *Devices worn are not comfortable during operations*

The answers to the questionnaire are summarized in Figure 9 above. Most participants reported a good confidence feeling during system's usage, and some of them also reported an operational advantage in performing the proposed tasks with respect to their usual operating modality. All the participants to the evaluation sessions have also been interviewed to better understand the motivations behind the answers provided.

QUESTION	I strongly agree	I agree	I do not know	I disagree	I strongly disagree
1	1	6	1	2	0
2	0	2	1	7	0
3	0	1	2	5	2
4	2	6	0	2	0
5	2	7	0	1	0
6	0	6	2	1	1
7	0	1	2	6	1
8	2	4	1	3	0

Table 1: *Scores reported after subjective system evaluation according to five-point Likert scale*

Most comments showed a general satisfaction about the finger-based interface, though most of them remarked the lack of a physical contact as something strange which is not easy get used to. Both the two visualization modalities were considered useful for improving the confidence and avoiding distraction errors during the operations, while, not surprisingly, the HMD caused a somewhat stressful experience to most users.

The subjective system evaluation highlights the potential of the proposed approach, though issues related to the hardware used might sometimes detract from the MR experience. According to questionnaire answers, the combination of augmentation and finger-based interface worked well, providing an intuitive interaction paradigm that proved to be suited to the application context. Overall, the MR aided servicing environment produced a valuable improvement in user's confidence during simulated interventions, which could eventually lead to a measurable reduction of time required to tasks completion.

2.2. Diminished Reality, an Alternative Visual Strategy for MR/AR

In previous Section 2.1., it has been remarked the importance of an effective visualization strategy to maximize the potential of virtual contents augmenting the visual field for different applicative scenarios. How and when to augment real objects and environment may have a great impact on the quality of user assistance provided. For instance, in some situations the scene observed should be possibly simplified rather than augmented. To this regard the concept of diminished reality might prove useful to simplify complex assemblies and to improve user confidence during operations. Generally, this term refers to removing real-world objects from a live video stream, as demonstrated in [19] and [20]. The common idea behind these works is to remove the unwanted real object and to reconstruct the portion of the scene occluded by it exploiting multi-angle capturing, a technique not always viable in a typical AR setup. Herling and Broll [21] also propose a diminished reality environment, asking the user to select the object to hide. However, this selection could be problematic in case of a high density of components like switches, screws, etc. (not infrequent in industrial environments). There are situations, indeed, in which adding extra information to the scene may lead to an even more confusing effect. In particular, this issue might arise in environments characterized by the presence of a large amount of interaction points (e.g., a control board, a rear panel of a complex device etc). In this case, showing further information in addition to those already present might be counterproductive. These are the main considerations behind this proposal of an alternative augmentation strategy, inspired by the concept of diminished reality and based on the selective occlusion of unwanted elements rather than on image based object removal. Two different examples of this kind of "diminished reality" are reported in the following sections 2.2.1. and 2.2.2., and they are both based on the Mixed Reality Engine described in Section 2.1.2. and applied to two different contexts: the industrial environment already presented and a home-environment, targeted to the broadest possible audience and not requiring dedicated hardware.

2.2.1. Selective removal of equipment features

As recalled before the MRE is able to support two different approaches to scene augmentation. Indeed, beside the "classic" strategy consisting in the visualization of different kind of virtual objects (e.g., arrows, labels, 3D models and so on) onto the captured scene, it can even remove distracting items from the field of view, leaving as visible only the elements required to perform the desired task. The goal in hiding part of the operating environment is to let the user focus on the physical elements on which to perform a particular task. From a more technical point of view, to the purpose of hiding real objects or part of them the engine renders an occluding (polygonal) surface on which can be applied either a diffuse map or an opacity map. By exploiting the device's formal representation (see section 2.1.3.), the engine either load the associated textured polygons (see Figure 8_bottom) or it builds up a procedural texture consisting of a black background featuring white "holes" of various shapes (e.g, circles, squares, polygons and so on) corresponding to the hotspots that should stay visible. All the necessary information to perform this task is available in the working environment database. Once the textured polygons are built, the engine renders it over the real device (see Figure 8 _top) occluding all the contextually not relevant hotspots. Both the more common augmenting and the diminishing strategies are meant to improve user's operational capabilities, however, there are contexts in which one is more suited than the other. Which of the two visualization methods should be used depends on the total number of hotspots present in the surrounding of the virtual contents visualized at a given step of the intervention. If the density of the hotspots exceeds a threshold then the diminishing modality is preferred. Anyhow, the user may always switch to the other modality in any moment by means of a specific Augmented/Diminished View toggle present in the visual interface. Finally, by combining both augmented and diminished reality a third hybrid visualization approach could be realized, providing a simplified view of the operating field in which only the elements left visible are augmented with additional info. Whatever the strategy adopted, scene augmentation or diminution is made possible thanks to a formal scene representation based on XML. The XML database consists of a collection of files providing the necessary information to correctly locate each relevant element of the working environment within the 3D space.

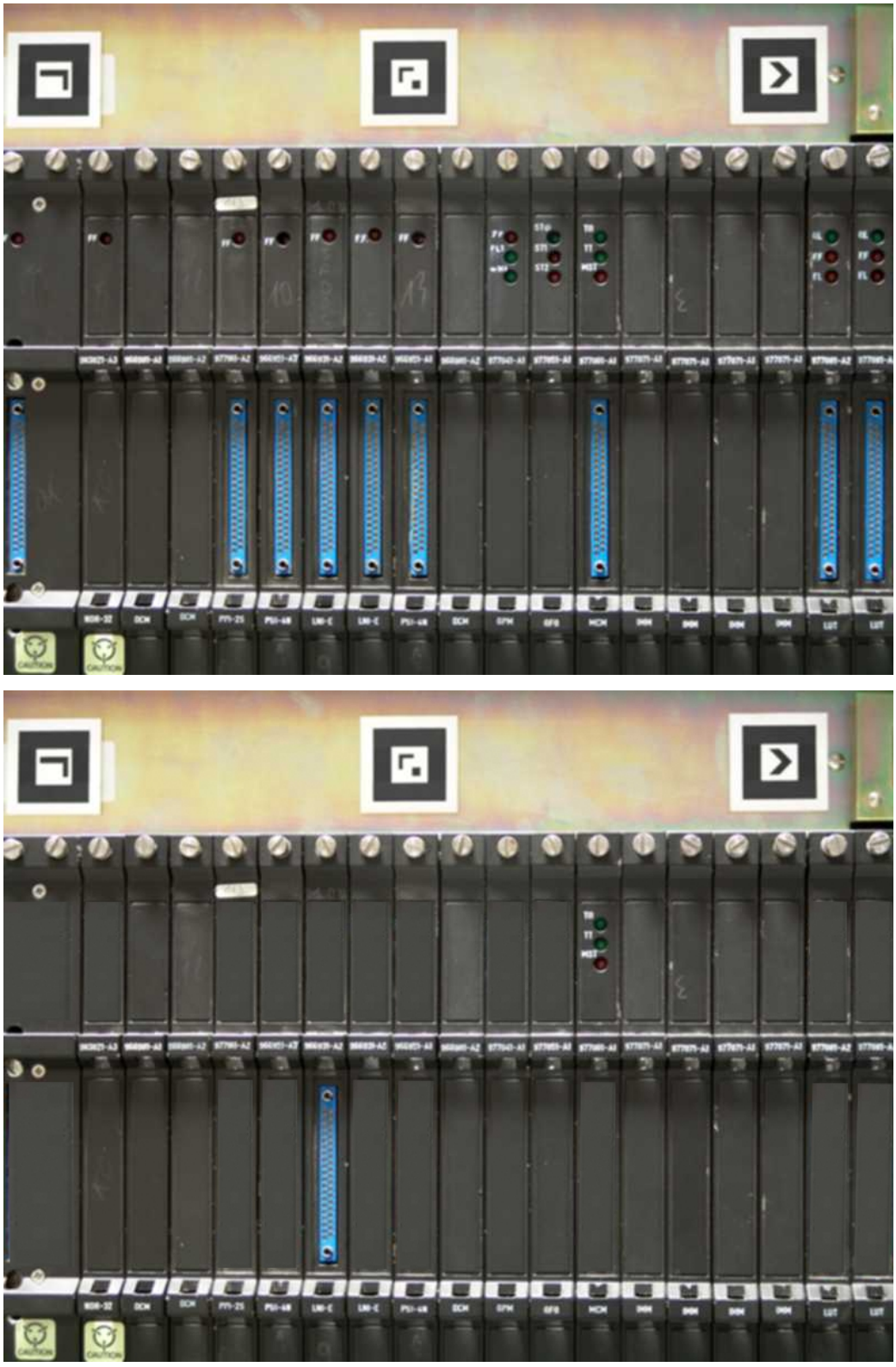

Figure 8: *Original frame (top) and diminished view (bottom). Only the required elements are left visible, while textured polygons hide potentially confusing items.*

2.2.2. A WYSIWYN approach to objects augmentation

In the last years, the growing diffusion of lightweight portable computing device like netbooks, tablets, and smartphones, featuring adequate processing power coupled with trackpad/touchpad interface, one or two webcams and eventually additional sensors (accelerometers, gps, gyroscopes, digital compass, etc.) has provided a low-cost platform to augmented reality applications, usually relying on more dedicated but also expensive and bulky technologies like see-through head mounted displays. There are several contributions, in literature, based on this technological premise. Counter Intelligence [22] is a proposal for a conventional kitchen augmented with the projection of information onto its objects and surfaces to orient users, coordinate between multiple tasks and increase confidence in the system. CyberCode [23] is a visual tagging system based on a 2D-barcode technology and provides several features not provided by other tagging systems. CyberCode tags can be recognized by the low-cost CMOS or CCD cameras found in more and more mobile devices, and it can also be used to determine the 3D position of the tagged object as well as its ID number. Chuantao et al. [24] present a contextual mobile learning system framework, enabling to learn mastering domestic and professional equipments using mobile devices like Tablet PC, PDA or Smartphone and exploiting RFID technology to achieve contextualization. Gausemeier et al. [25] describe an image based object recognition and tracking method for mobile AR-devices and the correlative process to generate the required data. The object recognition and tracking base on the 3D-geometries of the related objects.

The purpose of this application is to showcase how the AR architecture described in 2.1.3. can be successfully re-engineered for applications targeted to everyday objects and environments and requiring inexpensive hardware like a compact netbook or a tablet PC. In particular, the proposed system is focused on providing accurate augmentation of AV components by means of visual aids to ease the most complex procedure involved with the advanced use of this diffused hi-tech equipment, which hardly could be performed without referring to the user manual. The system supports either the typical “additive” augmentation paradigm (co-registered graphics like labels or 3D objects) or the so called “mediated” or “diminishing” approach to AR, consisting in the “What You See Is What You Need” (WYSIWYN) approach to selective removal

of part of the physical object to focusing user attention only on the visual features required in a particular operative context. One interesting aspect of this system is that a portion of the user/machine interface is the same augmented device. Its buttons, labels, handles and so on are the graphical interface widgets of which the user disposes to interact with the system. During any operation, the user can choose if sending the suggested commands by trackpad/touchscreen interface or directly on the device. In the former case the AR engine sends the command to the device through a serial communication interface acting as a middleware. In the latter case the system acts only as a virtual assistant.

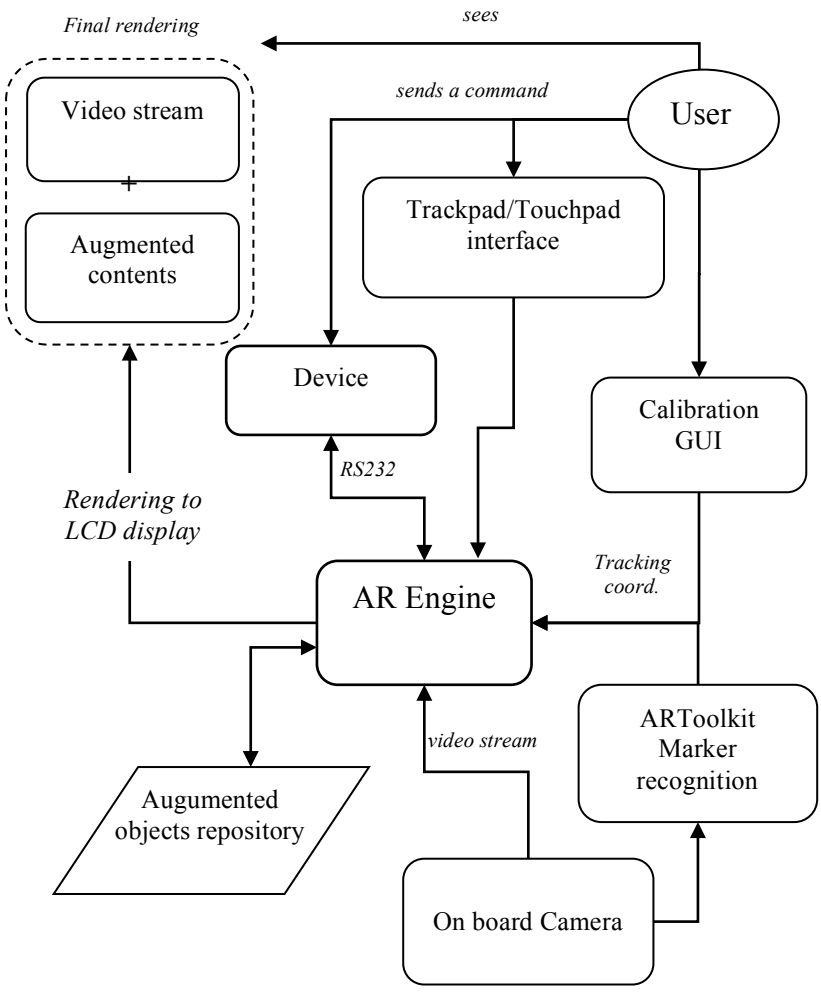

Figure 9. *The main system components.*

Figure 10. *Six markers surrounding an AV component. Each marker is a 4 cm. side square.*

The overall schematic view of the system is shown below in Figure 9. The AR engine is the main system component. It is built on Quest3D, a commercial authoring environment for real-time 3D applications featuring an "edit-while-executing" programming paradigm [26]. Each scene augmentation is applied to specific hotspots defined over the AV component by means of previous measurements and a XML database is used to record the precise hotspots locations. When the user points the camera toward the component, the AR engine load the corresponding XML file. The syntax of a device file consists essentially of a list of <hotspot> tags. An <id> tag allows the engine to recognize any hotspot from each other while a <shape> tag is exploited as approximation of the hotspot real shape to highlight it when required. A <details> tag provide short information about the function of the hotspot. Any hotspot tag has three child tags <position>, <rotation> and <size> which represent the relative hotspot's 3D transformation with respect to the multi-marker's reference system. The

AR engine, according to the aforementioned descriptors, builds up the virtual scene by means of a *DOM* XML parser. To find the required data in the application database a XML-based *XPath* query language is used. Combining the information from tracker and from XML database, the AR engine is therefore able to locate in every moment the user in the real world and to extract all the required augmenting contents from the repository. This repository consists of either 2D or 3D objects (eventually animated) such as text labels, graphics and parametric "occluding objects" used to selectively hide features of the real component. An example of a typical tag structure for a hotspot is listed below:

```
<hotspot>
<id>OnOffButt</id>
<label>On/off button</label>
<shape>circle</shape>
<details>This is a on/off button</details>
<position x="0.0" y="0.0" z="0"></position>
<rotation x="0" y="0" z="0"></rotation>
<size x="0.01" y="0.01" z="0.01"></size>
</hotspot>
```

For any augmented hotspot the engine allows the user to edit his/her preferred attributes. When the user selects an interested hotspot he/her can associate to the selection the required visual aid. These attributes are stored in an external XML file associated to the AV component. The engine also provides an editor which allows the user to edit his/her manual information. The editor generates for each component an XML compliant file which can be used for augmentation. This feature can be handy if the user wants a custom documentation written according to his/her needs. Besides the augmentation, the engine performs another crucial task that is the support to complex procedure like, for instance, 7.1 surround installation and setup. Each procedure can be represented as a finite deterministic automaton (DFA). According to this approach, a particular state represents a procedure step and its links define the execution order. DFA result very suited to this context providing all the elements (states, links) required to

represent both simple and complex procedure in a easy, verifiable and legible way. It results convenient to convert the DFA representation of a particular procedure in a XML file, exploiting the above mentioned *DOM* parser and query language. The user can switch to the next or previous step using the trackpad/touchpad. At each step the system might suggests the user how to operate the device, explaining the purpose of the operation and the involved hotspot. An example of a simple step is described below.

```
<step>
<label rgbText="default"
rgbBackground="default">Turn off the device</label>
<hotspot id="OnOffButt"></hotspot>
</step>
```

The “subtractive” augmentation capability is able to hiding part of the AV component to let the user focus on the elements of the physical interface required to perform a particular operation, thus avoiding the confusion due to information overload. This represents a mediated or diminishing approach to augmentation, and rather than adding visual contents, all hotspots are hidden except those involved in the current operation. It is strictly based on the techniques exposed in section 2.2.1. and adapted to be performed on a less powerful mobile computing platform (a notebook or even a tablet). Advanced AV components may be equipped with a RS232 bidirectional serial communication interface. For those components, the system provides access to any functionality also by means of the netbook/tablet touchpad instead of operating directly on the device. To the aim of testing the proposed system in facilitating the usage of a commercial AV component, it in a real applicative scenario: the augmentation of a Denon AVR-3805 surround receiver during two not trivial procedures (level equalization and input-output wiring) which typically force the user to refer to a voluminous manual. Generally any receiver/amplifier is characterized on the front panel by a variable number of buttons, switches, knobs etc. controlling for example, source selection, sound effects and so on.

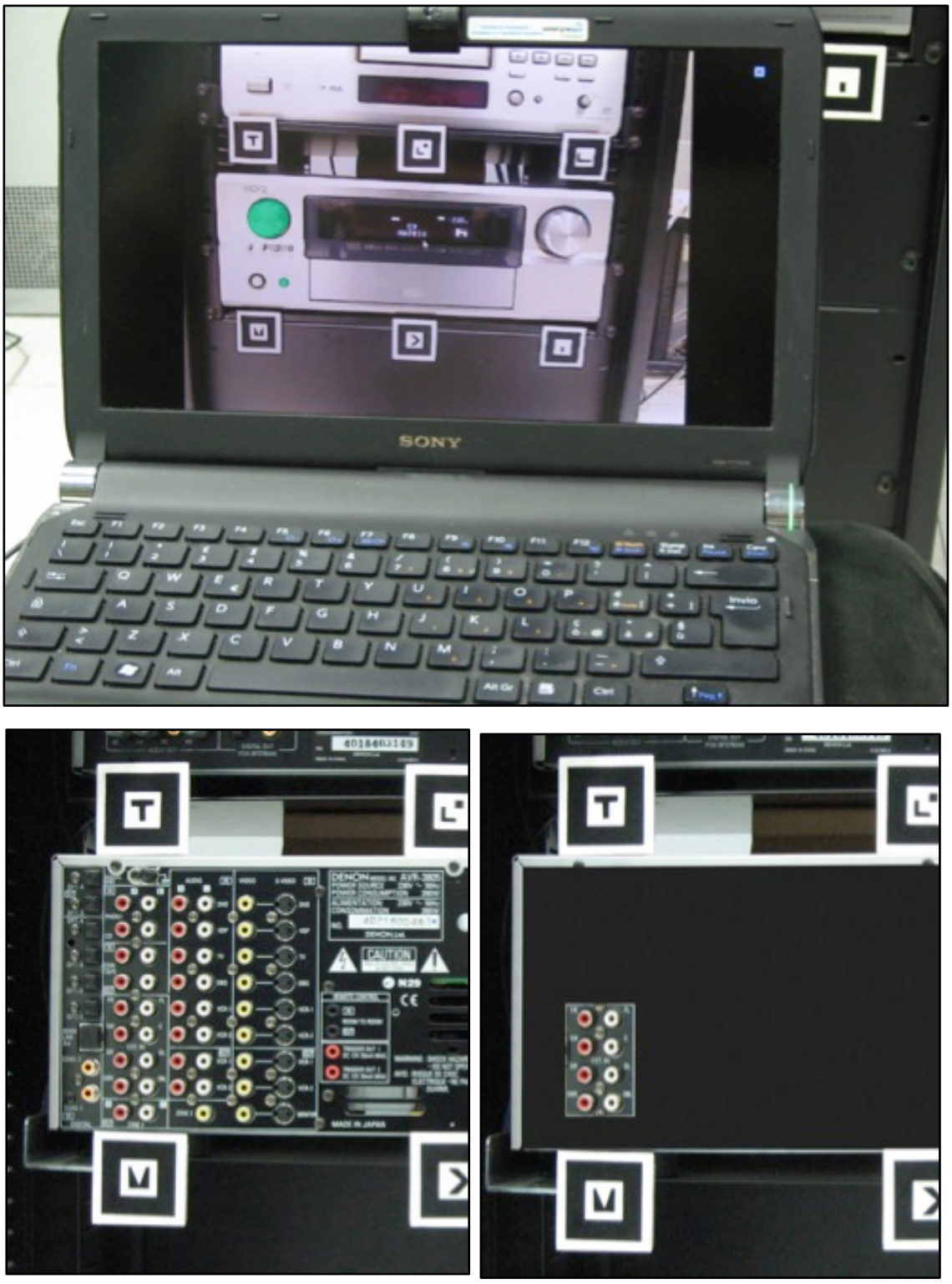

Figure 11 *Top: an additive augmentation of front panel highlighting a knob and a button. Bottom: subtractive visualization strategy showing the actual rear panel (left) compared to the mediated version (right) with only the connectors required.*

On the other side the rear panel is much less used but is crowded by a big number of I/O connectors whose correct usage depends on the comprehension of the particular wiring scheme chosen. The test-bed used for experiments is based on a Sony Vaio netbook featuring Intel dual core processor and Nvidia Quadro FX-500 series graphics board. The scene capturing is performed by a Logithech Camera C905 providing a video resolution of 800x600 pixel at 30 fps. This hardware setup has been capable to render an augmented scene featuring up to thousands of polygons providing a sustained frame rate always above 30 fps (limited to the camera's fps). During the AR assisted operations the user sees through the netbook's LCD display the virtual contents rendered over the captured scene. This design allows the user to focus on the relevant items to perform the required tasks. Clicking with mouse pointer on a specific device's hotspot, the system shows some short technical data.By double-clicking an additional window is shown, which contains the detailed information for the selected hotspot. Both additive and subtractive approaches to augmentation have been tested (see Figure 11) according to the considerations drawn above, so the additive approach has been tested on the frontal panel of the amplifier whereas the subtractive has been tested on the rear panel.

