

Computer Vision and Human-Computer  
Interaction: artificial vision techniques and use  
cases with creating interfaces and interaction  
models

Marco Comite

The 22th of April 2013



**Introduction** Here is described how Computer Vision could give improvements to Human-Computer Interaction. Starting from a brief description of computers and human beings, follows a description of computer interfaces and Computer Vision. Then there's a description of how a human being can be recognized by Computer Vision and follow some cases in which these techniques are applied. The last part is about describing Computer Vision for Human-Computer Interaction models based on the use cases.

Keywords: Computer Vision, Human-Computer Interaction, use cases, interaction models.



# Contents

<b>1 Computer, the Technology</b>	<b>9</b>
1.1 Introduction . . . . .	9
1.2 Architecture . . . . .	9
1.2.1 Mainframe . . . . .	10
1.2.2 Minicomputer, Microcomputer and Personal Computer . . . . .	10
1.2.3 Portable device . . . . .	10
1.2.4 Smart Thing . . . . .	11
1.3 Computer Interfaces . . . . .	11
1.3.1 Introduction . . . . .	11
1.4 Computer Vision . . . . .	12
1.4.1 Introduction . . . . .	12
1.4.2 Camera . . . . .	13
<b>2 Human, the Target</b>	<b>17</b>
2.1 Introduction . . . . .	17
2.2 Morphology . . . . .	17
2.3 Functional Morphology . . . . .	18
2.4 Recognition . . . . .	18
2.4.1 Introduction . . . . .	18
2.4.2 Head and Face . . . . .	18
2.4.3 Eyes . . . . .	19
2.4.4 Facial Expressions and Emotion Recognition . . . . .	23
2.4.5 Hand and Hand Gesture . . . . .	25
2.4.6 Body and Body Gesture . . . . .	26

<b>3</b>	<b>Human-Computer Interaction</b>	<b>33</b>
3.1	Introduction . . . . .	33
3.2	Computer Vision for Human-Computer Interaction . . . . .	33
<b>4</b>	<b>Use Cases</b>	<b>35</b>
4.1	Introduction . . . . .	35
4.2	People with Disabilities . . . . .	35
4.2.1	Introduction . . . . .	35
4.2.2	Blind and Visually Impaired People . . . . .	35
4.2.3	Deaf and Hearing Impaired People - Sign Language . . . . .	36
4.2.4	Autism Spectrum Disorder . . . . .	39
4.3	Entertainment . . . . .	40
4.3.1	Media Player . . . . .	40
4.3.2	Videogames . . . . .	40
4.4	Shopping . . . . .	45
4.5	Office . . . . .	48
4.6	Videoconferencing . . . . .	49
4.7	Virtual Input Devices . . . . .	52
4.8	Object-Computer Interaction . . . . .	53
4.9	Remote Control . . . . .	56
4.10	Wearable Visual Interface . . . . .	60
<b>5</b>	<b>Interaction Models</b>	<b>65</b>
5.1	Introduction . . . . .	65
5.2	Components . . . . .	65
5.2.1	Person . . . . .	66

5.2.2	Camera . . . . .	66
5.2.3	Device . . . . .	66
5.2.4	Devices . . . . .	67
5.2.5	Screen or Output . . . . .	67
5.2.6	All components . . . . .	68
5.3	Workflows . . . . .	68
5.3.1	Generic Information Flow . . . . .	68
5.3.2	Simplest Information Flow . . . . .	69
5.3.3	Object-Compute-Output . . . . .	69
5.3.4	Person-Compute-Person . . . . .	70
5.3.5	Full-Duplex Person-Compute-Person . . . . .	70
5.3.6	Acquire-Elaborate-Show . . . . .	70
5.3.7	Acquire-Elaborate-Project-Show . . . . .	72
5.4	Interaction Models Properties . . . . .	74
5.5	Interaction Models Properties Values . . . . .	74
<b>6</b>	<b>Conclusions</b>	<b>77</b>
<b>7</b>	<b>Acknowledgements (Italian)</b>	<b>79</b>





# 1 Computer, the Technology

## 1.1 Introduction

Nowadays the word “computer” means a lot of things. During the time has changed many shapes but the main purpose is always the same: help man to do things better and faster. Started as a machine to make calculation, passed through daily work tool to an assistant in our pocket, being useful, helpful and funny! When we say computer, it could be referred to a personal computer, to a mainframe, to a smartphone, ... to many things some of us have never seen and to others we are all used to live with. Next paragraphs introduce these different types of machines.

## 1.2 Architecture

Basically a computer is a device that can be programmed to carry out an output after have computed an input with some rules. Conventionally is composed by a processor, a volatile memory, a storage memory, input and output peripheral devices. The processor is the demanded part for doing calculation and for controlling the parts of the computer. The volatile memory is the temporary memory the processor uses to help itself with calculation or for doing the demanded job. The storage memory is a base for data to be maintained, for example hosting the input and the output of the calculation or some data to be simply stored for later use, also after have switched off and on the computer. Peripheral devices are everything is connected to the computer being helpful with increasing the productivity of computer’s jobs or the human interaction with it. During last decades computers have increased their speed and reduced

their dimensions. Let's see these steps.

### **1.2.1 Mainframe**

The mainframe maybe is the less indicated type of device to use to talk about Human-Computer Interaction, because of it's purpose. But it's useful to understand that more is a device portable, higer is the possibility of using an interaction via Computer Vision. But mainframe can be yet useful for Human-Computer Interaction thinking about the all new cloud-oriented services which are being developed, by which Computer Vision aided devices are coming out, for example the all new Google Glass, a device for Augmented Reality that uses a camera and a connection to give the user a better experience retrieving information from the Internet directly on a mini screen in front of an eye.

### **1.2.2 Minicomputer, Microcomputer and Personal Computer**

Before the explosion of smartphones market, PCs were very widespread. Latest models came up with frontal camera which aided the diffusion of testing Computer Vision Human-Computer Interaction, also thanks to the continuous reduction in price, but nowadays is not so diffuse Computer Vision as a service for Interaction with PCs.

### **1.2.3 Portable device**

Lately smartphones came up, small devices with powerful processors and cameras mounted on them. It seems this will be an aid for interactive interfaces with Computer Vision, for example is coming out a smartphone of Samsung wich can recognize if your gaze is on the screen, if isn't the screen is turned

off or the playing video is paused. When the gaze came back to the screen, all brings life again.

#### **1.2.4 Smart Thing**

With the term Smart Thing is intended all devices smaller than a PC but with almost-like features of a PC, for example dedicated devices we can find in our cars, our houses, ... specific devices for specific tasks, but intelligent enough to have the capability of communicate with us in *intelligent* ways.

### **1.3 Computer Interfaces**

#### **1.3.1 Introduction**

More the computer is smaller, more its presence in men's life is stronger. Reducing dimensions, first used types of interfaces are no more comfortable. Computers are objects that need someone or something that tells them what to do and when do it. To tell a computer what to do you have to give it a command and to do this you need to interact with it in some ways. With consumer market personal computers, since ever, the most common way to interact with is through a keyboard, a set of buttons that send to the computer some kind of electric impulses that are recognized by the demanded interface as specific commands: you push a letter button and you'll see its value in a word processing program. Analyzing further more what we use commonly we can also look at the mouse: a little device that stay in your hand and moves a little arrow on the screen. And what about the computers' returning information? The interface a common computer uses to communicate with us is the screen: a set of little lights that together can produce images and texts we can recognize and

work with. Looking back at the beginning of computer history we can see that input devices were a sort of holed papers that the computer was able to read and understand as commands. The output was also on a paper support, the computation result was printed on some sheets of paper so the human could read and understand it, because there were no keyboards and no screens as intended today. Coming back to our days, maintaining keyboard and mouse while decreasing the device's dimensions could be a problem. Thanks to the parallel increasing of the speed of the processor and other components, we can exploit other types of interface. Last years the most used new type of interface seems to be touch screen. Nowadays also cameras have made many improvements and exploiting processor speed is it possible to develop interesting Computer Vision-aided interfaces thanks to the fact that any computer or smartphone is configured with it. Are we going to assist to a new revolution?

## **1.4 Computer Vision**

### **1.4.1 Introduction**

Computer Vision is the transformation of data from a still or video camera into a decision or a new representation. For example the detection of a person in a scene or turning a color image into a grayscale one. Computer Vision is giving vision capability to a computer in order to achieve a sort of comprehension about the objects being part of the images. For doing this the image is converted into numbers the computer can understand to apply the needed algorithms. Later the numbers are converted again into images to be shown or simply when the algorithm terminates fires some other command.

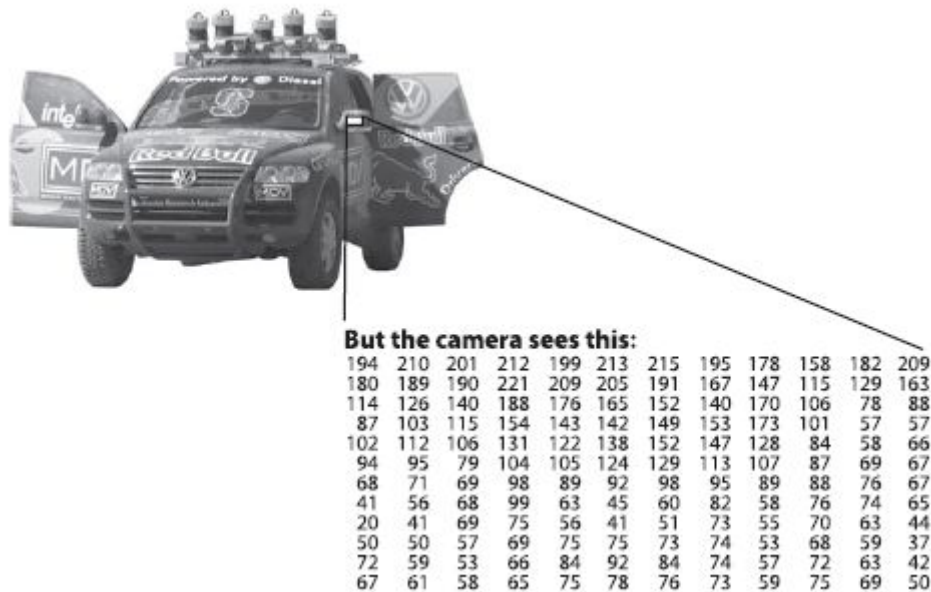


Figure 1: From image to numbers [11]

### 1.4.2 Camera

Vision begins with the detection of light from the world. Light from the sun or a light bulb travels through space until striking an object. Some of that light is absorbed, some is reflected and some of this one make its way to our eyes (or our camera) is collected on our retina (or our imager). The geometry of the ray's travel from the object to the retina, or imager, could be represented by a simple model. This model is the pinhole camera model. A pinhole is an imaginary wall with a little hole which blocks all the rays except ones that pass through the hole. In a physical pinhole camera this point is projected onto an image surface. As a result, the image on this plane is always in focus and the size of the image relative to the object is given by a single parameter of the camera: its focal length. In Figure 2  $f$  is the focal length of the camera,  $Z$  is the distance from the camera to the object,  $X$  is the length of the object and  $x$

is the object's image on the imaging plane.

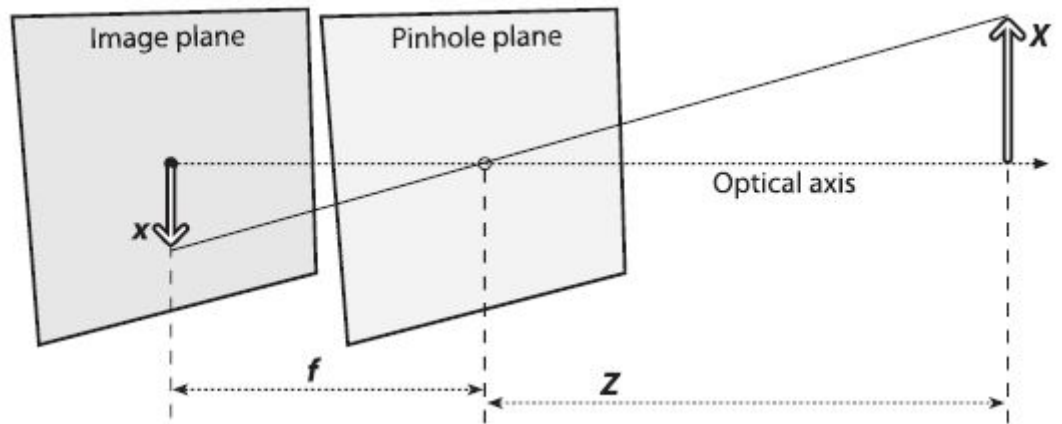


Figure 2: Pinhole camera model [11]

Now we can rearrange the above figure obtaining a more simple to manage camera model, see Figure 3.

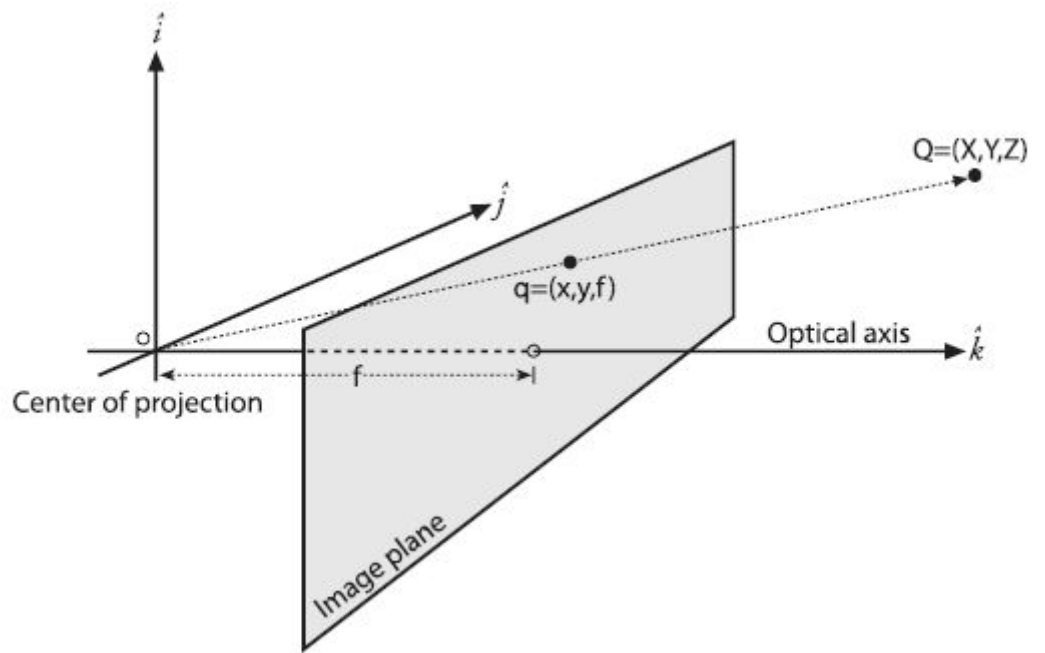


Figure 3: Rearranged pinhole camera model [11]





## **2 Human, the Target**

### **2.1 Introduction**

Most purposes of things men make are human oriented. Computers are an example, they were made to (protect when military first invented and) make men life better. To better understand how to make computers and humans communication and make it easy for the best interaction, we need to understand how men are made, obviously, by the other side, we are limited by technology in taking advantage of our discovering. Imagine what could happen if we could control every device with our mind. While awaiting for that type of technology being discovered, we can study at our best what we know to take advantage of the technology we already have to make our life better using computers.

