Saccadic Search with Gabor features applied to Eye Detection and Real–Time Head Tracking

F. Smeraldi O. Carmona J. Bigün

Swiss Federal Institute of Technology (EPFL)
Microprocessor and Interface Laboratory
CH-1015 Lausanne

Abstract

The Gabor decomposition is a ubiquitous tool in computer vision. Nevertheless, it is generally considered computationally demanding for active vision applications. We suggest an attention–driven approach to feature detection inspired by the human saccadic system. A dramatic speedup is achieved by computing the Gabor decomposition only on the points of a sparse retinotopic grid. An off–line eye detection application and a real–time head localisation and tracking system are presented. The real–time system features a novel eyeball–mounted camera designed to simulate the dynamic performance of the human eye and is, to the best of our knowledge, the first example of active vision system based on the Gabor decomposition.

Key words: Gabor Decomposition, Active Vision, Retinotopic Sampling, Saccadic Eye Movements, Head Tracking, Face Recognition.

1 Introduction

Gabor decomposition has long been known to be a powerful tool for pattern recognition tasks [6,3]. Its use in computer vision problems is motivated by strong biological analogies, since Gabor responses constitute a good model of the responses of the simple cells in the visual cortex [7,12]. Unfortunately, the calculation of Gabor filter responses imposes a heavy computational load which has so far made these features ill–suited for active vision applications. In this paper, we propose a bio–inspired approach to circumvent this problem.

The human eye explores a visual scene by performing a sequence of large "jumps", known as saccades, between the different points of interest, on which fixation is maintained for a short while [4,10]. Saccades have been shown to play a central role not only in the exploration of a scene, but in the underlying cognitive processes proper, where there appears to be a selection mechanism for filtering task relevant information [8,9]. Our approach to head and facial feature detection consists in performing an attention–driven search based on a model of saccadic eye movements. The algorithm is built around a log–polar retinotopic grid. Gabor decomposition is computed only on the points of the grid, thus allowing real–time performance.

In the first part of this paper we will report on simulation experiments in which the saccadic search is used to detect the eyes of subjects off–line but on real images (M2VTS database). We shall then describe a real–time setup in which the retinotopic grid is attached to a steerable eyeball–mounted camera, designed to mimic the dynamic performance of the human eye. The "artificial eye" so obtained is able to perform head localisation and tracking.

2 The retinotopic sampling grid

Central to our attentional strategy is the use of a sparse retinotopic sampling grid which is rigidly displaced on the images. The grid has log-polar geometry, meaning that the density of sampling points decreases exponentially with the distance from the centre. In our approach, we limit the computation of the Gabor decomposition to the points of the retinal grid, and require the grid to be displaced in order for other image regions to be considered. This sampling topology automatically implements a "focus of attention" concept, concentrating the computational effort on the current fixation point. Furthermore, by keeping the global number of fixations low, it is possible to perform the feature extraction by direct filtering in the image domain, without the need of a Fourier transform. This brings about a dramatic increase in efficiency.

In analogy with the operation of the visual cortex of primates and humans [5], we found it beneficial, at least during the finest part of the search (section 4.2), to tune our frequency decomposition so that it matches the variable sampling rate of the retina. We therefore associate high frequency Gabor filters to the fovea of the retina, while low frequency responses are extracted at the periphery, where the sampling rate is coarser. For our eye detection application, this allows a smooth integration of dense information from the centre of the eyes and global information from the outline of the orbit.

3 Log–polar frequency domain sampling

The log–polar mapping has also been applied to the design of the Gabor filters in the frequency domain. Standard complex valued Gabor functions in the frequency domain are scaled, translated and shifted versions of the following function:

$$\hat{\mathcal{G}}(\vec{\omega}|\sigma_x, \sigma_y, \omega_0) = \\ \exp\left(-\frac{(\omega_x - \omega_0)^2}{2\sigma_x^2}\right) \cdot \exp\left(-\frac{\omega_y^2}{2\sigma_y^2}\right)$$

The parameters σ_x, σ_y, ω_0 and the rotation parameter are chosen to cover the frequency plane as completely at possible.

