Face and Eyes detection to improve natural human-computer dialogue

Joseph Machrouh, Franck Panaget, Philippe Bretier and Christophe Garcia

***Abstract*— In this article, we present the architecture of a visual system able to detect and track a human face in video streaming in a context of human machine natural dialogue. The visual component will be integrated in a complete interaction system which must react fast enough to carry a normal conversation with the user. This system contains a face detection module coupled with a skin-colour tracking component and an eyes detection module. This system has been evaluated on video streaming and several image databases.**

I. INTRODUCTION

Face detection is a major research axis related to applications in the fields of biometrics, surveillance, man-machine interaction, animation and database indexation [11].

Many face detection techniques cannot process frontal faces that appear on a complex background [5] [18]. In realistic applications, faces appear on complex backgrounds and in varying positions. Face detection should achieve its task in different illumination conditions, face positions and camera distances. Some techniques can detect a face in realistic conditions but do not run in real time. In human-machine natural dialogue, the complete system has on average less than 1 or 2 seconds to react to a user's input message.

The goal of our research is to conceive a face detection system able to detect and track a human face in video streaming in a context of human machine natural dialogue. The visual component will be integrated in a complete interaction system which must react fast enough to carry a normal conversation with the user. It is also very important not to miss the user during the interaction. Two constraints have thus been taken into account: minimum computing time and low error rate.

In this paper, section II provides details on our vision component. Section III presents the experimental results. Finally, we present the application of our vision system in human-machine dialogue in section IV, followed by a conclusion where we expose some prospects for our work.

II. VISION COMPONENT

A. Face detection module

We use a face detection module, named "Convolutional Face Finder (CFF)" [8], coupled with a face tracking component.

Joseph Machrouh, Franck Panaget and Philippe Bretier are with France Telecom R&D, TECH/EASY labs, 2, avenue Pierre Marzin - BP 50702, 22307, Lannion Cedex, France (joseph.machrouh, franck.panaget, philippe.bertier)@francetelecom.com

Christophe Garcia is with France Telecom R&D, IRIS Labs, 35512, Cesson Sévigne Cedex France christope.garcia@francetelecom.com

1) CFF: CFF permits to locate the presence of someone in front of the camera (see figure 1). It uses a neural-based face detection scheme to precisely locate multiple faces of minimum size 20x20 pixels and variable appearance in complex real world images. Based on a specific architecture of convolutional neural layers, the system automatically synthesizes simple problem-specific feature extractors and classifiers from a training set of faces, without making any assumptions or using any hand-made design concerning the features to extract or the areas of the face pattern to analyze. A good detection rate of 90.3% with 8 false positives have been reported on the CMU test set, which is the best result published so far on this test set.

Fig. 1. face detection with CFF.

Even if this algorithm offers great accuracy, it is inefficient when the heads are rotated more than ± 30 degrees in image plane and turned more than ± 60 degrees. Moreover, the processing time increases with the number of faces detected in the image. But, in our context of natural human-computer dialogue, it is exactly when a face is detected that the other components of a dialogue system require CPU. In order to solve both problems, we developed another algorithm based on skin colour segmentation. CFF is used to detect a face in frontal position and to allow the skin colour algorithm to locate that face in the following images. In addition to its low computational cost, the skin colour-based face tracking permits to track a face whatever its orientation.

2) Skin colour regions: Most existing face detection systems use histogram colour for segmentation [12]. The skin colour model can be used for face localization [6] [13], tracking [3] and hand localization [1]. Others simply classify according to predefined scales [13]. The main difference between those systems is the choice of colorimetric space. The most used ones are HSV [12] [3] [10], YCrCb [12] [10], I1I2I3 [15] and TSL [19]. According to Terrillon et al. [20] who compared nine different colour spaces for face detection, TSL gives better results. Following Hsu [12], we choose YCrCb as it is perceptually uniform [17], it is widely used in video compression [10] and is similar to TSL for luminance and chrominance separation.

