BIOLOGICAL APPROACH FOR HEAD MOTION DETECTION AND ANALYSIS

A.Benoit, A.Caplier
LIS-INPG
46, avenue Felix Viallet
38031, Grenoble, France
benoit@lis.inpg.fr; caplier@lis.inpg.fr
www.lis.inpg.fr

ABSTRACT

This paper proposes a frequency method to detect head motion. Our method is based on the processing achieved by the human visual system. In a first step, a filter coming from the modeling of the human retina is applied. This filter enhances moving contours and cancel static ones. In a second step, the FFT of the filtered image is computed in the log polar domain as a modeling of the primary visual cortex. Head movements are related to the variation of the spectrum energy. They induce specific variations of that energy : the energy increases in case of moving head and is minimum in case of static head. This yields to an easy motion analysis : motion direction is related to the orientation of the maximum of the image spectrum energy and motion amplitude is related to the amplitude of the total energy spectrum. Moreover, this method allows to detect with reliability all head motion events (slow and fast motions) with the use of a robust temporal indicator which compares the current energy value of the spectrum with respect to the previous ones.
Both head motion detection and analysis are done in real time. No constraint about head motion is needed. The system is working under all type of lighting conditions since the retina filtering is able to cope with illumination variations. As a consequence, our detector is well suited for applications such as non verbal head gestures communication interpretation or vigilance surveillance system of drivers.

I. INTRODUCTION

The aim of the presented work is the detection and the interpretation of rigid head motion for a purpose of high level interpretation of non verbal head "gestures" involved in the human communication process (for instance, head nods are related to approbation). Here, we are focusing on the preliminary step of rigid head motion detection and analysis. Work about 2D motion estimation is generally based on a computer vision approach. Different methods [2] such as differential methods, block matching methods or frequency methods have been proposed in order to estimate a sparse or a dense optical flow to be associated to all the pixels of an image. A comparison of the performances of all these methods is done in [2]. In our opinion, the main drawback of such methods is that they yield to a low level optic flow field which is very difficult to interpret without a post-processing of motion segmentation. Estimation of 2D motion based on parametric models [3] gives a more compact representation of motion but interpretation is not always easy and depends on the model used for the projection 3D/2D [1] [9].
In this paper, we propose a different approach for head motion detection and analysis. The main idea is to develop an approach based on the modeling of the human visual system involving a spatio-temporal filtering step occurring at the retina level and a frequency analysis occurring in the primary visual cortex. This approach yields to the estimation of head motion direction and type and to the estimation of speed variations. As human beings, our method can cope with variations of illuminations and with motions of different amplitude. No hypothesis is needed about the motion type (we can deal with translations and rotations). A temporal integration of the detected head movements is then proposed in order to detect with accuracy the least occurring head motion. Any detected head motion is referred as head motion event. The analysis of the successive head motion events is necessary for a high level interpretation of head «gestures» (cf. several periodic vertical head motions are interpreted as approbation).
In section 2 the retina pre filtering is described. This gives a filtered image where moving contours have been enhanced and where illumination variations have been removed. The log-polar frequency spectrum of the pre filtered image is computed and analyzed in section 3 (modeling the process occurring at the primary visual cortex level). Finally, section 4 proposes a method to detect all head motion events.

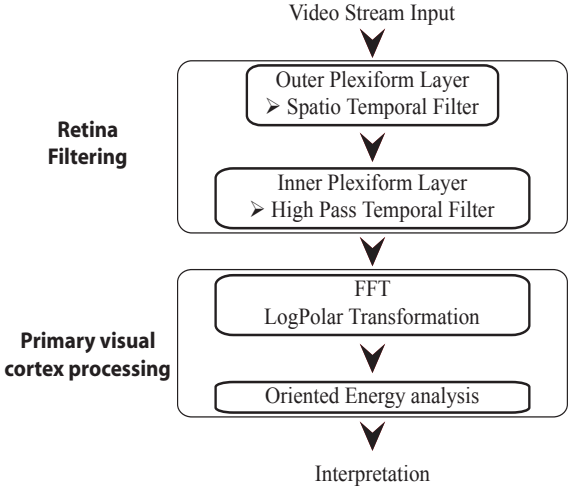

figure 1 : general algorithm

II. RETINA FILTERING

Figure 1 gives a general overview of the algorithm. The processing at the retina level consists in an efficient filtering made of two stages [4] :

- at the Outer Plexiform Layer (OPL), all the treatments are modeled by a non separable spatio-tempral filter (see Figure 2). This filter has a band pass spatial effect in low temporal frequencies which is responsible for contours enhancements. It has a wide band pass temporal effect for low spatial frequencies which smooths illumination variations. It has a low pass effect for high temporal frequencies and a low pass tendency for high spatial frequencies that minimizes spatio temporal noise.

