KATHMANDU UNIVERSITY JOURNAL OF SCIENCE, ENGINEERING AND TECHNOLOGY
VOL. 6, No. I, MARCH, 2010, pp 11-24

TRACKING FACIAL EXPRESSIONS BY USING STEREOSCOPY VIDEO AND BACK PROPAGATION NEURAL NETWORK

[1]R. Romero-Herrera, F. J. [2]Gallegos-Funes, [2]A. G. Juarez-Gracia, [1]J. López-Bonilla*
[1]Computing Higher School-ESCOM
[2]Superior School of Mechanical and Electrical Engineering National Polytechnic Institute,
Edif. Z-4, 3er.Piso, Col. Lindavista, 07 738 Mexico city,

*Corresponding author: jlopezb@ipn.mx
Received 6 December 2009; Revised 27 February, 2010

ABSTRACT
In this paper we propose a method to tracking facial expressions. A system with two cameras is used to capture stereoscopic video sequences. The frames are acquired and analyzed by matching two stereoscopic frames through a correlation method that performs image processing to obtain a resulting frame, and then it is processed to recognize a human face by using the Viola and Jones (VJ) method. The face is located via the Nitzberg operator and it provides the feature points of the eyes, eyebrows, nose and mouth, which are introduced into a Backpropagation neural network that is capable of learning and classifying different types of facial expressions that make a person, feel such as: surprised, scared, unhappy, sad, mad and happy. Finally, the result of this process is recognition of facial expressions.

KEYWORDS: Facial expression, Backpropagation neural networks; VJ method; Nitzberg algorithm.

INTRODUCTION
The emotions are stages of human experience mainly, as these exert a high impact on cognition, perception and everyday tasks such as learning, communication and even making rational decisions [1]. However, science and technology have ignored emotions when creating systems that are often frustrating for users because the effect has been misunderstood and it is difficult to measure [2]. The computer systems did not include emotions in their design, until few years ago. The main point of this work is to keep looking in order to maximize the interactivity among the user and the computer by seeking and identifying the computer user's emotions so that computer can react to user's feelings or vice verse [3]. This research area has been called "counting cash" and it is in full development at an international level [4].

Before the eighties the problem of face recognition had not received the necessary attention and is usually assumed that the face had already been detected, later came the first algorithms based on heuristics and anthropometric. Later in the nineties the development of detection algorithms faces began to grow up [5], by proposing a variety of techniques, from basic algorithms of edge detection algorithms to high-level compounds using advanced methods of pattern recognition. These detection techniques are three approaches: a) based on **facial features or local features**, which works by seeking certain elements of the face such as eyes, nose, mouth; b) **Holistic-based on image**. In this case the method works with an entire image or specific areas of it which features are extracted and they can represent the object sought;

c) **Hybrid approaches**. These methods use both the local and global information for detection and are based on the fact that the human perceptual system distinguishes both local and global characteristics of the face [5,6]. The mentioned approaches have emerged from different works such as references [7-9], on which information is used like skin color to make the detection. Results are around 90%. Reference [10] uses neural networks to segment the face detection in rates reaching between 77.9% and 90.3% for different network configurations, and [5] employs a Haar basis for feature extraction. TFE (Tracking Facial Expressions) is a system for studying and learning on facial expressions based on face tracking. The objective of the project is to conduct a methodology to recognize the facial expressions of a person (user) through a process of image processing and pattern recognition that describes them in order to learn through a neural network. The proposed method consists on the following stages:

a) Video acquisition. The proposed system captures 24 frames per second with a resolution of 320x320 pixels.

b) Preprocessing. In the preprocessing stage some features of the captured image are improved.

c) Feature extraction. The VJ method locates the face from the frame. Then, with the obtained information, the feature points of frame are found by the Nitzberg operator. These points bring specific areas, such as, the eyebrows, nose and mouth. We process these areas to find features that allow establishing a pattern for each area.

d) Recognition. The data from feature extraction is introduced in a Backpropagation neural network for training the net to recognize the facial expression of user when his face is captured with the cameras.