The result of CFF is a rectangle around the face. Using a simple histogram of this area, which enables us to extract

all shades of the detected face's colour, is not efficient enough for two reasons. Firstly, this rectangle contains eyes, eyebrows, hair, and glasses. Secondly, there might exist shade variations due to the camera's noise. We thus propose, on the one hand, to select a sub-area of the rectangle, where the probability of having skin coloured pixels is higher, and on the other hand, to represent skin colour distribution by a two-dimensional Gaussian law.

To avoid possible noise from non-skin coloured pixels, we use a priori knowledge of a human head's proportion and form to determine the area E to pick up skin colour samples (see figure 2a)

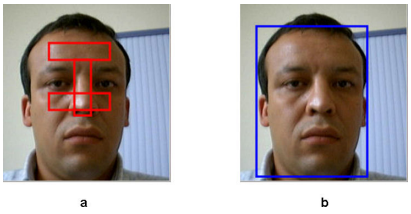

Fig. 2. (a) Skin colour is initialized in the area E, (b) face detection with skin colour.

For the initialization of our model we choose to represent skin colour distribution by a two-dimensional Gaussian law of parameters the mean μ and the covariance matrix Σ of all the normalized pixel components c in the area E.

$$p(c/skin) = \frac{1}{2\pi.|\sum|^{\frac{1}{2}}}.e^{-\frac{1}{2}(c-\mu)^T\sum^{-1}(c-\mu)} \quad (1)$$

$$\mu = \frac{1}{M}.\sum_{i=1}^{M} c_i \quad and \quad \sum = \begin{bmatrix} \sigma_{CrCr} & \sigma_{CrCb} \\ \sigma_{CrCb} & \sigma_{CbCb} \end{bmatrix} \quad (2)$$

Where M is the number of pixels and $c_i = \begin{pmatrix} Cr_i \\ Cb_i \end{pmatrix}$ is the colour vector of pixel i (Cr_i and Cb_i represent the Cr and Cb components of pixel i in YCrCb format) and $\sigma_{xx} = \frac{\sum_{i=1}^{M} x_i^2}{M} - \mu_x^2$ and $\sigma_{xy} = \frac{\sum_{i=1}^{M} x_i.y_i}{M} - \mu_x\mu_y$

With this method, only one scan of the area is necessary.

For the face tracking, we consider a pixel to be skin colored if:

$$p(c/skin) = \frac{1}{2\pi.|\sum|^{\frac{1}{2}}}.e^{-\frac{1}{2}(c-\mu)^T\sum^{-1}(c-\mu)} \geq \lambda \quad (3)$$

which in effect performs Mahalanobis' distance minimization.

Once the skin colour filtering is performed, we determine clusters of connected pixels (connected component analysis). Clusters and holes of area less than 0.5% of the frame area are respectively discarded and filled so that only a small number of clusters are considered for further analysis.

We then extract the characteristics from the clusters (surface, perimeter, compactness, average, variance...) to detect faces (see figure 2b).

B. Eye detection

In man-machine interaction, eye detection is the first step toward evaluation of head orientation and gaze direction [7] [14]. Our aim in eye detection is to recognize some communication gestures such as head nods.

Our approach is as follows: locating the face using the face detection module, estimating the rough position of the eyes and improving eyes localisation using the eye detection module, which operates a processing sequence based on eye region colorimetric specificities. In YCrCb space, the chrominance (Cr, Cb) and luminance (Y) information can be exploited to extract eye region. According to our experiment in many face databases, the area around the eyes has specific colorimetric values. Cb values are higher than Cr ones [12]. Concerning Y values, this area contains both high and low values.

The goal of the process is to accentuate the brighter and darker pixels of the eyes, initially through the chrominance (Cr Cb) and through the luminance (Y) as shown in figure 3.

