Motion Estimator Inspired From Biological Model For Head Motion Interpretation

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

ABSTRACT

This paper proposes a real time frequency method to estimate global rotation or translation and the corresponding direction of a moving head. Our method is based on the analysis of the image spectrum in the log polar domain. In this domain, spectrum analysis is easier (rotation motions are transformed into translations for example). But in order to make the log polar spectrum analysis easy, an efficient prefiltering stage inspired from the human biological model of the human retina is required. Indeed, after this pre filtering step, mobile contours are enhanced and static contours are removed, high frequency noise is eliminated and local and global variations of illumination are cancelled. The analysis of the log polar spectrum energy leads to the detection of motion type (rotation and translation) and to the estimation of the motion direction. The method is used for the interpretation of global head nods (up/down and right/left oscillations).

I. INTRODUCTION

As explained in [1], the fundamental question in image processing is no more "how are things" but , rather "what is happening ?". A low level image analysis has to be related to a high level interpretation of the scene. The aim of our work is to automatically compute the type of motion (translation or rotation) and direction of a moving head in order to interpret head nods.

Work has been done about 2D motion estimation. The paper [2] proposes a comparison of performances between block-matching methods, differential methods and frequency methods. All these methods yield to a dense 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 [9] [10]. Moreover, these method are sensitive to noise and illumination variations.

The aim of our work is not to estimate a dense low level optic flow but to detect the rotation axis or translation direction of typical rigid head motions involved in non verbal spoken communication. The idea is to process as human beings who are not exactly estimating the motion of each point in a scene. However, they can make a high level interpretation of the motion.

In section 2 the retina pre filtering is described. This yields to a filtered image with enhanced moving contours and removed static contours. The log-polar frequency spectrum of the pre filtered image is computed and analyzed in section 3. Finally, section 4 proposes a method to interpret global head nods (approbation or negation).

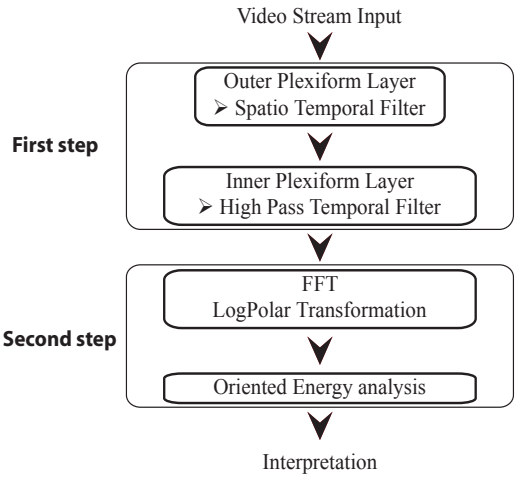


figure 1 : general algorithm

II. PRE FILTERING

Figure 1 gives a general overview of the algorithm. The first step consists in an efficient filtering stage.
Since the method is based on the analysis of the frequency response of the head moving contours, it is necessary to enhance such contours. But illumination variations can temporarily hide moving contours or modify their amplitude and noise has to be attenuated because it can corrupt these moving contours.

For this prefiltering step, the spatio temporal filter introduced in [4] and modelling the human retina behavior has been chosen. It is able to enhance moving contours, to remove static ones, and to cancel spatio temporal noise and illumination variations. An other advantage of this filter compared with a cascade of classic band pass filters is that process can be achieved in real time and more efficiently. The preprocessing step is composed of two stages, the first (II.1) enhances all contours and the second extracts only the moving ones (II.2).

II.1 Retina Outer Plexiform Layer pre filtering (OPL)

This filtering stage is the retina filter OPL (Outer Plexiform Layer) resulting from the modeling of the human retina behavior [4]. This is a non separable spatio temporal filter. The involved synaptic network is modeled by the circuit shown in figure 2 and its transfer function is :

$$G_B(z,ft) = \frac{1}{1+\beta_c+\alpha_c[-z^{-1}+2-z]+j2\pi f_t\tau_c} \cdot \frac{\beta_h+\alpha_h[-z^{-1}+2-z]+j2\pi f_t\tau_h}{1+\beta_h+\alpha_h[-z^{-1}+2-z]+j2\pi f_t\tau_h}$$

where

$$\alpha_i = \frac{r_i}{R_i}, \beta_i = \frac{r_i}{r_{fi}}, \tau_i = r_i.C_i$$

The i(k,t) correspond to the photo receptors inputs signals and the b(k,t) are the bipolar cells outputs. r_i, R_i are resistances and C_i are capacities that create the temporal effect.

