Using Hand as Support to Insert Virtual Object in Augmented Reality Applications

Mohamed Sakkari[1], Mourad Zaied[2], Chokri Ben Amar[2]
[1]Computer Science Department
University of Gabes
Faculty of Sciences of Gabes (FSG)
Cite Erriadh 6072, Zrig, Gabes
Tunisia
[2]Research Groups on Intelligent Machines
University of Sfax
National Engineering School of Sfax (ENIS)
BP1173, Sfax, 3038
Tunisia
sakkari.mouhamed@gmail.com, {mourad.zaied, chokri.benamar}@ieee.org

ABSTRACT: *Increased reality computer is an emerging field, which is progressing sharply. Its principle is to mix the real world with the virtual world. The real corresponds to concrete scenes related to a specific environment, whereas the virtual corresponds to synthetic scenes constructed by the computer with no real existence. It tries to encourage the user to interact with the real scene as if the virtual objects added are there for real. The augmented reality then aims at increasing our perception of the real world by adding virtual objects. It helps the user to execute new tasks in the real world in an innovative way. Therefore, to look realistic, the virtual objects must be properly aligned with real-world objects. In this work we have presented a system of augmented reality without markers. To achieve this augmentation, we have used the hand of the user. The system then has to identify it to ensure the realization of the augmentation.*

Keywords: Augmented Reality, Virtual objects, Hand tracking

Received: 8 August 2011, Revised 23 September 2011, Accepted 30 September 2011

1. Introduction

The Augmented reality [1-2, 3] aided by a computer is rapidly growing is an emerging field. Its principle is to mix the real and the virtual world. It aims at increasing our understanding of the real world, adding fictional elements, not visible a priori. According to Ronald Azuma is “an environment that includes both virtual reality and real-world elements. For instance, an AR user might wear translucent goggles; through these, he could see the real world, as well as computer-generated images projected on top of that world.” [4].Augmented reality refers therefore to the different methods to embed with a realistic way and in real-time [5] virtual objects in an image sequence. The increase or the augmentation in the real scene is an operation of image processing that is done in the background of a real video stream.

In augmented reality applying the knowledge of the position and orientation from the perspective of the camera (or camera pose) is essential because it ensures those of the increased or augmented scene. Indeed, the pose estimation allows the space-time

coherence to model a virtual camera through which a rendering of the virtual world is made with the same characteristics of the real camera. There by properly aligning the real and virtual worlds. There are two main classes of the pose estimation methods: the methods-based markers and method-without markers.

Methods based markers place artificial targets in the real scene in order to facilitate visual tracking and pose estimation. These markers are particularly simply detected containing codes that distinguish them. Their positions in the benchmark world are known a priori and their pose estimation follows the following scheme: detection of markers, 2D/3D mapping, and finally the calculating of the camera pose.

Methods without markers exploit the existing natural features in the real scene as corners, edges and line segments. The data extracted from the 2D image of the scene are mapped to the 3D data extracted from the 3D model of the scene.

To increase the user's real environment, we must have a camera fitted with respect to a landmark. Its calibration is to determine, geometrically, its optical properties, its position and its orientation. In other word, it is to calculate the pose (3 orientation parameters and 3 positional parameters) of the real camera to make it "coincide" with the virtual camera (the one used for perspective 3D rendering). To do this, the method most used for AR applications inside (prepared environment), is to place in the real scene markers that are used to calculate the 3D coordinates from three specific points recognized by the system. This technology can be used with a simple webcam or mobile phone. The process of increasing based on markers consists of a sequence of operations on each image of the video stream in order to identify the presence of a marker, and then to identify among the different markers loaded into the application (in the context of a multi-markers).

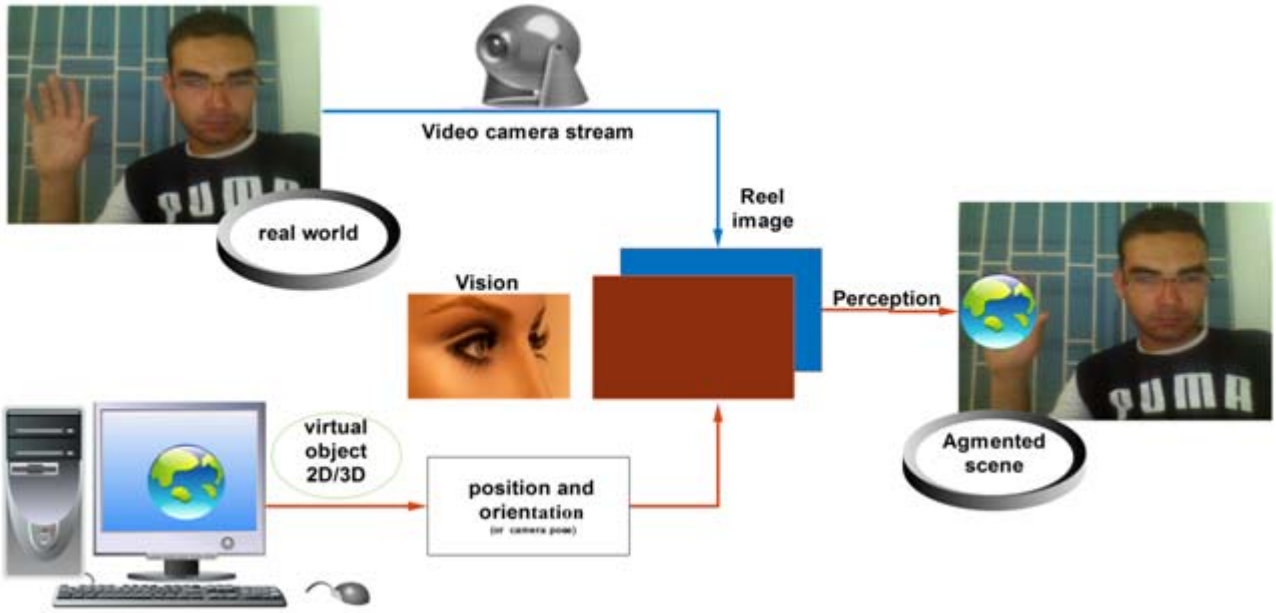

Figure 1 . Principle of increase

Although this technology is the most used and best known for augmented reality applications it has limitations:

• This technology is designed for augmented reality applications inside because it requires a prepared environment, thus a priori knowledge of the environment in which the user is going to progress.

• Another disadvantage of methods based on markers is to be restricted to their zone of visibility. Indeed, a marker can become undetectable when the camera moves away.

• The main disadvantage of this technology is the non-homogeneity and the change of the main characteristics of the scene.

Unlike the approaches based markers, in this work we want to allow the user to test an augmented reality system without being required to have systems of electromagnetic motion capture or markers .it is, therefore, to achieve an augmented reality system based on

computer vision. We are particularly interested in increasing the real scene-based component of this scene. The method that we are proposing is to use the hand of the user to achieve the increase.