2.3. Handling Occlusions While Interacting In Mixed Reality Environments

As the number of augmented and mixed reality applications available on a variety of platforms increases, so does the level of interaction required, possibly leading to the emergence of challenging visualization issues. To this regard, it is worth to note that the illusion of the co-existence of virtual objects in the physical world (the essence of MR paradigm) is typically made possible by so called video-based[1] see-through approach in which the rendering of virtual contents is superimposed onto the surrounding environment captured in real time by means of a proper transformation. This trick works well until the order of the planes to be composited is coherent to their distance from the observer (see Fig. 12_Left). But, whenever an object of the real world is expected to occlude the virtual contents, the illusion vanishes since the order of rendered planes does not lead to a correct visualization (see Fig. 12_Right). As a result, what should be seen behind a real object could be visualized over it instead,

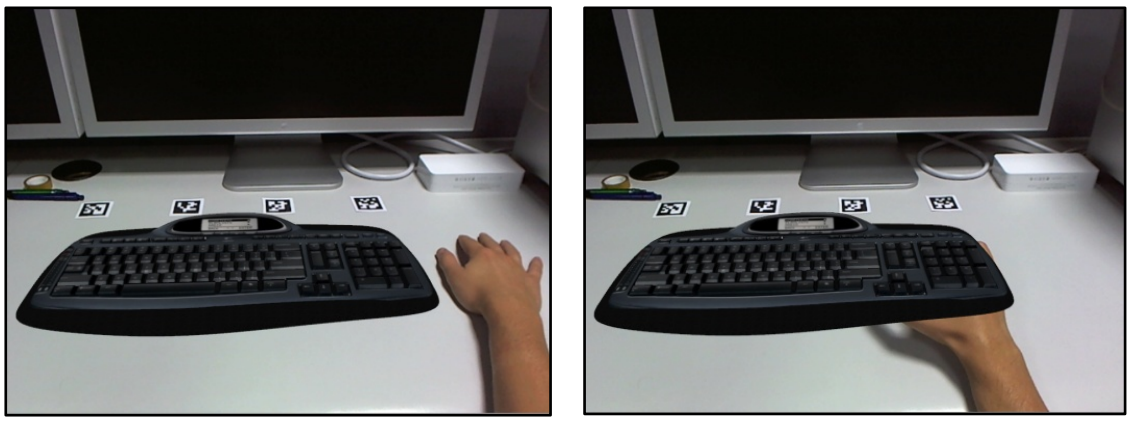

Figure 12. *Left: A virtual model of a keyboard rendered onto a captured frame of real environment to augment it. The hand positioned along the right side of the keyboard does not ruins the Mixed Reality illusion. Right: The same MR scene, but as the hand is positioned over the keyboard, it is occluded by the virtual content.*

generating a "cognitive dissonance" due to the loss of spatial coherence along the axis normal to camera plane that may compromise scene comprehension and, ultimately, the interaction capabilities during the MR experience. Hand occlusion in augmented reality is a challenging topic and scientific literature presents diverse approaches to it.

[1] Optical see-through is the other well known option for MR/AR, but besides being less diffused it is inherently less suited to support processing of environment visualization.

In particular, displaying occluded objects in a manner that a user intuitively understands is not always trivial. Furmanski et al. [27] in 2002 developed new concepts for developing effective visualizations of occluded information in MR/AR applications. They designed some practical approaches and guidelines aimed at evaluating user's perception and comprehension of the augmented scene and distances. Many researchers aimed at solving the incorrect occlusion problem by analyzing various tracking methods or by integrating vision-based methods with other sensors [28]. Lee and Park proposed to address this issue in AR environment introducing the usage of an Augmented Foam [29]. A blue foam mock-up is overlaid with a 3D virtual object, which is rendered with the same CAD model used for mock-up production. By hand occlusion correction, inferred by color-based detection of the foam, virtual products and user's hand are seamlessly synthesized. The advantage of the augmented foam is that it is cheap and easy to cut allowing to realize simple and complex shapes. On the other hand, it imposes that for all augmented objects has to be present in the scene the physical counterpart made of foam. A color-based similar approach is discussed by Walairacht et al [30]. They exploited the chroma-key technique to extract only the image of the hands from a blue-screen background merging the image of the real hands and the virtual objects with correct occlusion. Although chroma-key is particularly fast and efficient, it requires the use of a colored background that represents a not feasible solution in many environments. In addition, it does not provide any information about real objects in the scene and their spatial distances. Buchmann et al [31] also handled hand occlusions in augmented reality exploiting marker-based methods to determine the approximate position/orientation of user's hands and, indirectly, their contour to fix the visualization order. The disadvantages are the inconvenience to wear specific gloves featuring fiducials on each finger and the rough level of accuracy in the segmentation of the hand from the background. In the field of medicine, Fischer et al [32] exploited a Phantom tracker and anatomic volumetric models in order to support surgical interventions resolving occlusions of surgery tools. They presented a simple and fast preprocessing pipeline for medical volume datasets which extracts the visual hull volume. The resulting is used for real-time static occlusion handling in their specific AR system, which is based on off-the-shelf medical equipment. Depth/range cameras (e.g. the

Kinect by Microsoft) have also been proposed [33][34][35] to provide a real-time updated depth-image of the surrounding world that can be conveniently used to evaluate whether a pixel from the captured environment is closer to the observer than the corresponding rendered pixel of virtual content, or not.

This technique can lead to a more accurate result and also enables evaluating distances of real objects in the scene and their inter-occlusions with virtual objects. However, it requires additional hardware (usually an infrared pattern emitter and a dedicated infrared camera) and it should match the field-of-view of the see-through cameras, to works effectively. The generation of a disparity map by using stereo matching techniques [36][37] represent the most suited choice to correctly segment user's hands in AR environments. Results produced by this technique are comparable to the ones form depth cameras without requiring dedicated hardware, which is a central aspect of this study.

The following sections describe the proposed method to addressing effectively hand occlusion in many MR/AR interaction contexts without any additional hardware, apart from video see-through goggles enabling stereo-vision. In brief, the rendered virtual objects are composited onto the incoming video see-through streams according to a disparity map encoding real-to-virtual visualization order at a pixel level as a gray-scale image by means of stereo matching. The disparity map is generated by a belief propagation global algorithm [38] that exploits GPU's highly parallel architecture for speeding up required calculations and for enabling real-time applications. The performance of the algorithm is optimized by segmenting the input image between hand and not-hand regions via a skin-tone filtering in the HSV color space (less affected from lighting conditions than RGB space). The purpose of this segmentation is twofold. From the one hand it is possible to reduce the region of interest (that directly affects the computational cost of the disparity map) to a cropped region of the original frame, on the other hand the contour of the segmented hand region is used as a reference to improve the edge sharpness of the disparity map. Some ad-hoc improvements aimed at further reducing the computational cost of the original algorithm are discussed in the following section.

2.3.1. The approach at a glance

The proposed processing pipeline is shown in Fig. 13 The diagram highlights the main elements in the image-processing pipeline. The user wears a HMD with two embedded cameras enabling stereo vision. Two separated video streams, from left and right camera respectively, capture the real scene from a different perspective point. On each stream, a simple and fast skin detection technique detects the user's hands in the scene. The binary image is used to apply a vertical crop to the original frame that preserves the region, including the foreground and the background, where the hands appear. On that crop two disparity maps, the one for the real scene captured and the other for the rendered content, are generated by exploiting a stereo-matching with belief propagation technique. The disparity maps are used to estimate the position of the hands in the field of view with regards to the virtual scene.

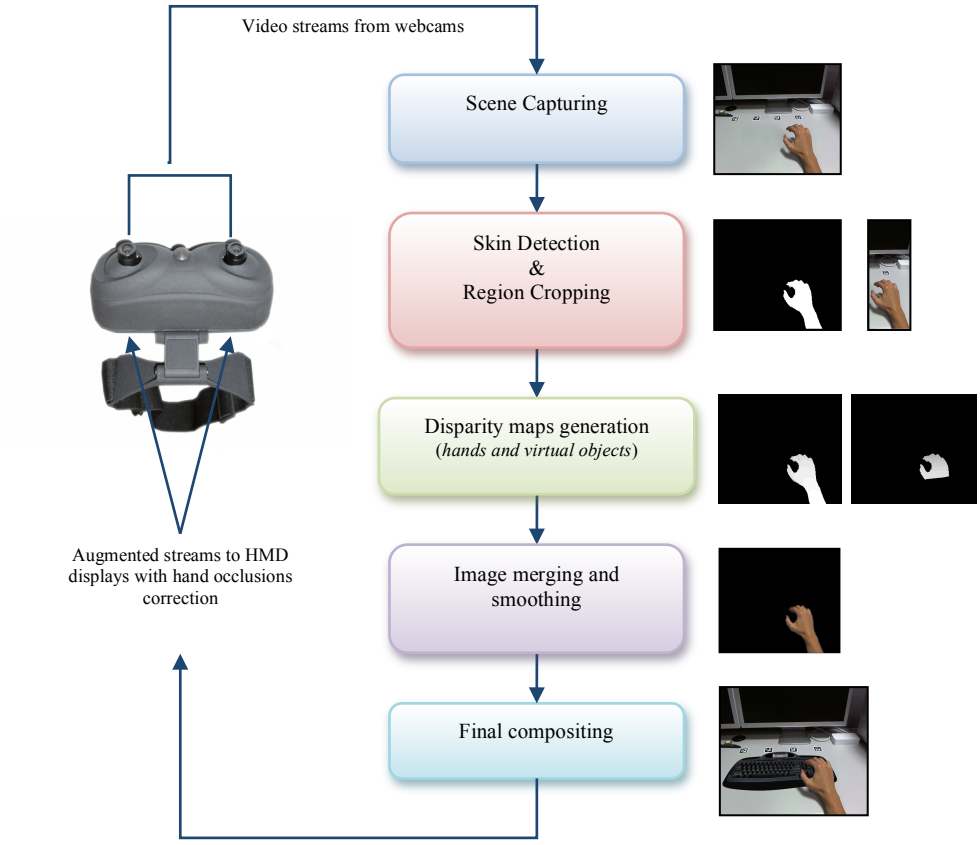

Fig. 13. *The overall architecture of the approach proposed.*

The occlusion correction is achieved by comparing them and combining the result with a skin-based segmentation of the hand. An edge blurring pass is applied to the segmentation in order to smooth the edges of the hand region. The combination of disparity map with blurred color-based segmentation of hands produces a cleaner approximation of the occlusions that can be applied as top-level layer of the augmented streams sent to the HMD displays.

2.3.2. Method description

The first step consists in capturing the scene observed by the user through the webcams mounted on the HMD. Since the HMD is intended for a stereoscopic vision, the streams from left and right camera capture the scene from a slight different point of view. Each of the two streams is therefore separately augmented by rendered virtual contents throughout the pipeline. Even though this implies a greater computational cost of the augmenting algorithm, it preserves the binocular vision of human eyes leading to a more reliable augmentation of the scene and the occlusion correction. Fig. shows one frame captured by one of the cameras mounted on the HMD while the user wears it.

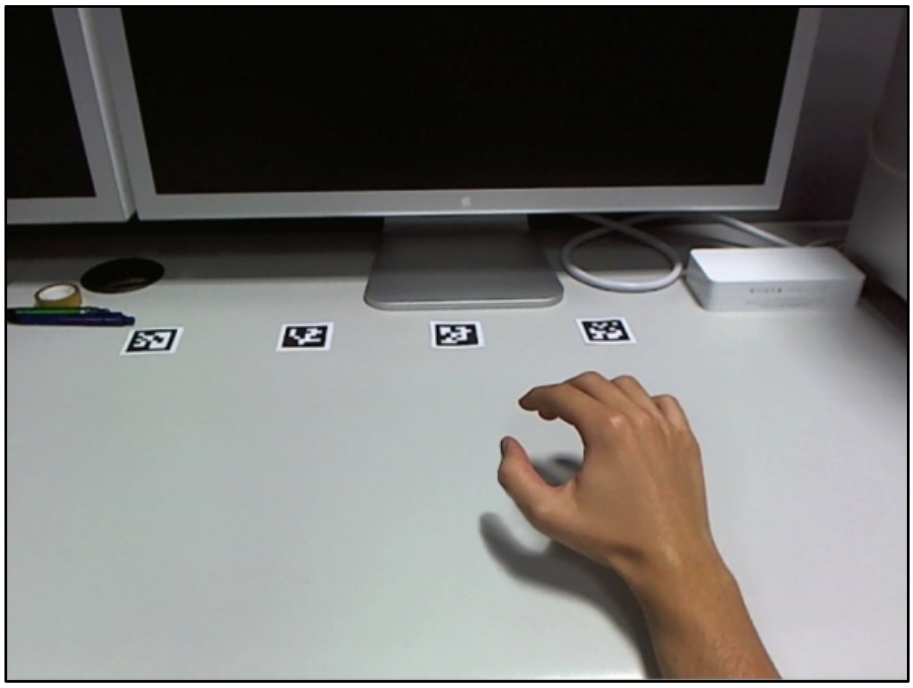

Fig. 14. *The scene captured by one of the cameras mounted on the HMD.*

To keep computational cost of the following steps slow, the frame is properly cropped so that the algorithm can focus the execution only on the relevant portion of the entire frame. The cropping is performed by a simple and fast skin color-based technique. It converts the video frame from the RGB to the HSV color space in order to have a simply way of filtering the color of naked hands by proper ranges of hue and saturation, thus leading to a gray-scale mask. Fast closure operators enable removing little irrelevant blobs in this mask and filling holes (if any) in main closed regions (the hands in our context). Every pixel inside the region boundaries (the hands' contour) is therefore set to full white (see Fig 15a). The intersection of this first mask with the rendered content's alpha channel (see Fig 15b) results in a new mask which limits the region on which the disparity maps of rendered content has to be computed (see Fig 15c). To this aim, stereo matching with belief propagation is therefore performed on these cropped regions. By processing only a limited region of the whole scene, it has been possible reducing the computational costs of this step, which is the most time consuming in the processing pipeline. Firstly the matching costs for each pixel at each disparity level in a certain range (disparity range) are calculated. The matching costs determine the probability of a correct match. Afterwards, the matching costs for all disparity levels can be aggregated within a cross-shape neighborhood window. Basically the loopy belief propagation algorithm first gathers information from a pixel's neighbors and incorporate the information to update the smoothness term between the current pixel and its neighboring pixels, and to iteratively optimize the smoothness term thus resulting in global energy minimization.

Each node is assigned to a disparity level and holds its matching costs. The belief (probability) that this disparity is the optimum arises from the matching costs and the belief values from the neighboring pixels. For each iteration, each node sends its belief value to all four connected nodes. The belief value is the sum of the matching costs and the received belief values. The new belief value is the sum of the actual and the received value and is saved for each direction separately. This is done for each disparity level. Finally, the best match is the one with the lowest belief values defined by a sum over all four directions resulting in the final hand(s) disparity map.

The main factor that affects every stereo-matching technique is the number of disparity ranges considered during the matching cost function. The more values are considered the more the disparity map is reliable but, the more the cost increases. Considering the main goal of performing a fast hands occlusion correction, the number of disparity ranges has been reduced. The rough disparity maps (obtained by composing it with the corresponding crop of the binary image from skin detection acting as alpha layer) has been refined by means of one pass of edge blur allows to smooth the edges of the color-based segmentation. The result is a smoother segmentation of user's hands (see Fig. 5d) that can be used for final compositing (Fig. 15e). For what concerns the rendered content, it would be simpler and faster to exploit the accurate depth info contained in the Z-buffer, but matching it coherently to the depth levels encoded in the hand(s) disparity map would be a not trivial task.

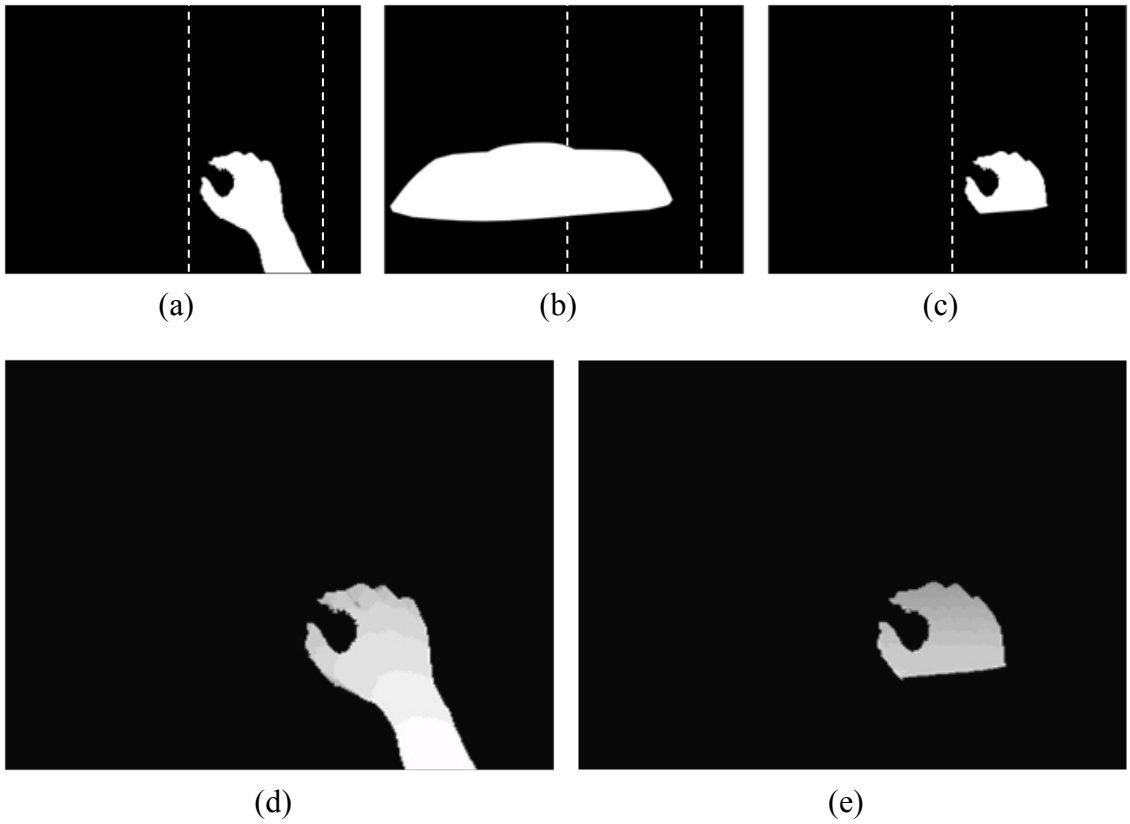

Fig. 15. *Skin detection with closure functions refinement (a). Alpha channel of augmented objects rendered onto the video stream (b). Disparity map of user's hand segmented from the scene by the skin color-based detection (d). Disparity map of the crop of the region where augmented virtual contents overlap the hand (e).*

Fig. 16. *Compositing and final result. The original real background shown in Fig. is composited according to the disparity map of the scene enabling a correct visualization and a meaningful interaction (note that hand's casted shadow is not currently handled by the proposed method).*

The final composited frame is obtained by comparing pixel-wise the gray level of the two disparity maps. The pixel whose gray level is lower than its homologous is considered not-visible from the observer's point of view and is discarded. Fig shows an example of the final result in which the hand of a user interacting in a MR environment is properly composited onto the augmented content.

2.3.3. Experimental results

Preliminary experimental trial of the proposed technique in a MR environment has been performed on a test-bed featuring an i7 Intel quad_core processor and an Nvidia GTX760 graphic board equipped with 1152 cores and 2 GB of VRAM. The user worn a Trivisio HMD that features stereo capturing by two embedded webcams (752x480 resolution, 6oFPS) and stereo vision by two 800x600 LCD displays (see Fig. 17).

Even though the method described in this thesis3.5 exploits time consuming algorithms, it meets the requirements of real-time application because it works only on a fraction of the whole captured scene. In addition, the improvement provided by utilizing graphics hardware acceleration makes possible to combine the time demands of stereo matching with typical marker-based tracking of the user on a stereo video stream. Table 2. summarizes the performance measured during the experimental session. In particular, the table shows the frame per second achieved by the proposed solution when the disparity maps are generated for 16 and 32 ranges of disparity values. During the experimental trial the user is free to move his/her hands thus implying a size of the crop of the scene that varies over time. During normal condition of interaction, the number of pixels of user's hand covers about 1/8 to 1/6 of the whole scene for over 60% of the experimental session. When the distance between user's hand and the point of view results shorter, e.g., the user brings his/her hands closer to the cameras, the stereo matching works on a wide crop of the scene leading to a drop in performances to the limit of a smooth real-time rendering. Future improvement of this method could take into account such issue providing an adaptive amount of disparity levels to consider during the matching cost function.