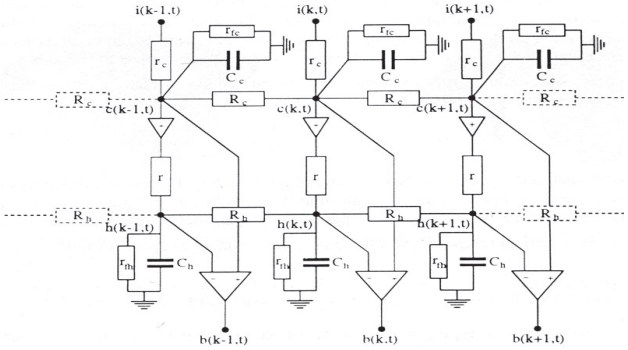


figure 2 : retina B electric model [4]

Figure 3 shows the resulting spatio temporal frequency response. This filter has a band pass spatial effect in low temporal frequencies, a wide band pass temporal effect for low spatial frequencies, a low pass effect for high temporal frequencies and it has a low pass tendency for high spatial frequencies.

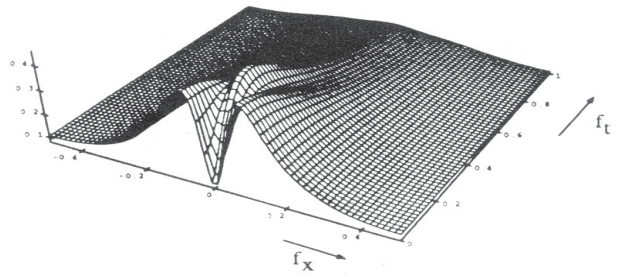
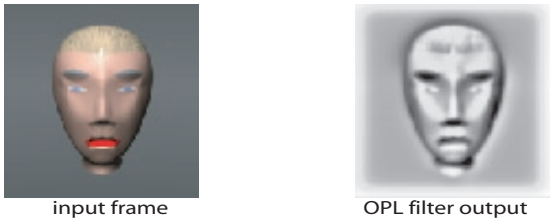


figure 3 : retina B transfer function [4]

Figure 4 illustrates the effect of the OPL filter on a head motion sequence in which the head tilts. The low spatial frequencies are attenuated, high frequency spatio temporal noise is cancelled and contours are enhanced.



input frame OPL filter output

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

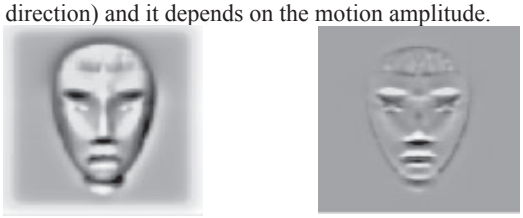
figure 4 : effect of the OPL filtering

OPL filtering requires only 10 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.
An other advantage of the OPL filter is that it is able to attenuate illumination variations which usually make motion estimation felt.

II.2 Retina Inner Plexiform Layer (IPL)

Human retina IPL filter is dedicated to the detection of moving stimulus. It is modeled by a temporal derivation operator [5]. This filter enhances moving contours and removes static ones. As a result, the spectrum of the filtered scene will only report power at the frequencies involved in the movement. Moreover, since the OPL stage attenuated the spatio temporal noise, its response after temporal derivation is minimized. Figure 5 illustrates the effect of the temporal derivative applied to the output of the OPL filter on the tilting head sequence. Moving contours perpendicular to the motion direction are accentuated while others are attenuated.

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 is contours perpendicular to the motion direction) and it depends on the motion amplitude.

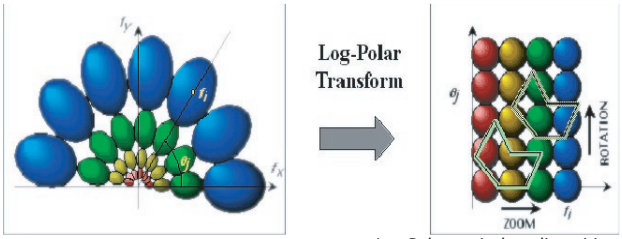


OPL filter output IPL output

figure 5 : effect of the IPL filter

III. FFT AND LOG POLAR TRANSFORMATION

After the pre filtering step, the processed frame contains only enhanced moving contours. The FFT in log polar domain is computed [8]. The log polar transformation allows to transform Cartesian zoom in a global translation along the frequency axis and cartesian roll into a global translation along the angle axis as illustrated in figure 6. The log polar transformation allows a large Cartesian spectrum of size M*N to be 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 computing 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.



cartesian domain filter's disposition Log Polar equivalent disposition

figure 6 : log polar transformation changes zoom and tilt into translations

Gabor filters are currently used to compute the log polar transformation but the problem is that the null frequency is transmitted and that in log scale, gabor filters response becomes asymmetric so that low frequencies are overweighted. Here, we propose to use GLOP filters (Log Polar Gabor Filters) introduced in [6] and defined by :

$$G_{ik}(f,\theta) = \frac{1}{\sigma\sqrt{2\pi}}\left(\frac{f_k}{f}\right)^2 \exp\left(-\frac{\ln\left(\frac{f}{f_k}\right)^2}{2\sigma^2}\right).\cos\left(\frac{1+\cos(\theta-\theta_i)}{2}\right)^{50}$$

Where the GLOP filter centered on frequency f_k in the θ_i orientation and scale parameter σ appears as a separable variable filter. Figure 7 shows a sample of 4 GLOP filters. These GLOP filters are symmetric in log scale and have a null response at the null frequency.