EMOTIONS

The emotions are reactions to events considered relevant to the needs, goals and concerns of each individual. These can also be defined as emotional psychological changes of behavior with cognitive components such as joy and fear.

People can have emotions depending on changes in the environment, for example, objects that appear or that move unexpectedly, noises and sounds are sharp enough to denote a feeling of fear. The emotions never can be isolated from each other as a person cannot be angry without being completely sad for example, states that a person is experiencing an emotion if this emotion is more present than the other emotions. P. Ekman suggests that there are six basic emotions (Joy, Sadness, Fear, Displeasure, Anger, and Surprise) plus a neutral one is the balance of six emotions. Based on investigations conducted by P. Ekman on the universality of facial expressions has decided to use the mechanism analysis of facial expressions to recognize that emotions are shown in the face of the user [10].

The use of facial expressions to recognize what the user feels and it will let us recognize only a limited number of emotions which are based on researches into the emotions that are considered. Core has decided to use six emotions.

The main problem in affective pattern recognition is to understand the correlation of emotion that can potentially be identified by a computer; these are behavioral and physiological expressions of emotion [11-14]. We can measure physical events, but we cannot recognize a person's thoughts. Research in

recognizing emotion is limited to correlate emotional expression that can be sensed by a computer, including such things as physiology, behavior, and even word selection when talking [15-17].
There is no a definitive model of emotions. Psychologists have been debating for years how to define them [11, 12]. The pattern recognition problem consists of sorting observed data into a set of states which correspond to several distinct (but possibly overlapping, or fuzzy) emotional states. Each emotional state is defined by a set of features [14]. Features may be just about anything we can measure or compute. Therefore, an important part of the pattern recognition process consists of identifying functions of those features which makes the difference of one state from another. Each state in this model is integrated into a larger scheme which includes other affective states that the user can move to and from. The transitions are defined by transition of probabilities. For instance, if we believe that a user in the affective state labeled as "anger" is more likely to make a transition to a state of "rage" than a state of "sadness", we need to adjust the conditional probabilities to reflect it [14-17].

STEREOSCOPIC VISION

People live in a three-dimensional environment, and our perception of space is created primarily by our two eyes, and their subsequent interpretations of what they see on the brain. The lens of the eye projected two images slightly different in our retinas, which are processed by the brain to a spatial representation [18]. Basically, to obtain two images of the same scene from two slightly different viewpoints, we establish points of correspondence between the two images that correspond to a single point of the taken scene by using a simple triangulation method. The distance of this point to the cameras is the process to find the resulting image of Fig.1.

Operation

1. In order to create a correspondence between points of two images, and then to obtain the distance to the camera point, before it is necessary to define the camera model to find the parameters that relate the coordinates of the image in pixels to correspond to any real world system, in metric units (calibration).
2. The earning of points that correspond in both images (matching) is the biggest problem of stereo vision because of its ambiguity, because an image point may correspond to any point in the other. To reduce these ambiguities we use sets of geometric and physical constraints in order to try to reduce the uncertainty. The most used model is the epipolar constraint, based on the geometry of the system, so we can transform parallel chambers, so that corresponding points are in the same line in both images.
3. Once the matching, the depth calculation is immediate. The inconvenience could be that the number of points for which correspondence was found to be low requires an interpolation to the desired depth map.

There are two basic types of matching algorithms:

1. Based on characteristics: It involves the extraction of points of interest in the image (usually edges), which carried out the match itself.

2. Based on area: carry out the correlation of gray levels in windows of different images, whereas in settings points corresponding patterns of intensity should be similar. The advantage of this method is that currently exist devices aimed at image processing capable of carrying out convolution and correlations in real time, with a yield much higher than general purpose processors.

