Proceedings of the 5th International Workshop on Advanced Motion Control (AMC '98)
Coimbra (Portugal), June 29th – July 1st, pages 684 – 687, IEEE Cat. Num. 98TH8354, 1998

Real–time Head Tracking by Saccadic Exploration and Gabor Decomposition

F. Smeraldi, O. Carmona, J. Bigün
Microprocessor and Interface Laboratory
Swiss Federal Institute of Technology (EPFL)
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 head detection and tracking 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. The real–time head localisation and tracking system presented features a novel eyeball-mounted camera designed to simulate the mechanical performance of the human eye. It is, to the best of our knowledge, the first example of active vision system based on the Gabor decomposition.

1 Introduction

The Gabor decomposition is a ubiquitous tool in pattern recognition [3, 1], and its use is motivated by strong biological analogies [4, 9]. Nevertheless, it is generally considered computationally demanding for active vision applications. In this paper, we present a real–time head detection and tracking system in which a bio–inspired approach is used to circumvent this problem. The human eye explores a visual scene by performing a series of large "jumps", known as saccades, between the different points of interest [2, 7]. Saccades play a central role also in filtering task–relevant information [5, 6]. We propose an attention–driven search strategy based on a model of saccadic eye movements. Input to the system is provided by a special eyeball–mounted camera designed to mimic the dynamic performance of the human eye. Motion detection is used to identify regions of interest in the image, thus "activating" the relevant points of a sparse log–polar retinotopic grid. A dramatic speedup is achieved by computing the Gabor responses only on such points, which can be done by direct image domain filtering. The camera is then centred on the point which matches a pre–computed head model the best; this procedure is iterated to achieve tracking.

2 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. The images were 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.

3 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 images. The field of view is then partitioned according to a 10×10 rectangular grid. Each

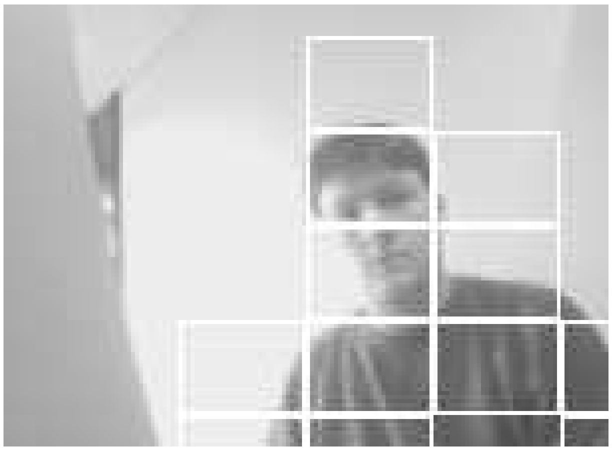

Figure 1: The output of the motion detection stage is a set of rectangles surrounding moving objects in the scene.

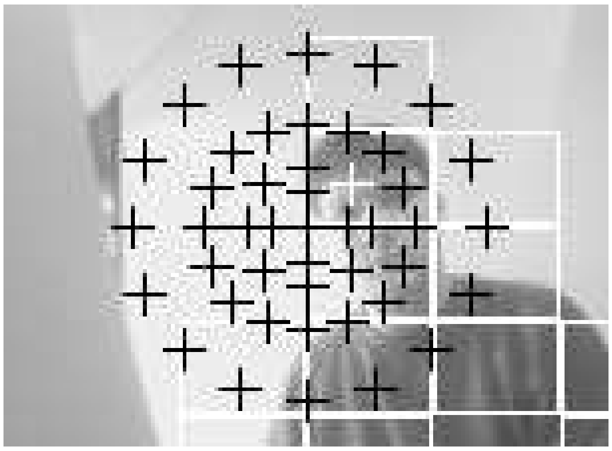

Figure 2: Gabor features are extracted only at the retinal points that happen to fall inside a region of motion. The retinal point at which the best match with the head model is found (marked white in the image) is selected as the target for the next saccade.

grid cell can be labelled as a region of motion or not according to the percentage of differing pixels it contains (figure 1). 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.