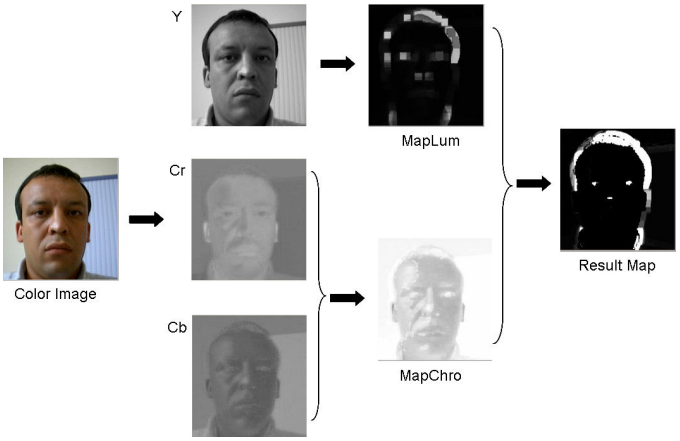

Fig. 3. EyeMap construction procedure.

First, we will try to emphasize eye brightness through chrominance. We note that around eye region we have $Cb > Cr$. This implies $Cr - Cb < 0$, so neg(Cr-Cb) will have saturated values (255) around the eyes.

$$MapChro = neg(C_r - C_b) \quad (4)$$

where $neg(x)$ is the negative of x (i.e. 255-x). And in a second time, we process through luminance Y: we dilate to propagate the high values and erode for the low ones. The division result will have high values around eye region.

$$MapLum = \frac{Dil(Y)}{Ero(Y)} \quad (5)$$

The result map, obtained by the AND operation of the two resulting maps MapChro and MapLum, shows isolated clusters at eyes location. A simple connected component analysis based on pixel connectivity (already performed in Face Detection) is sufficient to determine clusters (or components). Then, we consider the head position, inter-reticular

distance and eyes characteristics (compactness, shape) in order to choose among the different clusters to identify eyes (see figure 4).

Fig. 4. face and eye detection.

III. Results

We have evaluated our system on video streaming and 2 series of image databases:

- the first database, the head pose database[1] [9], consists of 15 sets of images. Each set contains 2 series of 93 images of the same person at different poses. There are 15 people in the database, wearing glasses or not and having various skin colours. The pose or head orientation is determined by 2 angles (h,v), which vary from -90 degrees to +90 degrees.
- The second database is a set of images collected on the World Wide Web, called www database. These colour images have been taken under varying lighting conditions and with complex backgrounds. Furthermore, these images contain multiple faces with variations in colour, position, scale, orientation, 3D pose and facial expression. This base was sorted in 5 subsets according to face pixel size.

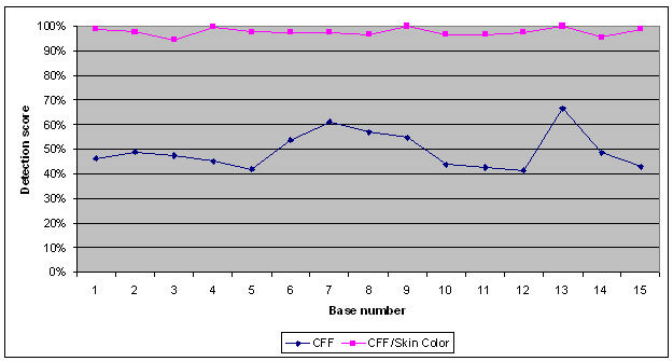

Fig. 5. face detection score rate.

The first test consists in applying the face detection in the head pose database, we show that the use of both algorithms (CFF and skin colour) can detect 98% of the faces when only 48% of the faces was detected when only CFF is used (see figure 5).

The second test consists in applying the face detection module and the eye detection module in the www database. Table I shows the total detection rate according face sizes. We can see that the rate detection is better when the face is larger.

[1]http://www-prima.inrialpes.fr/perso/Gourier/Faces/HPDatabase.html

TABLE I
EYE DETECTION RESULTS ON WWW DATABASE. DR: DETECTION RATE FP: FALSE POSITIVES)

Face size	39 × 38	80 × 91	147 × 166	179 × 204	205 × 250
DR	84.12%	87.04%	93.55%	93.75%	94.87%
FP	3.44%	4.01%	3.54%	3.51%	3.94%

Qualitative results are shown in figure 6. We can see that our algorithm can detect face and eyes despite variations in colour, position, scale, orientation, 3D pose and facial expression.