However, when only a small number of logarithmically spaced frequency channels is used, problems arise in obtaining a uniform coverage of the frequency plane. Given that the spacing between the centres of the filters increases exponentially, the symmetric Gaussian shape doesn't appear to be optimal, since it extends the same distance towards the (well sampled) central region of the frequency space as well as towards the loosely sampled periphery. For these reasons, we choose to substitute for the commonly used Gabor function a modified filter

$$\hat{\mathcal{G}}'(\vec{\omega}|\sigma_\rho, \sigma_\phi, \rho_0) = \\ \exp\left(-\frac{(\rho - \rho_0)^2}{2\sigma_\rho}\right) \cdot \exp\left(-\frac{\omega_\phi}{2\sigma_\phi^2}\right)$$

where $(\rho, \phi) = (\ln(|\vec{\omega}|), \tan^{-1}(\omega_y/\omega_x))$ is the conformal mapping of the frequency plane to log polar coordinates [1]. These filters have been previously used in texture analysis problems, showing high discrimination power [2]. We therefore construct a uniform grid of Gaussian filters in the log–polar frequency domain, which in turn yields the desired consistent and exponential coverage of the Fourier plane (figure 1).

4 Eye localisation

When human subjects explore a natural scene, they do not use their eyes to scan it in a raster–like fashion. They rather perform rapid jumps between regions of interest, which they fixate for about 0.3 seconds. All the relevant information is acquired during such fixations, although much of the time is

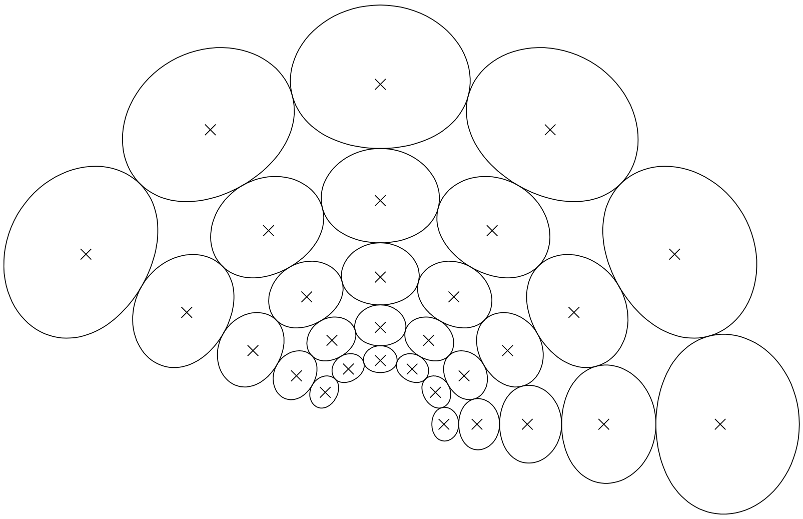

Fig. 1. Iso–curves of the Gabor filters created by uniform sampling of the frequency plane in log–polar coordinates. The crosses represent maxima, whose positions are slightly biased towards origin.

spent in deciding where the next saccade should be aimed. In 1957 Yarbus, who pioneered the study of the saccadic system, found that the stopping places of a subject's gaze exploring human faces were more densely distributed in the eye region [13,4]. This motivated us to use saccadic search for eye detection, even because of the relevance of such region for face recognition.

The procedure consists of three main steps. At first, local information driven saccadic eye movement is used to home the retinotopic grid on one of the eyes; following, the search is refined by pixel-wise displacement of the grid; finally, if detection is successful a saccade is performed to the assumed position of the other eye. During each of the above steps, several criteria are applied to check for the consistency of information. If a mismatch is detected, doubtful assignments are discarded.