4 Saccadic search

Finding the position of a head in an image requires, in our framework, computing Gabor feature vectors to be matched against the head model. Since the feature extraction step is computationally demanding, the use of complex tools such as the Gabor decomposition is currently considered impractical for active vision applications.

The approach we followed consists in selecting a priori a fixed subset of image points on which the Gabor decomposition can be computed [8]. Such a subset is organised in a log–polar retinotopic grid (figure 2) which is centred on the image and can be thought of as if it were "attached" to the camera, in the sense that it moves rigidly on the scene when the camera moves. The information contained in the grey level of the image pixels is accessible to the system only to the extent that it contributes to the response of a filter centred at one or another of the retinal points. Access to a more detailed description of a particular image area is obtainable only by centring the retina at that position. Since the number of points in the grid is small (29 in our experiments), the required responses can be computed directly by filtering in the image domain, with a reduced computational effort.

A further selection among the retinal points can be performed based on the information from the motion detection stage. We chose to compute the set of 24 Gabor responses only at the retinal points that fall into regions of motion. The Euclidean distance between these vectors and the head model is then computed, and the camera is moved so that the field of view is centred on the point at which the best match has been found. The central point of the retina is always considered in the computation regardless of motion, so that the system is allowed to keep fixating the same spot if nothing has changed in the scene.

Each frame is processed independently of the preceding ones, i.e. there is no time integration. Nevertheless, the structure of the retina, which is denser near the centre, helps to concentrate the computational effort in the region in which the head was found at the previous iteration. Also, as the system homes in on a head the resolution at which the Gabor decomposition is sampled automatically becomes higher, allowing for finer adjustments. On the other hand, motion detection can draw the attention of the system to the periphery of the retina, thus allowing for quick recovery from tracking failures or rapid detection of subjects entering the field of view.

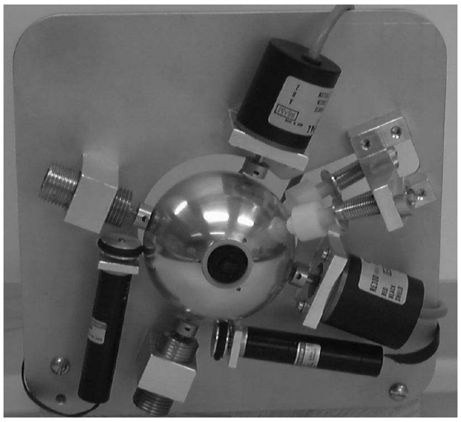

Figure 3: Front view of the Swiss Institute of Technology Vision Sphere. The two narrow black cylinders are the DC motors; the larger ones are the optical encoders.

5 Hardware setup

The algorithm was implemented on the Swiss Institute of Technology Vision Sphere (figure 3). 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 cards generates 20KHz pulse–width modulated 8–bit precision signals that drives 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 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.

Video Signal	Composite NTSC
Pixels	542 (H) $\times$ 492 (V)
Field of View	43^o (H) $\times$ 32^o (V)
Excursion	$\pm$ 40 ^{o}
Speed	500 ^{o}/s
Acceleration	40.000 $^o/s^2$
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.

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 $\times$ 480) each time it believed the head of a person to be centred in the image. A 110 $\times$ 130 "passport photo" subframe was then cropped at the image centre and stored for visual inspection. The subframe represents 5% of the image area. Its size has been computed from the average head size (75 $\times$ 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 4). 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 5). 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).

7 Conclusion

We have discussed a real–time head detection and tracking system based on saccadic exploration and the Gabor decomposition. This is, to the best of our knowledge, the first attempt to apply the Gabor decomposition in such an active vision task.

The systems displays a certain robustness to scale

Figure 4: A "passport photo" subframe covering 5% of the image size is grabbed each time the system believes a head is centred in the field of view.

Figure 5: 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).

changes, a good recovery speed and the ability to effectively discriminate between a head and other objects, as for example a hand waving in front of the camera. Robustness could have been increased by implementing a simple contrast test to rule out geometric patterns. Also, a more complex classifier could have been used as a post–processing stage to give the system feedback about the fact that the object being tracked actually is a head.

Acknowledgement

This work has been 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] 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.

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

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

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.