### **2.2 Morphology**

The best way to explore the subject of this work, the human, is to study it from a mechanical way without forget how complex the human body is. Looking at the human body entirely we can see that there are from person to person some common parts and other different, like the skin color and the lenght and shape of some parts. But many parts are similar, for example starting from the top we have a head, a trunk, two arms and two legs. The head have a round shape containing two eyebrows, two eyes, a nose, a mouth and a chin. Every one of these have a lot of details, that can change from person to person from color to shape to relative positions. But it's possible to find some boundaries in these differences. Going down we have the trunk. Then the arms, their parts. The hands and their components. There are a lot of differences from

one person to another, but basically a hand have five fingers, divided in more subparts. The same the legs, are composed of many parts. Human body is a set of connected parts, every human body is similar to others' except for some bounded differences.

## **2.3 Functional Morphology**

The human body is not static, it's dynamic. The human body has a lot of moving parts, from the head on the trunk to the fingers on the hand to the muscles on the face. We have moving fingers on moving hands on moving arms on moving trunk on a moving pair of legs. The human body could not be so easy to be tracked.

## **2.4 Recognition**

### **2.4.1 Introduction**

Our target is to have the possibility to give commands to a computer using what it can see with one or more cameras and have a feedback correctly related to them. The computer must be able to recognize objects and behaviours from images and videos it has at its input. At first we are going to study single human's body parts viewed as static objects, then seen as moving objects and then we try to interpret human behaviours to have the possibility to receive a coherent feedback from the computer.

### **2.4.2 Head and Face**

The head and face recognition has been studied a lot due to its many applications. Face recognition is mainly used to give the computer the capability of

recognizing the person standing in front of the camera. Face recognition suffers of the same issues of other recognitions: the pose, the appearance, age, lighting and expression [15]. Face recognition starts with face detection and if it's on a video face tracking is also necessary. Most systems use a of skin-tone and face texture to determine the location of a face and use an image pyramid to allow faces of varying sizes to be detected. Typically translation, scale and in-plane rotation for the face are estimated simultaneously, along with rotation-in-depth when this is considered. There are many ways to recognize a face and the choose is due to the application domain. In [18] are analyzed two ways for recognizing faces: the first one based on the computation of a set of geometrical features, the second one based on almost-grey-level template matching. Both are good methods but based on their experiments made using a database of images and letting a computer processing, they discovered that the best method, with their application is the one with templates giving perfect recognition while a result of 90 percent of correctness with the other. Nowadays to have a better recognition that can count on the use of both methods. Without see the details, for recognizing faces could be used neural networks, elastic template matching, Karhunen-Loeve expansion and many other techniques which are intended of determining principal parts of the face. [18].

### **2.4.3 Eyes**

The eye is another thing can be used as input for giving commands to a computer. Also to be more near to create computers capable of understand emotions, being ergonomics. For this purpose gaze is very important and is necessary to track the pupil. In [6] three algorithms used for eye pupil location are

described and tested. The basic use of tracked pupil is shown in Figure 4

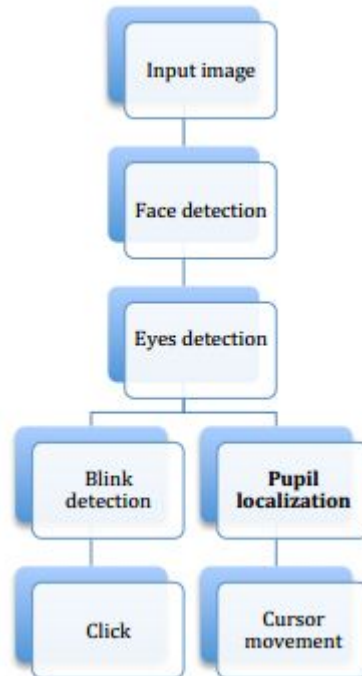


Figure 4: Basic use of eye tracking

The first algorithm is based on Cumulative Distribution Function (CDF), the idea is to filter the image obtaining only the pupil, see Figure 5.



Figure 5: Cumulative Distribution Function

Once the pupil is found it's possible to track it, see Figure 6

The second algorithm is Project Functions (PF). The idea of the method is similar to the one used in CDF algorithm, but in this case pixel intensities are

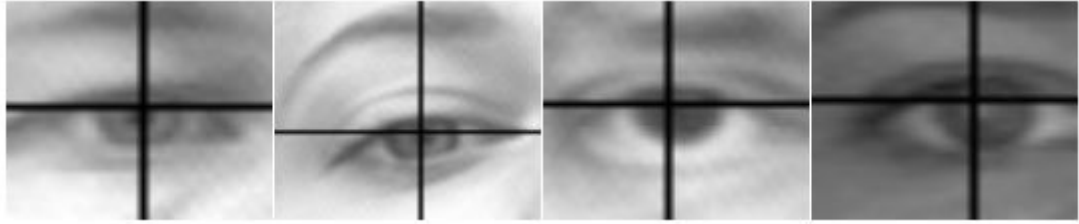


Figure 6: Cumulative Distribution Function result

projected on vertical and horizontal axes. Those projections divide the whole picture to homogenous subsets, see Figure 7.

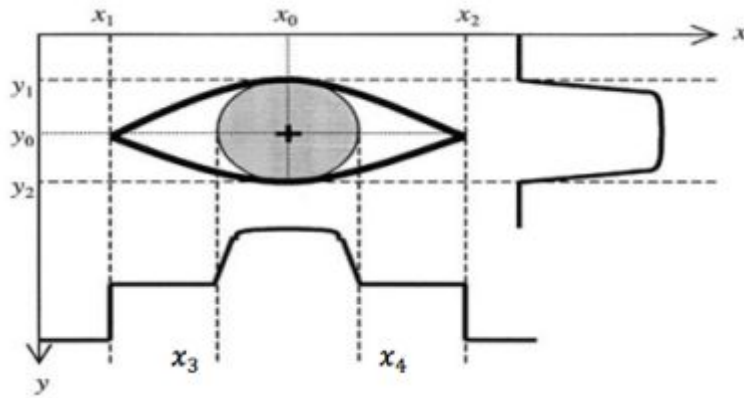


Figure 7: Project Functions

Follows the plot of vertical (A) and horizontal (B) General Projection Function (black) and its derivative (white) over a grey scale picture acquired with webcam, see Figure 8.

Later edges of iris are found, see Figure 9.

And the pupil is detected, Figure 10.

The third algorithm is Edge Analysis (EA). The target is make show up the borders of the eye as shown in Figure 11.

Later the eye is found at the cross of two strips as shown in Figure 12.

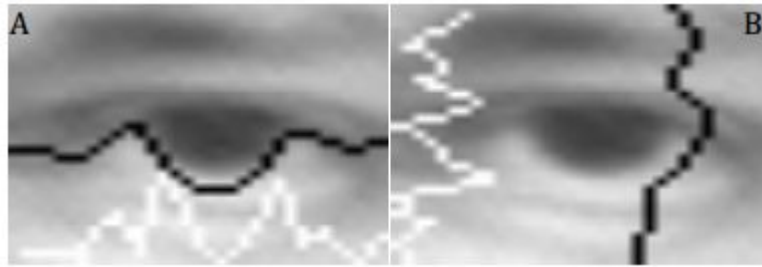


Figure 8: Project Functions



Figure 9: Project Functions

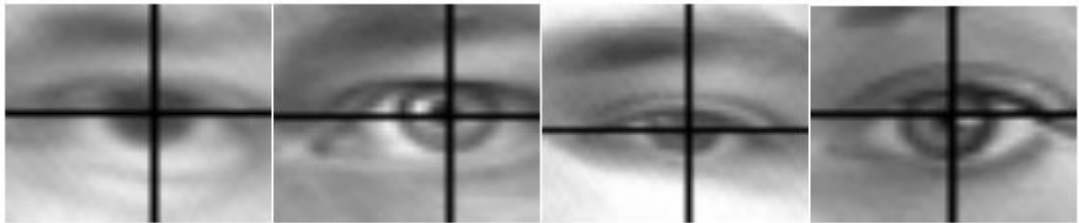


Figure 10: Project Functions



Figure 11: Edge Analysis

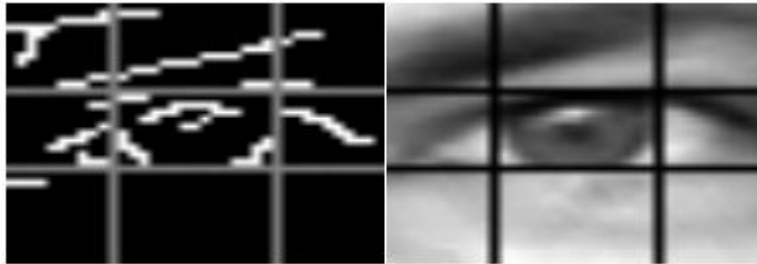


Figure 12: Edge Analysis

#### 2.4.4 Facial Expressions and Emotion Recognition

Once the face is recognized, it's important to reach another level of understanding of the human being: recognize emotions. First we have to categorize emotions as facial expressions, as in Figure 13



Figure 13: The emotion wheel, Plutchik, 1980

And we can see the differences of expressions in Figure 14.

The principal approaches in recognizing expressions are: target-oriented, the

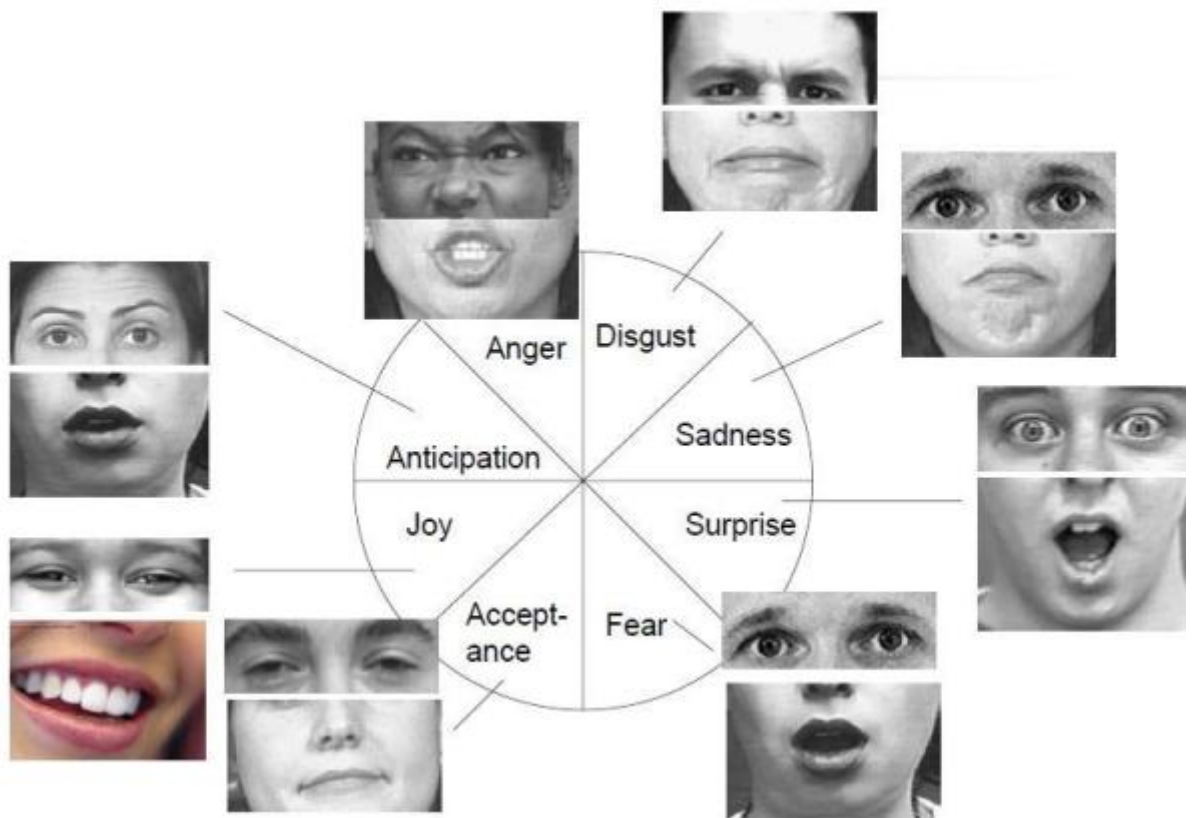


Figure 14: Expressions



use of a single image of a face at apex of expression, gesture-oriented, extract facial temporal information from a sequence of images, and transitional, use two images representing a face in its neutral condition and at apex of the expression, see Figure 15.



Figure 15: Expressions

#### 2.4.5 Hand and Hand Gesture

Although hand tracking algorithm has been widely used in virtual reality and HCI system, it is still a challenging problem in vision-based research area [8]. In [1] are presented main methods to recognize hand gesture. The first one is Appearance Based Approach which is a system that makes use of finger tips to reach the target of recognizing the hand. Another way is using the Model Based Approach which uses statistics information about skin color, and model base system for hand gesture recognition where the joints in fingers had one degree of freedom (DOF), effective joints had 2 DOF and spherical joints had 3 DOF. So fingers had 4 DOF while thumb had 5. Then is defined local coordinate systems with the origins on the joints (Figure 16). This system was interdependent on fingers movement. It's used fast fitting method and find the angles of each

fingers.