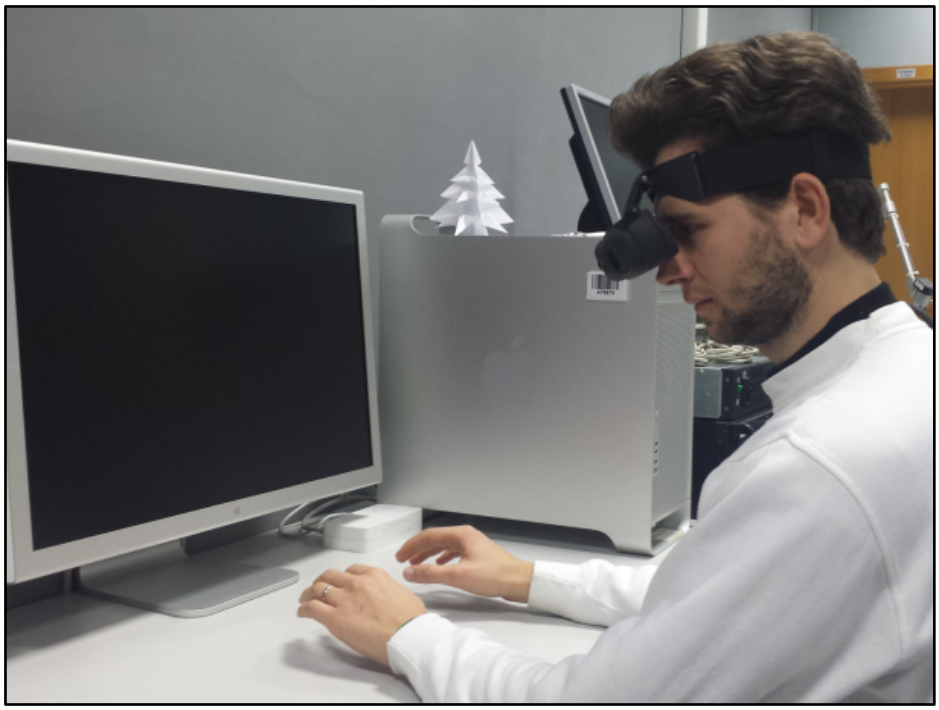

Fig. 17. *A user wearing the Trivisio HMD during the experimental trial.*

Even though these enhancements are inherently effective only on naked hand region, even not-naked arms can be reasonably handled by the disparity info alone. According to first users evaluations, the combination of augmentation and the detection of occlusions worked well, providing an intuitive interaction paradigm suited to a wide range of application contexts. Binocular scene capture and stereo rendering of virtual contents improve depth perception of real environment while stereo matching allows to estimate the distance from the observer and real/virtual objects in the scene.

Table 2**.** *Frame per second recorded during the experimental trial at different size of the cropping region of the scene.*

Crop size	**# disparity levels**	
(fraction of the whole scene, which	**16**	**32**
consists of 360960 pixels (752x480))	**FPS**	**FPS**
< 1/8 (~ 45120 pixels)	56	48
< 1/6 (~ 60160 pixels)	42	33
< 1/4 (~ 90240 pixels)	31	22
< 1/2 (~ 180480pixels)	25	12

Issues related to the hardware used (the reduced HMD's resolution/field-of-view, rough hands segmentation under rapid user's movements) have to be more carefully addressed to achieve a robust system behavior. In particular, the generation of the disparity maps for the hands when they occupy the most of the framed scene. As a further development of this technique, besides improving the quality of the disparity map, it would be interesting trying to address the incorrect visualization of the shadows casted by the hands when they should be projected onto a virtual object.

2.4. Visual Interaction in Mixed Reality by Means of Gestures

The typical AR paradigm implies that the user is able to perceptually merge the real environment around him/her with 2D/3D, static/ animated virtual contents he/she can interact with [39]. Such interaction can be hardly achieved through conventional mouse or keyboard-based devices even because, usually, AR is not experienced while seated at the desktop. To this regard, gesture based interfaces may represent a more suited and natural approach to interaction within mixed reality contexts. Though the underlying software and hardware technologies have been around since the early '90s [40], they were initially too expensive and often characterized by sub-optimal performance. More recently, most of the past issues has been addressed, making these devices much more affordable and reliable. The first uses of virtual/augmented reality in medicine, eventually empowered by gestural interfaces, go back to the mid-late 90s. Among the most significant examples, virtual surgical simulators, remote/tele surgery or even multimodal diagnostic imaging have to be cited. However, despite its potential advantages, the use of such kind of interfaces has been rather limited due to practical issues [41]. At present, one of the main research topics of gesture based interaction in medical imaging is represented by sterile contact-less approaches to operating-room practice. Conventional "material" interfaces like mouse and keyboards are difficult to sterilize. Possible workarounds are the presence of an assistant to operate the computer, or the use of a voice-recognition based control system. Both such possibilities might not be practical for the task at hand, therefore gestural interfaces can represent an appealing alternative and a feasible choice [42] for such kind of applications.

The approaches to gesture recognition can be roughly classified into three main groups. Some systems exploit computer vision procedures, which typically require an inexpensive hardware (one or more cameras) but that have to deal with possible occlusions problems. It is worth noticing that such problems are among the "hard" ones addressed by pattern recognition and computer vision research. A second group of systems aims at solving the problem at sensor level, by means of instrumented gloves and non-image based tracking systems. Of course, this second group is suitable only when sterile operation is not a constraint. A last group includes systems where an

ambient equipped with sensors “tracks” the users. Most applications included in this group only require a gross identification of user’s position and gestures. For a survey of technologies and approaches, see [43]. As a consequence of the ever increasing diffusion of virtual entertainment applications, the hardware dedicated to visual interaction (graphics boards, force feedback interfaces, instrumented gloves, gyroscopic sensors, accelerometers, tracking systems, head mounted displays, and so on) has become much more affordable than in the past. This fosters the research on multimodal interfaces [44] as the obvious candidate to enhance and integrate the functionalities and the interaction paradigm provided by the traditional WIMP (Windows, Icons, Menus, Pointers) approach [45]. Multimodal interfaces feature multiple sensors and i/o channels which can be combined in different schemes, whenever the WIMP metaphor is poorly applicable or can be enhanced. This condition is likely to happen when the (virtual) objects and the tasks to be performed on them are inherently 3D, thus requiring a (not always intuitive) combination of 2D actions as a workaround. In this case a gestural interaction within a more realistic 3D setting, performed by means of hands/fingers tracking, may represent a much more natural and effective approach to the task [46]. In the last years, multimodal interfaces have been proposed in many application domains [47]. Among them, it is worth to mention advanced VR based simulators, remotely controlled systems and virtual training environments, often combining different channels such as voice and gestures [48]. It is worth underlining that gesture based interfaces are well known in literature, since their usage has been often proposed even before the success of WIMP (see for example the gestural part of the pioneering system “Put That There” [49]). They have been especially investigated for settings when a "natural" interaction paradigm is an important requirement or in case conventional input devices are not a feasible choice. Medical simulation is one of the fields in which gestural interaction paradigm is best exploited. A visual approach to gesture recognition is generally preferred for this kind of applications because it does not require the user to interact with, or wear, any specific device, as the recognition is generally based on video acquisition and processing of gestures. In the medical field this has the additional advantage of avoiding introducing further objects into the medical environment, when this may cause troubles (e.g. for maintaining a sterile setting). Graetzel et al. [50] follow

this line of research, capturing color and depth information by means of a stereoscopic camera to the aim of robust, high-speed hand detection and tracking within a limited workspace in which hand gestures are interpreted as mouse commands (pointer movements and button presses). Another example of sterile operating room gesture based interface is represented by the Gestix system by Wachs et al. [51] which exploits simple user's gestures to perform image navigation and manipulation by associating them to commands based on their relative positions on the screen. Color-motion cues are exploited to track the user's hand while a finite state machine enables switching among available functions. A different approach on the same topics is proposed in [52] by means of instrumented gloves and magnetic motion tracking technologies. Tani et al. [53] discuss the use of a glove-driven interface in radiological workstations, and present a prototype that aims at integrating common functions, such as virtual manipulation and navigation control, with a basic gesture interface. The user can control the mouse by simply pointing to the screen and moving the hand, or can perform gestures which are conventionally associated with specific commands in the given context. The interaction with 3D medical data (synthetic polygonal models of human organs, or of anatomical districts, generated via different techniques, as well as voxel based representations of real diagnostic imaging produced by radiological workstations) represents one of the more challenging application of gestural interfaces, and understanding the human factors influencing such kind of interaction remains one of the challenging problems in computer graphics. In fact, most computer graphics application are designed to operate via the usual "point and click" paradigm, since it is claimed to be intuitive [54]. However, when complex 3D data manipulation is required, this is not necessarily true. Common functions like image rotation with respect to a given point, which are performed intuitively in a bidimensional space, become more complex in three dimensions, requiring a combination of multiple 2D transforms or a more powerful interaction paradigm. As a result, approaches focusing on recognition accuracy, which try to map mouse and keyboard operation onto gesture patterns in a 3D space, might not exploit the full potential of this interaction technique. In this section, a framework based on a floating interface for gesture-based Interaction is presented. It puts together a context adaptive head-up interface, which is projected in the central region of the user's

field of view, and gesture based interaction, to enable easy, robust yet powerful manipulation of virtual contents visualized onto the real environment surrounding the user. The interaction paradigm combines one-hand, two-hands and time-based gestures to select tools/functions among those available as well as to operate them. Conventional keyboard-based functions like typing are available too, and can be performed without a physical interface by means of a floating (virtual) keyboard layout. The aim is to set-up a sort of mixed interaction paradigm by which the user can switch from direct (virtual) manipulation operations to more conventional system interaction operations, without changing his/her gesturing space. Though the proposed implementation of the approach to gesture-based interaction is tightly related to a particular choice of tracking/gesture recognition technology, it is worth remarking that the focus of this research is on the interaction paradigm rather than on the specific gear adopted. A prototype application addressing the practice and the training to medical imaging is presented, including a report on usability evaluation.

2.4.1. Gesture recognition by means of multiple sensors

The proposed framework exploits gesture recognition and tracking within a mixed reality environment plus, in order to provide advanced patterns for human-computer interaction. A more natural and familiar way of managing objects and situations, e.g. through (virtual) direct manipulation, can improve global user performance and increase applications effectiveness. Of course, a careful design is required to obtain this goal. Despite the appropriateness of this kind of interaction, in some situations the need arises to also perform classical keyboard-and-mouse supported operations, such as selecting an object from a menu of models or entering parameters for a complex operation. On one hand, switching to a “real” device would break off the interaction flow in a disturbing way. In other words, the user might need to move or change position to reach a different place or equipment, so interrupting the task operation flow. On the other hand, the way the user is accustomed to perform such operations has to be considered. Consistency with familiar interaction patterns is a very important guideline, aiming at reducing the time and mental efforts needed to learn a new application [55]. For this reason, a virtual keyboard is projected onto the actual environment, so that the user can

comfortably switch from manipulation operations to system interaction operations without physically shifting his/her locus of attention/operation. The familiar point-and-click pattern is also re-proposed through fingers movements, to extend the range of available functions while keeping the number of basic gestures to a minimum. A schematic view of the whole architecture is shown in Fig. 18. Briefly, the three main system's components are responsible for Gesture Recognition, Interaction Control and Mixed Reality Environment respectively. User's hands capturing and tracking, as well as head tracking, represent the input channels while a stereoscopic HMD (Head Mounted Display) is the main output device. Each one of these components is detailed in the following subsections. Three main input channels are processed and synchronized by the proposed system: right hand, left hand and head.

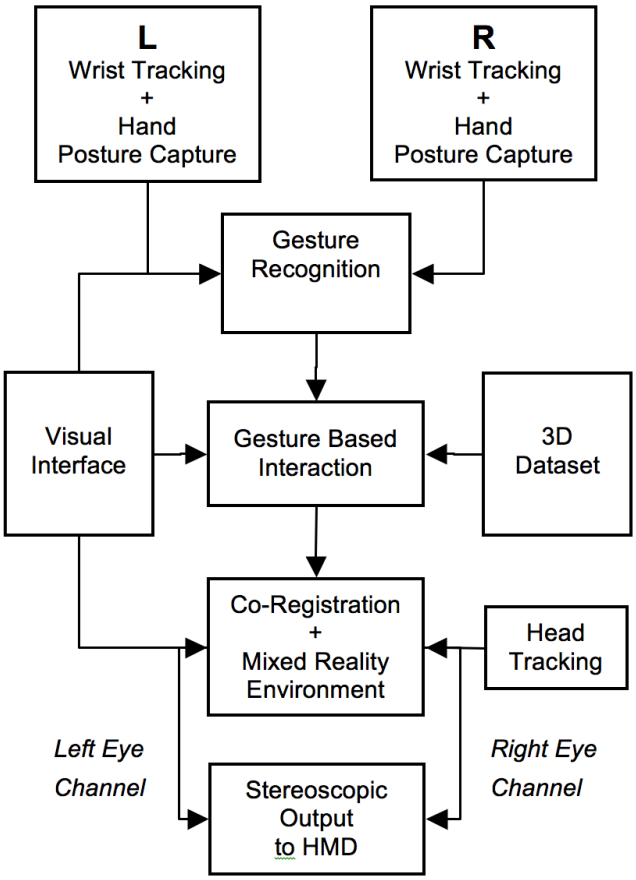

Figure 18. *A schematic view of the whole architecture.*

The first two channels are exploited to capture the position and orientation in 3D space of each finger for both hands, while the third channel is crucial to the purpose of enabling a mixed reality environment through virtual-to-real co-registration. In practice the system must adapt the position of virtual objects to the user's point of view, which can be inferred from the position of the head. In training systems, which are the main target of this proposal, sterile operation is not a constraint (i.e. it does not address a real operating room scenario). The main focus, indeed, is on the pattern of interaction, regardless of the specific involved technology. Therefore, it is reasonable to exploit wireless instrumented gloves and ultrasonic tracking devices to capture user data. Image based techniques are inherently contact-less, but potentially prone to inter-hands and inter-fingers occlusions, and solving the related problems was out of the scope of this work. On the contrary, instrumented gloves are technically more complex and expensive (hardware wise) but each single finger can be equipped with individual sensors to measure flexion and abduction, which are unaffected by those of any other finger. In this case a couple of wireless 5DT Data-glove 14 ultra have been used (Fig. 19), featuring fourteen 12 bit piezo-resistive sensors for the measurement of finger flexion and abduction. Besides outputting fourteen raw values, the gloves provide a binary (open/closed finger) value, resulting from the comparison of normalized joint flexion values to a threshold. This leads to 2^4 different combinations or gestures of the four tracked fingers (thumb is not considered). While this simplified data lacks the full precision sensors are capable of, it turns out handy as partially flexed fingers do not compromise gesture recognition. The four hand postures used in this study are shown in Fig. 20a: fist, pointing, index finger, bended index finger and flat hand. They have been chosen as they are among the simplest to perform for most users, and among the most used in natural interaction [56]. Motion tracking hardware IS-900/VET from Intersense co. is used to capture left/right wrists and head position and orientation, for a total of six DOFs (Degrees Of Freedom) for each channel. Since instrumented gloves do not provide any spatial information, the system calculates the exact position in 3D space of a particular fingertip by applying forward kinematics to hand-back position/rotation and finger flexion/abduction.

Figure 19. *Wireless instrumented gloves (Data Glove 14 ultra from 5DT) coupled with wrists/head wireless motion tracking (IS-900/VET Inertial Tracking from Intersense) worn during testing.*

The use of such data is presented in the following section, when interaction techniques are discussed. The IS-900/VET is based on ultrasonic tracking to sample motion data. A clear advantage of this technology over video-based solutions is represented by the wide and scalable capture volume which frees the user from the need to be positioned in a precise place within the camera field of view. Magnetic based tracking systems are also scalable, but they may be affected by electrical and magnetic fields, eventually present in a radiology facility. The precision of the measurement is in the range of a few millimeters for spatial position and within 0,5 degrees for angles, while sampling rate features up to 180 measures per second, a value which is more than adequate for gesture based applications. Raw data are preprocessed to filter capture noise by means of a high frequency cut and temporal averaging. Three data streams (left and right hand/fingers plus head) result from this process; the first two ones are exploited for gesture based interaction, while the third one is required for real-to-virtual co-registration as detailed in section 3.3. The main purpose of the gesture recognition module is to detect specific gestures by means of corresponding flexion-abduction patterns and therefore to trigger

the associated interaction activities. Gesture detection exploits timed automata [57], to recognize not only one-hand and two-hands gestures but also timed patterns, thus allowing the user to access a wider range of functions with a small set of easy-to-perform gestures. Moreover, taking time into account enhances the quality of user-system interaction when a feedback is required in a reasonable time, or when the time elapsed between elementary actions can influence the interpretation of their composition. The aim here is to augment the basic one-hand gestures through timed patterns, or via a combination of left and right hands for a simple yet more powerful interaction. The use of timed automata offers a further key benefit as it enables the system designer to formally verify the interaction model by means of well-established model checking procedures [58]. In the proposed architecture, just eight gestures are defined and recognized, as shown in Fig. 3. Four of them are basic (one-hand) gestures (Fig. 20a), two are defined through a two hand combination of the aforementioned basic gestures (Fig. 20b), while the last two (Fig. 20c) are obtained by a timed sequence of basic gestures (for instance, fist/flat-hand/fist, or double pointing). In any of the cases considered, recognized gestures are represented by a vector including gesture index, first hand x-y-z spatial coordinates, first hand yaw-pitch-roll angles, second hand x-y-z spatial coordinates, second hand yaw-pitch-roll angles.

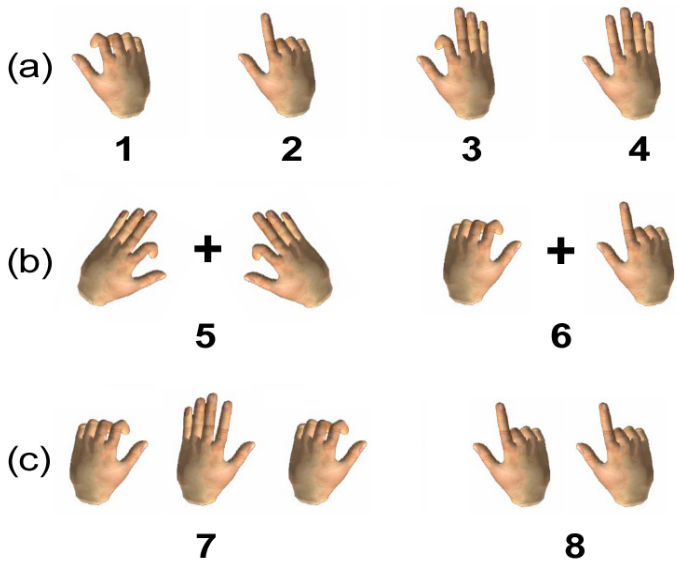

Figure 20. The eight gestures required to operate the system. They include one-hand (a), two-hands (b) and (c) time-based gestures.

2.4.2. Context-adaptive interaction approach

The Interaction Control module is fed by the output (i.e. the recognized gesture vector) of the previous gesture recognition stage, and enables each available functionality by translating gestures into actions through a timed automata architecture. The approach used is context-based, therefore gestures are evaluated depending on the current interaction status, thus allowing the same gesture to control different functions in different operational contexts (rotation, measurements, landmark assignment, etc.). Selection of operational modes and of related functions is accomplished by means of a virtual interface, This interface is displayed as a frame surrounding the application-dependent 3D content (see Fig. 6), eventually including text information concerning the ongoing operations (e.g.: numerical values for coordinates and angles, distances, etc.). From the user's point of view, the interface layout is perceived as it was floating in front of him/her in a close-by position, due to the stereoscopic rendering. Interface positioning along the depth of the visual field, is adjustable by means of a calibration procedure. According to this procedure, the user is requested to touch a sequence of small targets, which are positioned at various depths with the index fingertip, thus allowing an adaptation of the parameters that regulate the stereo effect. Interface design is mainly aimed at reducing the number of gestures required to operate it, therefore point-and-click interaction paradigm is replaced by its gestural adaptation. According to this philosophy, selection is triggered hitting an active area by index fingertip (see Fig. 3, gesture #2), an action or a confirmation is triggered by double hitting (see Fig. 3, gesture #8), a cancel/escape command is triggered by a fist/flat-hand/fist sequence (see Fig. 3, gesture #7). Whatever the gesture recognized, visual and acoustic feedbacks are provided to confirm the "pressure" of a key or to acknowledge a particular command, thus reducing wrong operations. If required, interface layout can be hidden at any time via a gesture toggle. The interaction design only requires one-hand gestures to operate the interface, but it provides support for two hands to achieve faster and more effective operations for more experienced users. For instance, one such user might type characters in a text field by both hands resulting much faster than a mouse-like character-by-character selection, and this would not require a physical keyboard. 3D

actions implemented so far, allow the user to rotate/move a virtual object, to place landmarks over its surface and to take distance measurements between landmarks. Object pan can be operated with any of the two hands in a straightforward fashion, while the user fully exploits and appreciates the advantage of two-hand gestures when rotating objects in 3D space . These operations are addressed by associating object rotation over three axes (though a user may lock one or two of them) with the rotation of a vector connecting the two points of contact between index finger and thumb on each hand (see gesture #5 in Fig. 3, and Fig.7). Object rotations performed according to this approach are much more accurate than the typical one-hand based ones, which are often implemented, for instance, in VR applications. Advantages include a greater control of rotation (two hands visualize more effectively the overall rotation), a more comfortable operation (it does not matter how each wrist rotates during interaction, as only the vector connecting the two hands is relevant), and a less jerky interaction, yet without losing responsiveness (a weighted average of both hands spatial information improves tracking). After rotation has been selected, the user can set the rotation handle for each hand (the anchor points used to interact with the model) by gesture #3 in Fig. 3 by simply moving the fingers along the object's surface. A valid handle location (i.e. one that lies within a valid region) is highlighted by a colored spot. If the gesture #5 (Fig. 3) is recognized, then the vector connecting the two handles is visualized and the interactive rotation is performed until this condition is true. Landmark positioning over an object surface may be accomplished according to two different operation patterns. In the first case, a rotation of the object is performed as explained before, to expose the location of interest, and then a landmark is placed by double hitting (gesture #8 in Fig. 3). A more intuitive, though less precise, operation pattern requires to perform the rotation of the object by a single hand, grasping the object (see gesture #1 in Fig. 3) in a position which acts as the pivot point, and double pointing the location on which the landmark has to be placed. Whenever a task involves positioning in 3D space, a precise calculation of the actual fingertip positions is performed through a combination of forward-kinematics applied to a 3D parametric hand model, which is adapted to the real user's hand measures during a calibration session. In this case the raw flexion values are

exploited for each finger. This setup procedure is performed only once, and may be saved and retrieved during system start up.

