Biologically inspired Cartesian and non-Cartesian filters for attentional sequences

H.I. Bozma, G. Çakıroğlu, Ç. Soyer *

Intelligent Systems Laboratory, Department of Electric Electronic Engineering, Institute of Biomedical Engineering, Bosphorus University, Bebek 80815, Istanbul, Turkey

Received 10 October 2001; received in revised form 24 June 2002

Abstract

The aim of this paper is to develop a rich set of visual primitives that can be used by a camera-endowed robot as it is exploring a scene and thus generating an attentional sequence—spatio-temporally related sets of visual features. Our starting point is inspired by the work of Gallant et al. on the area V4 response of the macaque monkeys to Cartesian and non-Cartesian stimuli. The novelty of these stimuli is that in addition to the conventional sinusoidal gratings, they also include non-Cartesian stimuli such as circular, polar and hyperbolic gratings. Based on this stimulus set and introducing frequency as a parameter, we obtain a rich set of visual primitives. These visual primitives are biologically motivated, nearly orthogonal with some degree of redundancy, can be made complete as required and yet implementable on off-the-shelf hardware for real-time selective vision-robot applications. Attentional sequences are then formed as a spatio-temporal sequence of observations—each of which encodes the filter responses of each fovea as an observation vector consisting of responses of 50 filters. A series of experiments serve to demonstrate the use of these visual primitives in attention-based real-life scene recognition tasks: (1) modeling complex scenes based on average attentional sequence responses and (2) fast real-time recognition of relatively complex scenes with a few saccades—based on the comparison of the current attentional sequence to the a priori learned average observation vectors.

Keywords: Attentional sequences; Selective perception; Visual primitives; Active vision; Real-time scene recognition

1. Introduction

The use of visual attention in robotic vision systems has been gaining increasing prominence (Abbott et al., 1993; Ballard, 1991; Ballard and Brown, 1992)—thanks largely to the foundation laid by vision scientists (Akins, 1996; Julesz, 1995; Kowler, 1990; Stark and Ellis, 1981; Treisman and Gelade, 1980). Attentive systems simulate the selective attention behavior of the eye and work in a continual loop of pre-attention and attention (Clark, 1999; Koch and Ullman, 1994). In pre-attention, simple computations are performed on the periphery of the visual field in order to determine where to fixate next and what will be the next fovea—the small region of highest acuity around

* Corresponding author. 
E-mail address: cagatay.soyer@onitron.com (Ç. Soyer). 
URL: http://www.isl.ee.boun.edu.tr.
the optical axis (Connor et al., 1997; Itti and Koch, 1999; Koch and Itti, 2000). Upon finding a new fixation point, the camera saccades in order to bring this new image to its fovea (Kapoula and Robinson, 1986; Malinov et al., 2000). Hence, a stream of fovea images is generated—a sample of which may be as shown in Fig. 1 for a particular scene. In attention, each fovea image is subjected to attentive processing. It has been proposed that attentive processing is sophisticated in nature—based on visual primitives that extract complex feature information (Parasuraman, 2000). As a result of attentive processing, a spatio-temporally related sequence of visual features—is generated (Soyer and Bozma, 1997; Soyer et al., 2002). One of key issues in this framework is that of visual primitives: What sort of processing is applied on the fovea once a fixation is made. In this paper, we are concerned with developing a set of biologically motivated visual primitives that can be used in attentional deployment in a robotic system.

The starting point of our work is based on the results presented in (Gallant et al., 1995, 1996) on the neural response in area V4 of the macaque monkey to the Cartesian and non-Cartesian gratings. The presence of cells tuned to conventional sinusoidal (Cartesian)—which in turn play a key role in natural shape representation—have been known since some time. However, their findings reveal cells with selectivity to other types of stimuli—circular, polar and hyperbolic (collectively called non-Cartesian) in particular. Together, Cartesian and non-Cartesian filters are thought to be related to intermediate stages of form processing. Inspired by these findings, we develop these filters for active vision-robot applications and experimentally show that they can be successfully utilized in scene recognition tasks.