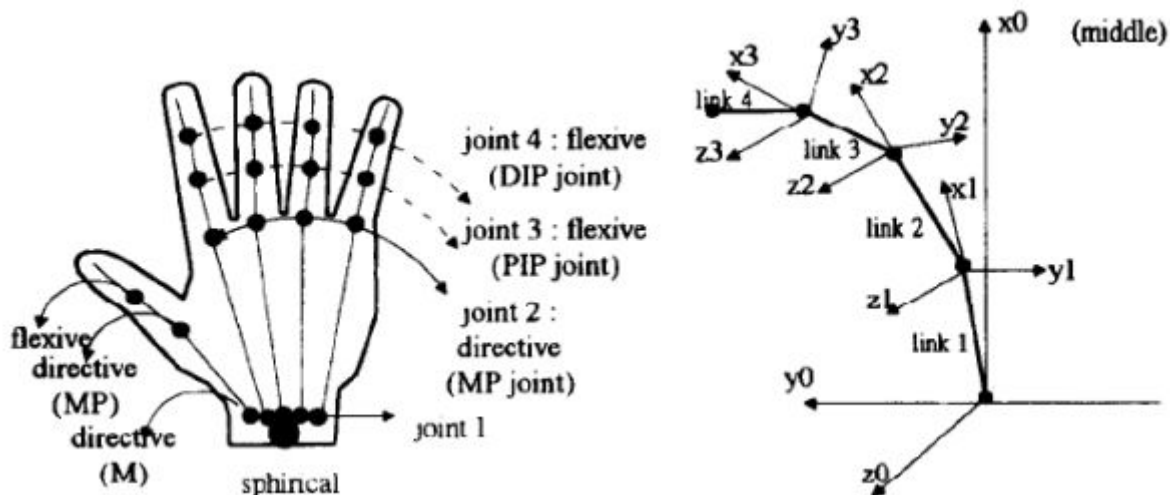


Figure 16: Model Based Approach

Other ways are using Artificial Neural Networks, Fuzzy Logic and Genetic Algorithms. The first approach consists of creating a neural network, that expands into the image of the hand and detect the raised fingers, see Figure 17.

The idea under the use of Fuzzy Logic is to develop a system which uses imprecise data. First a set of clusters is use on data, later a Finite State Machine recognizes the gesture. Genetic algorithms have also been used to recognize the hand, specifically for noise-removal, later Dynamic Bayesian Network is used for gesture segmentation and recognition, see Figure 18. The best point of genetic algorithms is that it work parallel on different points for faster computation.

#### 2.4.6 Body and Body Gesture

For recognizing body and body gesture, we can find a lot of methods. For example the first is the silhouette based feature extraction. First of all the silhouette is extracted, then some features are computed and later is possible

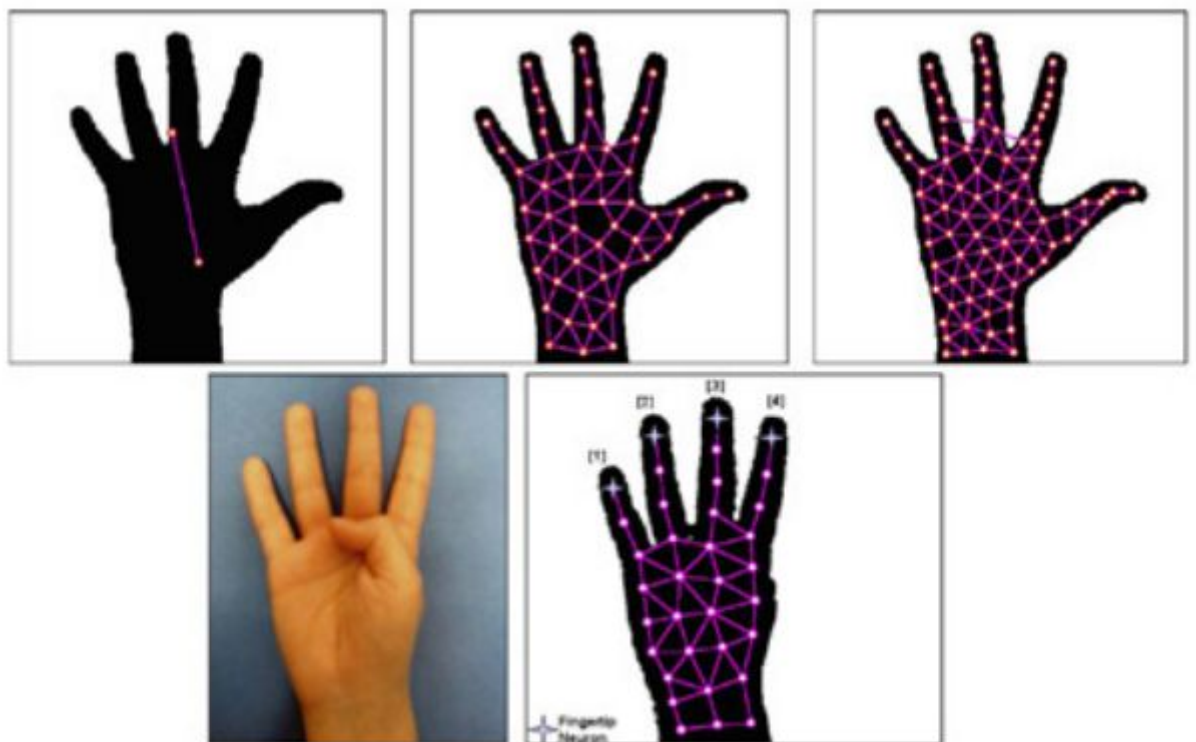


Figure 17: Artificial Neural Network

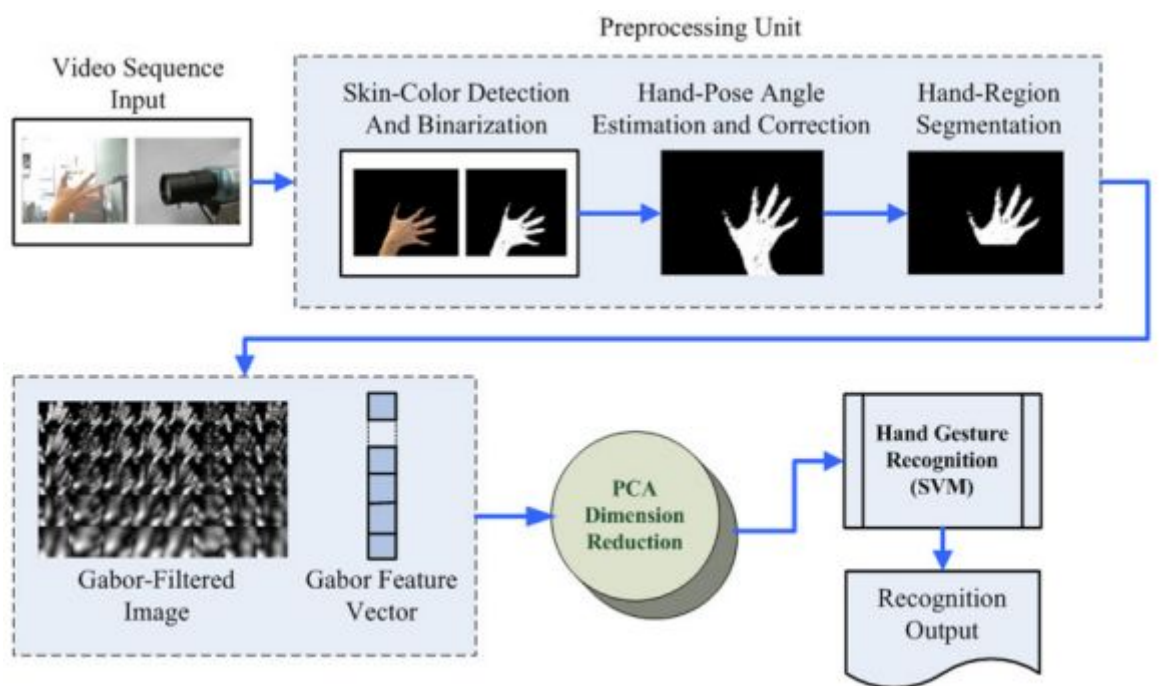


Figure 18: Genetic Algorithm

to detect the position of main parts of the body, see Figure 19.



Figure 19: Body recognition

The first step to recognize a body is to separate it from the background, can be used techniques as thresholding or statistical techniques, see Figure 20.



Figure 20: Background separation

To recognize is it possible base the search on colors, Figure 21 or tracking subregions, Figure 22.

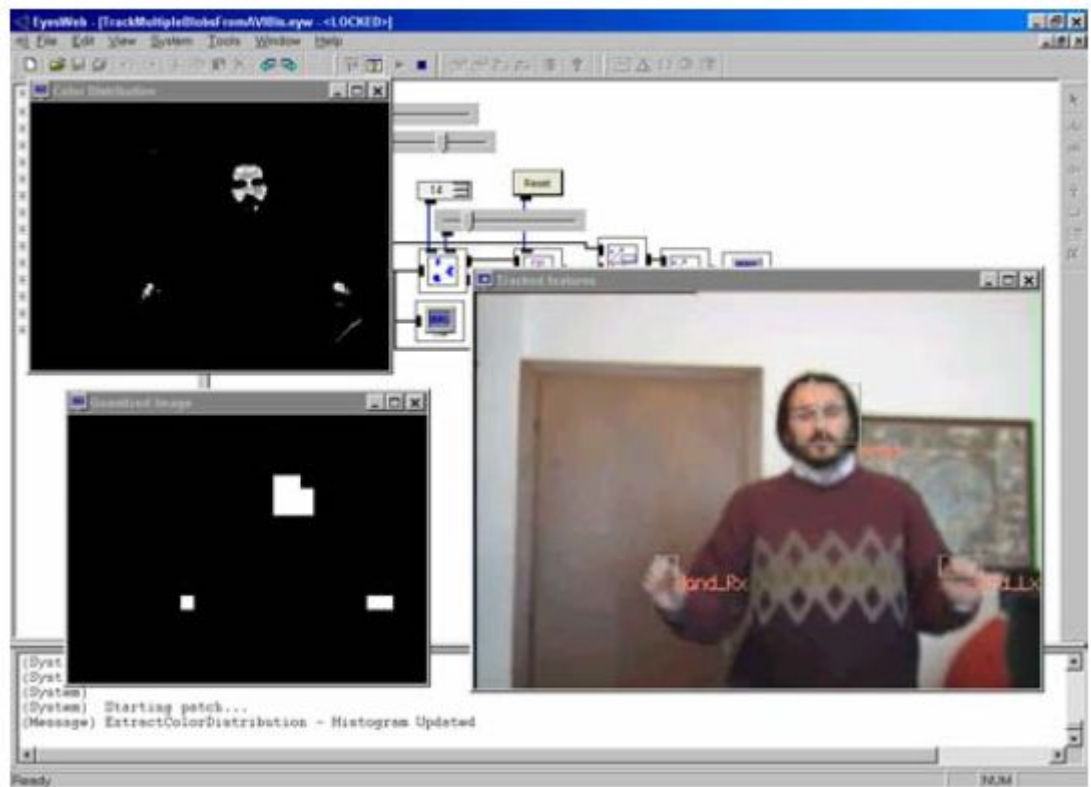


Figure 21: Color recognition

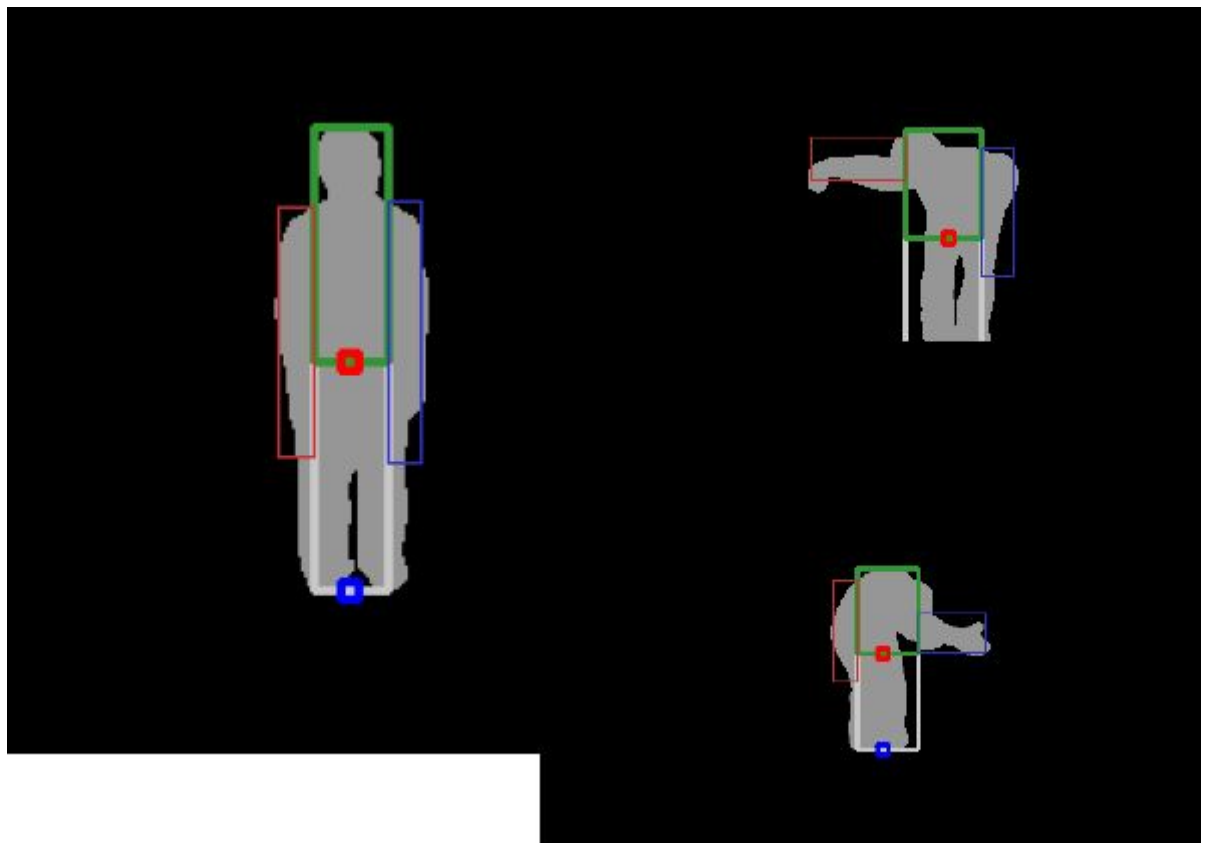


Figure 22: Subregion tracking





## **3 Human-Computer Interaction**

### **3.1 Introduction**

As seen before there has been always the need of interact more or less with machines. For this proposal about fifty years ago was born the science of Human-Computer Interaction. This science has the goal of study man and human behaviour to make communicating with machines more effective with reducing the possibility of misunderstanding in using a machine by human.