4.1 Saccadic search

A local, appearance–based description of the search target (the eyes) is constructed by averaging the Gabor responses from the centre of the eyes of the persons in the training set. The resulting feature vector e_{av} consists of six orientation-selective responses for each one of the five frequency channels employed [11].

At the beginning of the search, the retinal sampling grid is placed at a random position on the image and the Gabor feature vectors are extracted for each of its points. Each of these vectors is subsequently matched against the reference e_{av}. The point of the grid for which the Euclidean distance from e_{av} is minimal

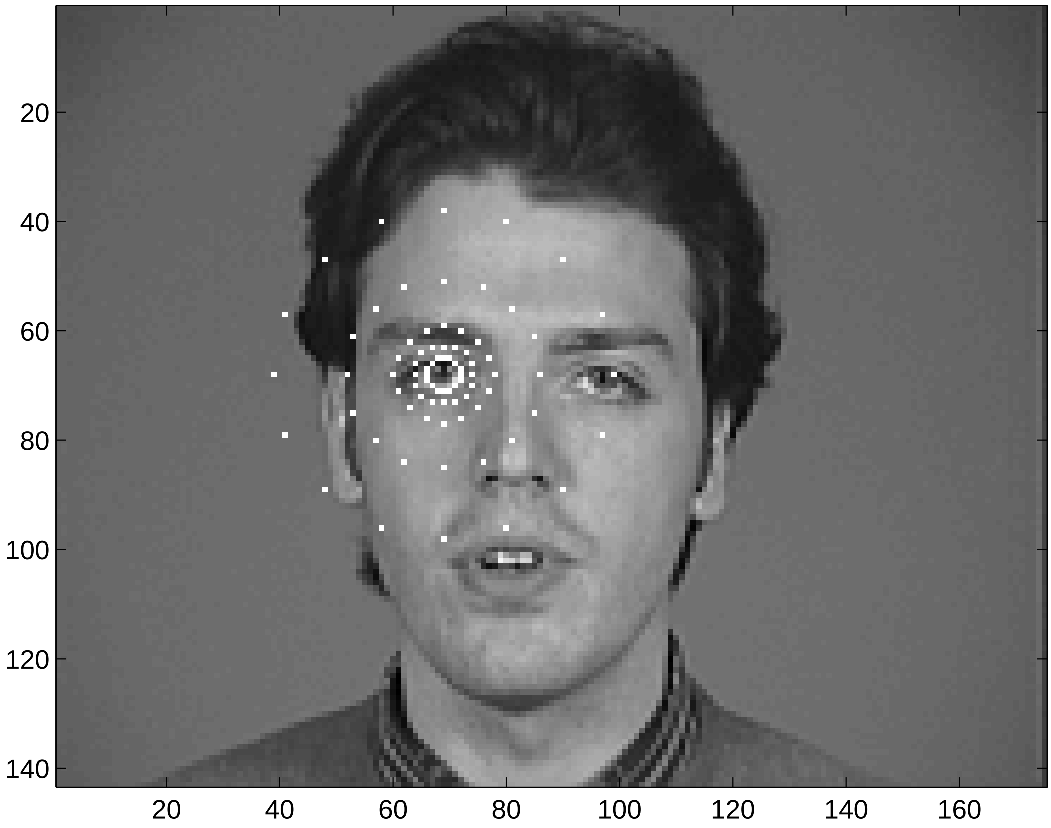

Fig. 2. The retinal sampling grid placed on a person's right eye for model creation.

is selected as the target for the next saccade. Saccadic search is assumed to have converged when saccades become shorter than a threshold. If no saccade target whose distance from e_{av} is reasonably low can be found (which can be the case if the search starting point happens to fall in a blank region of the image), the search is restarted from a random position.

4.2 Refining the search

A more complete description is obtained, for each eye, by placing the retinal sampling grid on the centre of each eye on the images of the training set (figure 2) and storing the Gabor responses from all of the retinal points. In order to reduce the sensitivity to positioning errors for small training sets, a relaxation procedure is used: user–supplied eye coordinates are employed to train a first version of the models, which is then used to perform a search on the training set itself. The eye coordinates thus detected are then used to retrain the system.