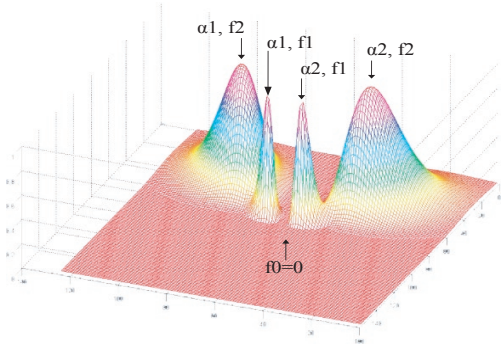


figure 7 : 4 GloP filters sample, placed on frequencies f1 and f2 with orientations α1 and α2, the null frequency f0 is centered

IV. GLOBAL HEAD MOTION INTERPRETATION

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

IV.1 Log polar spectrum interpretation

IV.1.1 Motion direction

The log polar spectrum reports the highest energy on the frequencies linked to the contours perpendicular to the motion direction. For example, the right part of figure 10 presents the log polar transform spectrum of an upward translating ring. Most energy is concentrated around 90° (vertical orientation). This energy is related to the motion of the horizontal part of the contours.

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 (see figure 8 a-b-c-d). 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 8 gives frames of a synthetic moving head and the corresponding cumulated energy curves. Figures 8-a, b and 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 8-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). Then, when achieving a complex rotation, these 2 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 8-d, 2 maximums appear which are related to the 2 rotations occuring at the same time. This complex motion can be analysed observing the amplitude variation of each maximum . In this case, tilt rotation (related to 182°) is faster than pan rotation (related to the 89° orientation).

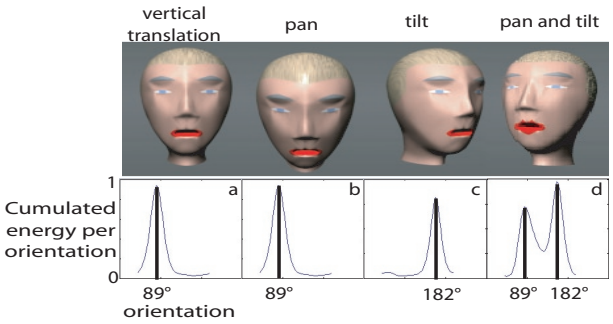


figure 8 : simple and mixed global head motion estimation
top : different motions of a synthetic head
bottom : cumulated energy per orientation

Finally, 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.

IV.1.2 Motion Type

Motion type (rotation or translation) is related to a moving or a static position of the maximum of the log polar spectrum. When an object rotates, rotations (roll, pan, tilt) can be considered in the log polar domain as global frequency translation (for roll, see section III) or localized frequency translations (for pan or tilt) along the rotation axis because of zoom effects in the spatial 2D projection of the 3D rotating object. In the case of pan and tilt, moving contours are compressed or dilated along the main rotation axis so that the associated spatial frequencies evolve. Figure 9 illustrates this effect on an horizontal rotation of a ring textured objet. The energy of the log polar spectrum is concentrated on the vertical contours (i.e. horizontal frequencies) because of the rotation orientation. Between frames 11 and 23, the object does a 25° horizontal rotation and the maximum energy translates from f_{11}=0.16 to f_{23}=0.22 normalized frequencies. On figures 9 and 10, white pixels corresponds to high energy values.

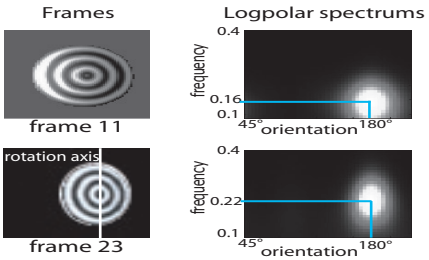


figure 9 : spectrum translation of a rotating object

When an object translates in front of the camera, there is no frequency change because contours are not modified. Figure 10 illustrates this effect with the same object translating upwards. Only horizontal contours give a response on the spectrum but there is no frequency translation of the energy spectrum.