### **3.2 Computer Vision for Human-Computer Interaction**

In the next chapters are discussed some cases in which Computer Vision aids Human-Computer Interaction.



## **4 Use Cases**

### **4.1 Introduction**

Here are presented some use cases in which is shown how widespread can be the use of Computer Vision in creating Human-Computer Interaction.

### **4.2 People with Disabilities**

#### **4.2.1 Introduction**

The best ways to use technology is when it can be used to help people. Here are presented some cases in which computers are used as a tool to help people in need to live better. In this case here are presented some cases in which computer vision can be used for helping blind, deaf and affected by autism people.

#### **4.2.2 Blind and Visually Impaired People**

Computers can be used as a tool for helping blind people. It can be used as an interface between the person and the environment trying to compensate human vision lack. The idea is to use computer vision within a device which can tell with voice or updating a braille device what is around the person. This can be done with one camera if the device is moveable with the person or with more cameras if thinking about a room or a house where to displace the cameras. This type of technology can be used for example for directing the person on a path or, thinking about a more intelligent device, for telling which objects are around the person itself.

An example of prototype can be found in [4] is described how the team de-

veloped a system for helping people to find daily necessities. They prepared a camera-based network to allow object match-recognition. They collected a dataset of daily necessities and applied a Speeded-Up Robust Features and Scale Invariant Feature Transform feature descriptors to perform object recognition.

Nowadays it's possible to find a lot of apps for smartphones in which using the device's technology blind people can be helped, for example being directed through a path letting people free to move by hearing hints about the position they are.

#### **4.2.3 Deaf and Hearing Impaired People - Sign Language**

Another aid offered by computer vision application can be the development of an interface for deaf people helping them, for example, allowing an easier communication with other people. The idea is making a device capable of recognizing the sign language to allow the consequent using of that information for many scenarios, for example allowing communication with people who doesn't know the sign language. An example can be found in [2] where it's presented a system for sign language recognition, see Figure 23.

The person moves the hands doing the right movements. The system records the gesture motion images applying the Background Subtraction method. Later the image is transformed to locate the skin region, see Figure 24.

To remove the noises produced in image process it's used Morphological and Connected Component method, see Figure 25.

Further, are also eliminated problems with recognizing process like a slanted hand gesture. Finally, it's used an Artificial Neural Network for recognizing the

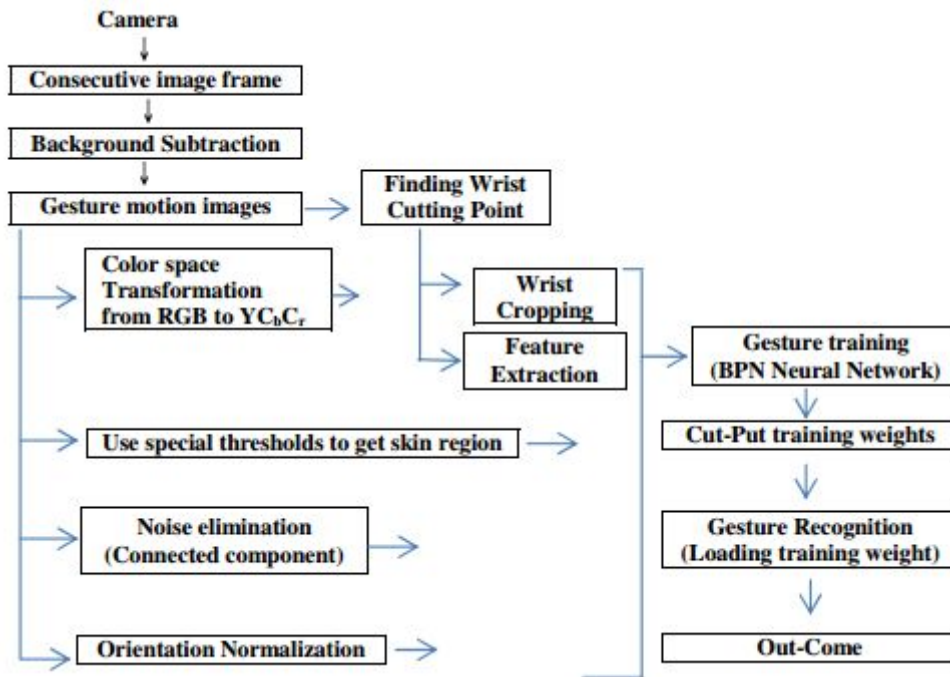


Figure 23: System Framework



Figure 24: Original image and after extraction result



Figure 25: Before and after Connected Component method

sign language. The out-come can be transmitted to handheld device such as smartphone. Three experiments were conducted on a quadcore Intel system. The first one is to determine and inspect how many features should be taken. Randomly were chosen 80 percent for training, the others for testing. Were taken 25\*25 features to be the feature size. Figure 26 is shown the results of gesture feature size grater than 25\*25. The recognition rate is about 94.63 percent in average and the processing time is only 39ms in each gesture.

Gesture feature size	Training time	Recognition rate
25*25	3539ms	94.2%
30*30	5549ms	94%
35*35	7784ms	93.7%
40*40	11024ms	94.1%

Figure 26: Different gesture size affect training time and recognition rate

Experiment 2 is to verify the result in different lighting condition. The result is good in different conditions, see Figure 27.











Luminance Test	Image	Result	Luminance Test	Image	Result
Bright			Complex Background		
Backlight			Complex Background2		
Warm color temperature			-	-	-

Figure 27: Different light condition affect the extraction result

Experiment 3 is to test recognition rate in this method: the gesture were randomly moving in front of the camera and the system would recognize 100

times for calculating the recognition rate. The result is shown in Figure 28. The recognition rate is about 89 percent in average and the processing time of 55ms in each gesture.

<b>Gesture</b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
<b>Reg. rate</b>	<b>88%</b>	<b>96%</b>	<b>88%</b>	<b>93%</b>	<b>87%</b>	<b>98%</b>	<b>82%</b>	<b>86%</b>	<b>90%</b>	<b>98%</b>
<b>Gesture</b>	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>E</b>	<b>F</b>	<b>G</b>	<b>H</b>	<b>I</b>	<b>J</b>
<b>Reg. rate</b>	<b>92%</b>	<b>94%</b>	<b>95%</b>	<b>97%</b>	<b>87%</b>	<b>91%</b>	<b>86%</b>	<b>81%</b>	<b>92%</b>	<b>84%</b>
<b>Gesture</b>	<b>K</b>	<b>L</b>	<b>M</b>	<b>N</b>	<b>O</b>	<b>P</b>	<b>Q</b>	<b>R</b>	<b>S</b>	<b>T</b>
<b>Reg. rate</b>	<b>81%</b>	<b>89%</b>	<b>80%</b>	<b>89%</b>	<b>81%</b>	<b>89%</b>	<b>80%</b>	<b>89%</b>	<b>80%</b>	<b>82%</b>
<b>Gesture</b>	<b>U</b>	<b>V</b>	<b>W</b>	<b>X</b>	<b>Y</b>	<b>Z</b>	<b>-</b>	<b>-</b>	<b>-</b>	<b>-</b>
<b>Reg. rate</b>	<b>98%</b>	<b>94%</b>	<b>100%</b>	<b>83%</b>	<b>100%</b>	<b>85%</b>	<b>-</b>	<b>-</b>	<b>-</b>	<b>-</b>

Figure 28: Recognition rate test

These results demonstrated the computer vision based hand gesture recognition can be applied to devices as portable and little as smartphones.

#### 4.2.4 Autism Spectrum Disorder

Computer vision can be a little part of a bigger project. The example is found in [5] where a team presents how the programmable robot NAO, a robot with cameras mounted in it, can acquire information about children suffering from Autism Spectrum Disorder (ASD) helping therapists with adopting new procedures in ASD therapy.

The overall context proposed in this paper is part of our long-standing goal to contribute to a group of community that suffers from Autism Spectrum Disorder (ASD); a lifelong developmental disability. [5] The objective of this paper is to present the development of our pilot experiment protocol where children with ASD will be exposed to the humanoid robot NAO. This fully programmable

humanoid offers an ideal research platform for human-robot interaction (HRI). This study serves as the platform for fundamental investigation to observe the initial response and behavior of the children in the said environment. The system utilizes external cameras, besides the robot's own visual system. Anticipated results are the real initial response and reaction of ASD children during the HRI with the humanoid robot. This shall leads to adaptation of new procedures in ASD therapy based on HRI, especially for a non-technical-expert person to be involved in the robotics intervention during the therapy session.

### **4.3 Entertainment**

#### **4.3.1 Media Player**

Detect gaze out of the screen so pause the video; scroll pages with eyes (Samsung Galaxy commercial applications)

Imagine you are watching a movie and you have the power to control the movie with your hands. In [9] is presented a system that allow you to play, pause, stop and do every other action you could do with a movie player, everything controlled by hand gesture. There is a first phase consisting in teaching the system the commands you want it recognizes associating them with the actions you want the media player will do. The main phase is the one in which is possible use the media player commanding it with your gesture.

#### **4.3.2 Videogames**

In 2003 Sony released EyeToy for PlayStation 2. EyeToy is a camera with a specific software that let the user to interact with videogames with gestures.



The software processes captured images using computer vision and gesture recognition allowing users to interact using motion and color detection. Some specific videogames were developed for this platform allowing users have fun with this new way to play games. EyeToy mainly used simple edge detection and color tracking and Digimask face mapping. Later in 2007 were released PlayStation Eye, the successor of EyeToy, adding more features to its predecessor like advanced facial recognition/analysis and CV-based head tracking. The facial technology can identify features such as eyes, mouth, eyebrows, nose, and eyeglasses, can read the shape of the mouth and detect a smile, also can determine the position and orientation of the subject's head and estimate the age and gender of the face.

Also Microsoft released its own recognition camera for its console, the Xbox. In 2006 was released Xbox Live Vision, a system similar to Sony's EyeToy. In 2010 arrived Kinect, a natural user interface that allows users to interact with games using their entire body. This device extended the capability of interaction with videogames allowing a greater variety of applications. These cameras are usable also with PCs opening new ways to user interaction using computer vision. This ease of access to there devices, also, help growing of mixed reality.

One of the hardest part to find out when developing a human interface with so many degrees of freedom as one based on computer vision is to provide a set of commands that is the most near to the natural way a human can do. In other words, to be effective, an interface must be shaped around the human most natural behaviour in order to fire the triggers for doing something. An example of how this type of approach can be afforded is in [14]. The team used a game and

the Wizard of Oz method with some children. The Wizard of Oz method consists of letting someone act in front of a system in the most natural way waiting for it to act the expected way. But making the system working is not the system itself but another human hidden from the first, the "wizard". Doing this the researcher can find out the most natural way the user uses to interact with the system letting the researcher understand the most used commands, statistically speaking. In the named paper [14] the team shows to some children a videogame in which the main character is moved by their movement. The truth is that a "wizard" moves it, but doing this the researcher made some statistics on how the children behaved in order to control the videogame. It's interesting because this type of approach can help on how develop a more natural-response system without using computer vision before, having all the problems about putting boundaries on the recognition of the actions, so the system can be developed after the study so the total result is completely user-oriented.

As for videogames engines that aid videogame developer to concentrate on the videogame itself and less on the graphics algorithms, in [7] is presented the Flexible Action and Articulated Skeleton Toolkit (FAAST). FAAST is a middleware to facilitate integration of full-body control with virtual reality applications and video games using OpenNI-compliant depth sensors (currently the PrimeSensor and the Microsoft Kinect). FAAST incorporates a VRPN server for streaming the user's skeleton joints over a network, which provides a convenient interface for custom virtual reality applications and games. This body pose information can be used for goals such as realistically puppeting a virtual avatar or controlling an on-screen mouse cursor. Additionally, the toolkit also provides

a configurable input emulator that detects human actions and binds them to virtual mouse and keyboard commands, which are sent to the actively selected window. Thus, FFAST can enable natural interaction for existing off-the-shelf video games that were not explicitly developed to support input from motion sensors. The actions and input bindings are configurable at run-time, allowing the user to customize the controls and sensitivity to adjust for individual body types and preferences. In the future, the developing team plans to expand FFAST's action lexicon, provide support for recording and training custom gestures, and incorporate real-time head tracking using computer vision techniques.

Another step to integration of Computer Vision with videogames is in [10]. This contribution presents an easy to implement 3D tracking approach that works with a single standard webcam. It describes the algorithm and shows that it is well suited for being used as an intuitive interaction method in 3D video games. The algorithm can detect and distinguish multiple objects in real-time and obtain their orientation and position relative to the camera. The trackable objects are equipped with planar patterns of five visual markers, see Figure 29. By tracking (stereo) glasses worn by the user and adjusting the in-game camera's viewing frustum accordingly, the well-known immersive "screen as a window" effect can be achieved, see Figure 30 even without the use of any special tracking equipment.