1.1 Related work

In [6] the author proposes an augmented reality system based on the user's hand, the proposed system can track the user's outstretched hand and use it as a model for increasing the virtual scene by 3D objects. A hand pose model is constructed in a one-time calibration step by measuring the fingertip positions relative to Each Other in the presence of ground truth scale information. The fingertips are detected using an algorithm based on the contour of the hand. The contour point with a high curvature value is sought as a candidate fingertip point.

A good System of Augmented Reality (SAR) is a system that keeps always a proper alignment between the virtual objects and real ones. This is possible thanks to the proper monitoring of the position and orientation of real objects. The tracking object is then an important phase in an SAR, the work [7] presents "*Flocks of Features*" a fast tracking objects method. The advantages of this method are its speed, its robustness against background noise (background noise), and its ability to track objects that undergo arbitrary rotations and large and fast deformations. In [7] the author shows the performance of "*Flocks of Features*" on the monitoring of the hands with a non-stationary camera in unconstrained environments, indoor and outdoor (in unconstrained outdoor environments).

The technique presented in [8] allows, from a stereo camera, to monitor the 3D position of both hands and face simultaneously and estimate the direction pointed by the head-hand axis. When it arrives into the field of the camera, the face of the user is automatically detected then monitored. Taking the face as a benchmark for the body, the hands are detected and tracked when the user advances them toward the screen. The system doesn't require any training or a specific calibration.

A tracking system with one hand in 3D space has been implemented [6, 8] using a stereo camera. In a ARS, the camera calibration is an important step. This is to extract the necessary information for the increase (augmentation) from scanned images generated by the camera In order to find the reference in the sequence and use it as a reference overlaying virtual objects, it is important to define the geometric correspondences between the real world and the images generated by the camera.

The objects are positioned in a frame associated to the scene. The goal is to determine the point of view [9,10] of the camera (position and orientation) in this benchmark .The Correspondence is to find the various transformations to be used in order to get the coordinates of the object, expressed in a system , specific to the object with its center as origin, to mark the image plane.

The rest of this paper is structured as follows: In Section II, general architecture of the approach is described in detail. In Section III, we show experimental results regarding the speed and robustness of the system. We discuss benefits and limitations of our implemented method. We present our conclusions and ideas for future work in Section IV.

2. Method Description

The proposed approach is to develop a system for augmented reality based on a real component of the scene without introducing a marker. To the increase we use the user's hand. The system must identify it in the first place then redirect to achieve the increase .Figure 2 shows the general architecture of the proposed system.

The process of the augmentation is, first, to identify and isolate the hand in the video stream from camera

2.1 Identification and isolation of the hand

This first step is to isolate the hand in the video. Many detection techniques exist. To identify and isolate the hand the proposed approach is based on the successive use of two techniques: The detection technique of the skin to eliminate the sections that do not have the color of the skin, and that of the neural network to determine areas which correspond to a hand. The main steps are summarized as follows:

- Detection of areas of skin color from an input image.
- Extraction of regions of skin
- Detection of the hand by neural network

2.1.1 Extraction of skin regions

There are several color schemes that can be applied to the detection of the skin; this variability depends on the color space adopted

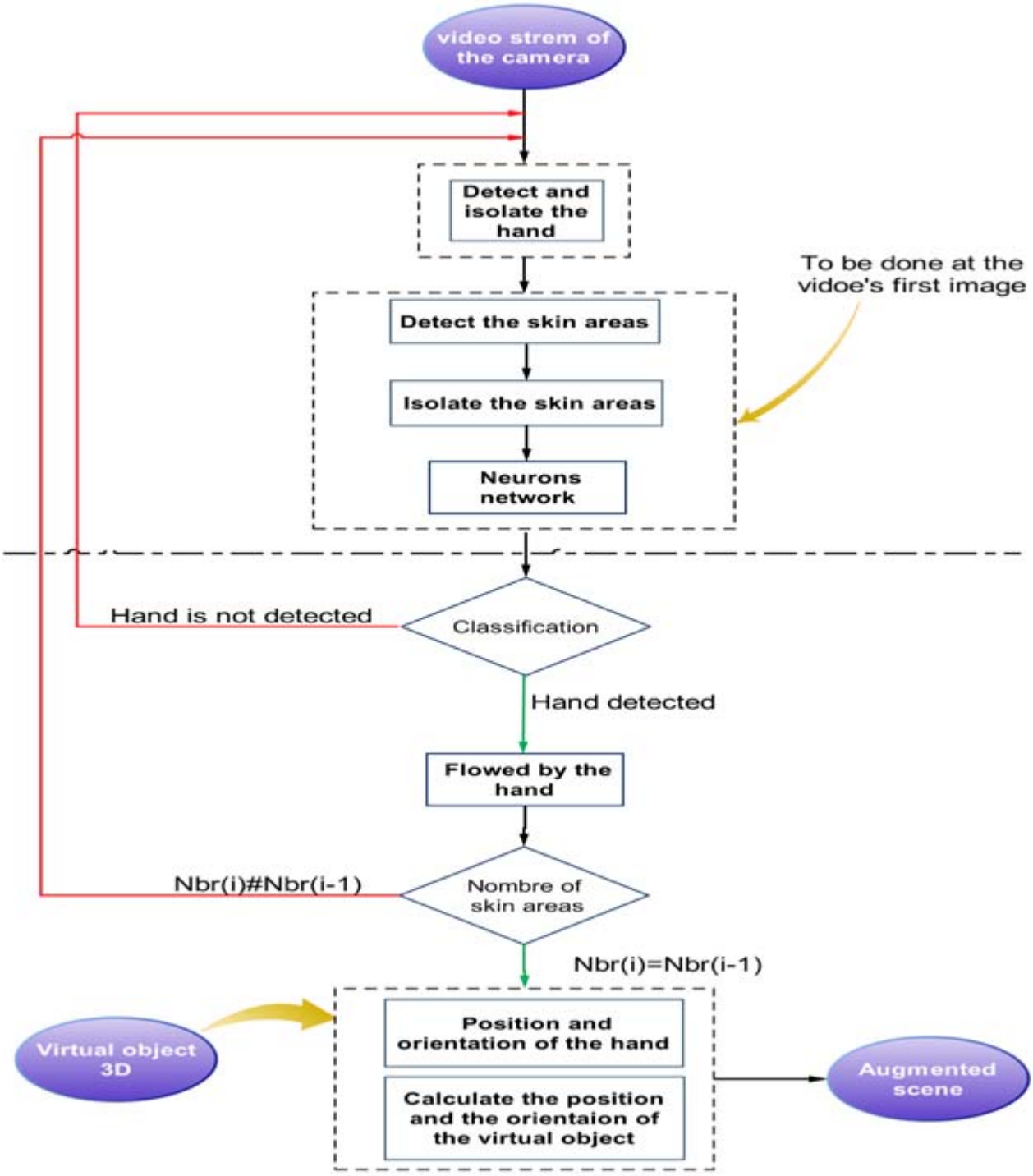

Figure 2. General architecture of the proposed system

for the representation of the chrominance pixels. The most used are: RGB space, normalized RGB, HSV and YCbCr. Several studies have shown that the distribution of skin color is limited to a small area plan chrominance. So we can use this property to detect the colored pixels in the color of the skin. We then select the YCbCr space to detect skin color. The equations of the transformation of the RGB space to YCrCb space are:

- Y = 0.299R + 0.587G + 0.114B
- Cb = 0.500R -0.419G - 0.081B
- Cr = -0.169R – 0.332G + 0.500B

In this space, the distribution area of skin color is fairly compact. Accordingly, we can determine whether each pixel has skin color or not. Detection of skin color using the YCbCr space requires no skin model, it simply checks a series of constraints to decide whether a couple of color (Cb, Cr) is a skin color or not. These constraints are:

$$Cb >= 77, Cb <= 127, Cr >= 133 \text{ and } Cr <= 173$$

A pixel of the input image that verifies the constraints set is part of the class "*skin color*", if not, it belongs to the class "*no skin color*". This phase plays a vital role as it reduces the test region and consequently the execution time is shortened.

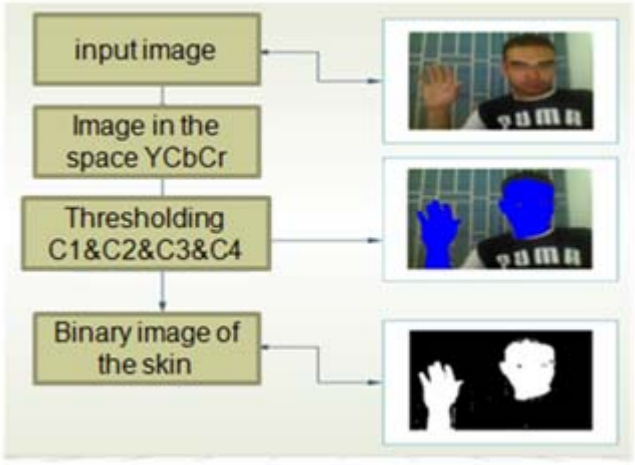

Figure 3. Illustration of the skin color detection process

2.1.2 Identify and isolate the hand

By using the result of the skin detection, we perform the extraction of regions that may contain a human hand. This process is summarized in the following steps:

- The separation of regions.
- The Calculation of the position and the air in each region.

The determination of the position is to determine the coordinates (X, Y) upper left corner of the area in question (denoted M for the first region in Figure 4 and N for the second region) relative to the input image.

To determine the air in the region simply calculate the difference between the coordinates (X1, Y1) of the lower right corner of the region and those (X, Y), as shown by the following formula: Width = X1 – X and Height = Y1-Y1.

This process determines for each Fream, the number of areas of skin colors and the position and size of each area. Each region is defined by two points: the first one is the upper left corner relative to the input image and the second one is the lower right corner.

2.1.3 The hand detection by neuron network

The input image system is a color image (in RGB space). But, to reduce the computational complexity, the input images of the neural network and training images are transformed to gray images. Before treatment of these images by neural network, we apply pre-treatment:

- Resizing of each area of skin color: the regions obtained in the extraction process results in areas of skin are of different sizes.

To be reviewed by the trained neural network, these regions are resized. While adding a strip of black pixels if the size of the region is less than 8000 (image size of 100 * 80 = learning), or by eliminating pixels well if the size of the region is greater than 8000.