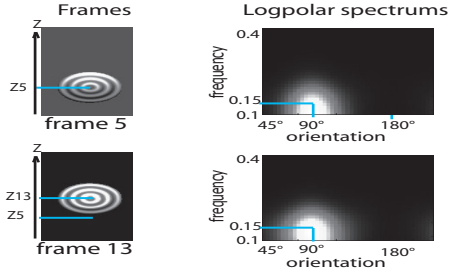


figure 10 : zspectrum evolution of a translating object

IV.2 Global head motion interpretation

The motion estimation algorithm is used to detect typical head motion such as horizontal or vertical head oscillations, currently used to express 'yes' or 'no'. This type of motion is defined by horizontal or vertical regular oscillations and periodic changes in motion direction. These motions can be difficult to estimate with common optic flow methods and feature point tracking because of their low amplitude and fast inversions. Moreover, disturbing events can introduce errors such as hands in front of the mouth for example.

Figure 11 shows one frame extracted from the video[1] of a person expressing 'yes' between frames 370 and 400 and 'no' from frames 400 to 462. The camera does not exactly face the person and the hand touches the mouth so that the hand could create parasite motion.

figure 11 : extract of the 'yes' and 'no' motions sequence

Figure 12 shows the temporal evolution of the maximum of the cumulated energy per orientation curve. The most relevant frequencies are located on the horizontal axis (90° +/- 3°) from frames 371 to 398 and in the vertical axis (180° +/- 3°) from frames 400 to 463 (note an estimation error near frame 415 because of a parasite hand motion). This corresponds to the presence of vertical and horizontal spatial movements in the video.

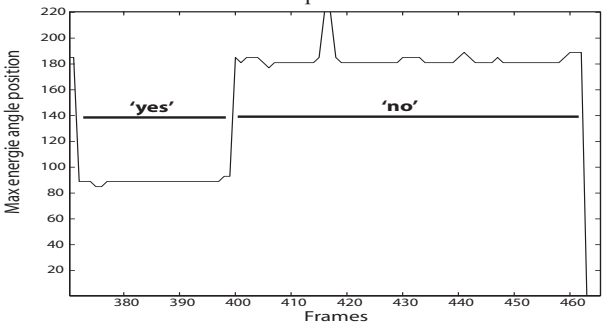


figure 12 : temporal evolution of the cumulated energy per orientation maximum

To extract the oscillation period, the total energy of the log polar transform is computed. Since only moving contours are responsible for non null energy in the spectrum, the total energy of the spectrum decreases very fast when motion stops. Figure 13 shows the temporal evolution of the total energy for the video sequence of figure 11. It shows that the total energy presents periodic minima to be related to each motion direction change. The period corresponds to the head oscillations. In this case, the mean frequency is 10 opposite motions per second. The algorithm is able to consider frequencies higher than 1.2Hz which represent a wide range of oscillations period expressing such attitudes.

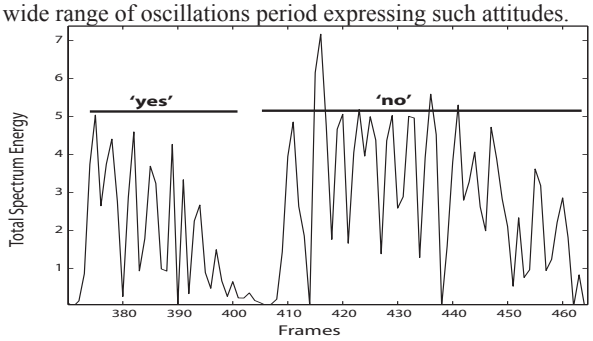


figure 13 : temporal evolution of total log polar energy spectrum

The performances of the head motion interpretation method have been evaluated in various test conditions: the system has 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 0.06), the algorithm is able to extract the expressed "yes' and 'no' with 90% success. Moreover, even in noisy conditions, the algorithm is able to extract these attitudes with 80% success when the face is 50% hidden, the frequency analysis only needs to get the main contours orientations of the face (i.e. some vertical and horizontal contours), not all face features are required. Finally, the algorithm works in real time, reaching up to 60 frames per seconds on a standard PC desktop Pentium 4 running at 3.0Ghz.

V. CONCLUSION

A real time method for global rigid head motion estimation has been proposed. The use of a pre filtering step inspired from the human retina behavior yields to a log polar spectrum easy to analyze : rotations are equivalent to oriented energy frequency translations, zooms are global frequency translations and 2D translations appear as localized energy drop that don't translate in frequency.

An example of high level head motion interpretation has been described. The proposed algorithm is well suited for global head motion estimation because a face presents vertical and horizontal contours. It also supposes that the scene has only one moving face in front of a static background. The behavior of this algorithm in the case of multiple motions of any rigid object with moving background is under study.

VI. REFERENCES

[1] A. Bobick and J. Davis "The Recognition of Human Movement Using Temporal Templates" *IEEE trans. On PAMI*, Vol.23, N°3, pp.257-267, March 2001

[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.

[10] 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

1. This video and other examples are available at http://www.lis.inpg.fr/pages_perso/benoit/