Talking about middleware, another tool that helps in developing videogames is [26], a computer vision library that allows an accurate real-time tracking of the entire human body as well as background removal using regular webcams. The technology allows development of videogames with cameras for multiple

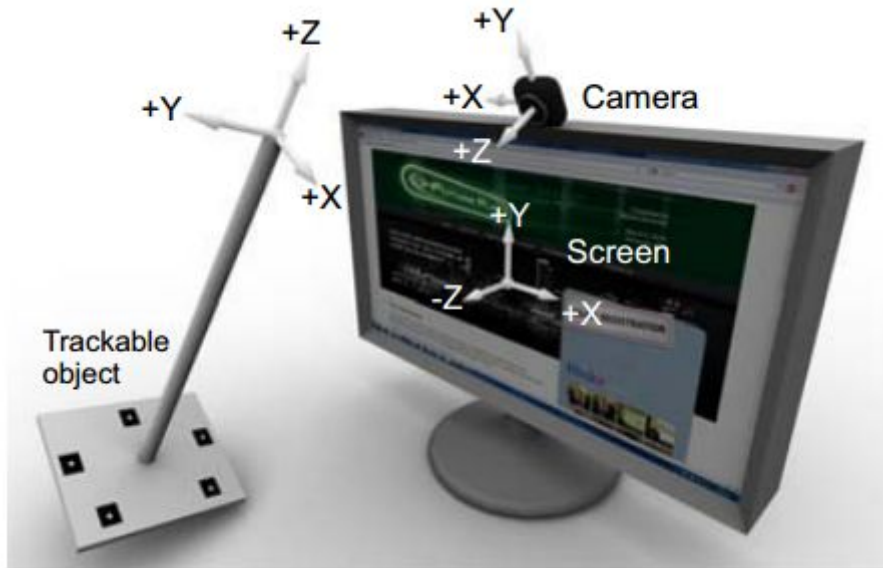


Figure 29: Trackable object

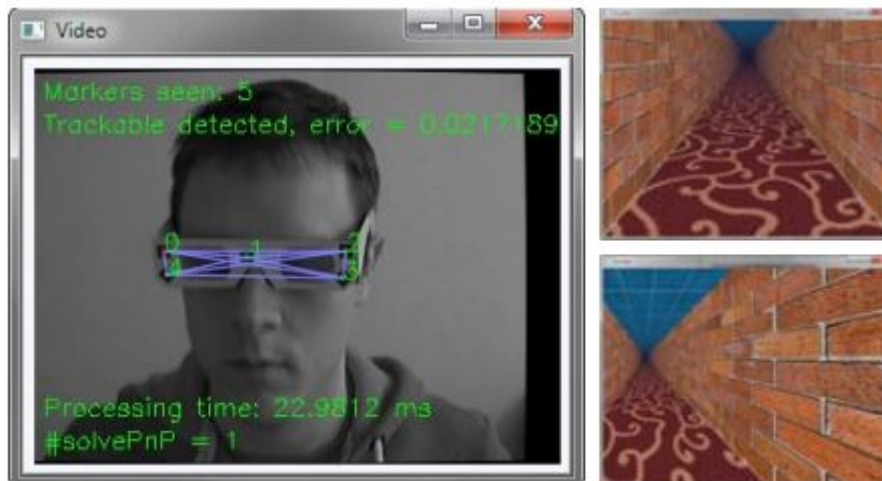


Figure 30: Example of "screen as a window"

platforms, even if the underlying camera technologies differ, such as one platform using a regular RGB camera and another using a 3D depth camera. An example in Figures 31 and 32.



Figure 31: FreeMotion Tracking



Figure 32: FreeMotion Game

#### 4.4 Shopping

Computer Vision could change also the way we go shopping. An example can be found in [12] in which it's possible to see an application of a so named Virtual Mirror. The idea is to have a camera looking at the feet of a customer which is

standing in front of a big screen. The customer can choose a pair of shoes from a virtual catalog and the screen shows the image of its feet wearing the chosen shoes in real time. It's possible to change the features of the shoes looking for the best combination that can satisfy him or her. In this way is it possible to customize the product before it is manufactured. In 33 a photo of a device of this type in action.



Figure 33: Virtual Mirror presented in [12]

The person stands in front of a screen with at the bottom a camera pointing at legs and feet level, see 34. Here is represented the scheme of the system: a camera, a screen, that can be touch to let the person customize the viewed

result at his/her feet.

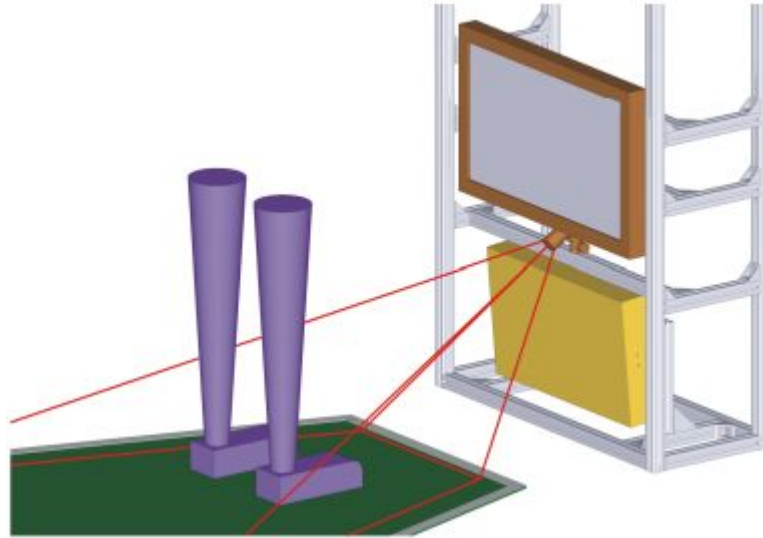


Figure 34: Virtual Mirror camera system presented in [12]

A 3D motion tracker estimates the position of and orientation for each foot using a model-based approach that is very robust and can be adapted to new shoe models. Also no markers are required on the shoes. After the system estimated the position of customer's feet, replace the image of the feet getting over the desired customer's shoes picture and showing it on the screen giving the effect of looking in a magic mirror which allow to dress every type of shoe without trying it for real moving freely the feet. The system uses an algorithm that tracks constantly the position of the feet allowing a sort of natural moving using the new shoes. In order to simplify the segmentation in a real environment with changing illumination and arbitrary colors of clothes, the floor under the customer is painted in green allowing the use of chroma keying techniques.

## 4.5 Office

Nowadays the office work is mostly demanded to using a PC or a laptop to accomplish main office activities. The type of work done using a PC regarding the office is limited by using input devices like mouse and keyboard to manage windows, documents, graphic objects, etc. There is a way in [3] in which is presented a system that let person to manage objects directly on the desktop without using devices but using only the hands. The main components of the system are cameras, projectors and a computer, as shown in Figure 35.

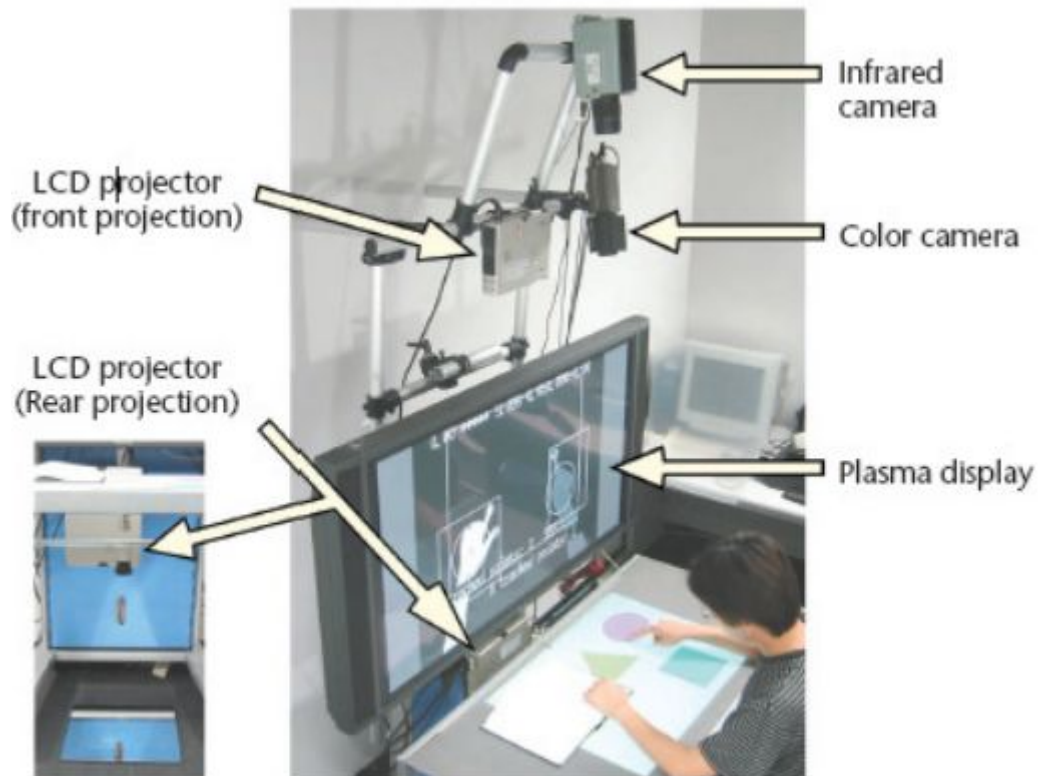


Figure 35: System presented in [3]

The computer elaborates a first set of objects projected by the projectors on the user's desktop. As the person moves his/her hands the system recognizes



the movement and have the capability of track if the person wants to resize or move the shown objects. The result is the camera acquire the information, the computer elaborates and the projector shows the resulted image allowing the user an enhanced experience.

Another example of using the desktop surface as input method is presented in [21]: *Inktuitive*. This time the idea is to make the person using a pencil to draw CAD projects, as shown in 36.

The system is composed by a projector that projects the image from under the table showing the image through a glass, ultra-sonic waves stationary receiver, infrared cameras and infrared emitter LEDs above a ultra-sonic waves transmitter pen. The system allows to detect 3D movements of the pen letting the user draw in a 3D space having an image projected on the desk and another on the front-screen. All this system can improve designers' projects-making.

In [25] is presented a device which allows the user to stand in front of a videowall and a Microsoft Kinect for using the body and the hand in order to manage an interface that allows to browse through the days schedule, speaker lineup, attendee directory and a real-time twitter feed, see 37

## **4.6 Videoconferencing**

The most important men invention is communication. Since when we can communicate from large distances we try to make the communication the most possible natural, trying to make long distance calls clear and doing the best to give people the feeling of speaking with a near person, see 38. To add a piece

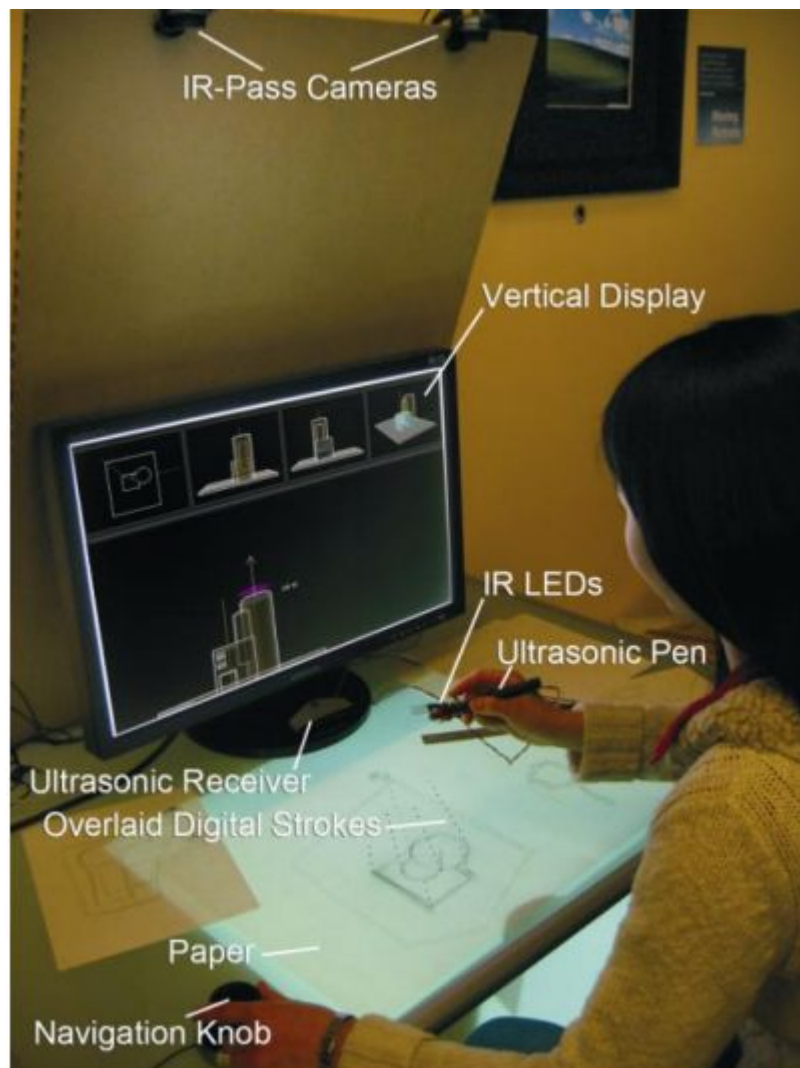


Figure 36: Inktuitive



Figure 37: Fresk's Computer Vision Interaction

to the puzzle of making a better communication we can refer to a system that tracks the face movement making the camera moving and zooming when the person moves from the stationary position. This was presented since the 90s, where was tested the possibility of tracking the position of people's faces. Yet at that time there were good results as in [17]. The system can adapt to different light conditions making it robust against environmental variability in real time.



Figure 38: Videoconferencing

## 4.7 Virtual Input Devices

Another interesting way to use Computer Vision is that it's possible to simulate existing real devices using your hands gesture simulating their presence to command a computer or something else. An example could be find in work of Pranav Mistry named Mouseless [19]: the user moves the hand like it is handling a mouse and uses it in a natural way. An infrared system detects the hand movement and simulates mouse moving on the screen and the clicks. An infrared laser makes a plane on a parallel level to the table while an infrared camera detects the interruptions made by the hand moving and the software converts these in mouse movement and clicks, see 39, 40 and 41

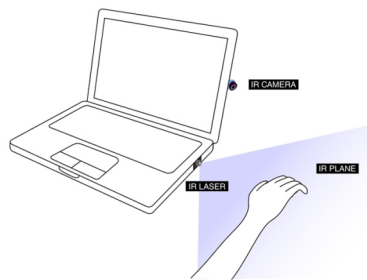


Figure 39: Mouseless system [19]

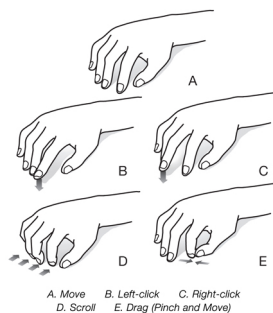


Figure 40: Mouseless recognized hand gestures [19]

Another application of Computer Vision from Pranav Mistry is a system



Figure 41: Mouseless [19]

that allows the user to interact with the computer screen like a touch screen but thanks to infrared technology you can move the mouse, click and doubleclick without touch the screen [20]. That's how it works: three infrared cameras project three layers parallel to the screen. An infrared camera detects the hand making holes in these planes and based on which one has been hit send the computer the signal of move, click or doubleclick. A picture in Figure 42.