- at the Inner Plexiform Level (IPL), process is dedicated to the detection of moving stimulus. This process is modeled by a temporal derivation operator [5]. As a consequence, this filter enhances moving contours and removes static ones. The amplitude of the contours response at the output of the IPL depends on the contours orientation w.r.t. the motion direction (the optimal case are contours perpendicular to the motion direction) and it depends on the motion amplitude.

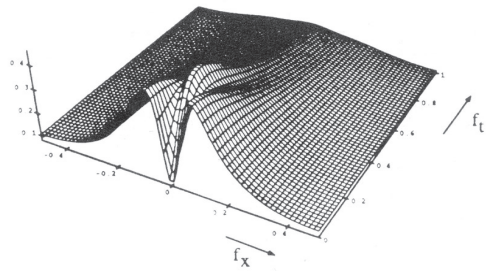

figure 2 : retina B transfer function [4]

Figure 3 illustrates the effect of the retina filtering on a head motion sequence in which the head pans. The OPL filter enhances all contours, attenuates the low spatial frequencies, and minimizes spatio temporal noise. The IPL filter attenuates static contours and enhances only moving ones (especially the contours perpendicular to the movement). As a result, the spectrum of the filtered scene will only report power at the frequencies of the contours involved in the movement.
An other advantage of the retina filter compared with a cascade of classic band pass filters is that process can be achieved in real time. OPL and IPL filterings require only 11 operations per pixel. If we approximate the effects of the OPL filter by a cascade of classic filters, the association of a band pass spatial filter (for contours extraction) and a low pass temporal filter (for spatio temporal noise cancellation) is necessary. Each filter would require at least 9 operations per pixel if standard 3*3 filters are used. Moreover, retina filter takes all the image information into account rather than localized information on small neighborhoods as it is done with classic filters.

Input frame OPL output IPL output

On the Outputs, gray pixels correspond to null response, black and white to negative and positive responses

figure 3 : retina filtering

III. PRIMARY VISUAL CORTEX PROCESSING

At the output of the retina filter, the FFT in the log-polar domain [8] is computed. This models the behaviour of the complex cells of the primary visual cortex which work as 2D oriented band pass filters [6]. The log polar transformation transforms Cartesian zoom in a global energy translation along the frequency axis and Cartesian roll into a global translation along the angle axis. As a result, the interpretation of the spectrum in terms of moving contours orientations analysis and global energy analysis is easy (see section 4). An other advantage of log polar transform is the reduction of computational complexity: a large Cartesian spectrum of size M*N is transformed into a reduced one defined by J angles per K associated frequencies. The more angles and frequencies there are, the more precision we get, but that involves higher computation time. As a compromise, we currently use a 45 angles per 45 oriented frequencies for 150*150 video frame size to get 4° angle resolution and fast computing time.

IV. RIGID HEAD MOTIONS DETECTION AND ANALYSIS.

We are focusing on global head motion estimation. Head motion is made of a rigid motion (global head motion) and of some non rigid motions (blinking, lip motion...). Rigid motions are supposed to be slower than the non rigid ones. Non rigid head motions are removed by tuning the parameters of the retina filter, its band pass temporal filter minimizes fast localized motions.

IV. 1 Motion direction detection

Because of the retina filtering, the log polar spectrum reports the highest energy on the frequencies linked to the contours perpendicular to the motion direction. Figure 4 presents the case of a head rotating in two different directions. The corresponding log polar spectrum shows that the energy is located near the rotation axis orientation, i.e. near the perpendicular of the motion direction.