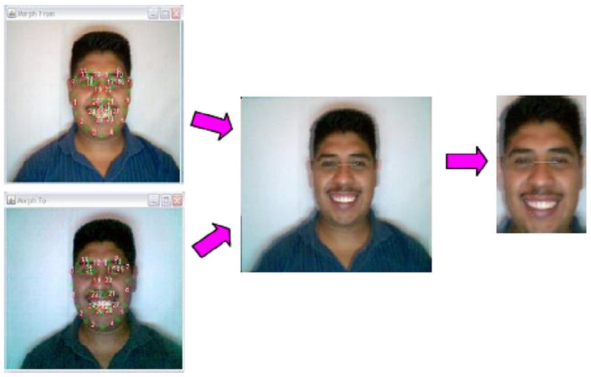

Figure 1.-Obtaining the resulting image of stereoscopic vision

VIOLA AND JONES (VJ) METHOD

The most essential stage of the proposed method is to find the face within the image because it is the region of image which contains the necessary information to recognize the emotional state of user.
There are many different techniques for the detection and location of the face in an image. Most of these techniques have problems such as the detection of a high number of false positives with no uniform background or a high processing time in the analysis of the image. Due to the requirements of real time processing in the proposed method we use the Viola and Jones (VJ) algorithm that let us to detect faces in video sequences of 24 frames per second with a conventional PC.

The VJ method employs a group of image features which contain light and dark rectangles distributed into the image to recognize the face. The main reason for using these rectangular features is that a system based on characteristics operates faster than a system based on pixels [19]. We use three kinds of features: two, three and four rectangles. These rectangular features (RCs- Rectangular Component) are ten different types as shown in Fig.2.

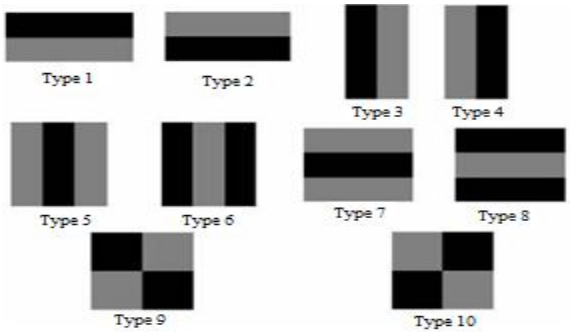

Figure 2. Set of ten rectangular features used in the process of locating faces

The use of RCs is placing them in a certain position within the image and calculates the difference between the amounts of pixels within the light and the dark side of the RC, fixing an integer value that it

must overcome a certain threshold to be considered as being on the facial feature that should be located. This training stage is to find the RCs that in a certain scale and position within a window of size 142x116 pixels, as well as its threshold, can pinpoint a feature of a human face, and this RC receives a name classifier. In Fig.3 we depict 16 classifiers on which each classifier in the figure locates a feature of the face.

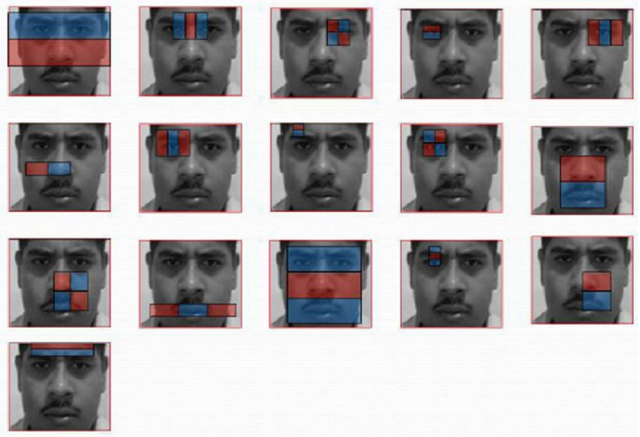

Figure 3. Sixteen rectangular features.

When we have all classifiers with their thresholds, these are placed in cascading to expedite the real time implementation. Each waterfall has a number of binder *n* and has a total of *m* levels of the waterfall. If there are many levels of the waterfall then it will take more processing time.

To determine the RCs very quickly at various scales using a representation of the image called "comprehensive picture". The comprehensive picture can be calculated with a couple of operations per pixel. Once obtained, any of the RCs can be calculated at any scale and location in a time constant. The image in a comprehensive position (*x, y*) contains the sum of the pixels above and left of (*x, y*), including it in the following way,

$$ii(x, y) = \sum_{x' \le x, y' \le y} i(x', y') \qquad (1)$$

where $ii(x, y)$ is the holistic picture, $i(x, y)$ is the original image (see Fig.4), x and y are the coordinates of point in the image, and x´and y´ their derivatives.