The two resulting "extended" eye models are used to distinguish left eyes from right eyes and to improve the precision of the localisation. A first comparison of the left and right eye models with the features currently "seen" by the retina is performed to state whether the spotted facial feature looks more like a left eye or a right eye. A gradient descent minimisation is successively performed by displacing the retina pixel-wise until the best match with the appropriate eye model is found.

The residual distance from the model is used to classify the detected feature as "eye" or "non–eye". The saccadic search is subsequently restarted in the expected direction of the other eye or, in the case that no eye has been found, from a random position.

Experiments have shown that the saccadic search may detect some erroneous local minima (e.g. the corners of the mouth, ear-rings or details in the hair). In order to discriminate such fake targets, the difference is computed between the candidate's distance from the attributed eye model and its distance from the alternate model. The ratio of this difference to the minimum distance, which we call the *asymmetry*, measures the amount to which the chirality of the detected feature contributes to the match. In our experiments, the asymmetry always turned out to be grater than 0.1 for correct matches, while it generally dropped of one or two orders of magnitude in the case of spurious identifications. The errors thus detected are treated by restarting the search from a random position.

4.3 Experimental results

The algorithm has been tested using a retinal sampling grid with 5 rings and 16 rays. The relation between the dimension of the retina and the size of the facial images is evidenced in figure 2.

The image database employed consists of forty frontal shots of twenty different persons[1] . The image resolution employed is 143 × 175 pixels. Differences between the shots of the same persons consist in tan changes, haircut, makeup, eyelid position, head position (heads are often slightly rotated) and slight scale changes. Several persons in the database wear eyeglasses.

Single shots from six persons have been used to extract the left and the right eye models. Repeated testing has then been performed on the whole database without any mismatch being found (figure 3). Information obtained from the outline of the orbit allows correct detection of the features even when the subject's eyes are closed (figure 4). In our trials we found the median of the number of fixation points to be 49 for the detection of both eyes, that is to

[1] This image database is a part of an audiovisual database, collected in the framework of the European person authentication project M2VTS.

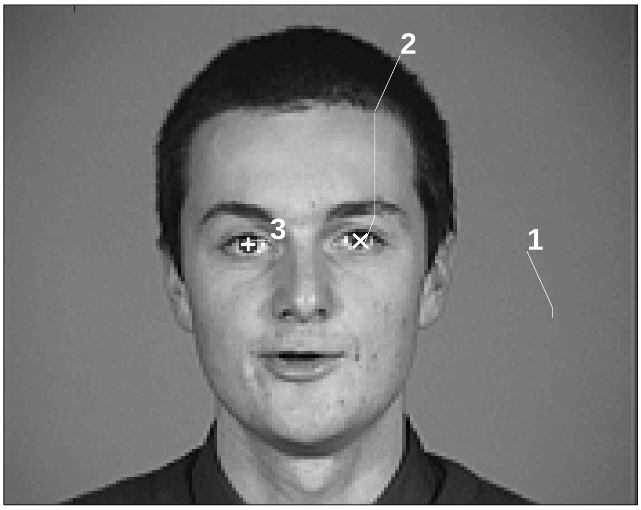

Fig. 3. The + and × signs denote the best match with the right and left eye models respectively. Numbers identify successive starting points for saccades. Eye detection required 51 fixations, the great majority of which are part of search refinement (only saccades are displayed here). Note how saccadic search 1 was considered uninteresting and therefore discarded. A random restart (2) then lead to detection of the left eye, after which saccadic search resumed (3) near the location of the right eye.