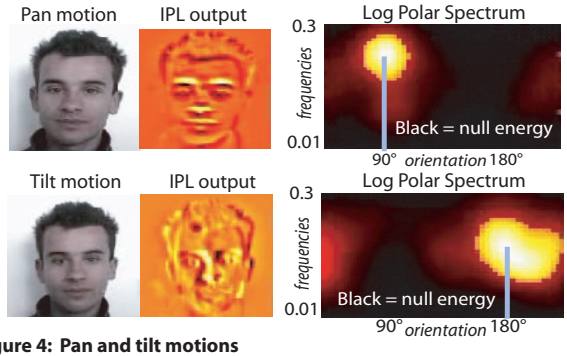

figure 4: Pan and tilt motions
the IPL output extracts contours perpendicular to the motion, their corresponding energies are maximum on the log polar spectrum

In order to estimate the motion direction, we sum the energy of the log polar spectrum for each orientation. This yields to a cumulated energy per orientation curve. On that curve, the abscissa of the maximum amplitude corresponds to the orientation of the most energized moving contours which are perpendicular to the motion direction. Figure 5 gives frames of a synthetic moving head and the corresponding cumulated energy curves. Figures 5-b and 5-c show that a single motion induces a single maximum on the cumulated oriented

energy per orientation curve. This maximum corresponds to the orientation of the displacement. In the case of multiple rotations (figure 5-d), the curve reports two maximums corresponding to the two involved rotation axis to be related to the 2 head main orientations (vertical and horizontal). When achieving a complex rotation, both orientations will report energy even if they are not exactly oriented along the motion direction. This is the well known aperture problem [7]. In our case, it becomes an advantage : in figure 5-d, 2 maximums appear which are related to the 2 rotations occurring at the same time. This complex motion can also be analyzed observing the amplitude variation of each maximum (see section IV.2).

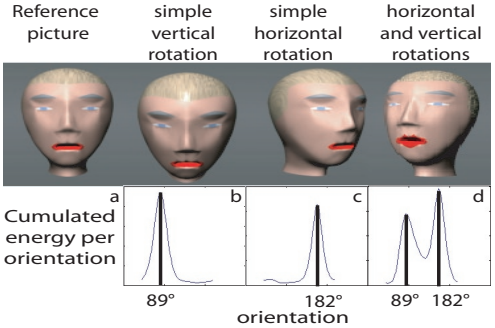

figure 5 : simple and mixed head motion estimation

The precision of the estimated orientation axis is influenced by the angle resolution of the log polar transformation and by the characteristics of the observed object. There is a higher precision if contours oriented perpendicular to the motion direction exist. This is the case with a moving head for its common motions.

IV. 2 Motion speed estimation

In order to get information about the motion amplitude, the total energy of the spectrum is computed by summing all the oriented filters energies. The observation of the evolution of the spectrum energy v.s. the motion amplitude shows that there is a linear relation between speed and total spectrum energy. Figure 6 shows the variations of the total energy of the spectrum when the speed is increasing along the sequence (10 different motions and 5 different speeds).

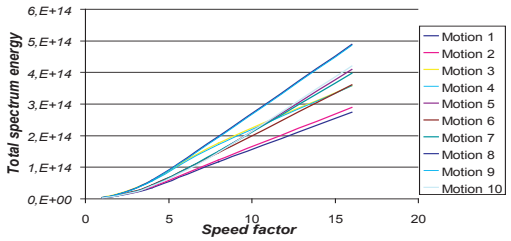

Figure 6 : Evolution of the total spectrum energy with the motion speed

Note that for head motions at very low speed, the energy evolution is no more linear. Since only moving contours contribute to the spectrum energy, this energy is null when no motion occurs.

IV. 3 Temporal evolution of spectrum energy and relative motion detection

Global head motions are set in a large range of amplitudes and speed. But slow motions can be as significant as fast ones in non verbal communication. For example during a conversation, head nods which are low amplitude movements often give more important communication information than other common larger and faster head motions (single head rotations done to look at something for example). The goal is to detect with high accuracy the least head motion. For example, if large motions are first detected, a slower motion occurring just after has also to be detected and has not to be confused with residual noise to avoid false detections. That is why we propose an adaptive motion detector.