2.4.3. Implementing the MR operating environment

According to the interaction paradigm previously defined, the above mentioned floating interface, including the keyboard, as well as the tridimensional contents have to co-exist in the visualization projected in the real environment, in the space surrounding the user. This task is accomplished by the Mixed Reality Environment module. In the present implementation, this module is based on Quest3D real time engine (refer to Fig. 21 for samples of the visual programming environment), while dynamic simulation is enabled by the open-source library Open Dynamics Engine (OpenDE, a.k.a. ODE) [59].

As mentioned previously, any mixed/augmented reality environment requires a precise co-registration of real and virtual objects. In other words, the objects in the real and virtual world must be properly aligned with respect to each other, or the illusion that the two worlds coexist will fail. To this aim, the user's head position and orientation (x, y, z, yaw, pitch, roll) are captured by the previously described motion tracking system and are exploited to obtain the desired effect. In fact, these data are used to transform the virtual content as seen from the user's point of view, and coherently to the reference system of the surrounding environment. As this system is designed to provide stereopsis (i.e. the impression of depth perceived when a scene is viewed through binocular vision) two rendering cameras (one for each eye) which match the exact position/orientation of user's eyes are feed at runtime, and each vertex of each virtual object to be displayed onto the real scene is transformed accordingly. The two resulting renderings (left and right) are therefore displayed through an optical see-through HMD helmet (a Cybermind Visette SXGA, see Fig. 22). With optical see-through HMDs, the real world is seen through half-transparent mirrors placed in front of the user' s eyes.

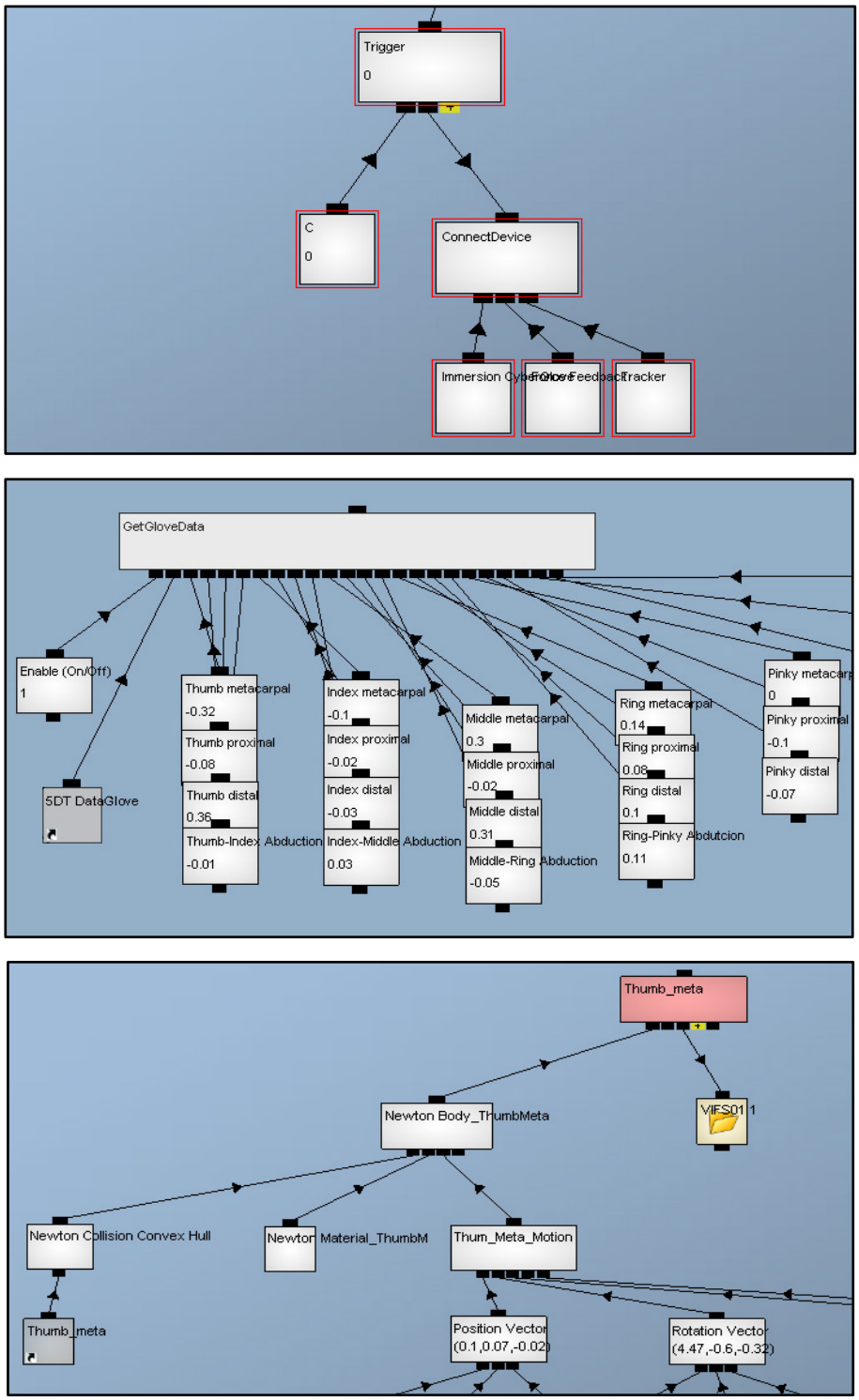

Figure 21. *A few examples of graph-based programming related to the gesture recognition component. (Top) Overall hand control. (Center) Data-glove and wrist tracker handling. (Bottom) Metacarpal thumb control.*

These mirrors also reflect the computer-generated images into the user's eyes, thereby optically combining the real and virtual world views. It is worth underlining that, due to the aforementioned nature of the viewing device, the overall (virtual + real) images as seen by the viewer during the experiments cannot be recorded. For this reason, Figures 23 and 24 hereafter have been simulated by overlaying the virtual content outputted by one channel of the visualization engine onto images of the working environment. The rendering engine has been adapted for optical see-through vision (for instance keeping into account the optical parameters of the HMD selected), but it could be adapted to video see-through as well. However, according to our experience, an optical see-through HMD is preferable over a video see-through solution (featuring comparable visual resolution) only if the quality of its optical combiners is high enough to provide a reasonably wide field of view. Unfortunately, this feature usually makes this equipment very expensive. As the overhead involved by this step is relevant, the suitability of such process depends on the real requirements of the application at hand. As a matter of fact, in most cases it could not be worth the effort.

Figure 5. *A close up view of the Cybermind optical see-through HMD worn to visualize the virtual contents onto the real environment.*

2.4.4. An application to medical imaging

This section presents a novel application of the framework described above to support interactive exploration of 3D medical datasets. With regard to the field of diagnostic imaging, it is very common to observe experts preferring to work on 2D sections of CT or NMR (for instance to place a landmark or to delimitate a region) rather than operating directly on a 3D view. The given explanation is that they do not feel confident about 3D environment and tools which are often available in commercial diagnostic systems, since these are usually considered appealing but not sufficiently reliable. The reason for this common belief possibly resides in the way these data are made accessible and on how complex is to interact with them through a bi-dimensional display and interface. Indeed, 3D visualization of diagnostic images is not inherently less accurate, as it is rendered on the basis of the (reliable) 2D sections, therefore the problem might rather be on the interface side. Operating on a 3D surface or volume requires a more powerful way to specify actions or locations in 3D space. This need is not well addressed by a usual 2D interaction paradigm, involving mouse or trackball. The system described in the previous section seems suited to address this problem and it is part of a wider project which aims at improving the usage of three dimensional data in medical imaging practice. Since from an early stage of this study, the main concern has been to assess if the proposed approach to 3D manipulation could match the requirements for a more accurate (therefore more useful) and intuitive way to deal with complex data. At the same time many conventional functions have to be accessible and easy to use, preferably in a way similar to the one experienced with common devices (mouse and keyboard). This compatibility is achieved by means of the floating interface and virtual keyboard which are visualized according to the aforementioned AR setting (see Fig. 23). Quoting Kölsch and Turk [60], “a virtual keyboard is a touch-typing device that does not have a physical manifestation of the sensing areas. That is, the sensing area which acts as a button is not a button per se but instead is programmed to act as one”. Moreover, as the user has not to move from his place or change his position to type commands to the computer, distractions and breakdowns in the gesture patterns are reduced to a minimum.

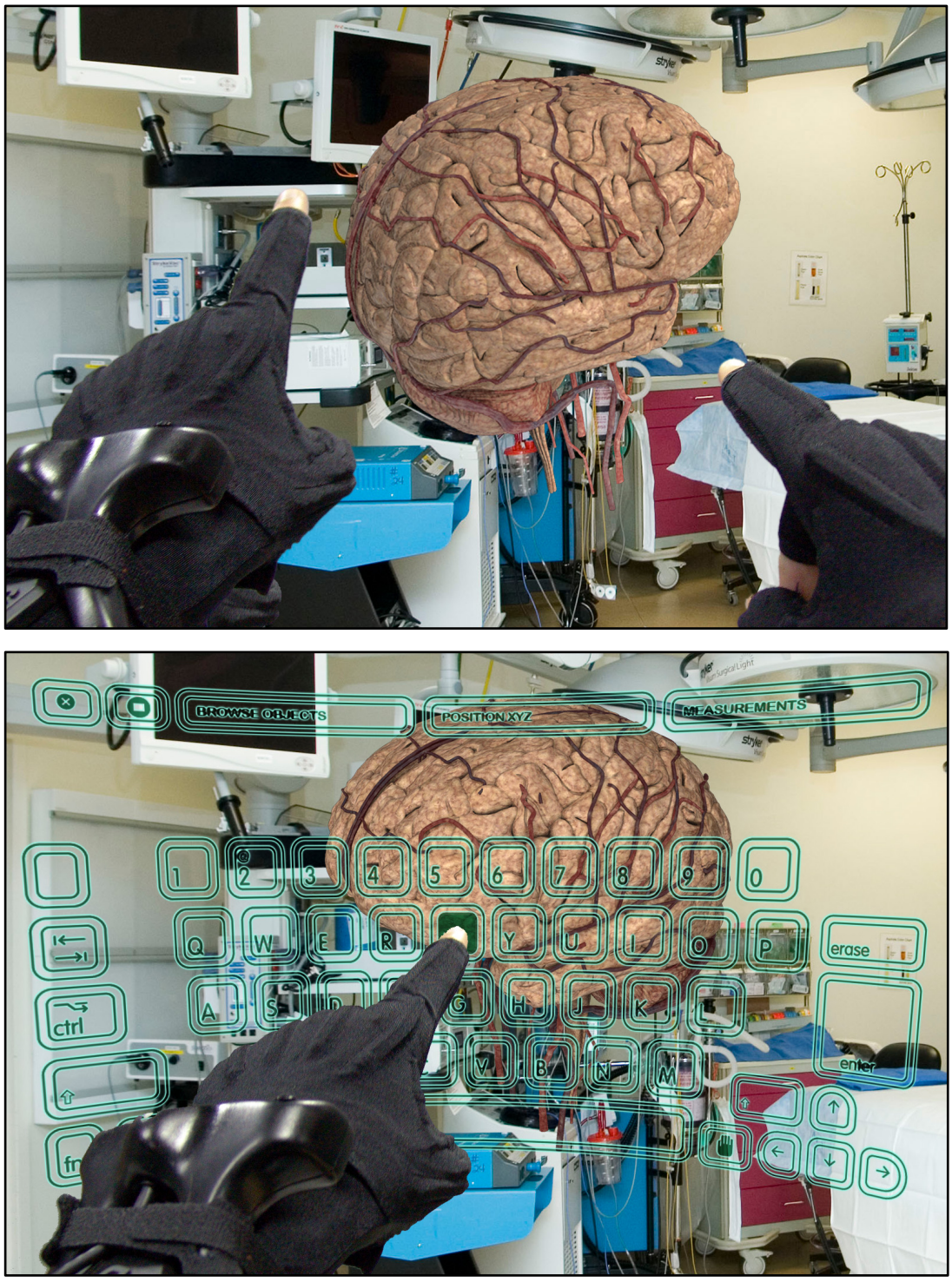

Figure 23. Sample images of the user's field of view (simulated) during gesture interaction within the mixed reality environment. The floating interface layout (which can be hidden (top) or shown (bottom) via a gesture toggle) is projected onto the central region of the field of view to enable the selection of the required functionalities as well as the interaction with virtual objects.

Due to the way the 3D models are computed (from CT or NMR), the exploration provided by the implemented system can also proceed by layers, allowing to explore the desired anatomical district at different depths.Such setting is intended to mainly address training or learning activities, but, with appropriate adaptations, it can also be exploited in operative contexts. The functionalities already implemented include: object browsing, two-hand operated object rotation/translation with respect to any axis, object transparency setting, landmark positioning and landmark-based measurements. For testing purposes, a library of anatomical 3D models has been used, since the visualization engine is currently suited to operate on polygonal-based objects rather than on voxel-based datasets, usually resulting from the processing of diagnostic images (see Fig. 7).

Figure 24. *User performing object rotation by means of two virtual handles, corresponding to the extremities of the green vector. This image shows the scene from the point of view of a third person, and is obviously simulated, due to the optical see-through design of the HMD. The brain model is the result of a true medical imaging processed to generate a 3D geometry, and further optimized (about 625,000 triangles). The shown color texture is fictional.*

As an alternative, it could be possible obtaining a polygonal mesh from the iso-surface resulting after a segmentation process applied to a voxel-based model. Regardless of the technique adopted, the inherent geometrical complexity of human organs often leads to a high polygon count for the 3D scene, with a value ranging from many tens of thousands to even millions of triangles, depending on the required level of detail. The real-time rendering hardware is based on a workstation including two quad-core Xeon Processors coupled with a Nvidia Quadro FX-5600 graphics board featuring 128 parallel cores and 1,5 GB of VRAM. During experiments, this hardware setup has been capable to render in stereo scenes featuring more than 5 million of polygons, at an output resolution of 2x1280x1024 pixels, with a frame rate always above 30 fps.

2.4.5. User evaluation study

The ISO 9241 standard is a milestone in usability definition, and defines it 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’’ [61]. Usability plays a significant role for the success of any interaction paradigm as well as for any application. Various usability evaluation techniques exist. Choosing among them is a trade-off between cost and effectiveness. A gross distinction is among techniques for user testing, and expert evaluation. The former requires to gather a significant sample of target users, and to derive useful information from their experience through a number of methods, as for example observing them during interaction, or asking them to speak aloud during operations to record their comments, or asking them to fill a questionnaire. Users are “expensive”, and especially for some very specialized applications they are difficult to collect. Moreover, a quite significant amount of effort is required to set up such experiments, and appropriate usability laboratories are also needed. On the contrary, expert evaluation is based on the experience of few usability experts, who test the application and analyze it according to well-established guide-lines. Some such methods, such as heuristic evaluation introduced by [62], are easier to administer than others and less costly. The other side of the coin in the case of heuristic evaluation is represented by the limitations arising from applying a small set of principles, the heuristics, to a wide range of systems. This was pointed out by various researchers, who

addressed this problem by developing more specific guidelines for particular system classes. In any case, the two approaches are in fact complementary.

Expert evaluation was performed in a first phase, with five usability experts, followed by a test with a group of fifteen users. It is worth underlining that a higher number of users would be desirable to increase the significance of evaluation results. However, the more specialised is the application to be evaluated, the more domain competencies are required by the users, the more difficult is to find a high number of them. As a matter of fact, it is important to be sure that possibly encountered problems are due to the application, and not to flaws in the user domain knowledge. A specific walkthrough was exploited [63]. Both experts and users have been asked to perform a list of tasks, before proceeding to fill the provided evaluation summary (an evaluation grid for the experts, a questionnaire for users, which are detailed in the following). Tasks performed by the users include:

- Load a new dataset within the visualization space.
- Select an anatomical object.
- Move an anatomical object on each of the three dimensions.
- Rotate an anatomical object along a freely specified pivot axis.
- Hide anatomical layers proceeding from the outside towards the inside.
- Show anatomical layers proceeding from the inside towards the outside.
- Change the transparency level of an anatomical layer.
- Show/Hide interface layout from the field of view.

A suitable evaluation grid was prepared to obtain comparable outputs from different experts. The grid was organized in four sections: graphic presentation, architecture (structure and navigation), functionality (suitability and correctness of functions), and support to the user.

Table 3. Items for experts evaluation

COMMUNICATION (GRAPHIC)
1 **- Quality of presentation**: The software product offers a user interface appropriate for the tasks, with appropriate colors, icons and a suitable layout of graphical elements. The arrangement of the elements in the interface, in relation to their function, allows the user to understand on the fly, without explanation, the function of the elements.
ARCHITECTURE (STRUCTURE AND NAVIGATION)
2 - **Quality of user dialogue:** does the software product present to the user a model of dialog which is easy to grasp by interacting with the application itself?
3 - **Simplicity of the structure of the tasks**: the tasks or activities that the user must perform correspond to a sufficiently simple structure of interaction, in relation to the user's need of processing information.
4 - **Support recognition rather than recall**: looking at the interface the user can figure out what to do, how can do it and, after performing an action, what has happened and what were the results.
5 - **Feedback**: the return information in response to the user action is timely and sufficient to make visible to the user the current state of the system, so as to avoid mistakes, misunderstandings and blocks during the interaction.
6 - **Prevention and recovery from interaction errors**: the interface guarantees the right level of flexibility to allow users to navigate freely without coming into blind alleys or in critical situations.
7 - **Consistency**: the syntax (language, input fields, colors, etc. ..) and semantics (behavior associated with the objects) in the dialogue are uniform and consistent throughout the entire software product.
FUNCTIONALITY (ADEQUACY AND ACCURACY OF FUNCTIONS)
8 - **Adherence to the user's language**: the language used at the interface is simple and familiar to the user and reflects the concepts, terminology known to him, and the content of the tasks involved.
EFFECTIVE SUPPORT PROVIDED TO THE USER
9 - **Efficiency and flexibility of use**: there are levels of support, tools and strategies of interaction for several different types of users.
10 - **Support and manual**: there are tools that can help assist the user in trouble.
OTHER: Additional comments from the evaluator

Each section includes specific items to account for. The items are presented by individual positive statements so as to avoid misleading interpretations, by describing how the interface should be. In commenting the compliance of the interface with each item, the expert can list encountered problems, positive aspects and suggestions. A list is presented in Table 3.Notice the presence of the "OTHER" item, to allow the expert to include any unanticipated comment.

In order to extract clear guidelines from merging the evaluation of the different experts, the evaluation results have been summarized and organized according to the Cognitive Dimensions (CD) framework [64][65]. Even if such framework is traditionally most exploited for visual languages, it could result extremely useful as high level 'discussion tools' (as defined by the author). As a matter of fact, CDs constitute a small and extremely clear set of terms providing a framework for a broad-brush assessment of almost any kind of cognitive artefact, even if it may not be trivial at first to see how to apply the framework to human-machine interactions, such as using a telephone or interacting with a video-surveillance application [66]. CD is adaptable to any stage of design, from the original idea to the finished artefact, and is also accessible to non-HCI experts. According to the guidelines in [67], from the original dimensions, all cognitively relevant, have been considered those which resulted to be most related to the kind of tasks at hand. In the same way, the notions of notation, environment and media upon which the analysis is based, is adapted to the context of the tested application and of the expected patterns of interaction.

The following ones were identified: a) the notation with the collection of virtual elements that can be manipulated, both pertaining to the interface and representing virtual objects related to the application at hand, b) the environment with the organization of virtual menus and gestures that can be assembled to shape interaction, i.e. to issue commands to the system to add, remove or change the visualization of virtual objects, and c) the medium with the overall virtual environment, as it is the place where the user can manipulate symbols. The following lines give a brief summary of some issues encountered in a first prototype according to such dimensions, and of possible solutions.

Abstraction: types and availability of abstraction mechanisms. The system does not allow the creation and management of abstractions in a straightforward way. As an example, in some contexts, it might be useful to create "scripts" of gestures based on user preferences and needs. However, this should require a kind of end-user programming, maybe to include the timed automata, which is out of the scope of this work.