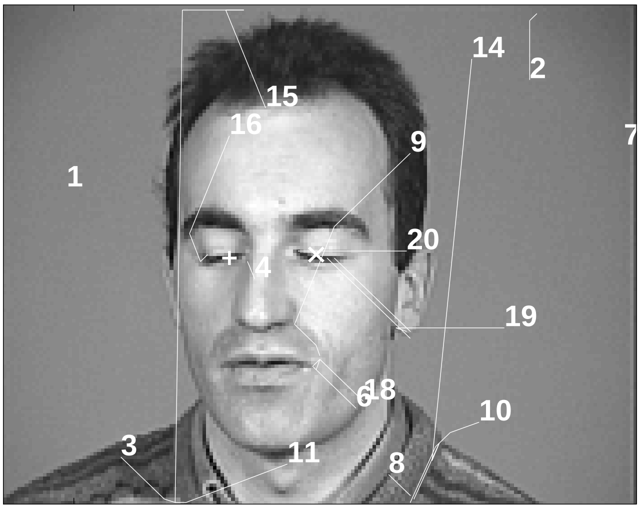

Fig. 4. Information from the outline of the orbit allows eye detection even if the person's eyes are shut. During this trial 99 fixations were made and 14 eye candidates located by saccadic search were rejected after comparison with the eye models.

say that the centre of the retinal sampling grid explores 0.2% of the image pixels. The number of fixations is considerably increased (typically 100) for subjects wearing glasses with strong reflections or having their eyes shut. This is mainly due to the fact that since the algorithm knows nothing about facial features other than the eyes, no alternative cues can be used to infer their spatial position when their visibility is low. Nevertheless, detection is always correctly accomplished at the end.

5 Real time head localisation and tracking

In order to demonstrate the flexibility and efficiency of the saccadic search algorithm we have implemented it into a real–time head localisation and tracking system. The retinotopic grid has been “attached” to a b/w steerable camera developed at our laboratory. The camera has a spherical mount and is specifically designed to mimic the dynamic performance of the human eye (figure 6).

The saccadic search algorithm described in section 4.1 can be readily adapted to the head detection task by substituting the eye model with an equivalent global description of the head. Details on this point are given in section 5.1. In order to increase performance, dynamic information from the image sequence can be used. This was achieved by complementing the algorithm with a motion detection stage (see section 5.2). A detailed description of the hardware can be found in section 5.3; finally, experimental results are discussed in section 6.

5.1 Modelling the head

A local, appearance–based description of the head has been obtained from a training set of 180 images of 6 different persons sitting in front of the (static) camera. Images have been acquired at a resolution of 320 × 240 pixels (half of standard SECAM resolution). The responses of 24 Gabor filters placed on the middle point between the eyes have then been computed for each image; the average response vector constitutes the head model.

The filters employed are organised in three logarithmically spaced frequency channels whose wavelengths range from 8 to 12 pixels; the orientation channels are 8. Due to the wide–angle optics adopted, the average distance between the eyes in a typical image is about 15–20 pixels, which is comparable with the wavelength of the filters. This is necessary if we want the filter responses to encode a signature of the whole head rather than of its subparts. Under these conditions, we can expect the map of the Euclidean distance between the head model and the Gabor vector responses extracted in all the points of an image containing a head to present a single pronounced minimum in correspondence of the head itself.

5.2 Motion detection

As the first step in the tracking procedure, two consecutive frames are acquired and the absolute value of their difference is computed. The difference is thresholded to identify pixels that differ significantly between the two im-