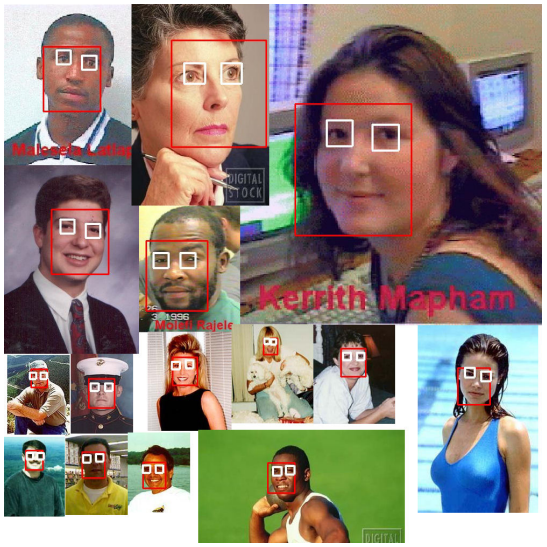

Fig. 6. Face detection and eye tracking results.

IV. Application

The vision component has been integrated into an Embodied Conversational Agent (ECA). An ECA is a virtual human that can dialog with humans by both understanding and producing speech (and written text), gestures and facial expressions. Existing ECAs differ in terms of behaviours and capabilities, but most of them require a fixed stereoscopic camera and/or several CPUs or computers to be able to run in "real time" (see [4] [2]).

Our ECA Nestor [16] integrates a continuous speech recognizer, a natural language interpreter, a dialogue manager, a natural language generator, an avatar player and the vision module. Nestor runs on a single laptop. In combining visual information with data obtained from speech recognition component, man-machine interaction is significantly improved.

For instance, Nestor initiates a dialogue by greeting a user who just appeared in front of the camera.

- *"Hello. My name is Nestor..."*

When the user disappears after Nestor has answered his request, it says:

- *"I can see that you are leaving, I hope you are satisfied with the information I gave you. Bye"*

When the user disappears and Nestor has not yet answered his request, it says:

- *"I can't see you anymore. Could you please come back in front of the camera?"*

And it does not listen anymore (the speech recognition is deactivated) until it sees the user again. If Nestor sees the user before a given time period, it continues the dialogue:

- *"I can see you again. I resume ..."*

If Nestor sees user after a given delay, it starts a new dialogue.

The eyes detection permits Nestor to modulate the dialogue according the face orientation. When the user does not look at the screen, the system closes all the modules and runs CFF to detect another(s) face(s) in the front of the camera. Visual information is also sufficient to adapt the dialogue depending on users' interaction. If the user takes notes, the system will then quote the list of answers slowly in order to leave him sufficient time to note. In the same way if the user looks at the screen without saying anything, the system repeats its question or details more its answer.

Fig. 7. Application of our system in an Embodied Conversational Agent.

V. Conclusion and future work

In this article, we described the architecture of the visual system integrated in our multimodal application. Consideration of the visual context improves significantly the interaction in a context of computer-human natural dialogue. Despite that image processing requires a lot of resources, the improvement justifies the computational time surplus.

Nevertheless, we have seen that we can save some resources without hurting performance when coupling a face detection module with skin colour tracking component.

This architecture has been tested and gives satisfying results with several people in front of the camera, and also on image databases.

In the future, this system will be able to detect the user's gaze direction. The current system can detect gestures but cannot yet recognize them. A learning phase for everyday communication gestures will start soon.

References

[1] S. Ahmad, "A usable real-time 3D hand tracker" *Proceeding of the 28th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, 1995, pp. 1257-1261.