Hidden dependencies: important links between entities are not visible. Some virtual commands trigger a sort of sub-menus. This is not evident in advance, and it is not always clear how to return to the upper-level item. As a solution, a special presentation of the virtual element might be considered.

Premature commitment: constraints on the order of doing things. This problem is not present in this approach. Apart for virtual sub-menus items, there is no predefined order to explore the virtual environment. The user is free to model interaction paths at his/her preference, and this is very important in training settings, where meta-cognitive abilities can be so spurred. Obviously, the lack of constraints is referred to the interaction flow, and not to the logic which might apply in the application domain at hand.

Viscosity: resistance to change. In this case the resistance is not implicit in the interaction, except when it is required by the application domain.

Visibility: ability to view components easily. This was not a problem for this framework. Whenever a sequence of decisions has to be taken, the preceding ones are always available. The only exceptions are some nested elements, i.e. the levels of some commands in the virtual interface menu, and the action of recalling the virtual interface after hiding it, which was not sufficiently prompted.

Closeness of mapping: closeness of representation to domain. Though simple to learn, the interaction style implies some familiarity with virtual settings. On the other hand, the gestures that were chosen to implement the interaction were considered quite natural and intuitive.

Consistency: similar semantics are expressed in similar syntactic forms. No relevant problems were found in relation with such dimension. Only some difficulties were encountered with timed gestures, as in such case the time elapsed between successive basic gestures may influence the compound gesture interpretation.

Error-proneness: notation invites mistakes. The clear interaction model helps in avoiding errors during objects manipulation, and in any case all actions are easily reversible.

*Hard mental operations: high demand on cognitive resource*s. Interaction within the framework were found easy to learn, except for an initial difficulty in precisely controlling the tracking devices.

Role-expressiveness: the purpose of a component is readily inferred. There is a clear separation between the virtual interface and the virtual objects making up the explored content. Therefore, it was always clear in what context the user was operating.

In any system evaluation procedure, user testing is of paramount relevance in confirming the validity and usefulness of the approach. To this aim, a user questionnaire was prepared to assess the perceived quality of the interaction after performing a number of tasks. The evaluation sessions involved fifteen users, ten among them were trainees in image based diagnostic while five were expert radiologists.

In the final questionnaire, the questions were presented using a five-point Likert scale, where respondents specify their level of agreement to a statement. In order to avoid any bias, some statements were in positive form and others in negative one. This was taken into account in the final assessment of results. Below is the list of the proposed statements.

1. Available gestures are easy to perform
2. Gestures are too many to remember them
3. Browsing by gestures is not intuitive
4. It is easy to select objects/place landmarks within the field of view
5. It is easy to perform object translation in any direction
6. It is easy to perform object rotation along any pivot axis
7. The type and number of available functions to interact with objects is not sufficient
8. Devices worn are comfortable during operations

The answers to the questionnaire are summarized in the Table 4 below:

The number of questions is low on purpose, since users are often negatively influenced by an excessive length of the list of questions (for tiredness, boredom, or loss of

concentration). As a matter of fact, even in the case of users questionnaire, a free comment field was included at the end to give the user the possibility to add comments and to underline issues, whre they wanted to provide a more detailed opinion. It is to say that, in general, processing free text comments is longer and harder, and that it is difficult to derive a quantitative evaluation. However, the limited number of users made this approach feasible in this case.

Many interesting comments were given by the users using such free text section. Overall, the participants reported a good confidence feeling during interaction. In particular, some of them reported to have clearly experienced an operational advantage in performing the proposed tasks, with respect to their usual bidimensional reference working environment. All users underlined the appreciated advantage of exploring 3D model instead of looking at 2D printed images. Moreover, all the participants to the evaluation sessions have also been interviewed to better understand the motivations behind the answers provided.

Overall the results show a general agreement about the whole interface and the interaction paradigm experimented. With regard to question #1, while a total of 11 testers out of 15 was able to perform the required gestures without relevant difficulties, 4 out of 15 expressed some concerns about their capacity to repeat effectively the gestures over time.

Table 4 Responses to users questionnaire

Question	I strongly agree	I agree	I do not know	I disagree	I strongly disagree
1	3	8	0	4	0
2	1	3	0	10	1
3	0	1	2	9	3
4	3	11	0	1	0
5	4	9	1	1	0
6	0	12	1	2	0
7	0	0	5	8	2
8	0	4	4	6	1

This kind of concern is somewhat reflected in the answers to question #2, where 3 participants considered the number of gestures available too high to remember. To this regard, the lack of familiarity with gesture based interface may be a possible explanation for these negative scores but, as none among the participants have ever used this kind of interfaces before, it is difficult to further elaborate on this result. Questions #4, #5 and #6 were aimed at assessing the operational effectiveness of the system. The answers are mostly positive with some limited difficulties experienced during object rotation along an arbitrary axis. Finally, it is interesting to note that answers to question #8 show a not adequate level of comfort experienced by some participants while wearing HMD and instrumented gloves. This problem is not new and, though modern helmets are much less bulky then their predecessors, they have to be still improved with regard to weight, resolution, contrast and particularly field of view to be accepted by a large audience. Nevertheless, the users were satisfied with the intuitiveness of the provided interaction (see question #3), and this was confirmed by the comments in the free text. This fact underlines that the provided natural kind of interaction with virtual artifacts is deemed as very satisfying, and this was the main goal. The approach is very promising, and it is likely that fast technological progress will soon allow to solve the emerged “ergonomic” problems.

3. Biometrics for Advanced Ambient Intelligence Environments

Information and Communication Technologies are increasingly entering in all aspects of our life and in all sectors, opening a world of unprecedented scenarios where people interact with electronic devices embedded in environments that are sensitive and responsive to the presence of users. Indeed, since the first examples of "intelligent" buildings featuring computer aided security and fire safety systems, the request for more sophisticated services, provided according to each user's specific needs has characterized the new tendencies within domotic research. The result of the evolution of the original concept of home automation is known as Ambient Intelligence [68], referring to an environment viewed as a "community" of smart objects powered by computational capability and high user-friendliness, capable of recognizing and responding to the presence of different individuals in a seamless, not-intrusive and often invisible way. As adaptivity here is the key for providing customized services, the role of person sensing and recognition become of fundamental importance.

This scenario offers the opportunity to exploit the potential of face as a not intrusive biometric identifier to not just regulate access to the controlled environment but to adapt the provided services to the preferences of the recognized user. Biometric recognition [69] refers to the use of distinctive physiological (e.g., fingerprints, face, retina, iris) and behavioural (e.g., gait, signature) characteristics, called biometric identifiers, for automatically recognizing individuals. Because biometric identifiers cannot be easily misplaced, forged, or shared, they are considered more reliable for person recognition than traditional token or knowledge-based methods. Others typical objectives of biometric recognition are user convenience (e.g., service access without a Personal Identification Number), better security (e.g., difficult to forge access). All these reasons make biometrics very suited for Ambient Intelligence applications, and this is specially true for a biometric identifier such as face which is one of the most common methods of recognition that humans use in their visual interactions, and allows to recognize the user in a not intrusive way without any physical contact with the sensor.

A generic biometric system could operate either in verification or identification modality, better known as one-to-one and one-to-many recognition [70]. In the

proposed Ambient Intelligence application the aim is to perform one-to-one recognition, as the goal is recognizing authorized users accessing the controlled environment or requesting a specific service.

The following sections, present and describe in detail a face recognition technique based on 3D features to verify the identity of subjects accessing the controlled Ambient Intelligence Environment and to customize all the services accordingly. In other terms to add a social dimension to man-machine communication and thus may help to make such environments more attractive to the human user. The proposed approach relies on stereoscopic face acquisition and 3D mesh reconstruction to avoid highly expensive and not automated 3D scanning, typically not suited for real time applications. For each subject enrolled, a bidimensional feature descriptor is extracted from its 3D mesh and compared to the previously stored correspondent template. This descriptor is a normal map, namely a color image in which RGB components represent the normals to the face geometry. Specific masks, automatically generated for each authorized person, improves recognition robustness to a wide range of facial expression and to facial hair as well.

3.1. Main approaches to 3D face recognition

As highlighted by various surveys [71] [72], face recognition represents a research topic for whom "the variety and sophistication of algorithmic approaches explored is expanding". Particularly for 3D face recognition, the main challenges result to be the improvement of recognition accuracy, a greater robustness to facial expressions and, more recently, the efficiency of algorithms. The various methods proposed so far can be categorised as holistic, if they perform face comparison at a global level; region-based, if they compare homologous regions between to faces; hybrid, if they exploits both the previous approaches and multimodal if they rely on both 2D and 3D features for the comparison, fusing togheter the results of both modalities of face matching. Many holistic methods are based on Principal Component Analysis (PCA) applied either to depth images [73] [74] or to both color and depth channels [75]. Other authors combine 3D and 2D similarity scores obtained comparing 3D and 2D profiles [76], or extract a

feature vector combining Gabor filter responses in 2D and point signatures in 3D [77]. Canonical sufaces have been exploited to mitigate the effects of facial expressions on recognition accuracy [78, 79]. Morphable models, and elastic registration have also been used [80], though the computational cost involved is relevant. Among region-based approaches, Xu et al. [81] aim to divide face in sub-regions using nose as the anchor, PCA to reduce feature space dimensionality and minimum distance for matching. They also proposed a method to face partitioning based on the intersection between spheres of increasing radius and the face scans [82]. Another major research trend is based on Iterative Closest Point (ICP) algorithm, which has been exploited in many variations for 3D shape aligning, matching or both. The first example of this kind of approach to face recognition has been presented from Medioni and Waupotitsch [83], while other authors [84] proposed to apply ICP to a set of selected subregions instead [85, 86]. Iso-geodesic stripes and 3D Weighted Walkthroughs (3DWWs) have been proposed in [87] proving to be accurate in terms of recognition and robust to intra-class variations. The methods belonging to the hybryd and multimodal categories aim at improving the precision of recognition by combining well established techniques like PCA, LDA and ICP and/or operating at a 2D and 3D level, to overcome the limits of the individual approaches. The work by Mian et al. [88] represent a good example of this approach, producing the best score on the FRGC v2.0 contest.

3.2. Technical issues in face recognition

As already recalled before, the research on face recognition conducted in the last two decades as witnessed by the various editions of face recognition contests organized in the last decade (the FRVT- Face Recognition Vendor Test [89] and the FRGC - Face Recognition Grand Challenge [90], produced a great number of algorithms and methodologies [91] [92] [93] [94] [95] [96]. The first objective was mainly in raising the upper limit of recognition accuracy in controlled conditions (one of the explicit goals of FRGC) also because most of the first publicly available reference datasets for face recognition like the FERET [97] and the YaleB [98] were acquired in studio with controlled settings and cooperative subjects.

However, after the first wave of approaches resulted in higher and higher recognition precision, the efforts were focused on improving performances in the presence of Pose, Illumination and Expression (PIE) variations, three issues extremely common in real world applications which could degrade significantly the accuracy of the results. That said, face recognition poses even more compelling challenges [99] that can be resumed as:

- Subject-sensor distance
- Image quality
- Unconstrained Pose/Expression/Illumination
- Partial occlusions and facial hair/ware

In face recognition applications, the average distance between the subject to be acquired and the sensing device is typically in the range of one meter or less. A common way to quantify the resolution of a face crop for recognition is to measure the distance between the pupils of the captured subject (for almost frontal shots). A distance of 20-30 pixels is generally considered a lower limit for reliable recognition, while 50+ pixels represent a more realistic measure. Let us consider the hardware related aspects of these requirements. By using a typical surveillance camera, capturing 640x480 pixels per frame and equipped with a "normal" lens providing a 45° FOV (Field Of View), there could be face crops with 55-60 pixels of intra-pupils spacing at about one meter of subject-camera distance, 27-30 pixels at 2 meters and only 10 pixels around 9 meters according to:

(1) $N_{IS}=Res_H/((2\sin(FOV_H/2)D_{C\text{-}S})/I_S)$

where N_{IS} is the inter-pupils spacing in pixels, Res_H is the horizontal resolution of captured frame, FOV_H is the horizontal field of view of the camera's lens, $D_{C\text{-}S}$ is the camera-subject distance and I_S is the average intra-pupils spacing (6,5-7,5 cm). The situation is even worse in case of wide-angle lenses (common in ceiling or elevated camera installations), as the resulting face crops would be much smaller and, probably,

difficult to use for recognition purposes. Obviously, it would be possible to get larger crops by increasing sensor resolution and/or decreasing the lens' FOV (higher focal length). In the first case this comes at the price of a more expensive device and of a much greater overhead both in terms of processing power/computing time required to analyze the larger frame and of a much greater video throughput (a 1280x720 sensor produces almost fourfold the pixels per frame outputted by a standard 640x480 device, while a Full HD 1920x1080 sensor has almost seven times more pixels). In the second case, decreasing the FOV means to use a greater fraction of the captured image, but this implies that the probability a subject will stay framed is much smaller for a fixed camera. To this regard the use of software controlled Pan, Tilt and Zoom (PTZ) cameras (possibly capable to follow the subject while zooming in/out when required to get more detailed face crop) could be of interest, though it would increase the complexity and the overall cost of the system.

With regard to the image processing approaches to cope with low resolution resulting from distant face capturing, most methods proposed so far exploits super-resolution techniques [100] [101] [102] which output recovered high resolution images from low resolution input samples, but often at the price of strong artifacts. Other methods exploits coupled metric learning [103], dictionaries [104] or multidimensional scaling [105], but all these techniques tend to degrade their performance in presence of PIE variations.

Whatever its size, if the captured facial image results of low quality the chances of a reliable identification are reduced. Factors like image noise and blur, indeed, may degrade the discriminant information in acquired images, thus lowering considerably the recognition performance. This is particularly true for face captured at a distance. Image noise is partly related to the characteristics of the typically small sized CCD/CMOS sensing device (in general, the smaller the physical pixel dimension the stronger the noise produced due to mutual interference effects, particularly in low light) but it is also due to compression artifacts, common in case of video capturing.

Blur (i.e. image defocusing) represents another cause of image degradation and loss of details. It results from incorrect focalization of the image on the focal/sensor plane (at a local and/or global level) due to optical defects or insufficient depth of field with

respect to the camera-subject distance. As depth of field decreases as focal length increases, it is easy to understand why using telephoto lens to zoom in subjects may result in more critical focusing and higher percentage of blurred frames. On the contrary, wide-angle lenses have very ample depth of field.

Apart from optical issues, image may be affected by another kind of defocusing effect, known as "motion blur" and related to the dynamic aspects of frame capturing. Motion blur is due to the insufficient frame capturing rate in presence of rapid subject or camera movements and it results in image blurring along the motion direction. Of course, the higher the capturing rate, the smaller the motion blur effect.

In outdoor scenarios, also weather conditions may involve additional factors eventually affecting image quality. Rain, fog, haze and snow may contribute in reducing image's dynamic range while increasing noise. In a typical scenario all the aspects mentioned before are inter-dependent at some extents, indeed face recognition could be performed at a distance which is likely to require medium-to-long focal length camera optics with limited depth of field and prone to exhibit motion blur artifacts in case the subject moves rapidly across the FOV width.

Furthermore, as larger imaging sensors require much costly and bulkier optics to achieve the same focal length and aperture, the sensor's size would probably be small, therefore featuring a greater average level of noise and a narrower dynamic range. From all these considerations, a clear need emerges for objectively quantifying image quality in facial images taken at a distance through a specific metric to assess their usability for recognition. To this aim, in [106] a signal-to-noise ratio estimator is proposed by correlating signal-to-noise estimation and noise level to the statistics of image edge intensity. Experimental evidence proves that this estimator is directly correlated to the recognition accuracy of a face recognition system so that it represents a trustable measure of image quality useful to filter out unsuitable incoming images.

Unconstrained Pose/ Illumination/Expression **(**PIE) variations represent one of the most active research topics for face recognition, as the combination of these factors often transforms a face's appearance so as to result less similar to its neutral version than to other faces belonging to completely different individuals. Consequently, uncontrolled pose, expression and illumination may easily lead to both false positive

and false negative recognition. Therefore, face recognition literature is rich of proposals specifically targeted to cope with pose variations by means of a number of approaches such as morphable models [107], tied factor analysis [108], stereo matching [109], albedo estimation [110]. Robustness to illumination variations has been approached by exploiting spherical harmonics [111], quotient image [112], total variation models [113]; stereoscopic images [114]; albedo [115] and dictionaries [116], while canonical image [117] iso-geodesic stripes [118] represents a few among the many methods developed toward the goal of expression invariant face representation. Though many advances have been registered over the years with regard to these topics, most of the solutions proposed perform at their best only in controlled situation and/or at a close distance. This is partly due to the operational constraints (e.g. 3D methods need a sensing technology often limited to indoor/close-range usage) and partly to the specific challenges faced for outdoor recognition, where the subject's freedom of movement and complexity of motion patterns is potentially much greater than in a typical indoor face verification scenario, the intensity and the effects of illumination can be extreme (direct sunlight, strong sharp shadows, environmental reflections, etc.) and facial expression may have an even stronger impact due to low resolution or long range issues in facial image quality.

A further challenge for face recognition in general and for face re-identification in particular is represented by the partial lack of information eventually due to occluding objects. These may be clothing elements (hats, scarves, sunglasses, etc.) worn by the subject to be identified, other persons or even architectural elements positioned between the subject-camera line-of-sight. In both cases the face descriptors resulting from capture may lack important features that possibly affect the reliability of recognition. Sparse representation [96] has proved to be suited to address this issue, while robustprincipal component analysis has also been exploited with interesting results [119]. Particular applicative scenarios might involve severe occlusion conditions, as the width of movements, the probability of a subject wearing sunglasses and the variety of objects and persons eventually hiding part of the captured face are much more relevant than in "conventional" face recognition. Beard and facial hair in general are (along with facial expressions) among the "user-factors" which may affect face recognition

performance. These variations of facial appearance have also been addressed by a number of papers but their impact on remote outdoor face recognition is greater for uncontrolled situations because average image quality, and resolution are likely to be lower.

3.3. Face signature by normal map

The basic idea behind the system proposed is to represent user's facial surface by a digital signature called normal map. A normal map is an RGB color image providing a 2D representation of the 3D facial surface, in which each normal to each polygon of a given mesh is represented by a RGB color pixel. To this aim, the 3D geometry is projected onto 2D space through spherical mapping. The result is a bidimensional representation of original face geometry which retains spatial relationships between facial features. Color info coming from face texture are used to mask eventual beard covered regions according to their relevance, resulting in a 8 bit greyscale filter mask (Flesh Mask). Then, a variety of facial expressions are generated from the neutral pose through a rig-based animation technique, and corresponding normal maps are used to compute a further 8 bit greyscale mask (Expression Weighting Mask) aimed to cope with expression variations. At this time the two greyscale masks are multiplied and the resulting map is used to augment with extra 8 bit per pixel the normal map, resulting in a 32 bit RGBA bitmap (Augmented Normal Map). The whole process (see Fig. 25) is discussed in depth in the following section. As the proposed method works on 3D polygonal meshes it is necessary to acquire actual faces first and to represent them as polygonal surfaces.

The Ambient Intelligence context, in which the face recognition algorithm has to operate, requires fast user enrollment to avoid annoying waiting time. Usually, most 3D face recognition methods work on a range image of the face, captured with laser or structured light scanner. This kind of devices offer high resolution in the captured data, but they are too slow for a real time face acquisition. Face unwanted motion during capturing could be another issue, while laser scanning could not be harmless to the eyes. For all this reasons a 3D mesh reconstruction from stereoscopic images based on [120] has been chosen since it requires a simple equipment more likely to be adopted in a real

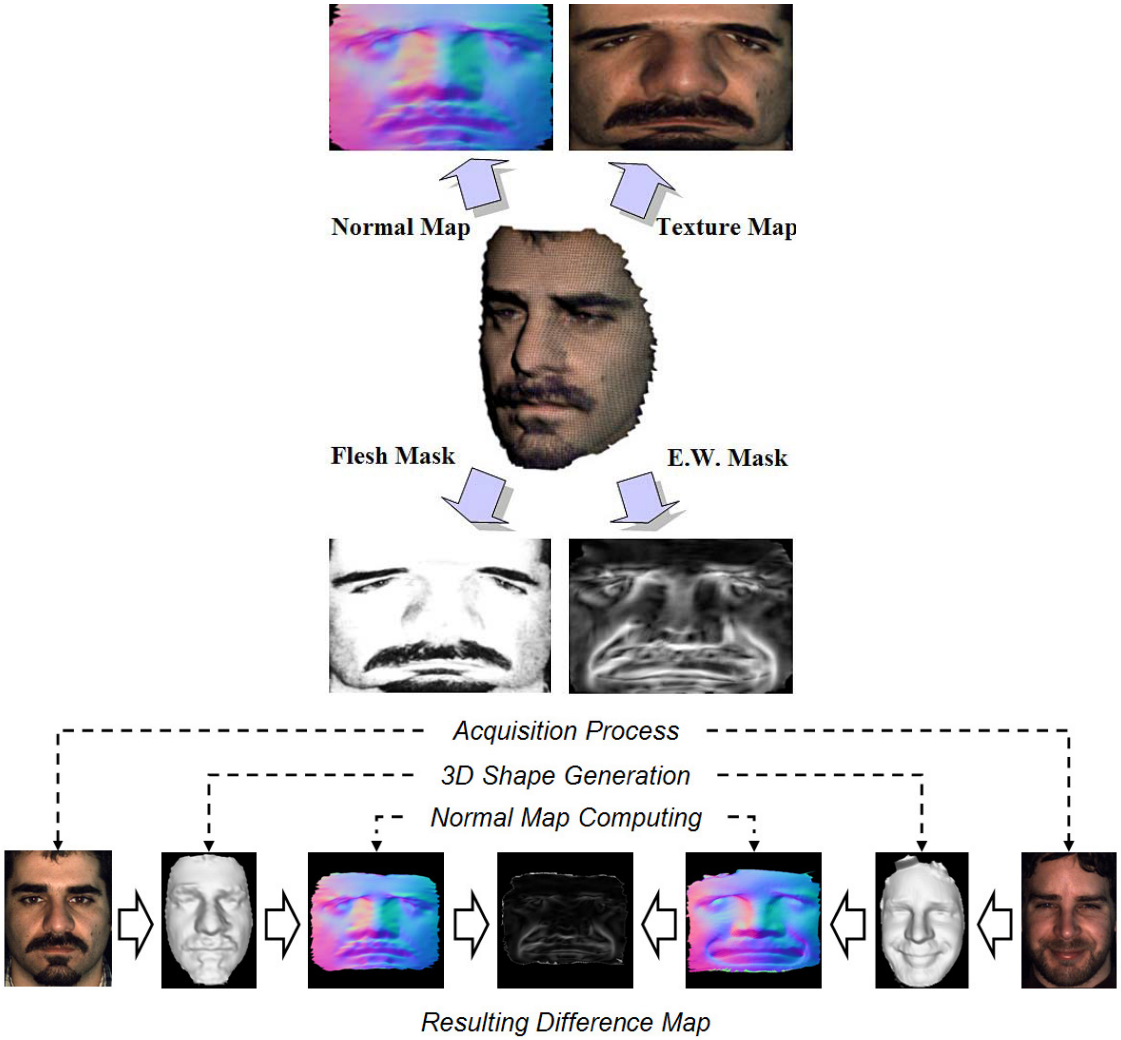

Figure 25. *Facial and Facial Expression Recognition workflow.*

application: a couple of digital cameras shooting at high shutter speed from two slightly different angles with strobe lighting. Though the resulting face shape accuracy is inferior compared to real 3D scanning it proved to be sufficient for recognition yet much faster, with a total time required for mesh reconstruction of about 0.5 sec. on a I7/3.4 Ghz based PC, offering additional advantages, such as precise mesh alignment in 3D space thanks to the warp based approach, facial texture generation from the two captured orthogonal views and its automatic mapping onto the reconstructed face geometry.