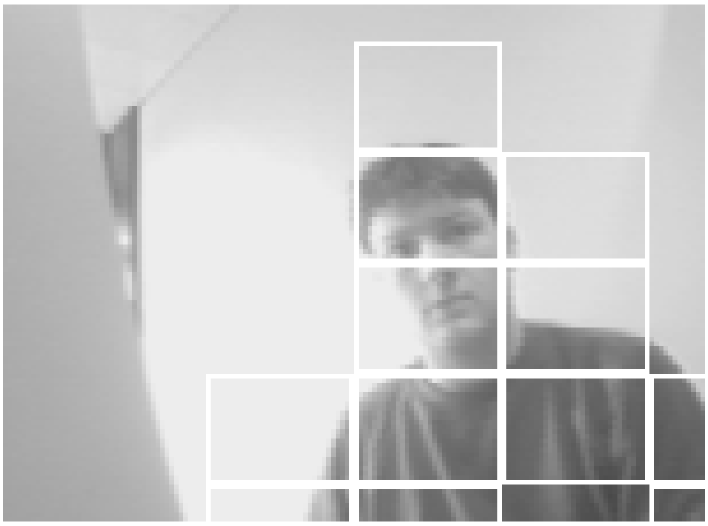

Fig. 5. The output of the motion detection stage is a set of rectangles surrounding moving objects in the scene.

ages. The field of view is then partitioned according to a 10 × 10 rectangular grid. Each grid cell can be labelled as a region of motion or not according to the percentage of differing pixels it contains (figure 5). Only the parts of the image that have been marked as regions of motion are considered for further processing. This step also allows to make sure that the camera has stopped moving before a frame is considered for processing. If more than 70% of the image appears to contain motion, the current frames are discarded and two new ones are immediately acquired.

5.3 Hardware setup

The algorithm was implemented on the Swiss Federal Institute of Technology Vision Sphere (figure 6). This system acquires high resolution images from a standard CCD camera that can tilt and pan under computer control. Its innovative principle is close to a "reversed" computer mouse: two orthogonally placed motors, replacing the encoders in a standard mouse, move the sphere. The position of the cameras is known using two orthogonally placed incremental optical encoders. Absolute desired camera positions are transmitted through a serial line to a Motorola 68336 micro-controller card. This card generates 20KHz pulse–width modulated 8–bit precision signals that drive the motors. The controller was elaborated by a robust–control pole–placing method. The low inertia of the Vision Sphere affords a fast response to reach any point of interest in the view space (table 1). Control is made easy by the fact that the two rotation axes and the optical axis of the camera intersect in a single point.