Using the next couple of recurrences:

$$s(x, y) = s(x, y\text{-}1) + i(x, y) \qquad (2)$$

$$ii(x, y) = ii(x\text{-}1, y) + s(x, y)$$

where $s(x, y)$ is the cumulative sum of the values of the pixels of the row, $s(x,\text{-}1)=0$, and $ii(\text{-}1,y)=0$, the integral image can be calculated in one pass on the original image [19].

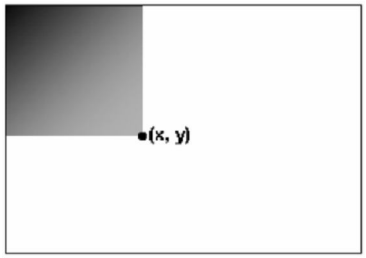

Figure 4. The value of a comprehensive picture at one point (*x, y*) is the sum of all pixels up and left

In Fig.5 the red rectangle shows the position in which the proposed method finds the location of the face with a size of 142x116 pixels by means of use of comprehensive classifiers.

Figure 5. Face Detector

POINT DETECTOR FEATURES

There are methods to derive maximum curvature points in an image directly using the values of self-image, these methods tend to define a measure called "corner" which is calculated for all points of the image. When this measure exceeds a certain threshold is considered that the point is a corner. Most of these methods used differential operators [20].

The Nitzberg's operator

Nitzberg proposes an operator to estimate the magnitude of the gradient and its orientation, as well, as evidence of the presence of corner points. It uses the information in the gradient at a particular neighborhood and combines it with a gradient operator to reduce noise. To calculate the partial derivatives in *x* and *y* throughout of image we use the following expressions [21]:

$$Ix(x,y) = (I(x+1,y) - I(x-1,y))/2$$
$$Iy(x,y) = (I(x,y+1) - I(x,y-1))/2 \qquad (3)$$
$$\nabla I(x,y) = \begin{pmatrix} Ix(x,y) \\ Iy(x,y) \end{pmatrix}$$

Here is defined as follows the 2x2 matrix:

$$Q(x) = \int dx' \rho(x-x') \nabla I(x') \nabla I(x')^T$$
$$Q(x) = \int dx' \rho(x-x') \begin{pmatrix} I^2 x(x') & IxIy(x') \\ IxIy(x') & I^2 y(x') \end{pmatrix} \qquad (4)$$

where $\nabla I(\mathbf{x})$ is a column vector and $\rho(\mathbf{x})$ is a function of decreasing weight with maximum value of 0, $Q(\mathbf{x})$ is the Nitzberg´s operator.

The weights $\rho(x)$ allow us to soften to a greater or lesser value. For simplicity, the role of weighting used in the experiment was selected by the following function [21]:
This descriptor contour allowed to analyze frames taken by a video capture system, which provides the result its frame represents the characteristic pixels of it [22].

$$\rho(x) = \begin{cases} \frac{1}{w^2} si - \frac{w}{2} \le x \le \frac{w}{2} y - \frac{w}{2} \le y \le \frac{w}{2} \\ 0 \quad \text{in other case} \end{cases} \qquad (5)$$

It is an estimate of the location of the eyes, nose and mouth which applies the Nitzberg's operator which helps to get only the most important information shown in Fig.6.

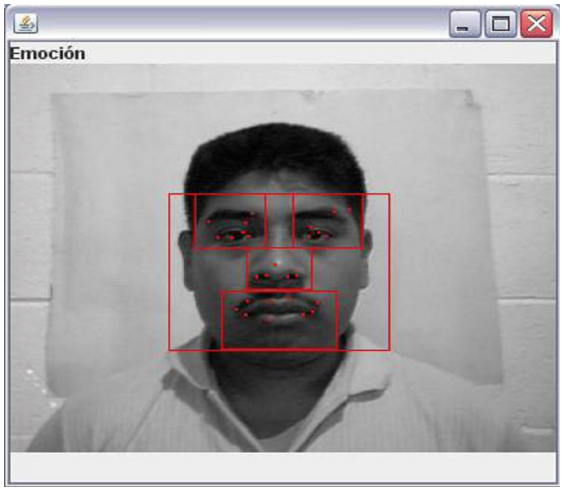

Figure 6. Implementation of Nitzberg's operator

NEURAL NETWORK.