#### 4.8 Object-Computer Interaction

It seems somehow in the future real life will be mixed to digital life. There is Virtual Reality with allows a person to be hosted in a virtual world made of colored lights. The same virtual world can acquire real objects from the real world and use them as input to relate them to some digital data already in the virtual world or the other side of the real world. Here an example: a table which has an image projected from the bottom letting someone to put objects over it. A camera records from the top recognizing objects the user let there. The

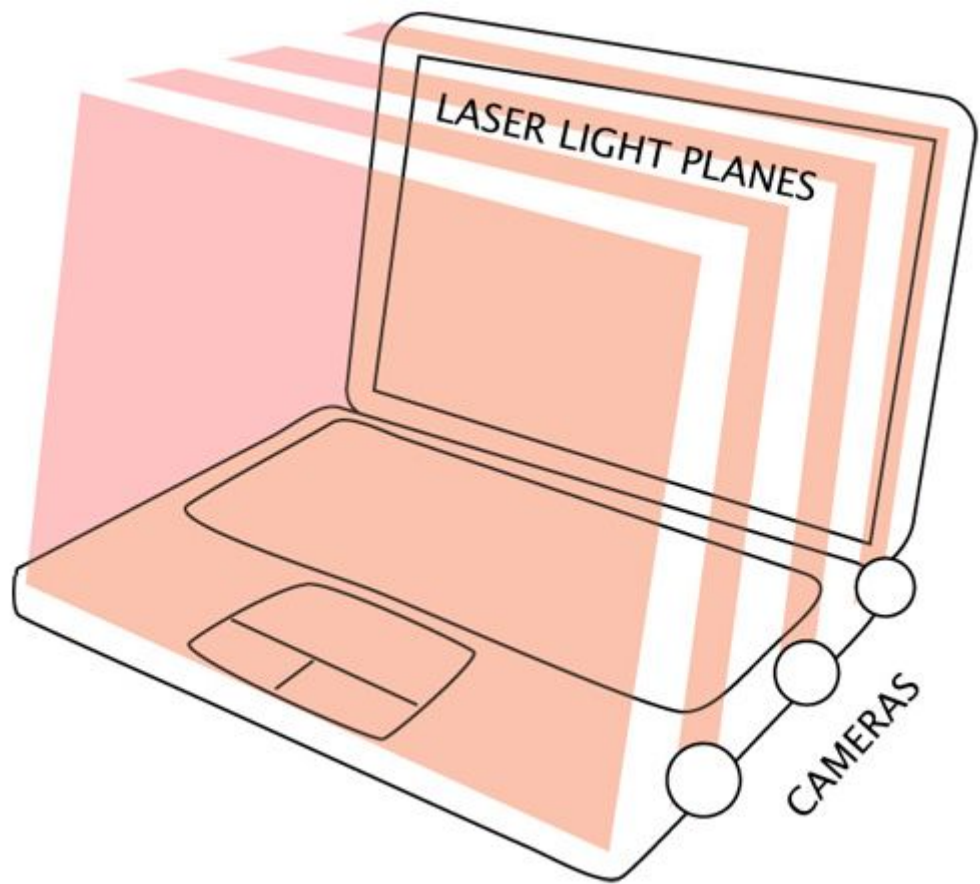


Figure 42: Precursor infrared layers [19]

computer recognizes the object and if it's smart enough to relate the objects information with something in its database or the internet, it will help the user giving him/her useful information or some kind of extra data regarding the real object. That's the mix of real world into virtual world using Computer Vision. And from virtual world rises more information, more data from an always more connected real world. There will be a time in which we won't be able to separate the two worlds in which one will depend in a strict way from the other? We'll see. By now here's presented Tapuma, Tangible Public Map [22]. An image is projected onto a table. The user can put over it some objects like a ticket or a mobile phone and a visual system recognizing these objects responds with information about them. See Figure 43.



Figure 43: Tapuma

## 4.9 Remote Control

In [16] we find an interesting application of Computer Vision: the remote control of home appliances. The system shows a menu which is controlled by hand gestures. In the paper's scenario there is a pie-menu on a screen enabling the possibility of remote control of home appliance such as TV sets and DVD players, see 44 and 45

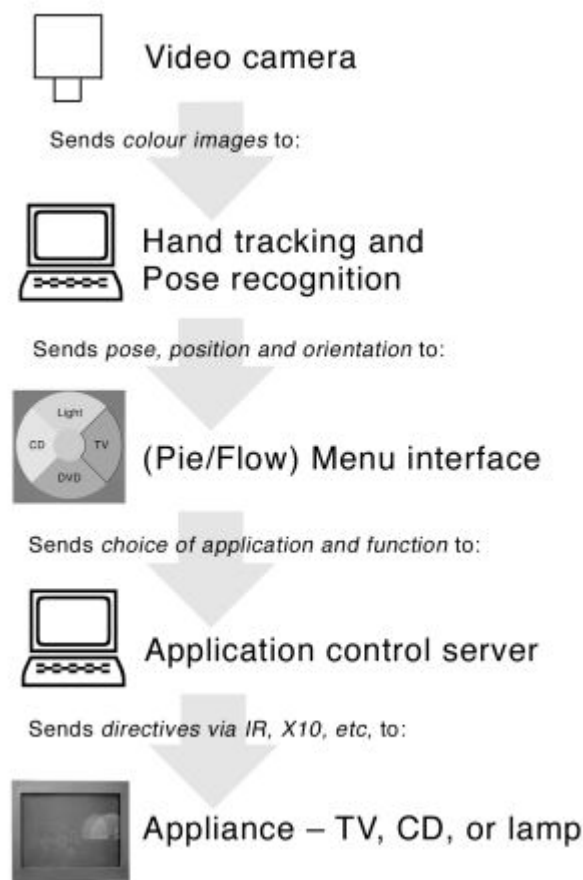


Figure 44: Remote Control System

There are some situations in which is needed the use of a robot to perform some critical operations without human intervention such as handling hazardous





Figure 45: Remote Control Demo

materials, dismantling bombs, mines, or nuclear facilities, and operating in inaccessible sites in rescue, undersea exploration and mining. One way of doing this is presented in [13]. A robot is remotely controlled by a human through the use of cameras. The person is pointed by two cameras which record and recognize the hand's movements. In another site there is a six degrees of freedom robot arm. The arm is pointed by cameras which give feedback of the movement to a screen the human has in his position. This system allows the human to control the robot by a Computer Vision system. The figure 46 shows the entire system, Figure 47 the human's site and 48 the robot's site.

Another type of remote control is presented in [23]. Using a smartphone you can point its camera to an object and if it's recognized starts communicating with it via wifi, if available on the target device. Once the communication is established the user can interact with the object, for example turning it on or off, as shown in 49.

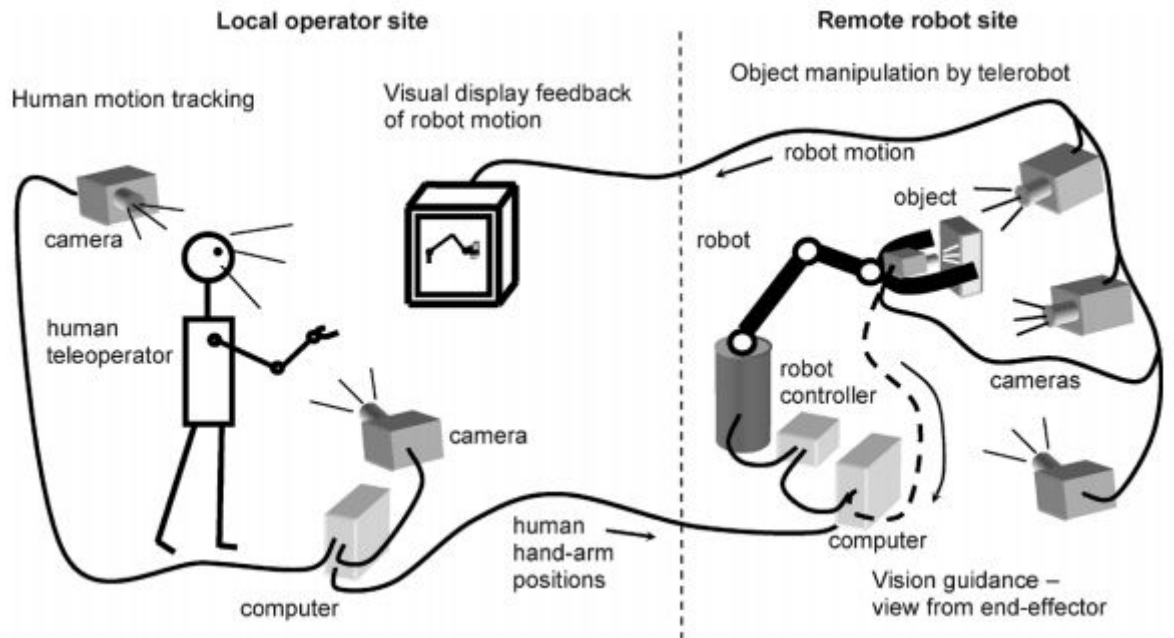


Figure 46: Robot Remote Control System

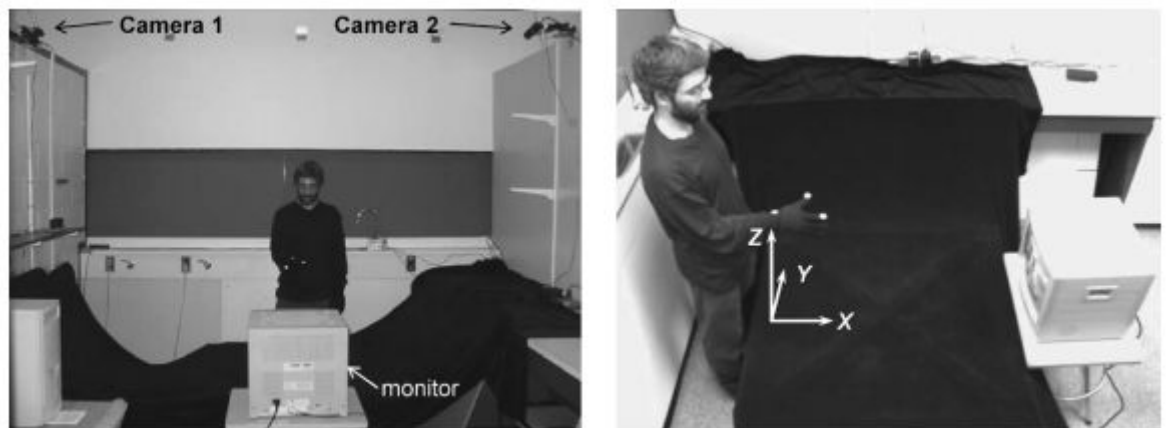


Figure 47: Robot Remote Control System

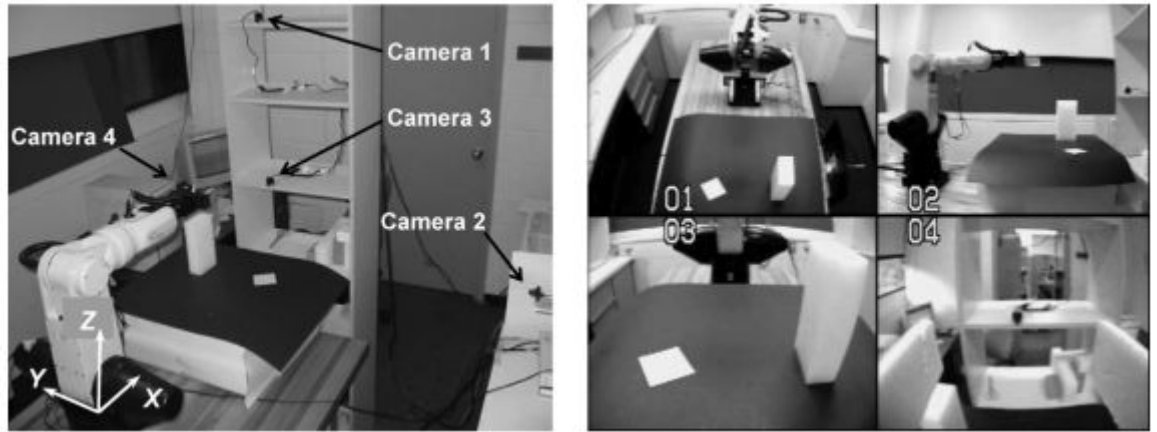


Figure 48: Robot Remote Control System



Figure 49: Teletouch

#### 4.10 Wearable Visual Interface

Exists yet a technology that allows us to do whatever we do nowadays like making a phone call, taking pictures and so on using only our fingers without handling any device. An example is SixthSense [24] which is a set of projector and camera. With it is possible to make phone calls, take pictures, edit pictures and so on only with gestures of the fingers. See Figures 50, 51, 52, 53 and 54