As the 3D polygonal mesh resulting from the reconstruction process is an approximation of the actual face shape, polygon normals describe local curvature of captured face which could be view as its signature. As shown in Fig. 26, these normals will be represented by a color image transferring face's 3D features in a 2D space. To preserve the spatial relationships between facial features, vertices' 3D coordinates are projected onto a 2D space using a spherical projection. It is now possible to store normals of mesh M in a bidimensional array N using mapping coordinates, by this way each pixel represents a normal as RGB values. The resulting array is referred as the Normal Map N of mesh M and this is the desired signature to be used for identity verification.

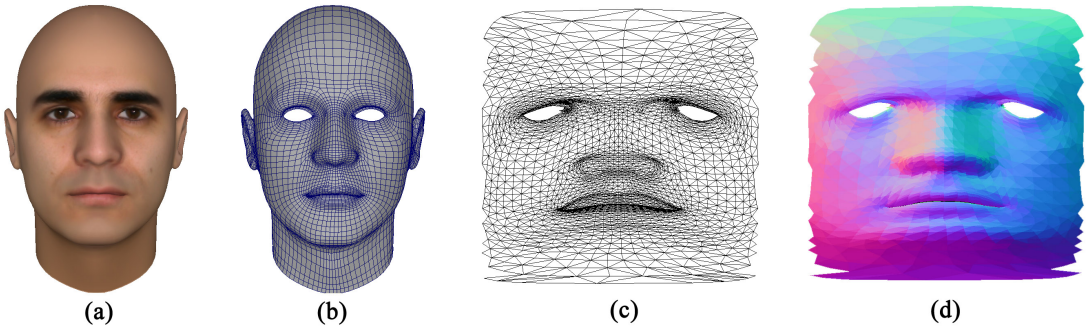

Figure 26. *(a) 3d mesh model, (b) wireframe model, (c) projection in 2D spatial coordinates, (d) normal map.*

To compare the normal map N_A from input subject to another normal map N_B previously stored in the reference database, has to be computed:

$$\theta = \arccos\left(r_{N_A} \cdot r_{N_B} + g_{N_A} \cdot g_{N_B} + b_{N_A} \cdot b_{N_B}\right) \qquad (2)$$

the angle included between each pairs of normals represented by colors of pixels with corresponding mapping coordinates, and store it in a new Difference Map D with components r, g and b opportunely normalized from spatial domain to color domain, so $0 \le r_{N_A}, g_{N_A}, b_{N_A} \le 1$ and $0 \le r_{N_B}, g_{N_B}, b_{N_B} \le 1$. The value θ, with $0 \le \theta < \pi$, is the angular

difference between the pixels with coordinates (x_{N_A}, y_{N_A}) in N_A and (x_{N_B}, y_{N_B}) in N_B and it is stored in D as a gray-scale color.

At this point, the histogram H is analyzed to estimate the similarity score between N_A and N_B. The resulting angles between each pair of comparisons (sorted from 0° degree to 180° degree) is represented on the X axis, while the Y axis represents the total number of differences found. The curvature of H represents the angular distance distribution between mesh MA and MB, thus two similar faces featuring very high values on small angles, whereas two unlike faces have more distributed differences (see Fig. 27). The similarity score is defined through a weighted sum between H and a Gaussian function G, as in:

$$similarity\_score = \sum_{x=0}^{k} \left(H(x) \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \right) \qquad (3)$$

where with the variation of σ and k is possible to change recognition sensibility. To reduce the effects of residual face misalignment during acquisition and sampling phases, the angle θ is calculated using a k × k (usually 3 × 3 or 5 × 5) matrix of neighbour pixels.

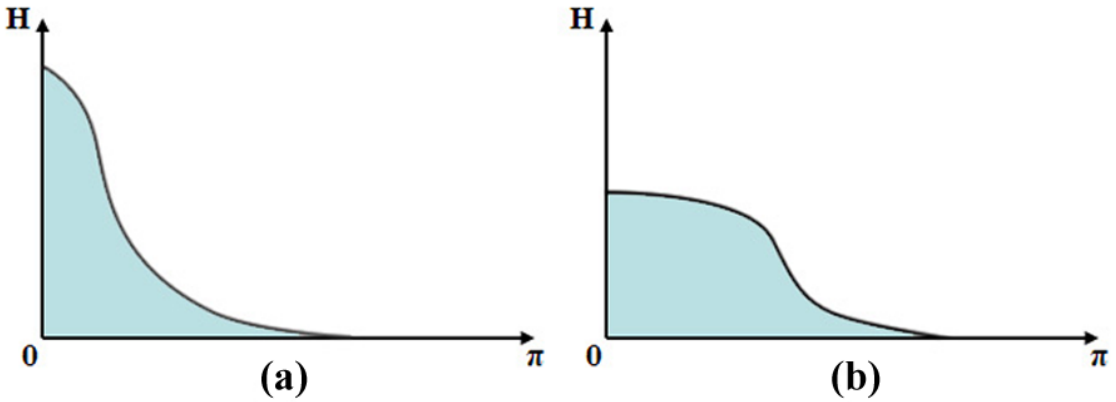

Figure 27. *Example of histogram H to represent the angular distances. (a) shows a typical histogram between two similar Normal Maps, while (b) between two different Normal Maps.*

The presence of beard with variable length covering a portion of the face surface in a subject previously enrolled without it (or vice-versa), could lead to a measurable difference in the overall or local 3D shape of the face mesh (see Fig.28). In this case the recognition accuracy could be affected resulting, for instance, in a higher False Rejection Rate FRR. To the aim of improving the robustness to this kind of variable facial features, the method relies on color data from the captured face texture to mask the non-skin region, eventually disregarding them during the comparison. Flesh hue characterization in the HSB color space is exploited to discriminating between skin and beard/moustaches/eyebrows. Indeed, the hue component of each given texel is much less affected from lighting conditions during capturing then its corresponding RGB value. Nevertheless there could be a wide range of hue values within each skin region due to factors like facial morphology, skin conditions and pathologies, race, etc., so defining this range on a case by case basis is required to obtain a valid mask. To this aim a set of specific hue sampling spots located over the face texture at absolute coordinates is considered. These spots are selected to be representative of flesh's full tonal range and possibly distant enough from eyes, lips and typical beard and hair covered regions.

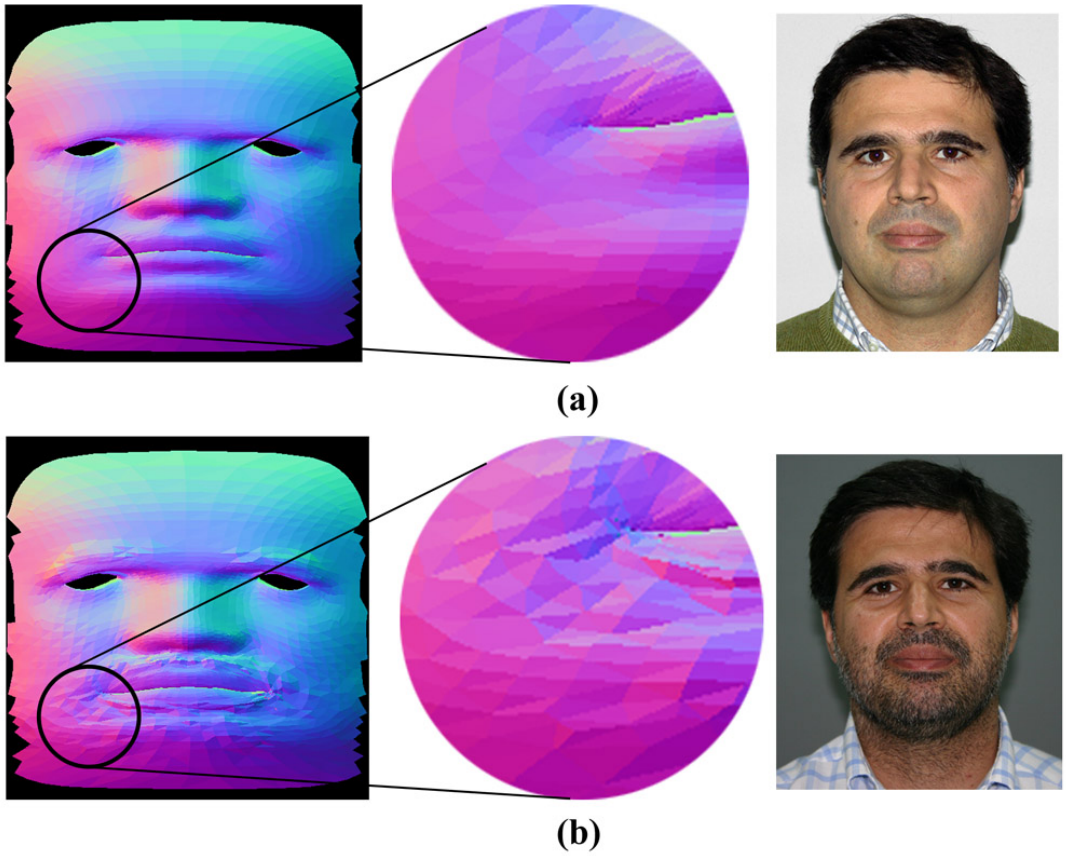

Figure 28. *Normal maps of the same subject enrolled in two different sessions with and without beard.*

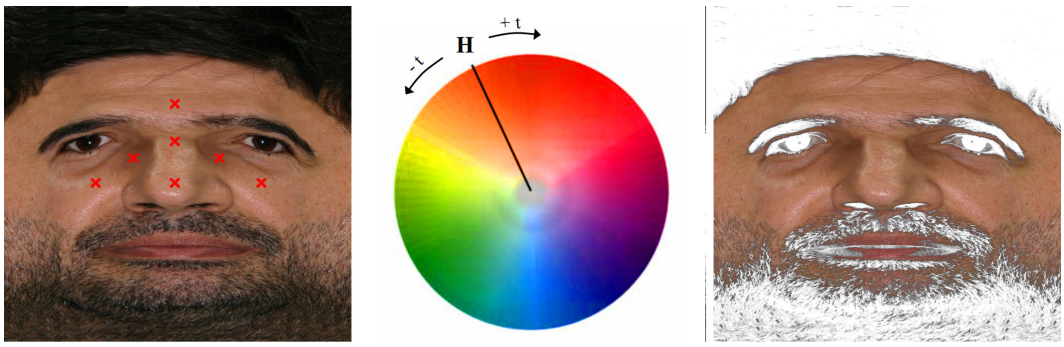

Figure 29. *Flesh Hue sampling points (a), Flesh Hue Range (b) non-skin regions in white (c).*

This is possible because each face mesh and its texture are centered and normalized during the image based reconstruction process (i.e. the face's median axis is always centered on the origin of 3D space with horizontal mapping coordinates equal to 0.5), otherwise normal map comparison would not be possible. A 2D or 3D technique could be used to locate main facial features (eye, nose and lips) and to position the sampling spots relative to this features, but even these approaches are not safe under all conditions. For each sampling spot not just that texel but a 5 x 5 matrix of neighbour texels is sampled, averaging them to minimize the effect of local image noise. As any sampling spot could casually pick wrong values due to local skin color anomalies such as moles, scars or even for improper positioning, the median of all resulting hue values from all sampling spots is calculated, resulting in a main Flesh Hue Value *FHV* which is the center of the valid flesh hue range. All texels whose hue value is within the range: $-t \leq FHV \leq t$, (where t is a hue tolerance which experimentally found that could be set below 10°) are considered belonging to skin region (see Fig. 29). After the skin region has been selected, it is filled with pure white while the remaining pixels are converted to a greyscale value depending on their distance from the selected flesh hue range (the more the distance the darker the value). To improve the facial recognition system and to address facial expressions an expression weighting mask (a subject specific pre-calculated mask aimed to assign different relevance to different face regions) is exploited. This mask, which shares the same size of normal map and difference map, contains for each pixel an 8 bit weight encoding the local rigidity of the face surface based on the analysis of a pre-built set of facial expressions of the same subject.

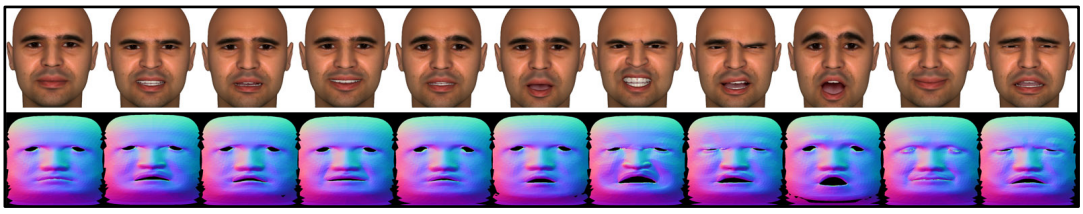

Figure 30. *An example of normal maps of the same subject featuring a neutral pose (leftmost face) and different facial expressions.*

Indeed, for each subject enrolled, each of expression variations (see Fig. 30) is compared to the neutral face resulting in difference maps. The average of this set of difference maps specific to the same individual represent its expression weighting mask. More precisely, given a generic face with its normal map N_0 (neutral face) and the set of normal maps N_1, N_2, …, N_n (the expression variations), the set of difference map D_1, D_2, …, D_n resulting from {`$N_0$ - $N_1$', `N_0 - N_2', …, `N_0 – N_n'} is calculated first. The average of set {D_1, D_2, …, D_n} is the expression weighting mask which is multiplied by the difference map in each comparison between two faces. Expression variations are generated through a parametric rig based deformation system previously applied to a prototype face mesh, morphed to fit the reconstructed face mesh [120]. This fitting is achieved via a landmark-based volume morphing where the transformation and deformation of the prototype mesh is guided by the interpolation of a set of landmark points with a radial basis function. To improve the accuracy of this rough mesh fitting a surface optimization obtained minimizing a cost function based on the Euclidean distance between vertices is applied. So each 24 bit normal map can be augmented with the product of Flesh Mask and Expression Weighting Mask normalized to 8 bit (see Fig. 31).

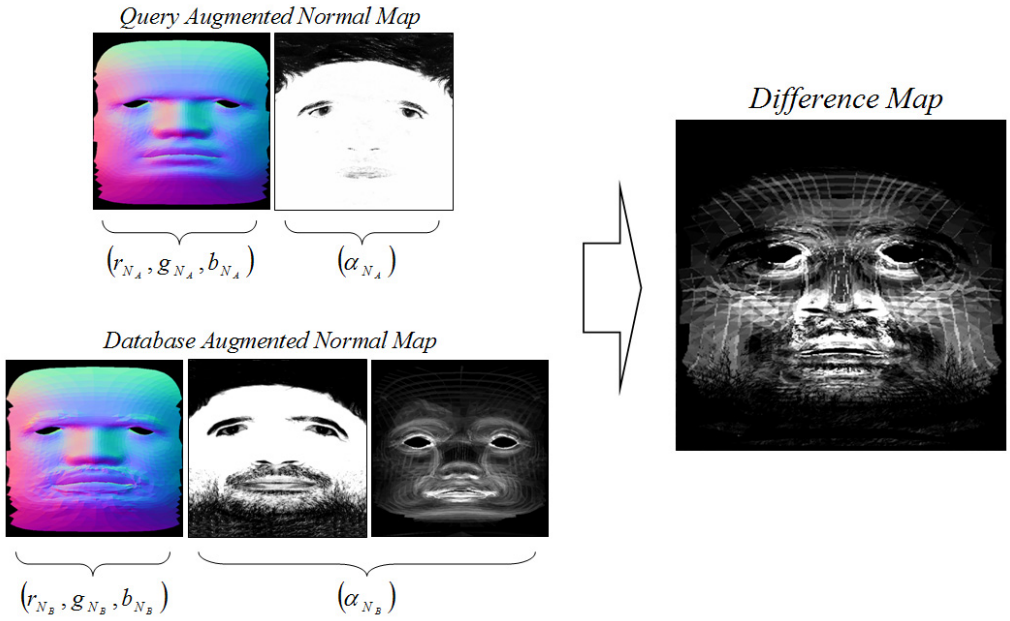

Figure 31. *Comparison of two Normal Maps using Flesh Mask and resulting Difference Map (c).*

The resulting 32 bit per pixel RGBA bitmap can be conveniently managed via various image formats like the Portable Network Graphics format (PNG) which is typically used to store for each pixel 24 bit of colour and 8 bit of alpha channel (transparency). When comparing any two faces, the difference map is computed on the first 24 bit of color info (normals) and multiplied to the alpha channel (filtering mask).

3.4. A Biometrics-empowered Ambient Intelligence Environment

Ambient Intelligence (AmI) worlds offer exciting potential for rich interactive experiences. The metaphor of AmI envisages the future as intelligent environments where humans are surrounded by smart devices that makes the ambient itself perceptive to humans' needs or wishes. The Ambient Intelligence Environment can be defined as the set of actuators and sensors composing the system together with the domotic interconnection protocol. People interact with electronic devices embedded in environments that are sensitive and responsive to the presence of users. This objective is achievable if the environment is capable to learn, build and manipulate user profiles considering from a side the need to clearly identify the human attitude; in other terms, on the basis of physical and emotional user status captured from a set of biometric features.

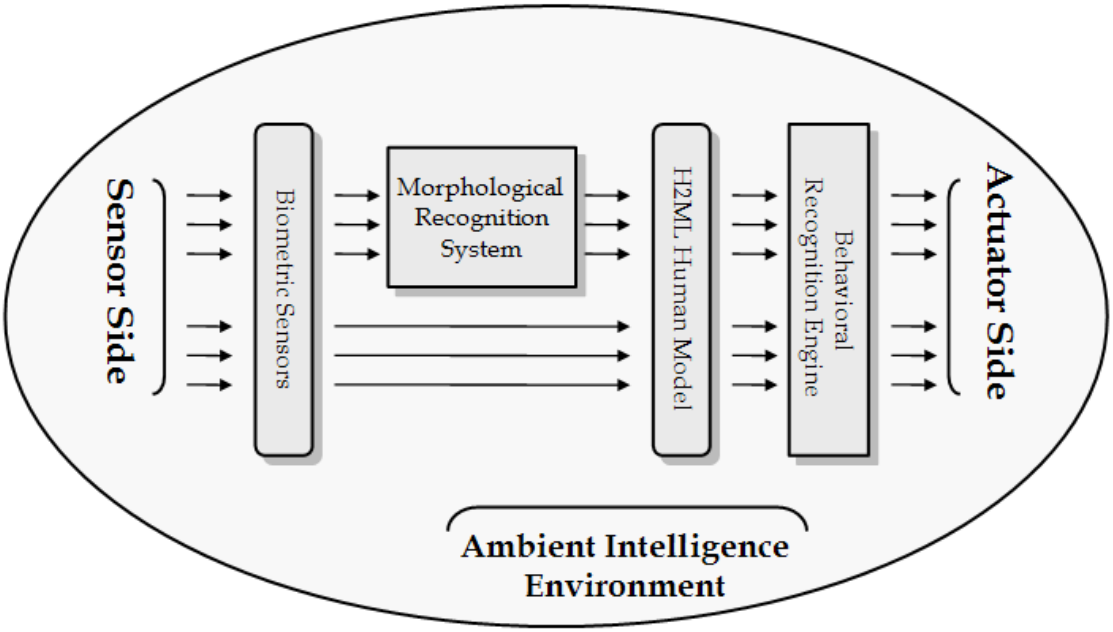

Figure 32. *Ambient Intelligence Architecture.*

To design Ambient Intelligent Environments, many methodologies and techniques have to be merged together originating many approaches reported in recent literature [121].In particular, a framework aimed to gather biometrical and environmental data, described in [122] is exploited to test the effectiveness of face recognition systems to aid security and to recognize the emotional user status. This AmI system's architecture is organized in several sub-systems, as depicted in Fig. 32, and it is based on the following sensors and actuators: internal and external temperature sensors and internal temperature actuator, internal and external luminosity sensor and internal luminosity actuator, indoor presence sensor, a infrared camera to capture thermal images of user and a set of color cameras to capture information about gait and facial features. Firstly *Biometric Sensors* are used to gather user's biometrics (temperature, gait, position, facial expression, etc.) and part of this information is handled by *Morphological Recognition Subsystems* (*MRS*) able to organize it semantically. The resulting description, together with the remaining biometrics previously captured, are organized in a hierarchical structure based on XML technology in order to create a new markup language, called *H2ML* (*Human to Markup Language*) representing user status at a given time. Considering a sequence of H2ML descriptions, the *Behavioral Recognition Engine* (*BRE*), tries to recognize a particular user behaviour for which the system is able to provide suitable services. The available services are regulated by means of the *Service Regulation System* (*SRS*), an array of fuzzy controllers exploited to achieve hardware transparency and to minimize the fuzzy inference time. This architecture is able to distribute personalized services on the basis of physical and emotional user status captured from a set of biometric features and modelled by means of a mark-up language, based on XML. This approach is particularly suited to exploit biometric technologies to capture user's physical info gathered in a semantic representation describing a human in terms of morphological features.