### 1.1. Related work

There has been tremendous progress in our understanding of the relationship between attention and visual coding—thanks largely to the foundation lay by early studies of vision scientists (Julesz, 1995; Treisman and Gelade, 1980). A common theme shared by these different propositions is the presence of encoder mechanisms—such as a set of filters or second-order statistic extractors (Caelli, 1995; Sagi, 1995). The focus has then been on the nature of these encoders.

One theory which has been supported by many findings—is that attention is directed to simple or complex features in the visual field (Julesz, 1995; Privetera and Stark, 2000; Treisman and Gelade, 1980). As such, one can conceive attention as a simple facilitation of certain set of neurons and thus highlighting particular features at a particular position in the visual field (Nakayama and He, 1995). This is referred to as Cartesian view of attention because it is specified in term of a viewer-centered geometric feature space. In order to generate computational models, researchers have then proceeded to mathematically model these receptive fields of the cells thought to be respon-
sive to Cartesian features. For example, real part of complex Gabon functions have been found to fit the receptive field weight functions found in simple cells in cat striate cortex (Jones and Palmer, 1987) and have then been used widely in camera-based vision systems (Privetera and Stark, 2000).

Interestingly, recent work suggests that attention mechanisms extend beyond a Cartesian view of attention to include surface representation. There is convincing evidence visual system constructs representations of surfaces and hence attentional deployment can also be based on surface based parsing (Nakayama and He, 1995). In the same spirit, Gallant and et al. have investigated surface representation in area V4 of the macaque monkeys (Gallant et al., 1995). For this purpose, a new stimulus set consisting of three categories of gratings has been introduced: conventional sinusoidal and newly developed circular, polar and hyperbolic gratings. The first category—referred to as Cartesian stimuli—is correlated with planar textured surfaces while the remaining—referred to altogether as non-Cartesian stimuli—correspond to textured spheres and saddles. The most striking finding is the highly selectivity of some V4 cells to textured surfaces while the remaining—referred to as Cartesian stimuli—is correlated with planar hyperbolic gratings. The first category—referred to as Cartesian stimuli—is correlated with planar hyperbolic gratings. The first category—referred to as Cartesian stimuli—is correlated with planar hyperbolic gratings. The first category—referred to as Cartesian stimuli—is correlated with planar hyperbolic gratings. The first category—referred to as Cartesian stimuli—is correlated with planar hyperbolic gratings. The first category—referred to as Cartesian stimuli—is correlated with planar hyperbolic gratings.

1.2. Mathematical formulation: pre-attention and attention

We assume that the visual processing is composed of a loop of pre-attentive and attentive stages which generate an attentional sequence—spatio-temporally related sets of features values (Soyer et al., 1996, 2000, 2002; Soyer and Bozma, 1997).

In the pre-attention stage, simple attentive features are computed from the periphery region in order to select the next fixation point and thus the next fovea to be fixed. Let \( I'_c \) represent the visual field image and \( I'_f \) represent the fovea image at time \( t \). Let \( C(I'_f) \) denote the set of candidate foveae—determined from the visual field. For each candidate fovea \( I'_f \in C(I'_f) \) an attention criteria \( a : I'_f \to \mathbb{R}^+ \)—a scalar valued function of interest based on the presence of simple features with low computational requirements—is computed. The candidate fovea maximizing these criteria is then selected as the next fovea:

\[
I'^{t+1}_f = \arg \max_{I'_f \in C(I'_f)} a(I'_f)
\]

When a selection is made, the optical axis of the camera is directed to bring that area into fovea. Such camera movements correspond to saccadic eye movements in humans. As a result, a sequence of foveae \( I_t = \{I'_1, \ldots, I'_t, \ldots\} \) is generated.

In the attentive stage, each fovea \( I'_f \) is subjected to detailed analysis in order to make an observation \( o' \) about the state of the fovea. In general, this analysis is much more computational than the pre-attentive stage and the visual primitives that are used can be rather complex. Let us suppose that this set consists of \( M \) different visual primitives and let the set of values of \( m \)th visual primitive be denoted by \( \Omega_m \). The flow of attentive processing is as follows:

1. Consider a filter bank \( F = \{f_{cw} \mid c = 1, \ldots, C; \omega = k\delta \omega, k = 1, \ldots, K \} \) where each \( f_{cw} \in F : I_f \to H'_m \) maps the fovea image \( I'_f \) to another image where \( H'_m(x,y) = f_{cw}(x,y) I'_f(x,y) \), \( m = 1, \ldots, M \) where \( m = (c-1)K + k \), \( \omega = k\delta \omega \) and \( M = (C-1)K \).
2. Choose an operator \( g_m : H'_m \to \Omega_m \) that acts on the outputs of these filter banks and maps each output to a value in its appropriate space \( \Omega_m \).
3. Each observation, \( o' : I'_f \to \Omega, \Omega = \Omega_1 \times \cdots \times \Omega_M \) then becomes a vector of \( M \) visual primitive values: \( o' = [g_1[I'_f], \ldots, g_M[I'_f]] \).

Thus, through the pre-attention attention loop, as a stream of foveae \( I_t = \{I'_1, \ldots, I'_f, \ldots\} \) is generated, so is an attentional sequence \( O^T = (o^1, \ldots, o^T) \).

Hence, an attentional sequence can be visualized to be a spatio-temporally related set of values of visual primitives—containing the critical visual data. Obviously, the choice of the visual primitives
is of utmost importance—if we are to use attentional sequences in visual tasks.

2. Visual primitives

The underlying possibilities of psychoneurological findings are that perhaps visual primitives are composed of both Cartesian and non-Cartesian filters (Nakayama and He, 1995; Sagi, 1995; Gallant et al., 1995, 1996). It has been argued that such a set of primitives would be capable of representing a broad range of shapes. Motivated by such a set of primitives would be capable of representing a broad range of shapes. Motivated by this, we have then set out to implement these as visual primitives for one of our robotic systems. In particular, 50 filters—six Cartesian orientations, concentric and radial filters and two hyperbolic filters for five frequencies—are implemented (Fig. 2).

2.1. Cartesian filters

The Cartesian filters $f_{c00} : \text{SO}(1) \times \text{SO}(1) \to [-1,1]$ can be modeled as follows. Note $\text{SO}(1) \triangleq [-\frac{\pi}{2}, \frac{\pi}{2}]$

$$f_{c00}(x,y) = \cos(\omega(2x + \beta y))$$

where

$$a = \sin((c-1)\pi/A)$$

$$\beta = \cos((c-1)\pi/A)$$

The subscripts c and $\omega$ are used to indicate the parametric role played by the orientation c and the frequency $\omega$ of the sinusoid. By taking $c = 1, \ldots, A$ and varying $\omega \in \{k \times \delta \omega \mid k = 1, \ldots, K\}$ so that Cartesian filters similar to those used in (Gallant et al., 1995) can be modeled. In our case, we use $A = 6, \delta \omega = 2, K = 5$.

2.2. Non-Cartesian filters

The non-Cartesian filters (Gallant et al., 1995) can also be modeled as a function of sinusoids, however the arguments of the cosine function now have a non-linear form. The concentric filters $f_{c} : \text{SO}(1) \times \text{SO}(1) \to [-1,1]$ can be mathematically modeled as $f_{c00}(x,y) = \cos(\omega(x^2 + y^2))$ where the $\omega$ parameter determines the frequency. By varying $\omega \in \{k \times \delta \omega \mid k = 1, \ldots, K\}$, a set of circular filters varying in frequency are generated.

The polar filters $f_{g0} : \text{SO}(1) \times \text{SO}(1) \to [-1,1]$ are given by the formula $f_{g00}(x,y) = \cos(\omega \arctan(y/x))$ where the parameter $\omega$ determines the frequency. By varying $\omega \in \{k \times \delta \omega \mid k = 1, \ldots, K\}$, again we can model a set of circular filters.

Finally the hyperbolic filter $f_{h00} : \text{SO}(1) \times \text{SO}(1) \to [-1,1]$ is given by $f_{h00}(x,y) = \cos(\omega y^2 - x^2)$ where $\omega$ parameter determines the frequency. The rotated $f_{h00}(x,y)$ form of the hyperbolic filters can be obtained by rotating $f_{h00}(x,y)$ function around origin by $\theta$ degrees $f_{h00}(x,y) = f_{h00}(\cos(\theta)x + \sin(\theta)y, -\sin(\theta)x + \cos(\theta)y)$. Taking $\theta = \pi/4$ and $\omega \in \{k \times \delta \omega \mid k = 1, \ldots, K\}$, we obtain a set of hyperbolic filters.