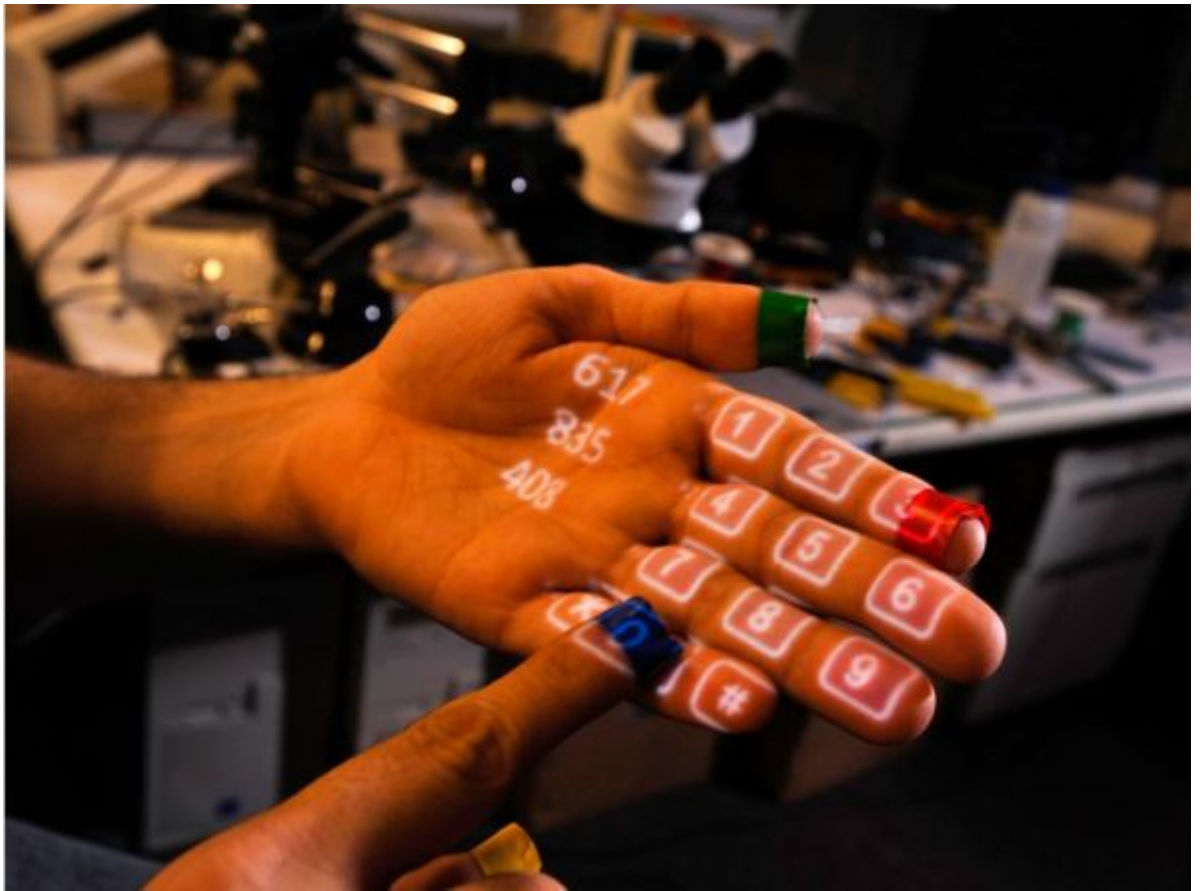


Figure 50: Phone call



Figure 51: Photos show



Figure 52: News watch



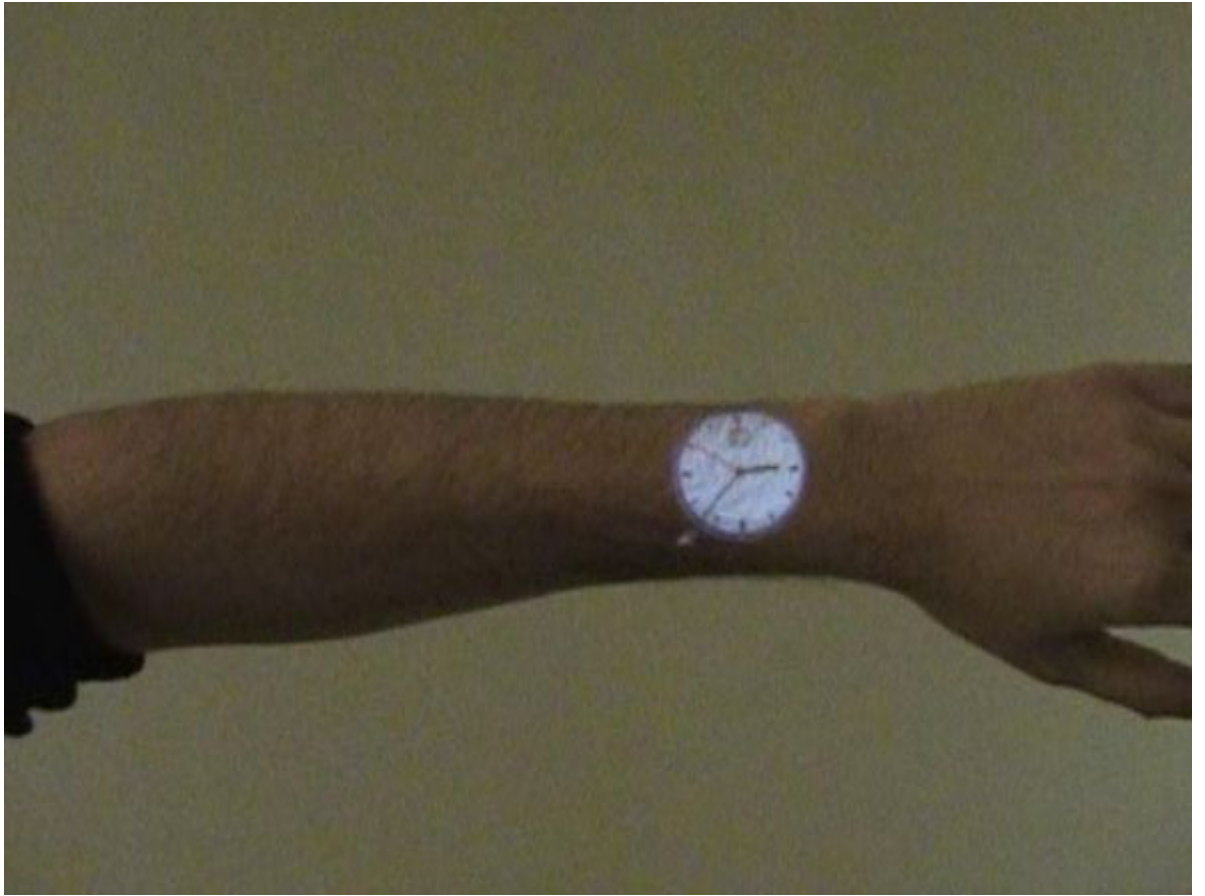


Figure 53: Watch



Figure 54: Information retrieval



## 5 Interaction Models

### 5.1 Introduction

Follow some generalizations based upon the use cases presented in the previous chapters. From the examples above is possible to extract some information: the actors of the scene, that could be a person or an object, the technology component like the computer, the screen or the projector and the way the information moves.

### 5.2 Components

The principal components of the scenarios above are people, objects, cameras, computing and communication devices, screens or output devices. Follows a list with a letter which will be used during the next paragraphs.

P = Person

O = Object

C = Camera (input)

D = Device (single/multiple computing/communication system)

S = Screen (output)

Explanation:

- P and O could be one or more filmed subjects
- C could be one or more cameras;
- D could be one or more computing and/or communicating devices;

### 5.2.1 Person

A person is a human, represented as in Figure 55.



Figure 55: Person

### 5.2.2 Camera

A camera is an acquisition device represented as in Figure 56

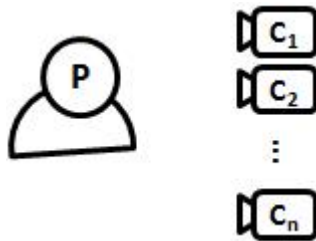


Figure 56: Cameras with Person

### 5.2.3 Device

A device is a computing and/or communication device, represented as in Figure

57

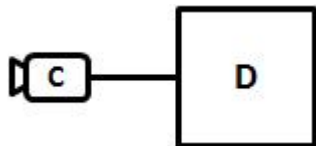


Figure 57: Device with Camera

#### 5.2.4 Devices

Devices are a set of objects Device connected also via a network. See Figure 58

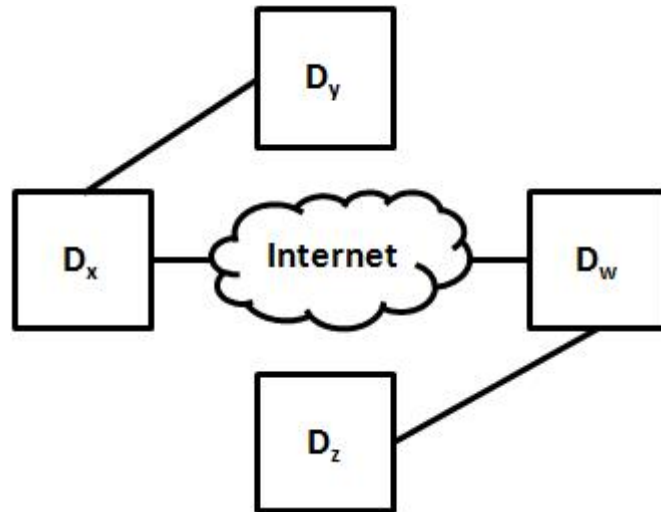


Figure 58: Devices

#### 5.2.5 Screen or Output

A screen or output device is something useful for human feedback connected to a device from which receive the image/output to give the human. See Figure 59



Figure 59: Screen with Device

### 5.2.6 All components

In Figure 60 a representation of a basic scenario with a human acquired by a camera connected to a computing device and to a screen.

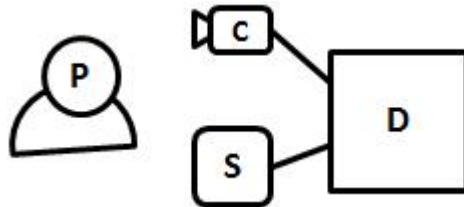


Figure 60: All components

## 5.3 Workflows

### 5.3.1 Generic Information Flow

Using the basic components described above and connecting the devices making a network we can develop a general scheme for a generic Human-Computer Interaction based on Computer Vision. Giving the flow of information it's possible to use the computing devices as nodes for computing with Computer Vision the image of the human/object in front of the connected camera and the screen/output device to let a feedback to a human. Cycling all this we have an interactive interface. This is feasible from any node to any node of the network. Doing this is represented a generic information flow as in Figure 61.

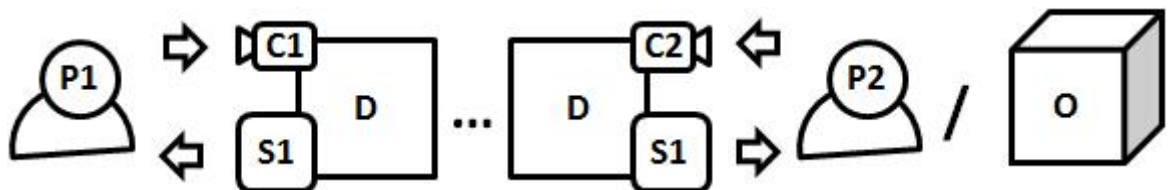


Figure 61: Generic Information Flow

### 5.3.2 Simplest Information Flow

In Figure 62 is represented the most simple way of doing Human-Computer Interaction based on Computer Vision: the human, the acquiring camera, a computing device and a screen for feedback. This represents the use case above about shopping.

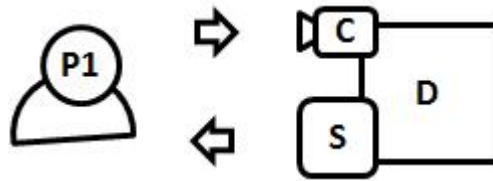


Figure 62: Simplest Information Flow

### 5.3.3 Object-Compute-Output

Another way of use the generic scheme for a particular scenario is letting the human pointing the camera to an object or the environment waiting for this being the triggering action for a Computer Vision work giving a response to the human. See Figure 63. A reference in the real world could be Teletouch presented above.

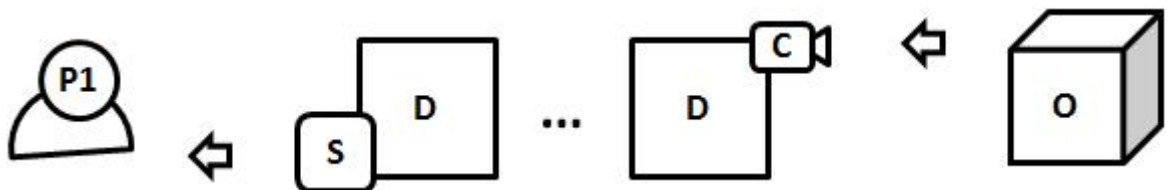


Figure 63: Object-Compute-Output

### 5.3.4 Person-Compute-Person

This case is similar to the above case but this time the pointed object is another person. See Figure 64. An example in the real world could be the recognition of sign language for another person.

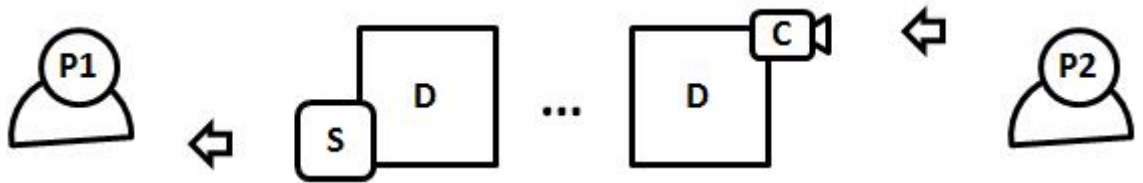


Figure 64: Person-Compute-Person

### 5.3.5 Full-Duplex Person-Compute-Person

This is an extension of the above case, in which both people can be seen by Computer Vision and receive a response from the twin device, see Figure 65. For this an example could be videoconference.