The temporal evolution of the total energy of the spectrum is analyzed to detect motions.When motion stops, no energy is reported on the spectrum, there is only residual spatio temporal noise. In the same way, when motion starts, the acceleration creates a global energy increase and deceleration involves a global energy decrease on the log polar spectrum. Figure 7 illustrates the evolution of the total energy of a video sequence in which a person expresses 'Yes' and 'No' with head nods. On the total energy curve, each energy maximum is equivalent to spatial motion maximum speed and each minima reports the motion stops. This example shows that each energy maximum can be different from a nod to the next one, nevertheless, they are as significant in non verbal communication. These nods are linked to a common expression, here 'Yes' or 'No' so they need to be detected as significant before being interpreted.

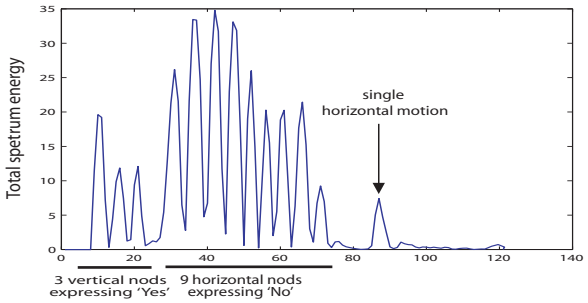

Figure 7: temporal evolution of the total spectrum

IV.3.a Noise level estimation

In order to avoid false head motion detection, we have to cope with residual noise. The noise level Enoise is computed at the beginning of the video acquisition : it is the mean of the residual noise level of the *n* first frames (currently *n*=20) in which no motion occurs.

We consider that the current energy E(t) is related to a significant head motion if E(t) > 2*Enoise. This criteria is not restrictive because even for slow head motions, related energies are more than 10 times the noise level even in noisy conditions. For example, in figure 7 each motion maximum report energy above 10 times the noise level (Enoise=0.15, note that the standard deviation of Enoise is σ_{noise}=0.01).

IV.3.b Motion level indicator

In order to interpret head motions, it is necessary to analyze sequences of head movements rather than motions independently. When a movement occurs, it has to be compared to the previous ones and the comparison must be adaptive to avoid detection errors. For example if an unknown high energy level introduced by a disturbing event is encountered, it has to be lowered to make the motion detector work for the next temporal events.

An adaptive indicator $E_1(t)$ is introduced, it can be considered as the output of an electric analog/continuous current converter applied to the total energy time evolution. $E_1(t)$ reaches each maximum energy value and decreases temporally with an 1/t curve tendency (capacity effect). Figure 8

illustrates this effect. When an energy maximum is reached, the indicator reaches this maximum and decreases slowly as an electric analog/continuous current converter with low pass filter does.
This indicator allows to estimate the current energy reliability level w.r.t. the last motion events (currently 0.5 second between events). If we consider the ratio :

$$\alpha_1(t) = E(t)^2/E_1(t)^2$$

$\alpha_1(t)$=1 when the current energy is high compared to the last energy values and $\alpha_1(t)$=0 when the current energy level is lower than the last energy values i.e. the last motions amplitude.
$\alpha_1(t)$ indicates the reliability level of the amplitude of the current motion compared to the very last motion events. On figure 8, the second graph shows the temporal evolution of the marker α_1. It is minimum when no motion occurs, maximum when motion increases and decreases when motion slows down. The main advantage is that $\alpha_1(t)$ values are only in the range [0; 1] so that thresholding is easy. A threshold level of 0.2 allows to detect all motions, even low and fast motions. The risk of false detections introduced by the noise level is minimized while considering total energy spectrum above 2*Enoise.