One of the main stages of the proposed method is to train a neural network. The neural networks are computational structures that can be trained to learn series of patterns from examples, and then providing results on new data [23]

Advantages

Adaptive learning: The ability to learn to perform tasks based on a training or experience in an initial.

Self-organization: a neural network can create its own organization or representation of the information it receives through a learning stage.

Fault Tolerance: the partial destruction of a network leads to a degradation of its structure, but some capabilities of the network can be retained, even suffer great damage.

Operation in real time: neural computations can be carried out in parallel, so machines are designed and manufactured with special hardware to get this capability.

They provide a free solution model.

Good capacity generalization.

It may be more efficient than statistical methods.

Appropriate to represent uncertainty.

Disadvantages

Training time can be high.

BACKPROPAGATION NEURAL NETWORK

In the Backpropagation neural network all the signals propagate forward, not backward linkages exist, and usually not self-concurrent, neither side, except in some cases with other types. The structure of the backpropagation neural network is presented in Fig.7 in order to idealize a model for the neural network, we define a Topology network

• Number of layers: less than four (facilitates the convergence of learning). For our application it was used only three layers.

• Size layers: input and output layers.

• Number of connections: connections between neurons in different layers.

The choice of learning algorithm depends on the problem to solve in that case is the recognition of the six basic facial expressions.

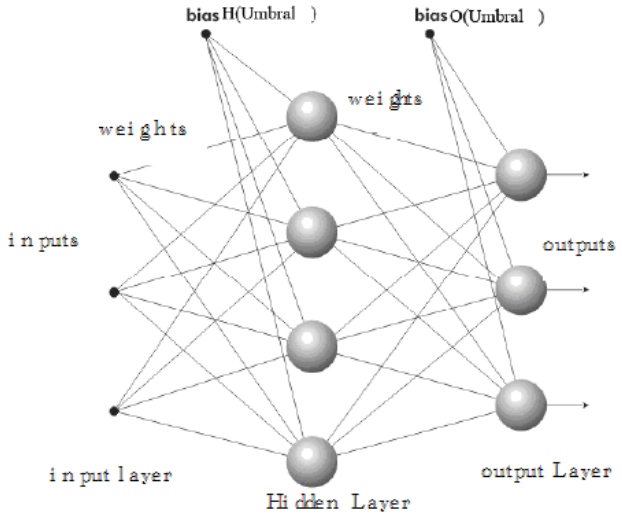

Figure 7. Backpropagation Neural Network.

The architecture for the implementation of learning algorithm is the network type perception multilayer with loop back propagation learning algorithm, which for the resolution of the problem has 22 inputs and 6 outputs as shown in Fig.8. All neurons in the layer directly dependent of the vector input formed by the resulting feature points (see Fig.9) that are obtained by each frame, sorted by left eye, right eye, nose and mouth. The neurons of the intermediate layer depend on the output of neurons in the input layer, and the neurons of the output layer depend of the outputs of neurons in middle layer. This is known as "feed forward".