Visual computations are carried out on a 200MHz Pentium PC equipped with

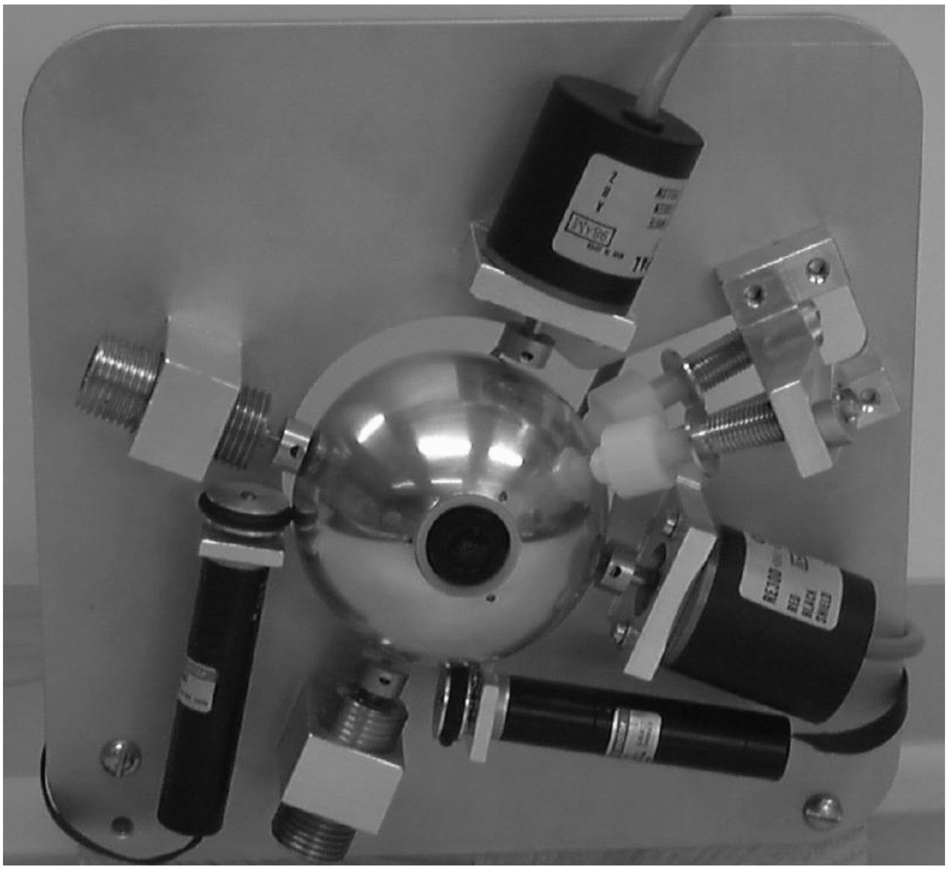

Fig. 6. Front view of the Swiss Federal Institute of Technology Vision Sphere. The two narrow black cylinders are the DC motors; the larger ones are the optical encoders.

Video Signal	Composite NTSC
Pixels	542 (H) × 492 (V)
Field of View	43° (H) × 32° (V)
Excursion	± 40°
Speed	500°/s
Acceleration	$40.000°/s^2$
Dimensions	$60 \times 170 \times 170$ mm^3
Power	12 V DC, 0.5 A up to 1,2 A

Table 1
Characteristics of the Swiss Institute of Technology Vision Sphere.

a PCI Matrox Meteor frame grabber. The average cycle time, including the time required by the camera to move, is about half a second; it is generally less for small corrections, since in that case the motion detection stage activates fewer retinal points and the camera settling time is shorter.

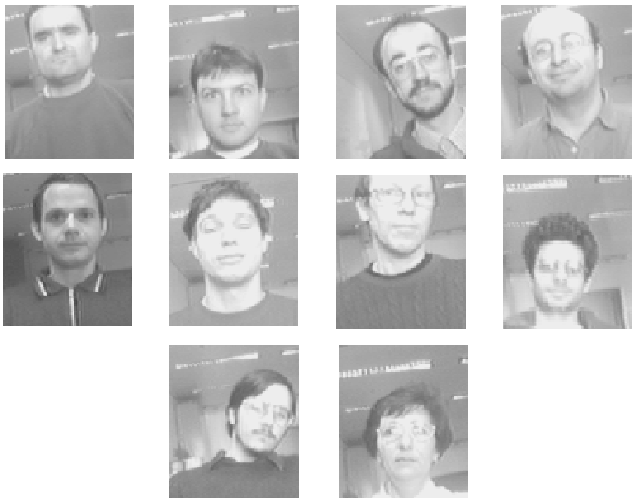

Fig. 7. A "passport photo" sub-frame covering 5% of the image size is grabbed each time the system believes a head is centred in the field of view.

6 Experimental results

The performance of the system has been tested by tracking the head of 10 persons sitting in front of the camera. Subjects were asked to move freely. The system was programmed to acquire a full SECAM resolution frame (640 × 480) each time it believed the head of a person to be centred in the image. A 110 × 130 "passport photo" sub-frame was then cropped at the image centre and stored for visual inspection. The sub-frame represents 5% of the image area. Its size has been computed from the average head size (75 × 90) by adding to each side the intrinsic system accuracy of 10 pixels (the radius of the inner circle of the retina) plus another 10 pixels of tolerance.

Out of 500 images acquired from the 10 persons, 446 (89%) turned out to represent the head as expected (figure 7). Three error typologies were found. In the first, subjects move outside the camera's range of operation. In the second, subjects move too quickly, causing the frames to be blurred. The third typology is erroneous head detection, which normally happens in the presence of sharp geometrical patterns with contrasting lines moving in the image (figure 8). This type of failure is partially corrected by the motion detection stage, and is due to the simplicity of the head model employed (represented by 24 responses at a single point).