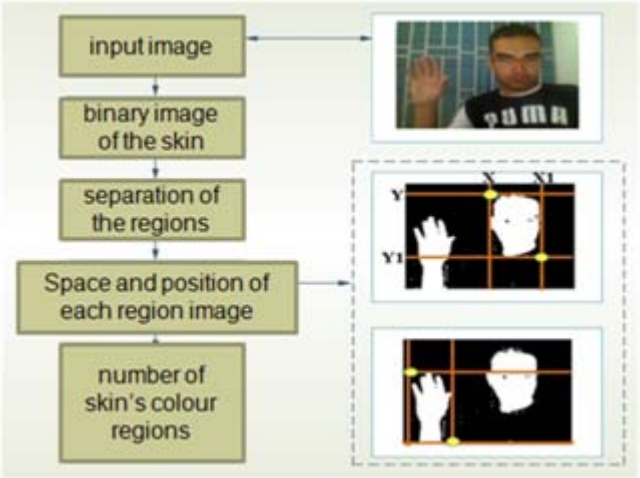

Figure 4. Extraction of skin regions

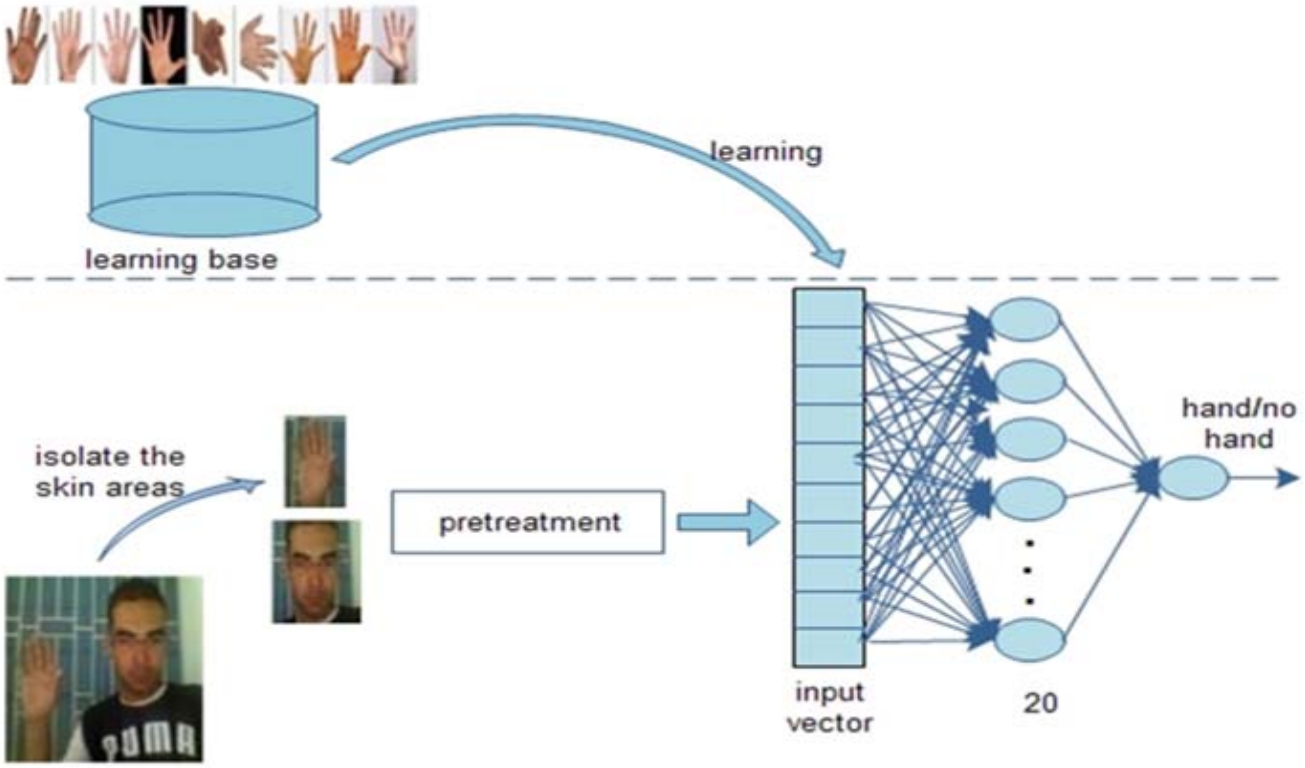

Figure 5. Hand Detection by the neural network

• Equate the histogram: histogram equation is a method of adjusting the contrast of a digital image using the histogram. It involves applying a transformation on each pixel of the image, and thus obtain a new image from Independent operation on each of the pixels. This transformation is constructed from the cumulative histogram of the original image.

The neural network is trained by the training images of size 100 * 80 to reduce complexity and execution time of the learning

phase. Images are resized and subsequently converted to grayscale. The neural network used is a layer formed of three PMC: input layer formed of 8000 neurons each neuron represents a pixel gray level, a hidden layer comprises 20 neurons and an output layer composed of two neurons .The hand detection by neural network is shown in the following figure:

3. Monitoring a hand-based dynamic reference image

To follow the hand in the video stream we propose a following method based on dynamic reference image. The principle of this method is as follows: at time (t = 0) the input image of the video is seen as reference image Im (ref).

The processes of separating the regions of skin color determine the number of regions of skin colors in the first Im (ref).

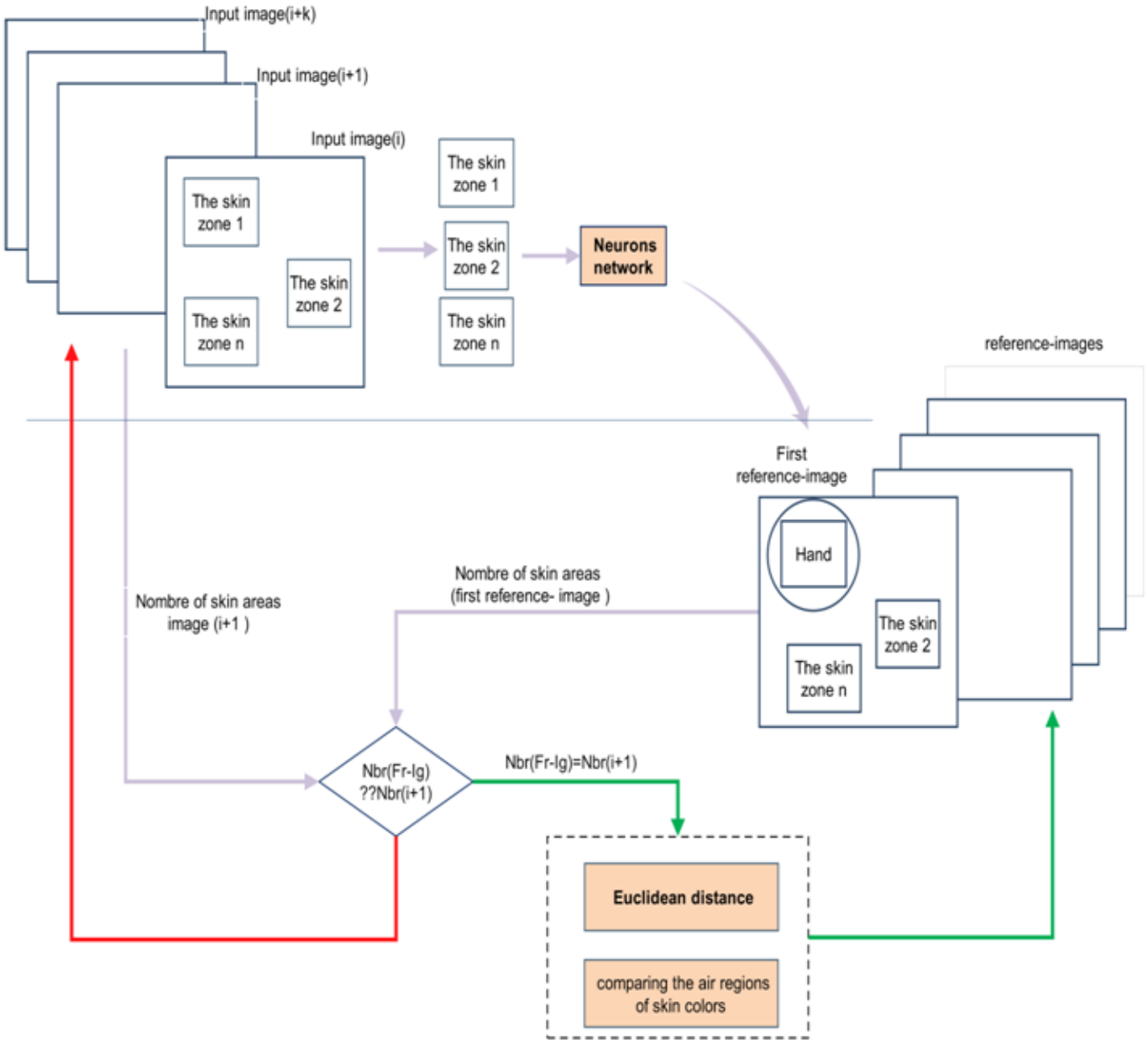

Figure 6. Tracking image-based dynamic reference

It also returns the position of each region relative to the reference image. For the first frame of the video stream neural network technique is called to determine the region that contains a hand.

The reference image is, therefore, defined as follows; Im (Ref): (number of regions, position of each region, each region class (hand, not hand)). At time (t =1) the reference image is no more Im (ref) = Im (t = 0) but Im (ref) = Im (t = 1). Figure 6 shows a diagram explaining the dynamic reference image.

A reference image is defined by the number of regions of skin color. Each region is defined by two points: upper left and lower right corner.