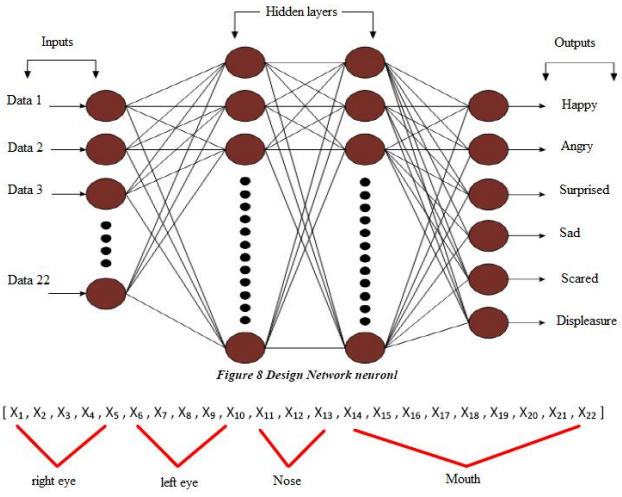

Figure 9. Input vector for Neural Network

By design rule of the pyramid is obtained by an approximation of the number of neurons in the hidden layers, thus the equation applies

$$H_1 = m\, r^2$$

$$H_2 = m\, r \qquad (6)$$

$$r = (n/m)^{1/3}$$

where H1 are Neurons in Layer 1, H2 are Neurons in Layer 2, m is number of network outputs, and n number of network inputs.

Performing the calculations for the first stage gave us an approximation for the first layer of:

r= 2.04, H1 = 20 and H2 = 40

RESULTS

Training stage: In order to train the neural network, we determine the input feature vector. The feature vector contains 4 features: the sum of the points in the right eye, the sum of the parts of the left eye, the sum of the parts of the nose and the sum of the parts of the mouth as shown in Fig.10.

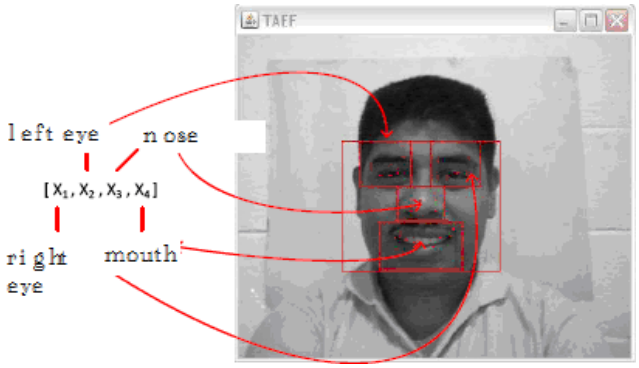

Figure 10. Vector Feature of four entries

Then, it became a study of points obtained in a series of 1000 frames per expression to normalize the number of points to take for each feature of the face, where they take a tabulation of points obtained by expression and determine the number of points to take for each feature. Test stage Once trained the neural network, took place a series of tests which were satisfactory, Fig.11, the percentage of false positives is 45% at each stage of rectangular features, so it was performed an analysis of expressions and see that there is great overlap between Fig.12 words each, so the neural network was not able to solve the problem by 100%, which led us to propose a new methodology for obtaining better results mentioned later. Finally, we discussed a sequence of frames, Fig.13, which was observed as the percentage varies from one of every emotion and it is the one that prevails because emotions cannot be isolated.

Figure 11. Final results for the identification of facial expressions (Happy, Scared, Sad).

Figure 12 Final results for the identification of facial expressions (Displeasure, Angry, Surprised)

Figures 11 and 12 show, that is impossible to isolate an emotion. However, one can observe the predominant expression on his face and in the graphs.

He could think that some expressions have similar faces, for example, fear and surprise. But it is also true that two expressions may have the same emotional intensity and is reflected in the faces and the bars.

Table 1. % detection face

	Person 1	Person 1	Person 1	Person 1
Number of Images	25	25	25	25
Number of correct recognitions	24	24	23	23
Number of faults	1	1	2	2
% detection	96 %	96%	92%	92%

Table 2. % detection face expressions for happy.

	Person 1	Person 1	Person 1	Person 1
Number of Images	25	25	25	25
Number of correct recognitions	20	20	20	20
Number of faults	5	5	5	5
% detection	80%	80%	80%	80%

It reached a face detection rate of 94% on average, Table 1. As far as the recognition of face expressions results by above are obtained of 80%, Table 2. This depends on the training of the neuronal network.