2.3. Filter response functions

The response of each filter is generated based on the convolution results. Let $g_{m} m = 1, \ldots, M$ where $m = (c-1)K + k, \omega = k\delta \omega$ and $M = (C-1)K$ represent the response function to filter $f_{c00}$. Experimentally, for Cartesian filters $f_{c00} \in F, c = 1, \ldots, 6$, the best choices for $g_{m}$ are experimentally determined to be those that disregard mean intensity levels. In our case, we define $g_{m}$ as the variance of the convolution result as: $g_{m} = \text{var}(f_{c00} L')$. Again, experimentally, for filters, $f_{c00} \in F, c = 7, \ldots, 10$, (Circular, Polar, Hyperbolic and Rotated Hyperbolic), the best $g_{m}$ choices are those that consider the greatest magnitude

![Fig. 2. Cartesian and non-Cartesian filter bank. The x-axis is c = 1, ... , 10. Note that [-1, 1] range is mapped to black–white grayscale.](image-url)
response. In our case, \( g_m \) is defined as: \( g_m = \max(f_{\text{co}} t_f)^2 \).

2.4. Optimal filter size

In order to determine optimal filter size, we have conducted a series of experimental studies using \( D \) different images. In this case, \( D = 5 \). Let \( Q = \{(i, j) \mid i, j = 1, \ldots, D, i \neq j\} \) denote the image pairs. The \( f \times f \) center region of each image is used to generate foveae varying in size—where \( f \in F = \{40 + 10l; l = 0, \ldots, 6\} \). Next a set of Cartesian and non-Cartesian filter banks consisting of size \( s \times s \) is generated—where \( s \in S = \{9 + 2l; l = 0, \ldots, 10\} \). In order to determine optimal filter size, we try to find \( s' \) that yields the maximum amount of differentiation between all pairs of images as follows:

1. First, we apply each \( s \times s \) filter bank to each \( f \times f \) fovea of each image \( j \) in order to generate the observation vector \( o_{ij} \).
2. Next, for all, \( (i, j) \in Q \), Euclidean distance \( d_{ij}(s, f) = |d^{s,f} - o^{s,f}| \) is computed. For each filter and fovea size \( (s, f) \) combination, the minimal \( d(s, f) \) value is computed as:
   \[
   d(s, f) = \min_{(i, j) \in Q} \frac{d_{ij}(s, f)}{\sum_{(x, y) \in Q} d_{xy}(s, f)}.
   \]
   Note that \( d(s, f) \) provides a measure of the worst-case differentiation for a filter the particular filter and fovea size \( (s, f) \) combination.
3. Following, for each \( s \times s \) filter bank, compute the average minimum normalized distance
   \[
   d(s) = \frac{1}{|F|} \sum_{f \in F} d(s, f).
   \]
4. The optimal filter size \( s' \) is defined as the filter size yielding the greatest average minimum normalized distance as \( s' = \arg\max_{s \in S} d(s) \).

Fig. 3 shows the filter size \( s \) vs. \( d(s) \) as well as the standard deviations where it is observed that \( d(s) \) ranges from 0.053 to 0.068. Smaller filter sizes yield higher \( d(s) \) values and in order to have a balance between maximum differentiation and high resolution, we choose \( s' = 15 \). Note that this is independent of the fovea size.

2.5. Filter cross responses

Gallant et al. report that V4 Cartesian cells give high response to the Cartesian gratings and V4 concentric cells have high response to the circular gratings, etc. In the next stage of our study, we investigate the responses of each filter to varying fovea stimuli—of Cartesian or non-Cartesian nature with varying frequency. For example, responses to Cartesian texture are shown in Fig. 4. In this graph, the squares show the responses to the filter given the type (x-axis) and the frequency (y-axis). The size of the squares increase as the response increases in magnitude. It is observed that the highest response to this texture comes from the same type filter. A similar observation is made for
the cases of concentric, polar and hyperbolic stimuli—samples of which are shown in Figs. 5–7. In summary, it is observed experimentally that the filters are consistent in the sense of giving the highest response to stimulus resembling itself—with some level of redundancy in the nearby filters.