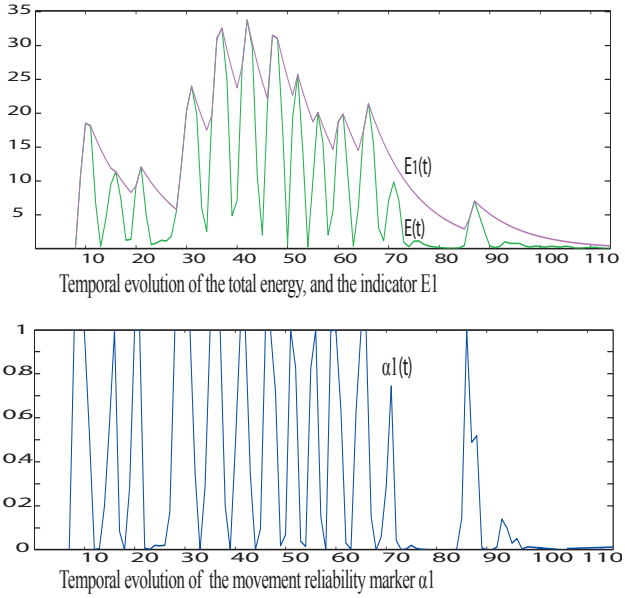

Figure 8 : spectrum total energy temporal evolution during head motions

V. PERFORMANCES AND APPLICATIONS

This motion event detector is currently used in a global head motion analyzer, a first application consists in detecting head nods used to express 'Yes' or 'No'. Such detection requires a reliable nod detector able to detect each nod motion and analyse the corresponding orientations to identify approbation or negation. The performances of this motion event detector have been evaluated in various test conditions : it detects movements events up to 100% success in standard office lighting conditions with the head occupying from 20% to 100% of the captured frame (currently 150*150 pixels). In low light conditions or noisy captured frames (Gaussian white noise of variance 0.05), the algorithm is able to detect the head movements with 90% success.
Finally, the algorithm works in real time, reaching up to 80 frames per seconds on a standard PC desktop Pentium 4 running at 3.0Ghz on which a webcam is installed. The algorithm analysis the webcam output and detects head nods in real time. An other advantage is that no initialisation step is required, the person moves freely in front of the camera, this biological approach shows the capacities of the human visual system to be adaptive and able to cope with various illumination and motion conditions.

VI. CONCLUSION

A real time method for global rigid head motion detection has been proposed. The algorithm inspired from the biological model of the human visual system show its efficiency in terms of motion detection and analysis : the use of the retina filter prepares the data and yields to a log polar spectrum easy to analyze.
An example of real time head motion detection has been described, the orientation of the motion events are detected with the analysis of the log polar spectrum. The proposed algorithm is well suited for global head motion estimation. This motion detector is integrated into a more complete algorithm able to detect motion type and direction of a moving head which yields to head movement interpretation.

VII. REFERENCES

[1] J. Xiao, T. Moriyama, T. Kanade, and J. Cohn. "Robust Full- Motion Recovery of Head by Dynamic Templates and Re-registration Techniques"*International Journal of Imaging Systems and Technology*, Vol. 13, pp. 85 - 94,September, 2003
[2] Barron J.L., Fleet D.J. and Beauchemin S.S., "Performance of Optical Flow Techniques", *International Journal of Computer Vision,* Vol. 12, No. 1, pp. 43-77, 1994.
[3] Odobez J.M., Bouthemy P. "Robust Multiresolution Estimation Of Parametric Motion Models", *Journal of visual Communication and Image Representation*. Vol 6 N°4 pp348-365 december 1995
[4]Beaudot W.H.A., "The neural information processing in the vertebrate retina: A melting pot of ideas for artificial vision", *PhD Thesis in Computer Science*, INPG (France) december 1994
[5] J. Ritcher&S.Ullman. "A model for temporal organization of X- and Y-type receptive fields in the primate retina". *Biological Cybernetics*, 43:127-145,1982.
[6] N.Guyader "Categorisation basée sur des modèles de perception. approche (neuro) computationnelle et psychophysique". *PhD thesis in Computer Science*, INPG (France) July 2004.
[7] A. Torralba, J. Hérault. "An efficient neuromorphic analog network for motion estimation". *IEEE Transactions on Circuits and Systems-I. Special Issue on Bio-Inspired Processors and CNNs for Vision*. Vol. 46(2): 269-280. 1999
[8] Oliva A., Torralba A.B., Guérin-Dugué A., Hérault J., (1999) "Super-Ordinate representation of scenes from power spectrum
shapes", *CIR-99, The challenge of image retrieval*, Newcastle, march 1999.
[9] Françoise Prêteux and Marius Malciu. "Model-based head tracking and 3D pose estimation". *In Proceedings of SPIE Conference on Mathematical Modeling and Estimation Techniques in Computer Vision*, pages 94-110, San Diego, USA, July 1998.