A point is defined by these coordinates relative to the reference image. The region that contains the hand is followed with the method of Kppv: the k nearest neighbor. The method consists of calculating the Euclidean distance between the position of the region containing the hand in the former im (ref) and the positions of the skin color regions detected in the new reference image

To increase the performance of the chosen method, we introduced another imposition to track it. It consists of calculating the difference between the area of the region containing the hand in the old reference image and the air regions of skin color in the new image.

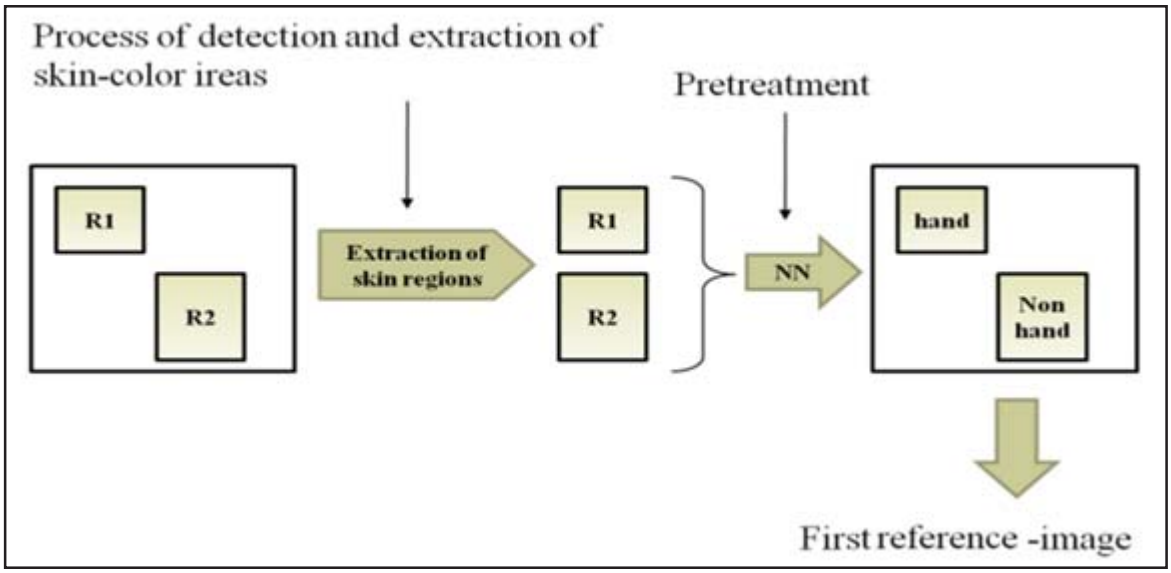

Figure 7. First reference image

For each current image the process returns the number of color area of the skin, the position of each region containing one hand and the positions of other regions of skin color and the air of each.

The current image that contains a hand is a reference one. To determine the hand position in the frame (i +1) the tracking process is to calculate first the Euclidian distance between the hand area in the frame (i) and zones containing a color in the frame (i +1).Euclidean distance is given by:

$$\sqrt{(x - x(j))^2 + (xy - y(j)^2}$$

With:

- j = 1... n, n is the number of zones detected in frame (i +1).
- x, y: the position of the regions that contains the hand detected in the frame (i)
- x (i), y (j): the position of the regions of skin colors detected in the frame (i +1)

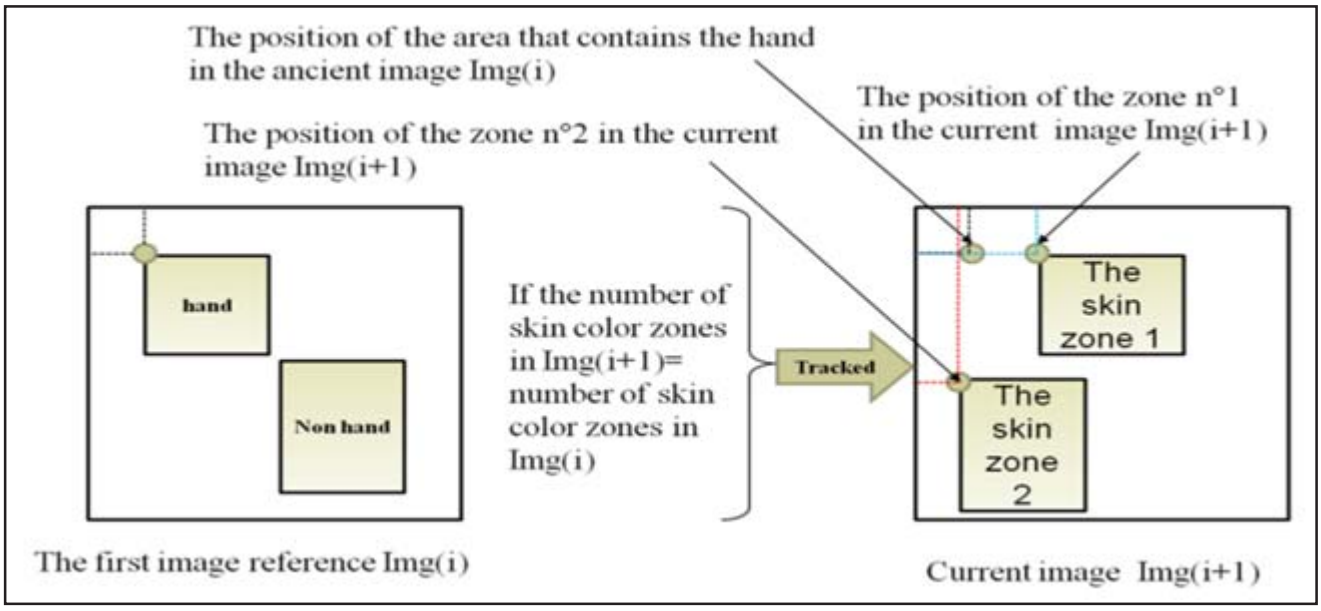

Figure 8. position of the skin zones

Comparing the Euclidean distances to determine the position of the hand in the frame (i +1) is therefore enough. To argue for tracking a second comparison consists in comparing the air regions of skin colors in the frame (i +1) in the air of the region that contains the hand in frame (i).