2.6. Intensity and rotational sensitivity

We first investigate the effect of mean intensity value on these filters. This is studied via taking five random 256-color grayscale images, considering all the $40 \times 40$ candidate foveae on these images and determining the percentage deviation of filter outputs in response to uniform change of added mean intensity. Ten different levels of intensity—ranging from darker mean to brighter mean are considered. During these experiments, pixel values that changed beyond the $[0, 255]$ interval are simply set to their saturation levels. The average percentage change in filter outputs are as shown in Fig. 8. As expected, outputs are relatively insensitive to uniform brightness changes.

We also investigate the rotational sensitivity of these filters. For this study, we consider the five random images used in the mean intensity study. Next, all images are rotated by 10-degree steps using bicubic interpolation. The central $40 \times 40$ fovea region of the original and the rotated image are compared with respect to their responses. The comparison is based on the Euclidean distance between the observation vectors of the original and the rotated fovea. Fig. 9 shows the mean, standard deviation and maximum of mean-percentage-changes for the sampled rotated (60, 120, 180, 240, 300 and 360) versions. From the figure we can see that mean and standard deviation of percentage changes in some of the filter outputs are lower than the other filters. This result can be used in some

---

**Fig. 5.** Responses of filter set to another Cartesian stimulus $f_{68}$ with $c = 6$ and $\omega = 8$.

**Fig. 6.** Response of filter set to polar $f_{8,10}$ stimuli.

**Fig. 7.** Response of filter set to a hyperbolic stimulus.

**Fig. 8.** Percentage deviation of filter outputs in response to uniform mean intensity change.
problems to minimize the effect of rotation by observing the outputs of selected filters.

3. Experimental results

This work has been implemented on APES—a mobile robot with active vision developed in our laboratory (Soyer et al., 1996, 2000), Fig. 10. Using four stepping motors it can translate and rotate its body and direct its cameras to the visual stimuli by pan and tilt motions. Its visual processing is based on attentive vision as explained in Section 1.2, Fig. 11. Its visual field is taken to be $120 \times 120$ and its fovea size is $40 \times 40$—which are experimentally determined as being optimal with regards to balancing the tradeoff between high resolution and low computational complexity. Filter sizes are taken to be $15 \times 15$.

3.1. Attentional sequences

In a series of experiments we investigate the responsive behavior of these filters. The experiment consists of looking at the three scenes: A wall with painting, a notice board and a coat rack. Sample first 10 saccades generated while looking at these scenes are as shown in Fig. 12. It must be noted that the attentional sequences generated on the same scene at different viewing times are never exactly the same. This is due to changing instantaneous camera motions and thus saccade paths with varying lighting conditions. We then look at
the response of these filters after each saccade. The filter functions $g_m, m = 1, \ldots, M$ for each saccade of each sample scene saccade are as shown in Figs. 13–15 respectively. As observed, the observation sequence $O^{T}$ as observed as a time signal changes from fovea to the next. Furthermore, these observations also change from scene to scene. Thus, we expect that observation sequence should provide sufficient information regarding the contents of the scene.

Finally, the time statistics of the approach allow real-time applicability. The computation of an observation vector requires 50 convolution operations followed by 30 variance and 20 maximum calculation operations. In our measurements on a Pentium IV 1300 MHz machine, using $40 \times 40$ foveae and $15 \times 15$ filters, a single convolution followed by a single variance/max operation costs 1.8 ms on the average. Thus a single observation vector can be constructed in 90 ms on the average—which is approximately one tenth of a second. As this is sufficient for our current application, no special attention has been paid to using more computationally efficient implementations. However, let us note that additional speeding can be achieved by resorting to methods such as convolution in frequency domain by using FFT.

3.2. Application: scene recognition

In a series of experiments we investigate the performance of these filters in a recognition task. The task is defined as recognizing the current scene—which is one of the previously three considered scenes—the painting, the notice board and the coat rack. APES uses the first set of runs in each scene to learn about the scene. For each scene, $N_r = 5$ runs are considered and at each run, APES makes $N_s = 35$ saccades before stopping. The scene model is generated simply by taking the average observation vector $T_i, i = 1, 2, 3$ for each attentional sequence for each run as:

$$T_i = \frac{1}{N_s \times N_r} \sum_{r=1}^{N_r} \sum_{s=1}^{N_s} o^s(I^r_i)$$

Fig. 16 shows mean filter responses for each scene. If we compare the average observation vectors to those of individual saccades for the particular scene, we observe that they tend to emphasize the more shared filter responses and attenuate the deviations.