[2] N.O. Bernsen, M. Charfuelàn, A. Corradini, L. Dybkjaer, T. Hansen, S. Kiilerich, M. Kolodnytsky, D. Kupkin and M. Mehta, "First prototype of conversational H.C. Andersen" *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI)*, Lecce, Italy, 2004, pp. 458-461.

[3] G.R. Bradski, "Computer Vision Face Tracking for Use in a Perceptual User Interface", *IEEE Workshop on Applications of Computer Vision*, Princeton, NJ, 1998, pp. 214-219.

[4] J. Cassell, T. Bickmore, L. Campbell, H. Vilhjàlmsson and H. Yan, "More than just a pretty face: conversational protocols and the affordances of embodiment", *Knowledge-based systems*, Vol. 14, 2001, pp. 55-64.

[5] R. Chellappa, C. L. Wilson and S. Sirohey, "Human and machine recognition of faces: A survey", *Proceeding IEEE*, Vol. 83 (5), 1995.

[6] J. Cai and A. Goshtasby, "Detecting human faces in color images", *Image Vision Computing*, vol. 18, 1999, pp. 63-75.

[7] G.C. Feng and P. Yuen, "Multi-cues eye detection on gray intensity image", *Pattern Recognition*, vol. 34 (5), 2001, pp. 1033-1046.

[8] C. Garcia and M. Delakis, "Convolution Face Finder: A Neural Architecture for Fast and Robust Face Detection", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 26 (11), 2004, pp. 1408-1423.

[9] N. Gourier, D. Hall and J. L. Crowley, "Estimating Face Orientation from Robust Detection of Salient Facial Features", *Proceeding of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures* ,Cambridge, UK.

[10] C. Garcia and G. Tziritas, "Face Detection Using Quantized Skin Color Regions Merging and Wavelet Packet Analysis" *IEEE Transaction on Multimedia*, vol. 1 (3), 1999, pp. 264-277.

[11] E. Hjelmas and B.-K. Low, "Face detection: a survey", *Computer Vision and Image Understanding*, Vol. 83, 2001, pp. 263-274.

[12] R.L. Hsu, M. Abdel-Mottaleb and A. K. Jain, "Face Detection in Color Images", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24 (5), 2002, pp. 696-706.

[13] J. Kovac, P. Peer and F. Solina, "Human Skin Colour Clustering for Face Detection" *In Zajc, Baldomir, editor, EUROCON 2003 - International Conference on Computer as a Tool*, Ljubljana, Slovenia, 2003.

[14] R.T. Kumar, S.K. Raja and A.G. Ramakrishnan, "Eye detection using color cues and projection functions" *Proceeding of International Conference on Image Processing*, vol. 3, 2002, pp. III-337-III-340.

[15] P. Menezes, L. Brethes, F. Lerasle, P. Dans and J. Dias, "Visual Tracking of Silhouettes for Human-Robot Interaction", *International Conference on Advanced Robotics (ICAR01)*, vol. 2, Coimbra, 2003, pp. 971-976.

[16] D. Pelé, G. Breton, F. Panaget and S. Loyson, "Let's find a restaurant with Nestor A 3D embodied conversational agent on the web", *AAMAS Workshop on embodied conversational characters as individual*, Australia, 2003.

[17] C. A. Poynton, *A technical Introduction to Digital Video*, John Wiley & Sons, 1996.

[18] A. Samal and A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: a survey", *Pattern Recognition*, Vol. 25, 1992, pp. 65-77.

[19] F. Tomaz, T. Candeias and H. Shahbazkia, "Improved Automatic Skin Detection in Color Images", *Proceeding of VIIth Digital Computing: Techniques and Applications, Sun, C. Talbot, H., Ourselin, S. and Adriaansen, T. Eds*, Sydney, 2003, pp. 419-427.

[20] J.-C. Terrillon, M.N. Shirazi, H. Fukamachi and S. Akamatsu, "Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images", *proceeding of the fourth international conference on automatic face and gesture recognition, IEEE Computer Society*, Grenoble, France, 2000, pp. 54-61.