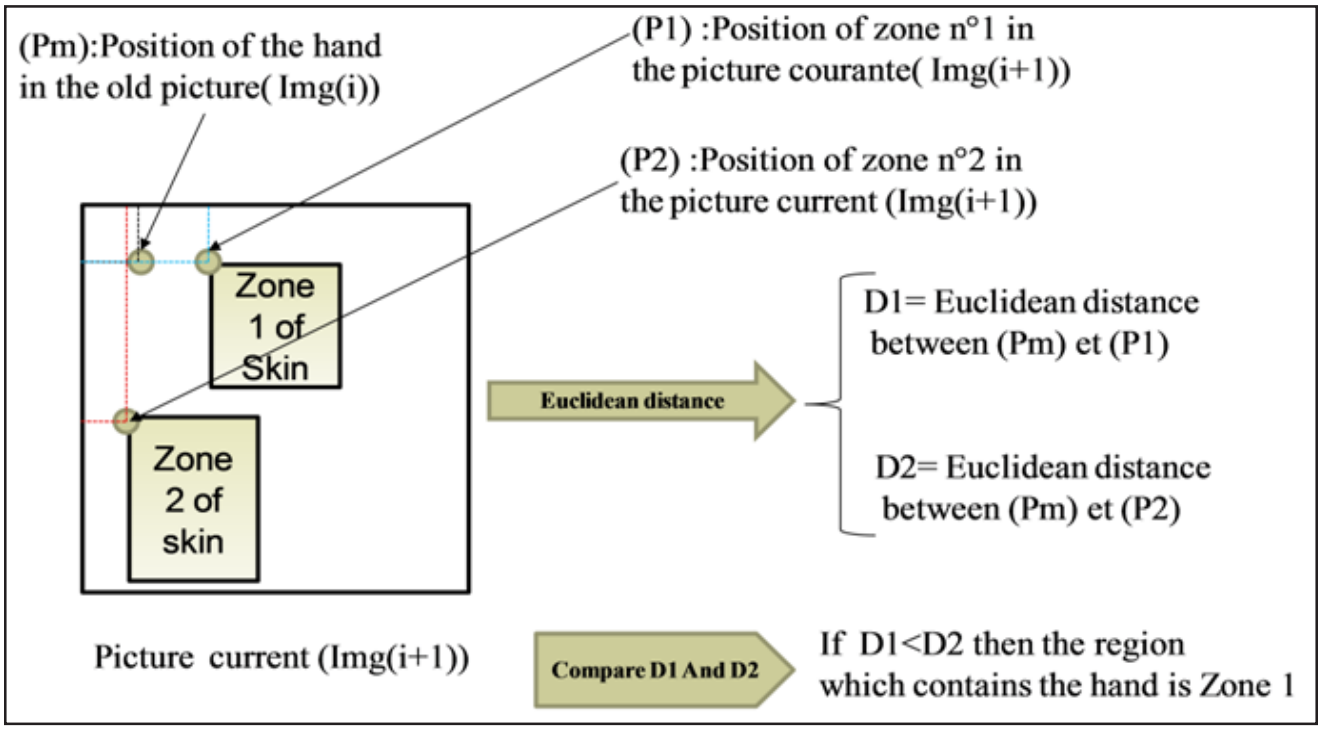

Figure 9. Comparing the Euclidean distance

Once identified and followed the hand is the image (i +1) that is the new reference image. This process is repeated for each frame (i + n). If the number of zones of skin colors detected in the current frame (i+1) is different to the number of zones in the image (i) it is necessary to repeat the hand detection again to see if new hands are present.

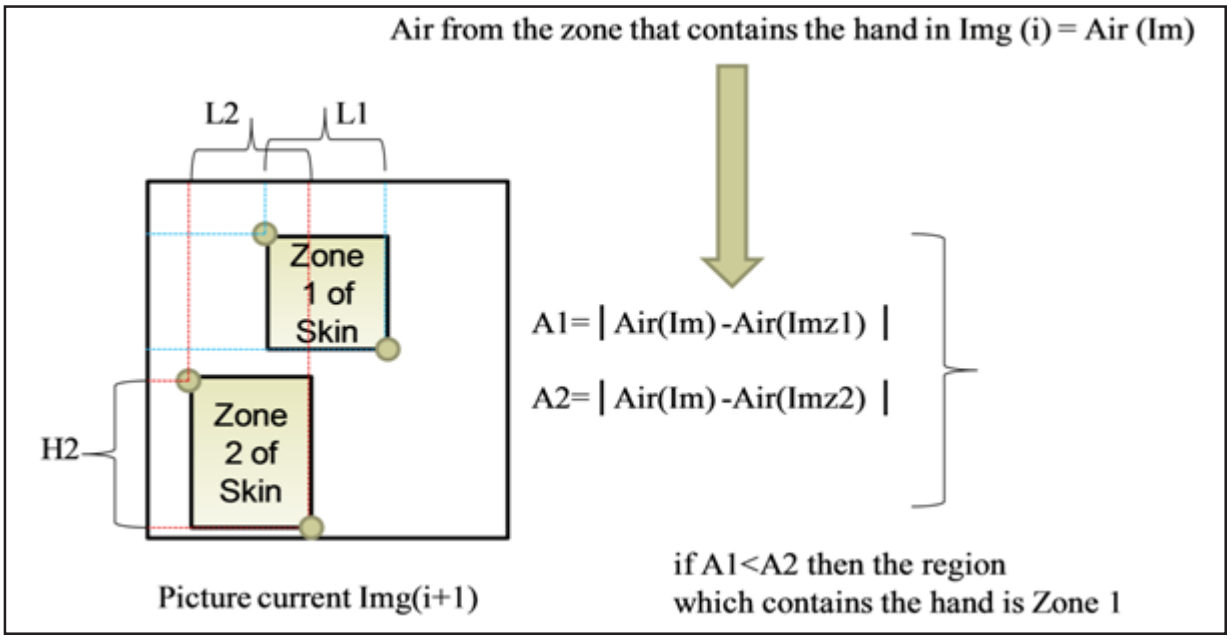

Figure 10. comparing the air regions of skin colors

5. Synthesis of image

This part is to generate 3D virtual objects by computer that will be embedded in the actual scene after identifying the reference and determining the parameters of the camera that data held in the scene. We use this application in virtual objects VRML. 'Virtual Reality Modeling Language "(abbreviated as VRML) is a specialized programming language in the representation of virtual worlds in 3 dimensions. It is an interpreted language and international standard ISO VRML files usually have the extension. Wrl. VRML programs