CONCLUSIONS

The high detection rates and low times processing show the effectiveness of the combination of techniques employed in the recognition of affective states. The facial expressions are the result of the combination of different emotional states, where two states may have the same intensity and in some other cases a single state may be the preponderant. The use of two cameras improves the results obtained with one camera, especially for tracking face, as is usually done.

REFERENCES

1. R.W. Picard 1995. “Affective Computing”; http://affect.media.mit.edu/publications.php

2. J. Klein. R.W Picard and J. Riseberg, 1997. “Support for Human Emotional Needs in Human Computer Interaction.” In notes of “Toward and HCI Research and Practice Agenda Based on Human Needs and Social Responsibility” Workshop Proc. CHI Conf. on Human Factors in Computer System, Atlanta, GA.

3. R.W. Picard, 2001 “What does it mean for a Computer to “Have” Emotions?” Chap. in “Emotions in Human and Artifacts.” Ed. by R. Trappl. P. Petta and S. Payr.

4. R.W. Picard, E. Vyzas, and J. Healey, 2001. “Toward Machine Emotional Intelligence: Analysis of Affective Physiological State,” IEEE Transactions Pattern Analysis and Machine Intelligence, 23, No.10, pp.1175 – 1191.

5. Ramírez, Carlos Alejo, Pérez, Manuel David, 2009. Detección de caras y análisis de expresiones faciales. Univ. de Sevilla.

6. http://www.sav.us.es/formaciononline/asignaturas/asigpid/apartados/textos/recursos/deteccionfacial03/doc.doc

7. Lecumberry, R, Federico, 2005. Cálculo de disparidad y segmentación de objetos en secuencias de video. Tesis de Maestría en Ingeniería Eléctrica; Universidad de la República de Montevideo, Uruguay.

8. Li, Yadong, Goshtasby, Ardehir y García, Oscar, 2000. Detecting and Tracking human face in videos, Wright State Univ.

9. Ferris, Rogério, Emidio de Campos, Teófilo y Marcondes, Cesar 2000. Detection and Tracking of facial features in video sequences. Lecture Notes in Artificial Intelligence, vol. 1793, 197-206.

10. Rowley, Henry, Baluja, Shumeet, and Kanade, Takeo, 1998. Neural Network – based face detection; IEEE.

11. Paul Ekman and Wallace V Friesen, 1969. “Unmasking Face: A Guide to Recognizing Emotions from Facial Expressions”.

12. A. Pease, 1981. Body Language, Sheldon Press.

13. T. Dalgleish, M. Power, 2000. Handbook of Cognition and Emotion, Wiley.

14. R.W. Picard, http://web.media.mit.edu/~picard/, Massachusetts Institute of Technology, 2007

15. R. W. Picard, 1997. Affective Computing, MIT Press, 1997.

16. Group of Emotive Calculation, http://affect.media.mit.edu/, Massachusetts Institute of Technology, 2007

17. Group of Emotive Machines, http://www.emotionalmachines.com/, MIT, 2007

18. Computer Science and Artificial Intelligence, http://www.csail.mit.edu/index.php, MIT, 2007

19. http://www.mendozajullia.com/papers/Visi%C3%B3n%20estereos%C3%B3pica.pdf

20. Viola, Paul y Jones, Michael 2001. “Robust Real-Time Object Detection”, Int. Conf. on Statistical and Computational Theories of Vision –Modeling, Learning, Computing and Sampling, Vancouver, Canada.

21. Cazorla Quevedo, Miguel Angel 2000. “Un enfoque bayesiano para la extracción de características y agrupamiento en visión artificial”, Univ. de Alicante, Tesis Doctoral.

22. M. Nitzberg, D. Mumford, y T. Shiota, 1993. Filtering, Segmentation and Depth. Springer-Verlag.

23. http://descargas.cervantesvirtual.com/servlet/SirveObras/01604852658925935210035/003960_3.pdf

24. Martín del Brio Bonifacio, Sanz Molina, Alfredo, 2001 “Redes neuronales y sistemas difusos”. Alfa Omega, Grupo RA-MA. R. Hilera, José y J. Martínez Víctor. “Redes neuronales artificiales”, Alfa Omega, Grupo RA-MA.