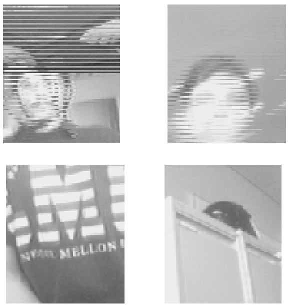

Fig. 8. The top two images show errors due to blurred frames. The two at the bottom represent errors induced by sharp geometric patterns or head–shaped objects (the helmet).

7 Conclusions

We have presented an attention driven search strategy mimicking the behaviour of the human saccadic system. The main feature of this algorithm is the log–polar sampling of the Gabor decomposition. We have discussed two applications: eye detection on static images and real–time head detection and tracking. The same algorithm has been applied in the two cases, with only minor modifications. We believe that the main advantages of our approach are its generality and the dramatic reduction of the information processed in order to perform the task. This is crucial to allow the use of such a powerful mathematical tool as the Gabor decomposition in active vision applications. Our setup constitutes, as far as we know, a novelty in that sense. Head localisation and tracking has been the privileged application for this research; however, further work remains to be done on the tracking of small objects (e.g. real–time eye detection) and the smooth pursuit of large objects in slow motion.

8 Acknowledgements

This work has been partly supported by the VIRSBS project within the European IT-LTR programme. We thank Mr. M. Bachmann and Mr. S. Menot for their help with programming and hardware.

References

[1] J. Bigun. Speed, frequency, and orientation tuned 3-d gabor filter banks and their design. In *Proceedings of International Conference on Pattern Recognition, ICPR, Jerusalem*, pages C–184–187. IEEE Computer Society, 1994.

[2] J. Bigun and J. M. H. du Buf. N-folded symmetries by complex moments in gabor space. *IEEE-PAMI*, 16(1):80–87, 1994.

[3] B. Duc, S. Fischer, and J. Bigun. Face authentication with sparse grid gabor information. In *IEEE Proc. of ICASSP, Munich*, volume 4, pages 3053–3056, 1997.

[4] D.H. Hubel. *Eye, brain and vision.* Scientific American Library, 1988.

[5] L. Maffei and A. Fiorentini. Spatial frequency rows in the striate visual cortex. *Vision Res.*, 1977.

[6] B. S. Manjunath, C. Shekhar, and R. Chellappa. A new approach to image feature detection with applications. *Pattern Recognition*, 31:627–640, 1996.

[7] G. A. Orban. *Neuronal operations in the visual cortex.* Studies of brain functions. Springer, 1984.

[8] J. B. Pelz. *Visual representations in a natural visuo-motor task.* PhD thesis, Carlson Center for Imaging Science, Rochester Institute of Technology, 1995.

[9] R. P. N. Rao, G. J. Zelinsky, M. M. Hayhoe, and D. H. Ballard. Eye movements in visual cognition: a computational study. Technical Report 97.1, National Resource Laboratory for the Study of Brain and Behavior, Department of Computer Science, University of Rochester, 1997.

[10] J. D. Schall, D. P. Hanes, K. G. Thompson, and D. J. King. Saccade target selection in frontal eye field of macaque. I. Visual and premovement activation. *The Journal of Neuroscience*, 15(10):6905–6918, 1995.

[11] F. Smeraldi, A. Makarov, and J. Bigün. Saccadic search with gabor features applied to eye detection. Technical Report 98/256, Swiss Federal Institute of Technology, Computer Science Department, CH-1015 Lausanne, January 1998. ftp://lamiftp.epfl.ch/pub/smeraldi/gaboreye.ps.gz.

[12] R. P. Würtz. Building visual correspondence maps — from neuronal dynamics to a face recognition system. In *Proceedings of the International Conference on Brain Processes, Theories and Models.* MIT Press, November 1995.

[13] A. L. Yarbus. *Eye movements.* Plenum, New York, 1967.