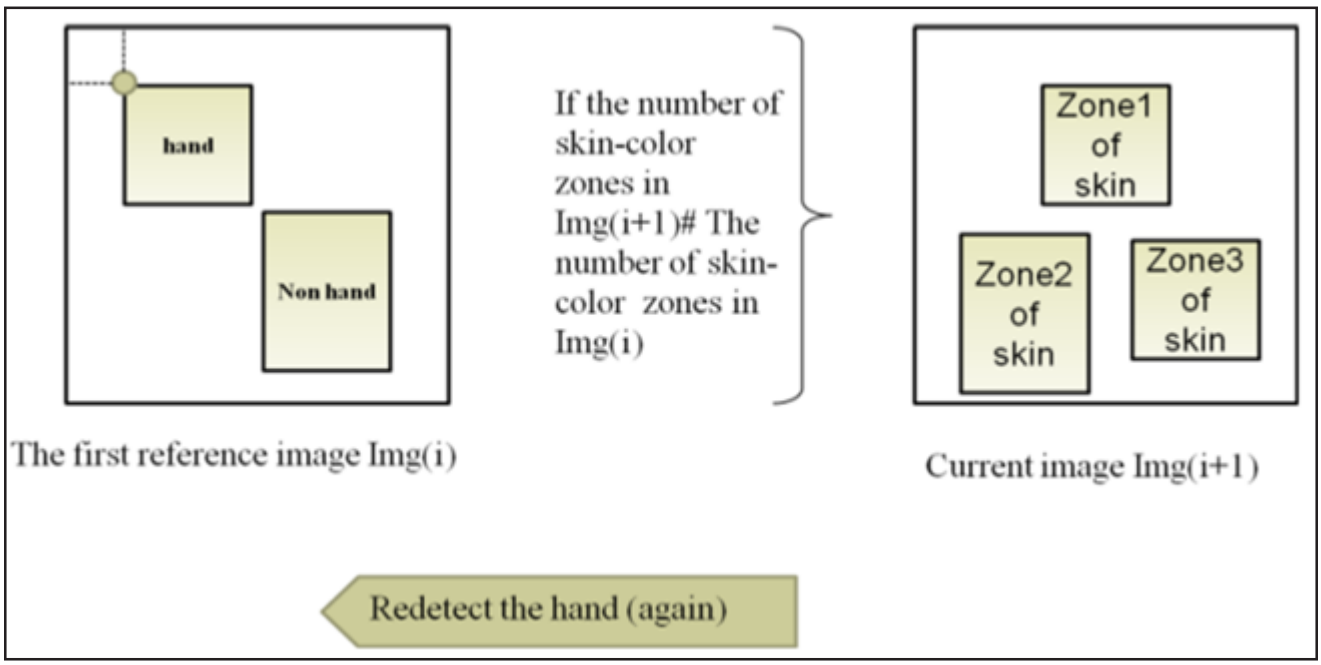

Figure 11. Case 2 number of zones of skin colors detected in the current frame (i +1) is different to the number of zones in the image (i)

can describe simple shapes (points, lines, polygons) or complex (spheres, cubes, cones, cylinders ...), text, images, animations, lighting, sound, hyperlinks, and their arrangement in space, their texture, their color.

The aim of the proposed system is to allow the user to test the no markers augmented reality and to keep at any time the interaction between the user and the virtual object embed in the scene. The interaction of the object of synthesis and the user is as follows:

- The first step involves determining the center of gravity of the region that contains a human hand.
- The second step aims at embedding (insert) the purpose of synthesis in the scene on the hand component. It is therefore sufficient to increase the real scene by the virtual with the following settings:

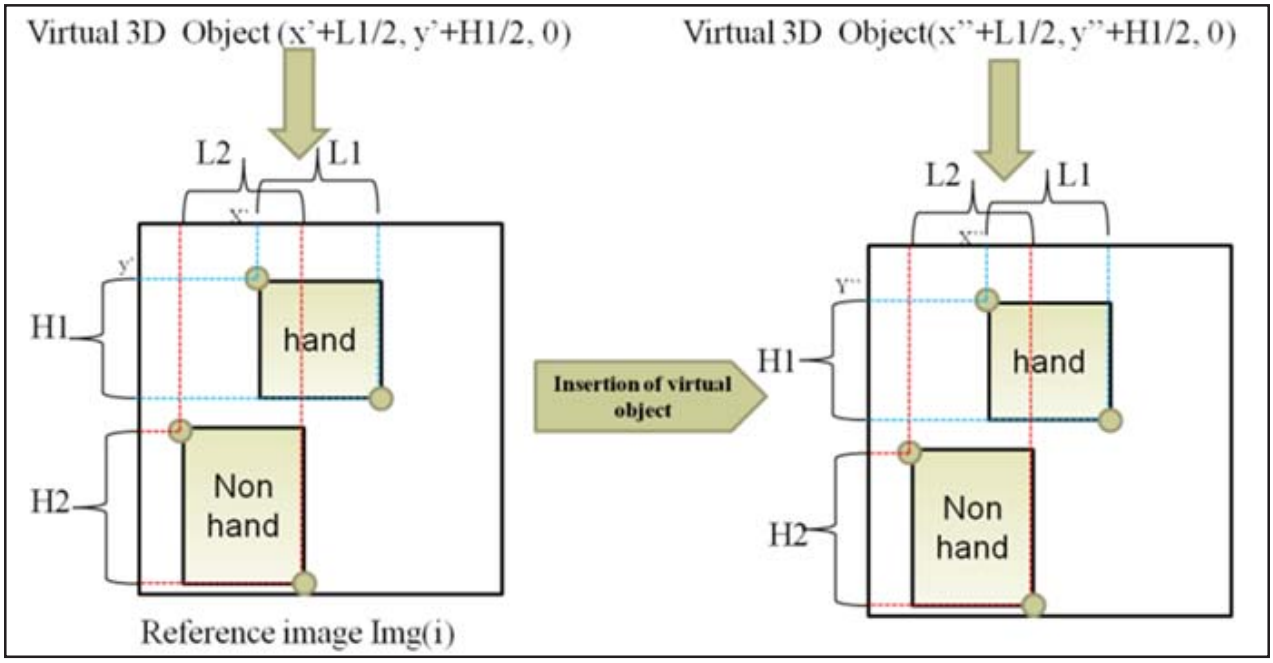

Figure 12. Insertion of the virtual object

The manipulation of the virtual object consists on, therefore, translating the object from its former position relative to the image to the new position in the current frame i +1. The error rate of Class 1 and 3 shows the robustness of the skin color detecting method. Even with images with colors very close to skin color method shows its effectiveness and the overall error rate is still acceptable.

5. Results

5.1 Detection and extraction of regions of skin color

The principle measurement of detection and extraction of skin color regions is to a sequence of test operation on various classes of test images. Three classes of test are available for tests: C1 contains images of simple background that is to say, a background with a single color different from the color of the skin, a second class C2 contains images with background similar to skin color, and the third class C3 contains images with textured background (randomly selected). The test set is formed by 100 frames for each class. The performance evaluation of detection and extraction of the color of skin is to calculate the error rate provided by each test on these three classes. Thus we define three types of detection: A detection low (not acceptable), an acceptable detection, and a good detection. Denotes a low detection sensor that provides an error rate above 30% class. The detection is acceptable if the error rate is between 15% and 20%. A good detection is the one that has an error rate class by less than 15%. The error rate detection and extraction of regions of the color of skin is given by the following formula:

- Error rate per class = (number of images, including detection of skin colors is 100%) / (number of images per class)

Detecting skin color is 100% if the number of areas returned by the process of detection and extraction of skin color is equal to the true

number of regions in the image. An overall error rate is defined as:

➢ Overall error rate = (number of images whose detection is low) / (number of images of the test base).