3.5. Experimental Results

As one of the aims in experiments was to test the performance of the proposed method in a realistic operative environment, a 3D face database was built from the face capture

station used in the domotic system described above. The capture station featured two digital cameras with external electronic strobes shooting simultaneously with a shutter speed of 1/250 sec. while the subject was looking at a blinking led to reduce posing issues. More precisely, every face model in the gallery has been created deforming a pre-aligned prototype polygonal face mesh to closely fit a set of facial features extracted from front and side images of each individual enrolled in the system.

Indeed, for each enrolled subject a set of corresponding facial features extracted by a structured snake method from the two orthogonal views are correlated first and then used to guide the prototype mesh warping, performed through a Dirichlet Free Form Deformation. The two captured face images are aligned, combined and blended resulting in a color texture precisely fitting the reconstructed face mesh through the feature points previously extracted. The prototype face mesh used in the dataset has about 7K triangular facets, and even if it is possible to use mesh with higher level of detail this resolution resulted to be adequate for face recognition. This is mainly due to the optimized tessellation which privileges key area such as eyes, nose and lips whereas a typical mesh produced by 3D scanner features almost evenly spaced vertices. Another remarkable advantage involved in the warp based mesh generation is the ability to reproduce a broad range of face variations through a rig based deformation system. This technique is commonly used in computer graphics for facial animation [123] and is easily applied to the prototype mesh linking the rig system to specific subsets of vertices on the face surface. Any facial expression could be mimicked opportunely combining the effect of the rig controlling lips, mouth shape, eye closing or opening, nose tip or bridge, cheek shape, eyebrows shape, etc. The facial deformation model used is based on [124] and the resulting expressions are anatomically correct.

The 3D dataset of each enrolled subject has been augmented through the synthesis of fiften additional expressions selected to represent typical face shape deformation due to facial expressive muscles, each one included in the weighting mask. The fifteen variations to the neutral face are grouped in three different classes: “good-mood”, “normal-mood” and “bad-mood” emotional status (see Fig. 33). For the first group of experiments, a database of 235 3D face models in neutral pose (represented by “normal-mood” status) each one augmented with fifteen expressive variations was obtained.

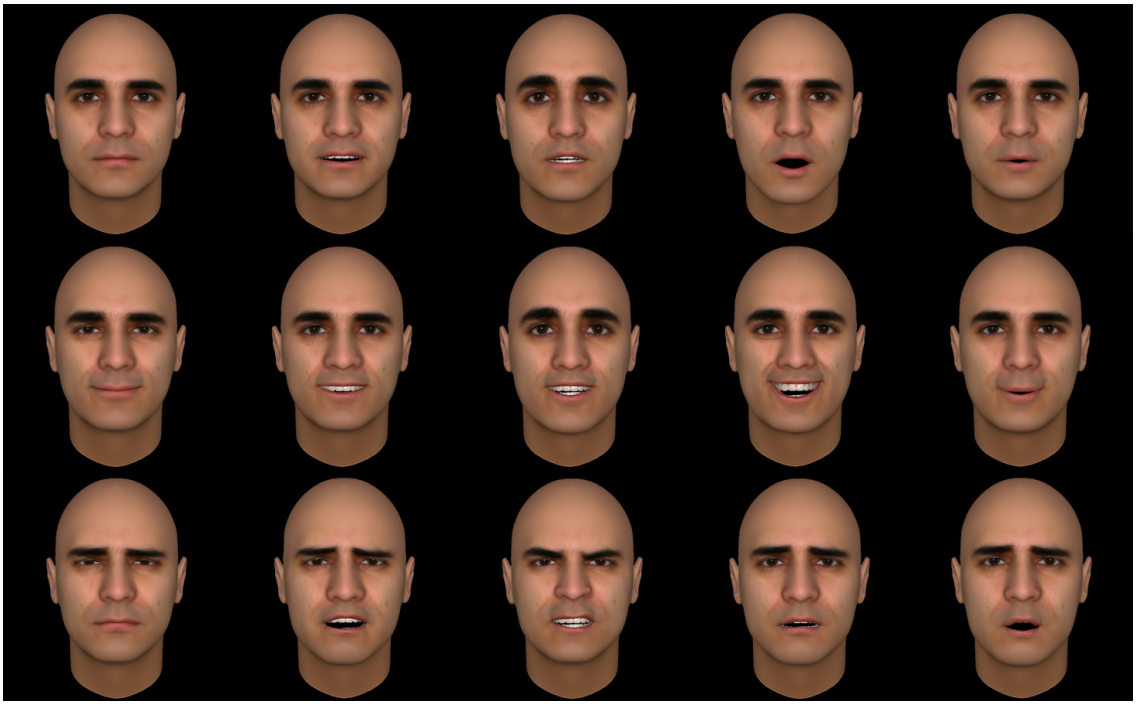

Figure 33. *Facial Expressions grouped in normal-mood (first row), good-mood (second row), bad-mood (third row).*

Experimental results are generally good in terms of accuracy, showing a Recognition Rate of 100% using the expression weighting mask and flesh mask, the Gaussian function with σ=4.5 and k=50 and normal map sized 128 × 128 pixels. These results are generally better than those obtained by many 2D algorithms but a more meaningful comparison would require a face dataset featuring both 2D and 3D data. To this aim a PCA-based 2D face recognition algorithm [125] [126] has been experimented on the same subjects. The PCA-based recognition system has been trained with frontal face images acquired during several enrolment sessions (from 11 to 13 images for each subject), while the probe set is obtained from the same frontal images used to generate the 3D face mesh for the proposed method. This experiment has shown that our method produce better results than a typical PCA-based recognition algorithm on the same subjects. More precisely, PCA-based method reached a recognition rate of 88.39% on gray-scaled images sized to 200 × 256 pixels, proving that face dataset was really challenging. Figure 10 shows the precision/recall improvement provided by the expression weighting mask and flesh mask. The results showed in Fig.34-a were

achieved comparing in one-to-many modality a query set with one expressive variations to an answer set composed by one neutral face plus ten expression variations and one face with beard. In Fig. 34-b are shown the results of one-to-many comparison between subject with beard and an answer set composed of one neutral face and ten expressive variations. Finally for the test reported in Fig. 34-c the query was an expression variation or a face with beard, while the answer set could contain a neutral face plus ten associated expressive variations or a face with beard. The three charts clearly show the benefits involved with the use of both expressive and flesh mask, specially when combined together.

The second group of experiments has been conducted on FRGC dataset rel. 2/Experiment 3s (only shape considered) to test the method's performance with respect to Receiver Operating Characteristic (ROC) curve which plots the False Acceptance Rate (FAR) against Verification Rate (1 – False Rejection Rate or FRR) for various decision thresholds. The 4007 faces provided in the dataset have undergone a pre-processing stage to allow our method to work effectively.

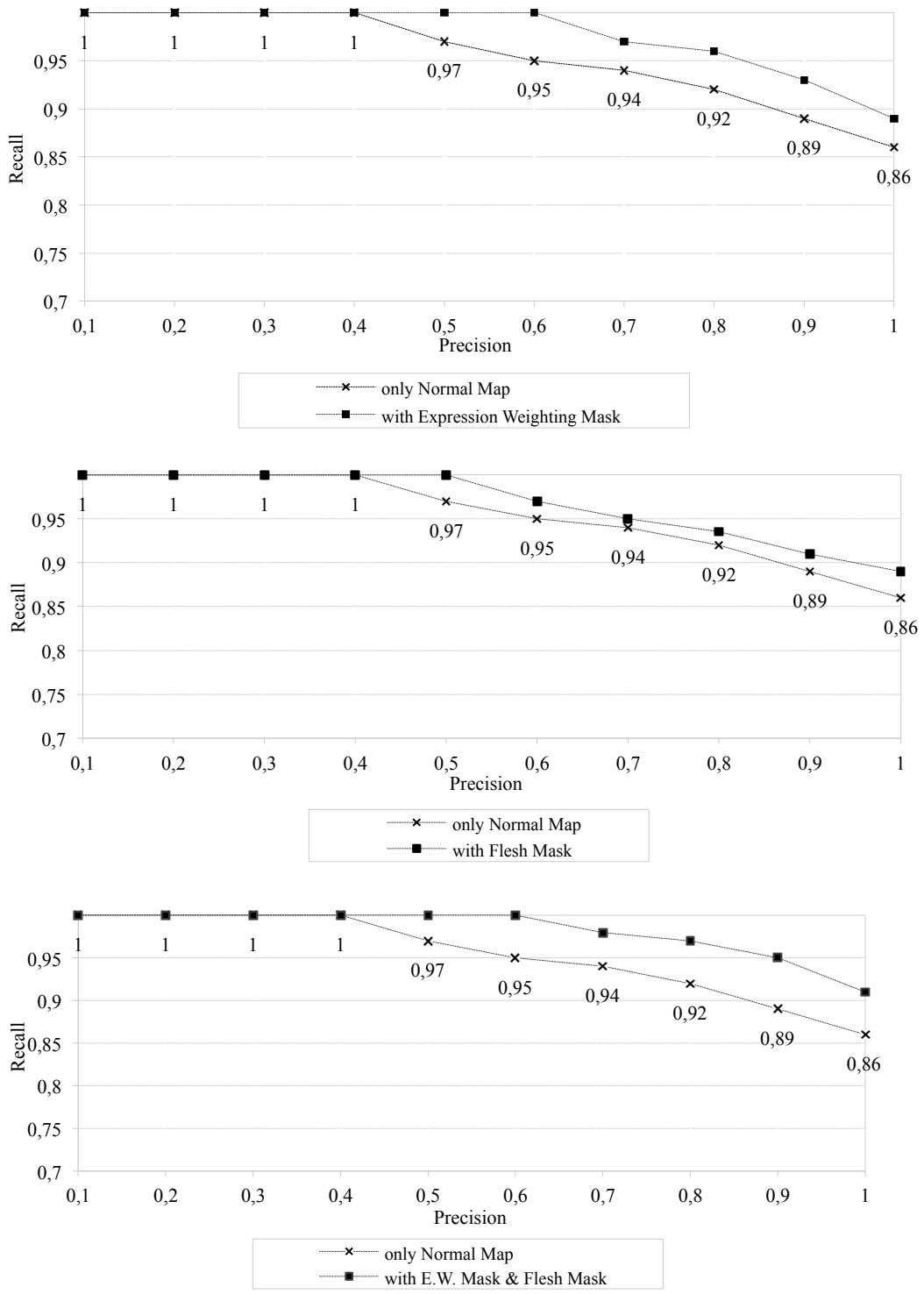

Figure 34. *Precision/Recall Testing with and without Expression Weighting Mask and Flesh Mask to show efficacy respectively to (a) expression variations, (b) beard presence and (c) both.*

The typical workflow included: mesh alignment using the embedded info provided by FRGC dataset such as outer eye corners, nose tip, chin prominence; mesh subsampling to one fourth or original resolution; mesh cropping to eliminate unwanted detail (hair, neck, ears, etc.); normal map filtering by a 5 × 5 median filter to reduce capture noise and artifacts. Fig. XX shows resulting ROC curves with typical ROC values at FAR = 0.001. The Equal Error Rate (EER) measured on all two galleries reaches 5.45% on the our gallery and 6.55% on FRGC dataset. Finally, the method has been tested in order to evaluate statistically the behaviour of method to recognize the "emotional" status of the user. To this aim, a one-to-one comparison of a probe set of 3D face models representing real subjective mood status captured by camera (three facial expressions per person) versus three gallery set of artificial mood status generated automatically by control rig based deformation system (fifteen facial expression per person grouped as shown in Fig. 35) has been performed.

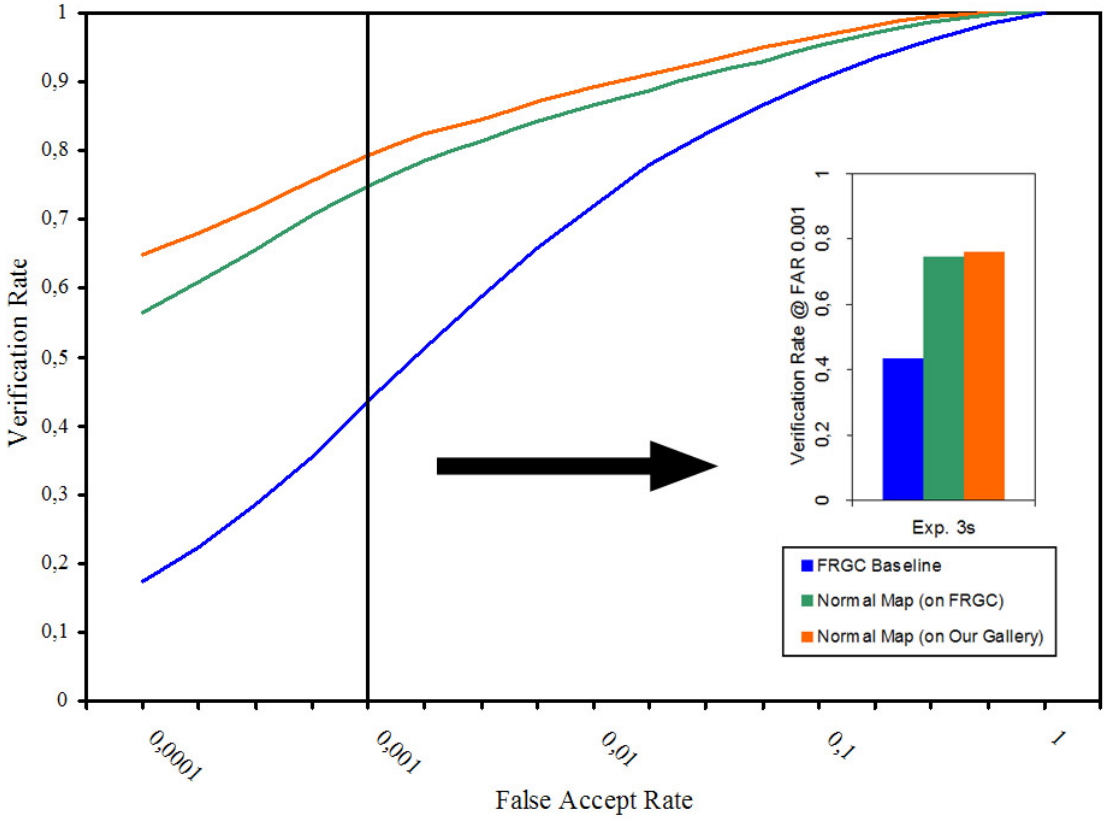

Figure 35. *Comparison of ROC curves and Verification Rate at FAR=0.001.*

As shown in Table 1, the results are very interesting, because the mean recognition rate on “good-mood” status gallery is 100% while on “normal-mood” and “bad-mood” status galleries is 98.3% and 97.8% respectively (probably, because of the propensity of the people to make similar facial expressions for “normal-mood” and “bad-mood” status). Ongoing research will implement a true multi-modal version of the basic algorithm with a second recognition engine dedicated to the color info (texture) which could further enhance the discriminating power.

Recognition Rate		
“normal-mood”	“good-mood”	“bad-mood”
98.3%	100%	97.8%

Table 1. *The behaviour of method to recognize the “emotional” status of the user.*

BIBLIOGRAPHY

[1] Van Krevenlen, D.W.F. and Poelman, R., "A Survey of Augmented Reality Technology, Applications and Limitations", The International Journal of Virtual Reality, 9(2), pp. 1-20, 2010.

[2] W. Friedrich, "ARVIKA-Augmented Reality for Development, Production and Service," Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR '02), 2002, pp. 3–4.

[3] K. Kiyokawa, M. Billinghurst, B. Campbell, and E. Woods. An occlusion-capable optical see-through head mount display for supporting co-located collaboration. ISMAR 2003, pp. 133–141

[4] Shimoda H., Maeshima M., Nakai T., Bian Z., Ishii H., and Yoshikawa H.: Development of a Tracking Method for Augmented Reality Applied to Nuclear Plant Maintenance Work. In the Proceedings of the ACM symposium on Virtual reality software and technology ACM New York, NY, USA 2006.

[5] Mendez E., Kalkofen D., Schmalstieg D.: Interactive Context-Driven Visualization Tools for Augmented Reality. In the Proceeding of ISMAR '06. Proceedings of the 5th IEEE and ACM International Symposium on Mixed and Augmented Reality.

[6] K. Pentenrieder, C. Bade, F. Doil, and P. Meier. Augmented reality-based factory planning -an application tailored to industrial needs. In ISMAR'07: Proc. 6th Int'l Symp. on Mixed and Augmented Reality, pp. 1–9, Nara, Japan, Nov. 13-16 2007. IEEE CS Press. ISBN 978-1-4244-1749-0.

7] Schall G., Mendez E., Kruijff E., Veas E., Junghanns S., Reitinger B., Schmalstieg D.: Handheld Augmented Reality for Underground Infrastructure Visualization. Personal and Ubiquitous Computing Journal. Volume 13 Issue 4, May 2009. Springer-Verlag London, UK.

[8] De Crescenzio F., Fantini M., Persiani F., Di Stefano L., Azzari P., and Salti S.: Augmented reality for aircraft maintenance training and operations support. IEEE Computer Graphics and Applications, 31(1):96–101, 2011.

[9] Wagner D., Langlotz T., and Schmalstieg D.: Robust and unobtrusive marker tracking on mobile phones. In Proceedings of 7th IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'08), Sept. 15–18 2008.

[10] Maidi M., Preda M., Van Hung Le: Markerless Tracking for Mobile Augmented Reality. Proceedings of International Conference on Signal and Image Processing Applications (ICSIPA2011), pp 301 - 306.

[11] Comport A.I., Marchand E., Pressigout M. and Chaumette F.: Real-Time Markerless Tracking for Augmented Reality: The Virtual Visual Servoing Framework. IEEE Trans. on Visualization and Computer Graphics 12, 4 (2006) 615-628.

[12] Fua P. and Lepetit V.: Vision Based 3D Tracking and Pose Estimation for Mixed Reality. Emerging Technologies of Augmented Reality: Interfaces and Design. IGI Global, 2007 (pp 1-22).

[13] Wei W., Yue Q., QingXing W.:An Augmented Reality Application Framework for Complex Equipment Collaborative Maintenance. Springer Berlin / Heidelberg 2011. ISBN: 978-3-642-23733-1 Volume 6874 pp. 154-16

[14] Wang R.Y., Popovi J.: Real-Time Hand-Tracking with a Color Glove. Published in Journal ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2009. Volume 28 Issue 3, August 2009 ACM New York, NY, USA.

[15] Mistry P., Maes P.: SixthSense – A Wearable Gestural Interface. In the Proceedings of SIGGRAPH Asia 2009, Sketch. Yokohama, Japan. 2009.

[16] Liu C., Huot S., Diehl J., Mackay W.E., Beaudouin-Lafon M.: Evaluating the Benefits of Real-time Feedback in Mobile Augmented Reality with Hand-held Devices. CHI'12 - 30th International Conference on Human Factors in Computing Systems - 2012 (2012)

[17] van Krevelen D.W.F.and Poelman R.: A Survey of Augmented Reality Technologies, Applications and Limitations. The International Journal of Virtual Reality, 2010, 9(2) pp 1-20.

[18] ARtoolkit online reference manual (http://www.hitl.washington.edu/artoolkit/documentation/)

[19] Enomoto A., Saito H.: Diminished Reality using Multiple Handheld Cameras. ACCV'07 Workshop on Multi-dimensional and Multi-view Image Processing, Tokyo, Nov., 2007.

[20] Jarusirisawad S., Hosokawa T., Saito H.: Diminished reality using plane-sweep algorithm with weakly-calibrated cameras. Progress in Informatics, No. 7, pp.11–20, (2010).

[21] Herling J., Broll W.: Advanced Self-contained Object Removal for Realizing Real-time Diminished Reality in Unconstrained Environments. In the Proceedings of 9th IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 2010.