APES was then randomly made to look at scene $i, i = 1, 2, 3$. For each run, it computes the average of observation vector of $A^i(n)$ first $n$ saccades defined as:

$$A^i(n) = \frac{1}{n} \sum_{k=1}^{n} o^k(I^r_i) \quad n = 1, \ldots, N_s$$

We then assessed the discrimination power of our visual primitives by computing Euclidean distance $d(A^i(n), T_i) = |A^i(n) - T_i|$ between average observation vectors for each scene and that of the current scene. If the current run is on scene $i$ and if $T_i$ is a good model of our scene, then the Euclidean distance $d(A^i(n), T_i)$ should be minimal in comparison to those between $d(A^j(n), T_i), j \neq i$ as follows:

Fig. 12. Sample first 10 saccades of painting (top), board (middle) and coat rack (bottom) scenes.
Fig. 13. Filter responses for each of the saccades for the sample "painting" scene saccade set.
Fig. 14. Filter responses for each of the saccades for the sample “board scene” saccade set.
Fig. 15. Filter responses for each of the saccades for the sample “coat rack scene” saccade set.
Fig. 16. The average filter responses for each of the three scenes: (top-left) painting scene; (top-right) board scene; (bottom-left) coat rack scene.

Fig. 17. Experiment results: Plots show the averaged distance (bounded by standard deviation) of all test runs of each class to the target observation vector (target scene). The tabulated values show that recognition errors for each scene type decrease with the number of saccades.
$$i = \arg \min_{y=1,2,3} d(A'(n), T_y), \quad \exists n \geq N_s,$$

Obviously, the number of saccades $N_s$ made is critical here. For instance, at the beginning after just one saccade, we would expect that on average the $d$ values for all the scenes would be of comparable magnitude unless of course the first fovea contains an extremely prominent scene feature. However, as we start making more saccades and hence more observations, we expect that $d(A'(n), T_i)$ becoming noticeably smaller than the others. Fig. 17 shows the experimental results for five runs on each scene respectively. It is observed that after the first saccade, as expected, all the $d(A'(1), T_y) \ y = 1,2,3$ values are close to each other. However, with more saccades, they start becoming differentiable. After four saccades or so, they converge. Furthermore, once convergence occurs, it is observed that average and variance of $d(A'(n), T_i)$ yields the smallest values and are considerably smaller in magnitude in comparison to $d(A'(n), T_j) j \neq i$. Interestingly, the table in Fig. 17 shows that after the 4th saccade, the scenes are correctly identified so that we do not need to make all of 35 saccades to distinguish the scenes. We have repeated our experiments in similar scenes— and the $d$ values converge to certain levels after a number of saccades—depending on the scene complexity.

4. Conclusion

This paper develops a rich set of visual primitives that can be used by a camera-endowed robot as it is exploring a scene and thus generating an attentional sequence—spatio-temporally related sets of visual features. The visual primitives are biologically inspired—based on the work of Gallant et al. on the area V4 response of macaque monkeys to Cartesian and non-Cartesian stimuli. First, based on these stimuli set, a filter bank is developed and implemented digitally. Next, this filter bank integrated to the attentive system of our robot which then deploys it during its attentive processing stages in order to generate an observation. Each observation consists of the responses of all the filters in this bank. Thus, as the robot is looking around, an attentional sequence—as a spatio-temporal sequence of observations is generated. We then present a series of experimental results which show that (1) the robot can use average observation vectors to model complex scenes and (2) it can recognize relatively complex scenes even after a couple of saccades via comparing the current attentional sequence to the a priori learned scene models. In our ongoing work, we are investigating several issues including adaptive visual field and fovea sizes and application in visual search tasks.

5. Uncited references


Acknowledgements

Burak Bitlis has done the preliminary studies. We would like to express our thanks to Professor Gallant of University of California, Berkeley who has kindly helped us regarding material presented in some of his papers. This work has been supported by Bogazici University Research Fund grant #01A202D.

References