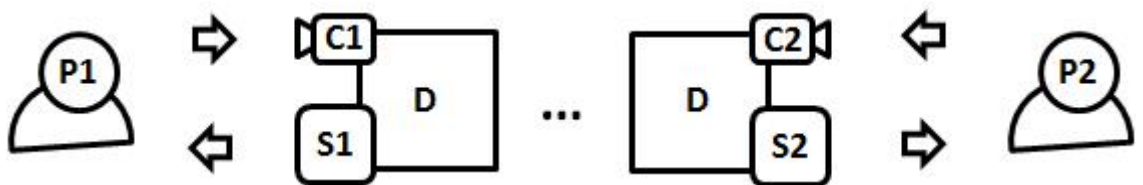


Figure 65: Full-Duplex Person-Compute-Person

### 5.3.6 Acquire-Elaborate-Show

Here is presented a specific case in which a person is directly interacting with a device which is fired by Computer Vision. In this case the output is shown to the person onto a screen. See Figure 66

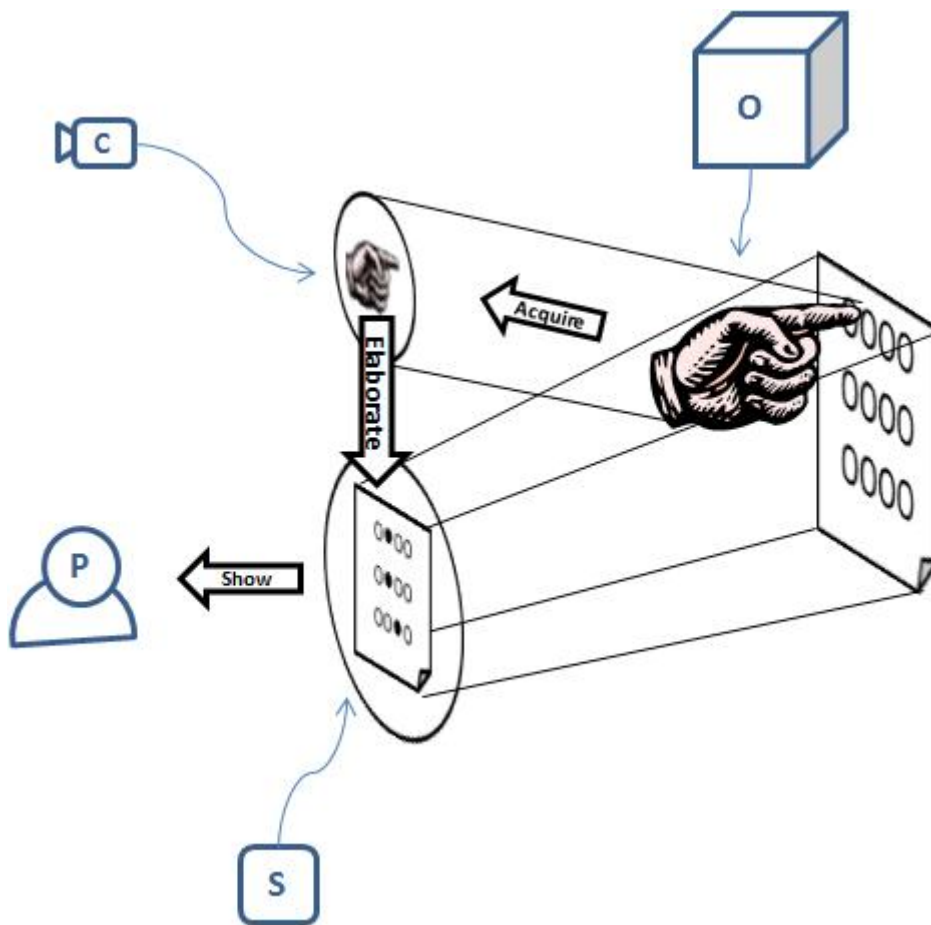


Figure 66: Acquire-Elaborate-Show

### 5.3.7 Acquire-Elaborate-Project-Show

The case of Figure 67 is projecting the feedback output directly onto the acquired part of the scene. An example for this could be SixthSense presented above.

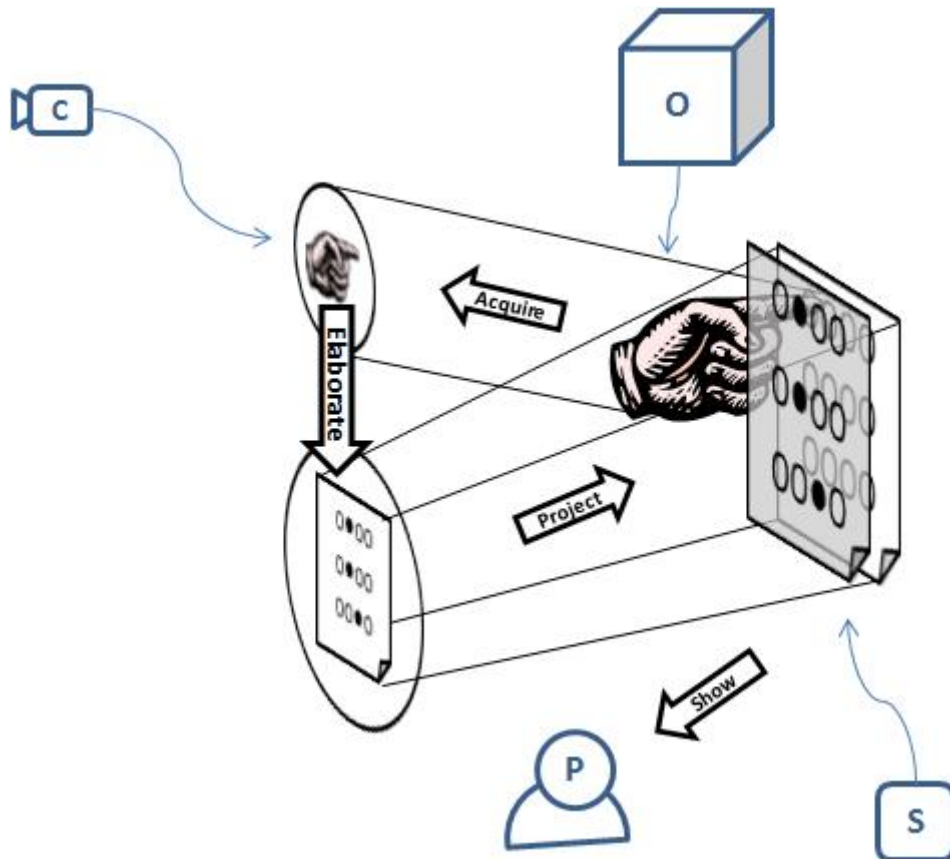


Figure 67: Acquire-Elaborate-Project-Show

Last but not least example is Figure 68 in which the projected surface is not above, but at the bottom, without disturbing the acquired scene. An example for this could be Tapuma presented above.



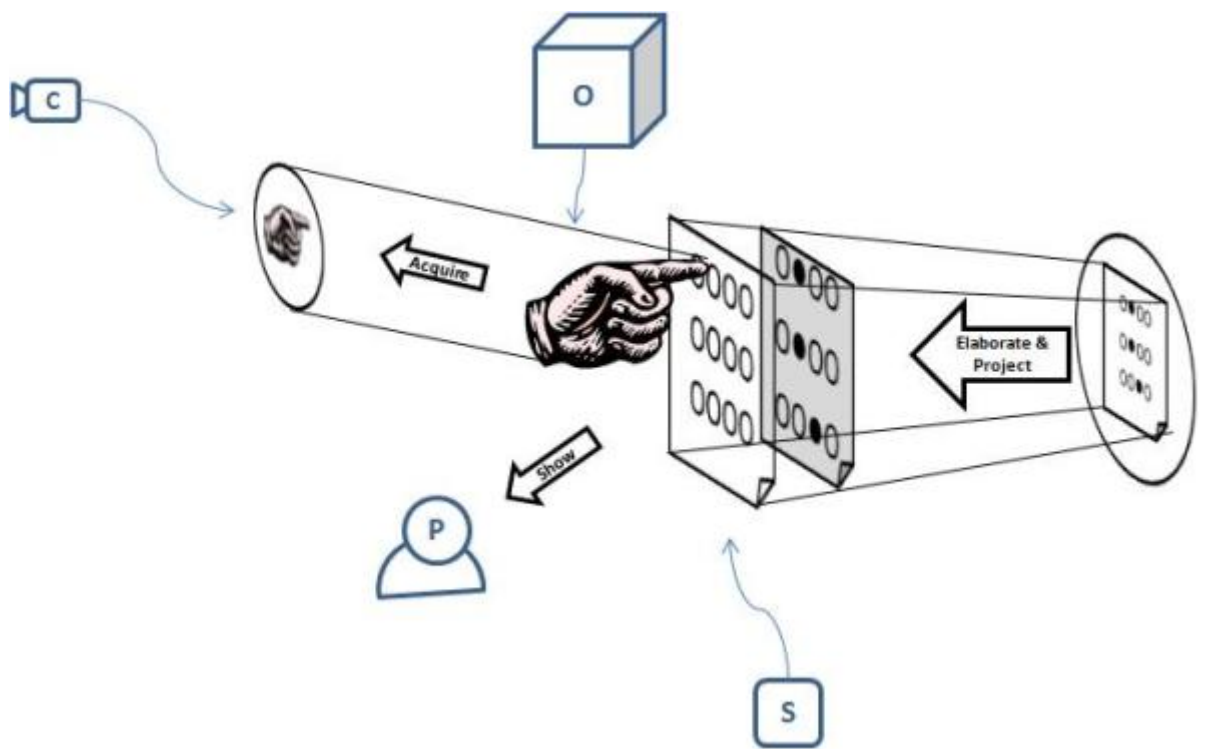


Figure 68: Acquire-Elaborate-Project-Show

## 5.4 Interaction Models Properties

Based on use cases above it was made the Figure 69. It represent a categorizing of the use cases using following drivers: if the camera is stationary or can be moved, the type of filmed subject, the recognition trigger that makes fire Computer Vision, if the output is shown to the acquired subject and the output type

Use Case	Camera is Stationary	Filmed Subject	Recognition Trigger	Output to Filmed Subject	Output Type
Blind and Visually Impaired People	Y/N	Environment	Objects	Give information	Virtual
Deaf People - Sign Language	Y	Person	Body Part	Do something	Virtual / Real
Autism Spectrum Disorder	Y	Person	Behaviour	Nothing	Virtual
Media Player	Y	Person	Body Part Behaviour	Do something	Virtual
Videogames	Y	Person	Body Part Behaviour	Do something	Virtual
Shopping	Y	Person	Body Part	Give information	Virtual
Office	Y	Environment / Person	Body Part	Do something	Virtual
Videoconferencing	Y	Person	Body Part	Nothing	Virtual
Virtual Input Devices	Y	Person	Body Part Behaviour	Nothing	Virtual
Object-Computer Interaction	Y/N	Environment	Object	Give information	Virtual
Remote Control	N	Object	Object	Nothing	Virtual / Real
Wereable Visual Interface	Y	Environment	Object	Give information	Virtual

Figure 69: Interaction Models Properties

## 5.5 Interaction Models Properties Values

From Figure 69 of the previous paragraph it's possible to represent the information derived as a table with in columns the scene's attribute and in rows the possible values. Doing this is possible see a categorized view of Human-Computer

Interaction with Computer Vision that can be useful for driving scenario set up or many other the choose of the right algorithms. See Figure 70.

Attribute	Value1	Value2	Value3
Camera is Stationary	Yes	No	
Filmed Subject	Environment	Person	
Recognition Trigger (Static)	Body	Body Part	Object
Recognition Trigger (Dynamic)	Body Behaviour	Body Part Behaviour	
Output to Filmed Subject	Give Information	Do Something	Nothing / To other person
Output Type	Virtual	Real	

Figure 70: Interaction Models Properties Values



## 6 Conclusions

Here we have seen the actors of the play, the computer and the human and how many ways are available for communicating, how many situations can afford from the exploitation of new technologies in all days life. Here were presented many use cases in which human is aided by using Computer Vision in interacting with or using computer in order to accomplish some duties or for having better communication with other people. With the increasing of speed of devices and decreasing of dimension an cost is it possible do what some years ago maybe could be thought as impossible. And nowadays we can think, who knows, that with all these possibilities we can achieve a good level also in known fields but not so widespread in our life as in Perceptual User Interface, Ubiquitous Computing, Augmented Virtuality, Augmented Reality and Hybrid Human-Computer Interaction to make life better.



## **7 Acknowledgements (Italian)**

Ringrazio i miei genitori, per la pazienza e la presenza in ogni circostanza.

Ringrazio mia sorella, la mia ragazza e gli amici per avermi supportato (e sopportato!) in questi anni.





## References

- [1] Ankit Chaudhary, J. L. Raheja, Karen Das and Sonia Raheja, *Intelligent Approaches to interact with Machines using Hand Gesture Recognition in Natural way: A Survey*, 2013
- [2] Hsi-Chieh Lee, Che-Yu Shih and Tzu-Miao Lin, *Computer-Vision Based Hand Gesture Recognition and Its Application in iPhone*, 2013
- [3] Shardul Agravat and Prof. Gopal Pandey, *Human Computer Interaction using Computer Vision*, 2013
- [4] Chucai Yi, Roberto W. Flores, Ricardo Chinchá and YingLi Tian, *Finding objects for assisting blind people*, 2013
- [5] Syamimi Shamsuddin, *Initial response of autistic children in human-robot interaction therapy with humanoid robot NAO*, 2012
- [6] Michal Ciesla and Przemyslaw Koziol, *Eye Pupil Location Using Webcam*, 2012
- [7] Evan A. Suma, *FAAST: The Flexible Action and Articulated Skeleton Toolkit*, 2011
- [8] Zhi-geng Pan, *A real-time multi-cue hand tracking algorithm based on computer vision*, 2010
- [9] Siddharth Swarup Rautaray and Anupam Agrawal, *A novel human computer interface based on hand gesture recognition using computer vision techniques*, 2010

- [10] David Scherfgen, Rainer Herpers and Timur Saitov, *Single webcam 3D tracking for video games*, 2010
- [11] Gary Bradski and Adrian Kaehler, *Learning OpenCV*, O'Reilly, 2008
- [12] P. Eisert, P. Fechteler and J. Rurainsky, *3-D Tracking of Shoes for Virtual Mirror Applications*, 2008
- [13] Jonathan Kofman, Xianghai Wu, Timothy J. Luu and Siddharth Verma, *Teleoperation of a Robot Manipulator Using a Vision-Based Human-Robot Interface*, 2005
- [14] Johanna Hoysniemi, Perttu Hamalainen and Laura Turkki, *Wizard of Oz Prototyping of Computer Vision Based Action Games for Children*, 2004
- [15] Andrew W. Senior and Ruud M. Bolle, *Face recognition and its applications*, Chapter 4, IBM T.J. Watson Research Center, 2002
- [16] S. Lenman, L. Bretzner and B. Thuresson, *Computer Vision Based Hand Gesture Interfaces for Human-Computer Interaction*, 2002
- [17] Martin Hunke and Alex Waibel, *Face Locating and Tracking for Human-Computer Interaction*, 1994
- [18] R. Brunelli and T. Poggio *Face Recognition: Features versus Templates*, 1993
- [19] Pranav Mistry, *Mouseless*, <http://www.pranavmistry.com/projects/mouseless/>
- [20] Pranav Mistry, *Precursor*, <http://www.pranavmistry.com/projects/precursor/>
- [21] Pranav Mistry, *Inktuitive*, <http://www.pranavmistry.com/projects/inktuitive.html>

- [22] Pranav Mistry, *Inktuitive*, <http://www.pranavmistry.com/projects/tapuma/>
- [23] Pranav Mistry, *Teletouch*, <http://www.pranavmistry.com/projects/teletouch.html>
- [24] Pranav Mistry, *Teletouch*, <http://www.pranavmistry.com/projects/sixthsense/>
- [25] Fresk, *Computer Vision Interface*, <http://fresklabs.com/projects/computer-vision-interface>
- [26] Virtual Air Guitar, *FreeMotion*, <http://www.virtualairguitar.com/technology/>