[22] L. Bonanni, C.H. Lee, T. Selker, "Counter Intelligence: Augmented Reality Kitchen." Long paper in Extended Abstracts of Computer Human Interaction (CHI) 2005, Portland, OR.

[23] Y. Ayatsuka , J. Rekimoto, Active CyberCode: a directly controllable 2D code, CHI '06 extended abstracts on Human factors in computing systems, April 22-27, 2006, Montréal, Québec, Canada

[24] Y. Chuantao, B. David, R. Chalon, A contextual mobile learning system for mastering domestic and professional equipments, Proceedings of the IEEE International Symposium on IT in Medicine & Education, 2009. ITIME '09, Seoul, Korea, pp. 61 – 66, 2000

[25] J. Gausemeier, J. Fruend, C. Matysczok, B. Bruederlin , D. Beier, Development of a real time image based object recognition method for mobile AR-devices, Proceedings of the 2nd international conference on Computer graphics, virtual Reality, visualisation and interaction in Africa, February 03-05, 2003, Cape Town, South Africa

[26] Quest3D visual development software: http://quest3d.com/

[27] Furmanski, C., Azuma, R., & Daily, M.: Augmented-reality visualizations guided by cognition: Perceptual heuristics for combining visible and obscured information. In:

Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR 2002), pp. 215-320. IEEE (2002).

[28] Shah, M. M., Arshad, H., & Sulaiman, R.: Occlusion in augmented reality. In: Proceedings of the 8th International Conference on Information Science and Digital Content Technology (ICIDT 2012) pp. 372-378. IEEE. (2012).

[29] Lee, W., & Park, J.: Augmented foam: a tangible augmented reality for product design. In Mixed and Augmented Reality. In: Proceedings of the Fourth IEEE and ACM International Symposium on Mixed and Augmented Reality (2005), pp. 106-109. IEEE (2005).

[30] Walairacht, S., Yamada, K., Hasegawa, S., Koike, Y., & Sato, M.: 4+ 4 fingers manipulating virtual objects in mixed-reality environment. Presence: Teleoperators and Virtual Environments (2002), pp.134-143. MIT Press Journal (2002).

[31] Buchmann, V., Violich, S., Billinghurst, M., Cockburn, A.: FingARtips: Gesture Based Direct Manipulation in Augmented Reality. In: Proceedings of the 2nd International Conference on Computer Graphics and Interactive Techniques (GRAPHITE 2004), pp. 212-221. ACM (2004).

[32] Fischer, J., Bartz, D., & Straßer, W.: Occlusion handling for medical augmented reality using a volumetric phantom model. In Proceedings of the ACM symposium on Virtual reality software and technology (2004), pp. 174-177. ACM. (2004).

[33] Gordon, G., Billinghurst, M., Bell, M., Woodfill, J., Kowalik, B., Erendi, A., & Tilander, J.: The use of dense stereo range data in augmented reality. In: Proceedings of the 1st International Symposium on Mixed and Augmented Reality (2002), p. 14-23. IEEE Computer Society (2002).

[34] Seo, D. W., & Lee, J. Y.: Direct hand touchable interactions in augmented reality environments for natural and intuitive user experiences. Expert Systems with Applications (2013), Volume 40, Issue 9, pp. 3784–3793.

[35] Kanade, T., & Okutomi, M.: A stereo matching algorithm with an adaptive window: Theory and experiment. Pattern Analysis and Machine Intelligence, IEEE Transactions on, (1994), pp. 920-932.

[36] Medioni, G., & Nevatia, R.:. Segment-based stereo matching. Computer Vision, Graphics, and Image Processing, (1985), pp. 2-18.

[37] Yang Q., Wang L., Yang R., Wang S., Liao M., NisterD.: Real-time global stereo matching using hierarchical belief propagation, in: The British Machine Vision Conference, 2006, pp. 989–998.

[38] Humenberger, M., Zinner, C., Weber, M., Kubinger, W., & Vincze, M.: A fast stereo matching algorithm suitable for embedded real-time systems. Computer Vision and Image Understanding, 114(11), pp 1180-1202. (2010)

[39] Poupyrev I, Tan DS, Billinghurst M, Kato H, Regenbrecht H, Tetsutani N, (2002) Developing a generic augmented-reality interface, Computer, Volume: 35, Issue: 3, 2002, pp. 44-50

[40] Azuma R, Baillot Y, Behringer R, Feiner S, Julier S, MacIntyre B, (2001) Recent advances in augmented reality, IEEE Computer Graphics and Applications, Volume: 21, Issue: 6, 2001, pp.34-47

[41] Krapichler C, Haubner M, Lösch A, and Englmeier K, (1997) "Human-Machine Interface for Medical Image Analysis and Visualization in Virtual Environments", IEEE conference on Acoustics, Speech and Signal Processing, ICASSP-97. Vol 4, pp. 21-24.

[42] Kohler M and Schroter S. (1998). A Survey of Video-based Gesture Recognition - Stereo and Mono Systems. Technical Report 693, Informatik VII, University of Dortmund.

[43] Karam, M. (2006). A framework for research and design of gesture-based human-computer interactions. P.h.D. Thesis. University of Southampton. Available at http://eprints.soton.ac.uk/263149/ (July 2012)

[44] Jaimes A, Sebe N, (2007). Multimodal human-computer interaction: A survey. Computer Vision and Image Understanding 108, 2007, pp. 116-134

Kaltenborn K.-F., Rienhoff O. (1993) Virtual Reality in Medicine. Methods of information in medicine. Vol. 32, N 5, 1993, pp.407-417

[45] Van Dam A, (1997) Post-WIMP user interfaces. Communications of the ACM, Volume 40 , Issue 2, pp. 63-67

[46] Oviatt S, Darrell T, Flickner M, (2004) Multimodal Interfaces that Flex, Adapt and Persist. Communications of the ACM, Vol. 47 N. 1, pp. 30-33

[47] Oviatt S, (2003) Multimodal Interfaces, In: J.Jacko & A. Sears (eds). The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications. Lawrence Erlbaum: New Jersey

[48] Reeves LM, Lai J, Larson JA, Oviatt S, Balaji TS, Buisine S, Collings P, Cohen P, Kraal B, Martin JC, McTear M, Raman TV, Stanney M, Su H, Ying Wang Q, (2004) Guidelines for Multimodal User Interface Design. Communications of the ACM, vol. 47 no. 1, 2004, pp. 57-59

[49] Bolt R. (1980). "Put-that-there": Voice and gesture at the graphics interface. SIGGRAPH Comput. Graph. 14, 3 (July 1980), pp. 262-270.

[50] Graetzel C, Fong T, Grange S, and Baur C, (2004). A Non-Contact Mouse for Surgeon-Computer Interaction. Technology and Health Care, IOS Press, vol. 12, no. 3, pp. 245-257.

[51] Wachs J, Stern H, Edan Y, Gillam M, Feied C, Mark Smith, and Handler J, (2007) Gestix: A Doctor-Computer Sterile Gesture Interface for Dynamic Environments. A. Saad et al. (Eds.): Soft Computing in Industrial Applications, Springer, ASC 39, pp. 30–39.

[52] Wachs J, Stern H, Edan Y, Gillam M, Feied C, Mark Smith, and Handler J, (2006) "A Real-Time Hand Gesture Interface for Medical Visualization Applications". In: Tiwari A., Rajkumar R., Knowles J., Avineri, E. Dahal, K.(eds) *Applications of Soft Computing : Recent Trends*. Springer Verlag, Germany, Series: Advances in Soft Computing, Volume 36/2006, pp.153-162,

[53] Tani BS, Maia RS, Von Wangenheim A, (2007). A Gesture Interface for Radiological Workstations. Proceedings of the Twentieth IEEE International Symposium on Computer-Based Medical Systems.

[54] Duke DJ, (1995). Reasoning About Gestural Interaction, ACM/Eurographics 95, Volume 14, Number 3, pp. 55-66

[55] Dix A, Finlay J, Abowd G, Beale R, (2004). Human-Computer Interaction. Third Edition. Prentice Hall

[56] Stern H, Wachs J, Edan Y, (2006). "Optimal Hand Gesture Vocabulary Design Using Psycho-Physiological and Technical Factors," 7th International Conference on Automatic Face and Gesture Recognition, FG2006.

[57] Alur R and Dill DL, (1994). A theory of timed automata. Journal of Theoretical Computer Science, 126(2):183–235.

[58] Alur R, Courcoubetis C and Dill DL, (1990). Model checking for real-time systems. In Proceedings of the 5th Annual Symposium on Logic in Computer Science, IEEE Computer Society Press, New York, 414-425.

[59] Smith, R. (2000) "ODE: Open Dynamics Engine", (www.ode.org).

[60] Kölsch M and Turk M. (2002) Keyboards without Keyboards: A Survey of Virtual Keyboards. In Proceedings of Workshop on Sensing and Input for Media-centric Systems, 2002

[61] International Organisation for Standardisation (1998) ISO 9241: Software Ergonomics Requirements for office work with visual display terminal (VDT), Geneva, Switzerland

[62] Nielsen J (1993) Usability Engineering. Academic Press, Cambridge

[63] Sutcliffe AG, Deol Kaur K, (2000). Evaluating the usability of virtual reality user interfaces, Behaviour and Information Technology, Volume 19, Number 6, 1 November 2000 , pp. 415-426 [64] Green TRG, (1989) Cognitive dimensions of notations. In Proc. HCI 89. , pp . 443 – 460, Cambridge University Press , Cambridge

[65] Green,TRG and Petre M, (1996). Usability Analysis of Visual Programming Environments : A 'Cognitive Dimensions' Framework. Journal of Visual Languages and Computing, 7, 1996, 131 – 174

[66] Bottoni P., De Marsico M., Levialdi S., Ottieri G., Pierro M., and Quaresima D. (2009). A Dynamic Environment for Video Surveillance. In: T. Gross et al. (Eds.): Proceedings INTERACT 2009, Part II, LNCS 5727, pp. 892–895.

[67] Green, T. R. G. (1991) Describing information artifacts with cognitive dimensions and structure maps. In D. Diaper and N. V. Hammond (Eds.) Proceedings of "HCI'91: Usability Now", Annual Conference of BCS Human-Computer Interaction Group. Cambridge University Press.

[68] Aarts, E. and Marzano, S. (2003). The New Everyday: Visions of Ambient Intelligence, 010 Publishing, Rotterdam, The Netherlands.

[69] Maltoni, D.; Maio D., Jain A.K. & Prabhakar S. (2003). Handbook of Fingerprint Recognition, Springer, New York.

[70] Perronnin, G. and Dugelay, J.L. (2003). An Introduction to biometrics and face recognition, Proceedings of IMAGE 2003: Learning, Understanding, Information Retrieval, Medical, Cagliari, Italy, June 2003.

[71] Bowyer, K.W.; Chang, K. & Flynn P.A. (2004). Survey of 3D and Multi-Modal 3D+2D Face Recognition, Proceeding of International Conference on Pattern Recognition, ICPR, 2004

[72] Jafri, R. and Arabnia, H. R. (2009). A Survey of Face Recognition Techniques. Journal of Information Processing Systems. 5, 2, 41--68.

[73] Hesher, C.; Srivastava, A. and Erlebacher, G. (2002). "Principal component analysis of range images for facial recognition" in Proc. International Conference on Imaging Science, Systems, and Technology, Las Vegas, NV, June 2002.

[74] Pan, G.; Han, S.; Wu, Z. & Wang, Y. (2005). 3D face recognition using mapped depth images, Proceedings of IEEE Workshop on Face Recognition Grand Challenge Experiments, June 2005.

[75] Tsalakanidou,F.; Tzovaras D. and Strintzis, M. (2003). "Use of depth and colour eigenfaces for face recognition" Pattern Recognition Letters, vol. 24, no. 9-10, pp. 1427–1435, June 2003.

[76] Beumier, C. and Acheroy, M. (2000). Automatic Face verification from 3D and grey level cues, Proceeding of 11th Portuguese Conference on Pattern Recognition (RECPAD 2000), May 2000, Porto, Portugal.

[77] Wang, Y. and Chua, C.-S. (2005). "Face recognition from 2d and 3d images using 3d gabor filters" Image and Vision Computing, vol. 11, no. 23, pp. 1018–1028, October 2005.

[78] Bronstein, A.M.; Bronstein, M.M. and Kimmel, R. (2005). Three dimensional face recognition International Journal of Computer Vision, vol. 64, no. 1, pp. 5–30, August 2005.

[79] Bronstein, A. M.; Bronstein, M. M. and Kimmel, R. (2006). Robust expressioninvariant face recognition from partially missing data in Proc. European Conference on Computer Vision, Gratz, Austria, May 2006, pp. 396–408.

[80] Amberg, B.; Knothe, R. and Vetter, T. (2008). Shrec'08 entry: Shape based face recognition with a morphable model in Proc. IEEE International Conference on Shape Modeling and Applications, Stoney Brook, NY, June 2008, pp. 253–254.

[81] Xu, C.; Wang, Y.; Tan, t. & Quan, L. (2004). Automatic 3D face recognition combining global geometric features with local shape variation information, Proceedings of Sixth International Conference on Automated Face and Gesture Recognition, May 2004, pp. 308–313.

[82] Xu, D.; Hu, P.; Cao, W. and Li, H. (2008). Shrec'08 entry: 3d face recognition using moment invariants in Proc. IEEE International Conference on Shape Modeling and Applications, Stoney Brook, NY, June 2008, pp. 261–262.

[83] Medioni,G. and Waupotitsch R. (2003). Face recognition and modeling in 3D, Prooceding of IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG 2003), pages 232-233, October 2003.

[84] Chang, K.I.; Bowyer, K.W. & Flynn, P.J. (2005). Adaptive rigid multi-region selection for handling expression variation in 3D face recognition, Proceedings of IEEE Workshop on Face Recognition Grand Challenge Experiments, June 2005.

[85] Faltemier, T. C.; Bowyer, K. W. and Flynn, P. J. (2008). A region ensemble for 3d face recognition, IEEE Transactions on Information Forensics and Security, vol. 3, no. 1, pp. 62–73, March 2008.

[86] Faltemier, T.C. Bowyer, K.W. ; Flynn, P.J. "Using multi-instance enrollment to improve performance of 3d face recognition," Computer Vision and Image Understanding, vol. 112, no. 2, pp. 114–125, November 2008.

[87] Berretti, S.; Del Bimbo, A.; Pala, P. (2010). 3D Face Recognition using iso-Geodesic Stripes, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 32, Issue 12, December 2010, pages 2162-2177.

[88] Mian, A. S.; Bennamoun, M. and Owens, R. (2007). An efficient multimodal 2d-3d hybrid approach to automatic face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 11, pp. 1927–1943, November 2007.

[89] Phillips, P. J., Scruggs, T., O'Toole, A. J., Flynn, P. J., Bowyer, K. W., Schott, C. L. and Sharpe M. (2007). FRVT 2006 and ICE 2006 large-scale results. National Institute of Standards and Technology. Gaithersburg, MD.

[90] Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., and Worek, W. (2006). Preliminary face recognition grand challenge results, in Proc. *International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, April 2006, (pp. 15–24).

[91] Belhumeur, P., Hespanda, J., Kriegman, D. (1997). Eigenfaces versus fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Machine Intell.* 19 (7), 711–720.

[92] Etemad, K., Chellappa, R., (1997). Discriminant analysis for recognition of human face images. *J. Optical Soc. Amer.* 14, 1724–1733.

[93] Wiskott, L., Fellous, J.-M., Krger, N., Malsburg, C.V.D. (1997). Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Anal. Machine Intell*. 19, 775–779.

[94] Moghaddam, B. (2000). Bayesian face recognition. *Pattern Recognit.* 33, 1771–1782.

[95] Bartlett, M., Movellan, J., Sejnowski, T. (2002). Face recognition by independent component analysis. *IEEE Trans. Neural Networks* 13 (6), 1450–1464.

[96] Wright, J., Ganesh, A., Yang, A., Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Machine Intell*. 31, 210–227.

[97] P.J. Phillips, H. Wechsler, J. Huang, P.J. Rauss, The FERET database and evaluation procedure for face-recognition algorithms, Image and Vision Computing, 16 (5) (1998), 295-306.

[98] Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J. (2001). From few to many: Ilumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Machine Intell.* 23 (6), 643–660.

[99] Chellappa, R., Ni, J., Patel, V. M. (2012). Remote identification of faces: Problems, prospects, and progress, *Pattern Recognition Letters,* 33, (pp. 1849–1859)

[100] Gunturk, B., Batur, A., Altunbasak, Y., Hayes, M.H.I., Mersereau, R. (2003). Eigenfacedomain super-resolution for face recognition. *IEEE Trans Image Process*. 12 (5), 597–606.

[101] Jia, K., Gong, S. (2005). Multi-modal tensor face for simultaneous super-resolution and recognition. In: Tenth *IEEE Internat. Conf. on Computer Vision ICCV*, vol. 2, (pp. 1683–1690).

[102] Hennings-Yeomans, P., Baker, S., Kumar, B. (2008). Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, CVPR 2008, (pp. 1–8).

[103] Li, B., Chang, H., Shan, S., Chen, X. (2010). Low-resolution face recognition via coupled locality preserving mappings. *IEEE Signal Process. Lett.* 17 (1), 20–23.

[104] Shekhar, S., Patel, V.M., Chellappa, R. (2011). Synthesis-Based Low Resolution Face Recognition. *International Joint Conference on Biometrics*, Washington, DC, (pp 1-6).

[106] Tistarelli, M., Li, S.Z., Chellappa, R. (2009). *Handbook of Remote Biometrics: for Surveillance and Security*, 1st ed. Springer Publishing Company Inc.

[107] Blanz, V., Vetter, T. (2003). Face recognition based on fitting a 3D morphable model. *IEEE Trans Pattern Anal. Machine Intell.* 25, 1063–1074.

[108] Prince, S., Warrell, J., Elder, J., Felisberti, F. (2008). Tied factor analysis for face recognition across large pose differences. *IEEE Trans. Pattern Anal. Machine Intell.* 30 (6), 970–984.

[109] Castillo, C., Jacobs, D. (2009). Using stereo matching with general epipolar geometry for 2D face recognition across pose. *IEEE Trans. Pattern Anal. Machine Intell*. 31 (12), 2298–2304.

[110] Biswas, S., Chellappa, R. (2010). Pose-robust albedo estimation from a single image. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, (pp. 2683–2690).

[111] Basri, R., Jacobs, D.W. (2003). Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Anal. Machine Intell.* 25 (2), 218–233.

[112] Wang, H., Li, S.Z., Wang, Y. (2004). Generalized quotient image. In *Proc. Internat. Conf. Computer Vision and Pattern Recognition*, (pp. 498-505) .

[113] Chen, T., Yin, W., Zhou, X.S., Comaniciu, D., Huang, T.S. (2006). Total variation models for variable lighting face recognition. *IEEE Trans. Pattern Anal. Machine Intell*. 28 (9), 1519–1524.

[114] Zhou, S.K., Aggarwal, G., Chellappa, R., Jacobs, D.W. (2007). Appearance characterization of linear Lambertian objects, generalized photometric stereo, and illumination-invariant face recognition. IEEE Trans. Pattern Anal. Machine Intell. 29 (2), 230–245.

[115] Biswas, S., Aggarwal, G., Chellappa, R. (2009). Robust estimation of albedo for illumination-invariant matching and shape recovery. *IEEE Trans. Pattern Anal. Machine Intell.* 29 (2), 884–899.

[116] Patel, V.M., Wu, T., Biswas, S., Phillips, P.J., Chellappa, R. (2011). Illumination robust dictionary-based face recognition. In: Proc. *Internat. Conf. on Image Processing*.

117] Bronstein, A. M., Bronstein, M. M., Kimmel, R. (2006). Robust expressioninvariant face recognition from partially missing data, in Proc. *European Conference on Computer Vision,* Gratz, Austria, May, (pp. 396–408).

[118] Berretti, S., Del Bimbo, A., Pala, P. (2010). 3D Face Recognition Using Isogeodesic Stripes, *IEEE Trans. on Pattern Analysis and Machine Intelligence* , 32 (12), 2162 – 2177

[119] Candès, E.J., Li, X., Ma, Y., Wright, J. (2011). Robust principal component analysis, *Journal of the ACM 5*8 (3).

[120] Enciso, R.; Li, J.; Fidaleo, D.A.; Kim, T-Y; Noh, J-Y & Neumann, U. (1999). Synthesis of 3D Faces, Proceeding of International Workshop on Digital and Computational Video, DCV'99, December 1999

[121] Basten, T. & Geilen, M. (2003). Ambient Intelligence: Impact on Embedded System Design, H. de Groot (Eds.), Kluwer Academic Pub., 2003.

[122] Acampora, G.; Loia, V.; Nappi, M. & Ricciardi, S. (2005). Human-Based Models for Smart Devices in Ambient Intelligence, Proceedings of the IEEE International Symposium on Industrial Electronics. ISIE 2005. pp. 107- 112, June 20-23, 2005.

[123] Blanz, V. & Vetter, T. (1999). A morphable model for the synthesis of 3D faces, Proceedings of SIGGRAPH 99, Los Angeles, CA, ACM, pp. 187-194, Aug. 1999

[124] Lee, Y.; D. Terzopoulos, D. & Waters, K. (1995). Realistic modeling for facial animation, Proceedings of SIGGRAPH 95, Los Angeles, CA, ACM, pp. 55-62, Aug. 1995.

[125] Moon H. and Phillips, P.J. (2001). Computational and Performance Aspects of PCA-Based Face-Recognition Algorithms, Perception, vol. 30, pp. 303-321, 2001.

[126] Martinez, A. M. and Kak, A. C. (2001). PCA versus LDA, IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2), 228-233.