The following table shows the results obtained after a series of test operations on the three classes based testing and error rates

	images with bad detection	images with correct detection	Error rates per class
Class1	2	98	2%
Class2	21	79	21%
Class3	12	88	12%
		Over all error rate =11.33%	

Table 1. Error rate

The error rate of Class 1 and 3 shows the robustness of the skin color detecting method. Even with images with colors very close to skin color method shows its effectiveness and the overall error rate is still acceptable.

5.2 Evaluation of hand detection per neural network

Learning neural network is started with a base of 100 images. Each size is 100 * 80 pixels. 70 images from the entire training set are a human hand at various conditions: hand in several forms, with different lighting, with shade and without shade. The second class contains images with a color close to skin color with 15 images containing faces and 15 containing human body parts, wood ...

An initial performance test is done on this basis to calculate the estimated rate of correct classification and misclassification rates. The classification rate is given by:

➢ Classification rate = number of detected hand) / (total number of hand in the mage).

The misclassification rate is given by:

➢ Rates of poor classification = (number of classification incorrect) / (total number of detection).

The number of incorrect classification is the number of objects classified as hand but are not hands.

If the number of hands in the image is 0 and the number of misclassification than 0 then the misclassification rate is equal to100%.

The first test is made on 25 frames results achieved a misclassification rate equal à100.pour improved the classification performance it is necessary to increase the training set and each time we repeat the test again. At each addition of training images in calculating the misclassification rate and the rate of correct detection. For the first increase in the training set was added 20 pictures and then we repeat this process a good rate until we obtain the detection .the tests are therefore to 100.120, 140, 160, 180 training images.

The results obtained after 5 tests on the same basis show the learning performance for 180 training images. The poor detection rate is becoming lower and the rate of correct classification is based on raising the number of training images.

We define the overall rate of correct detection as follows:

➢ Overall rate of correct detection = (number of images with good detection rate = 100%) / (number of test images)

The following table shows the change in overall rate of correct detection as a function of number of training images:

number of images of the test base	100	120	140	160	180
Overall rate of correct detection	44%	48%	60%	72%	83%

Table 2. Overall rate of correct detection a function of number of training images

5.3 AR with scoring mark vs. AR without scoring mark

In this section we compare the technique of AR without markers through our proposal and markers for the AR using ARToolkit [11]. The table below shows the results of increasing the real scene by virtual objects obtained through ARToolkit and our proposal keeping the same test condition changing each time the distance between the camera and the hand for our application and between the camera and Markers in the case of ARToolkit.

distance in cm / methods		10	20	30	40	50	60	70	80	100	120
ARToolkit	rate	1	1	1	1	1	1	1	0	0	0
Our App.		1	1	1	1	1	1	1	1	1	1

Table 3. Performance increase

The increase rate is 1 when the scene is enhanced by an object; but it is 0 when there is no enhancement.

Figure 12. Insertion of the virtual object

6. Conclusion and perspectives

In this work we presented an augmented reality system without a marker. To the augmentation we appeal to the user's hand. The system must therefore first identify it and then track it to achieve the augmentation.

To isolate the hand in the video, a process of detecting the hand in the scene is proposed and it is made up of three parts: the skin color detection in YCbCr space, the separation of these video regions and the treatment of these areas by a neural network checked to determine the region containing the human hand.

After detecting and isolating the hand it is necessary to track it at this level. We propose a tracking method of a dynamic reference image based.

The object following is an important phase in a AR system. Indeed, a good AR system is a system that maintains at all times proper alignment between the virtual objects and real objects in the scene. For each movement made by a user of the virtual objects system

must follow the position and orientation of the real objects in the scene. This possible, through effective monitoring of real objects in each view. Object following is then an important step in the process of increasing.

Figure 13. Inserting a virtual object on the user's hand

To improve this work we suggest the inclusion of a module allowing us to know the hand's gestures and to let the 3d object interact on the basis of these gestures.

We can use the techniques of speech recognition for augmented reality. We are working now on commanding virtual objects using speech.

8. Acknowledgment

The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB program.

References

[1] Bowskill, J., Morphett, J., Downie, J. (1997). A Taxonomy for Enhanced Reality Systems. Submitted to the International Symposium on Wearable Computing (ISWC '97), IEEE Computer Society .

[2] Azuma, R. T, Baillot, Y., Behringer, R ., Feiner, S., Julier, S., MacIntyre, B. (2001). Recent advances in augmented reality. IEEE COMPUT GRAPH, 21 (6) 34 - 47.

[3] Milgram, P., Kishino, F. (1994). A Taxonomy of Mixed Reality Visual Displays.IEICE Transactions on Information Systems E77-D (12), 1321-1329.

[4] Azuma, R. T. (1997). A Survey of Augmented Reality. Presence: Teleoperators and Virtual Environments, 6(4) 355-386

[5] Jethwa, M., Zisserman, A., Fitzgibbon, A.W (1998). Real-time Panoramic Mosaics and Augmented Reality.On-Line Proceedings of the Ninth British Machine Vision Conference (BMVC).

[6] Lee,T., Hollerer, T. (2007). Handy AR: Markerless Inspection of Augmented Reality Objects Using Fingertip Tracking.11th Int'l Symposium on Wearable Computers (IEEE ISWC). Four Eyes Lab,http://ilab.cs.ucsb.edu/index.php/publications

[7] Kölsch,T.,Turk, M. (2004). Fast 2D hand tracking with flocks of features and multi-cue integration. In Vision for Human-Computer Interaction, p. 158.

[8] Mayol,W.W, Davison, A. J.,Tordoff, B. J., Molton, N. D, Murray, D. W. (2004).Interaction between hand and wearable camera in 2D and 3D environments. In British Machine Vision Conference(BMVC).

[9] DEHAIS,C. (2008).Contributions pour les applications de réalité augmentée.Suivi visuel et recalage 2D. Suivi d'objets 3D représentés par des modèles par points.Doctorat of the TOULOUSE University..

[10] CHALON, R. (2004). Réalité Mixte et Travail Collaboratif : IRVO, un modèle de l'Interaction Homme – Machine. Doctorat of the Lyon University.

[11] Billinghurst, M., Cheok, A ., Prince, S., Kato, H. (2002). Real World Teleconferencing. IEEE Computer Graphics and Applications.

Author Bibliography

Mohamed SAKKARI received the Master degree in Computer Science and Multimidia From the Higher Institute of Computer and Multimedia of Gabes Tunisia in 2011.Assistant in the Computer Science Department,Faculty of Sciences of Gabes, Tunisia.

Dr. Mourad ZAIED received the Ph.D degree in Computer Engineering and the Master of science from the National Engineering School of Sfax respectively in 2008 and in 2003. Assistant professor in the Department of Electrical Engineering of the National Engineering School of Gabes

Prof. Chokri BEN AMAR received the M.S. and PhD degrees in Computer Engineering from the National Institute of Applied Sciences in Lyon, France, in 1990 and 1994, respectively.