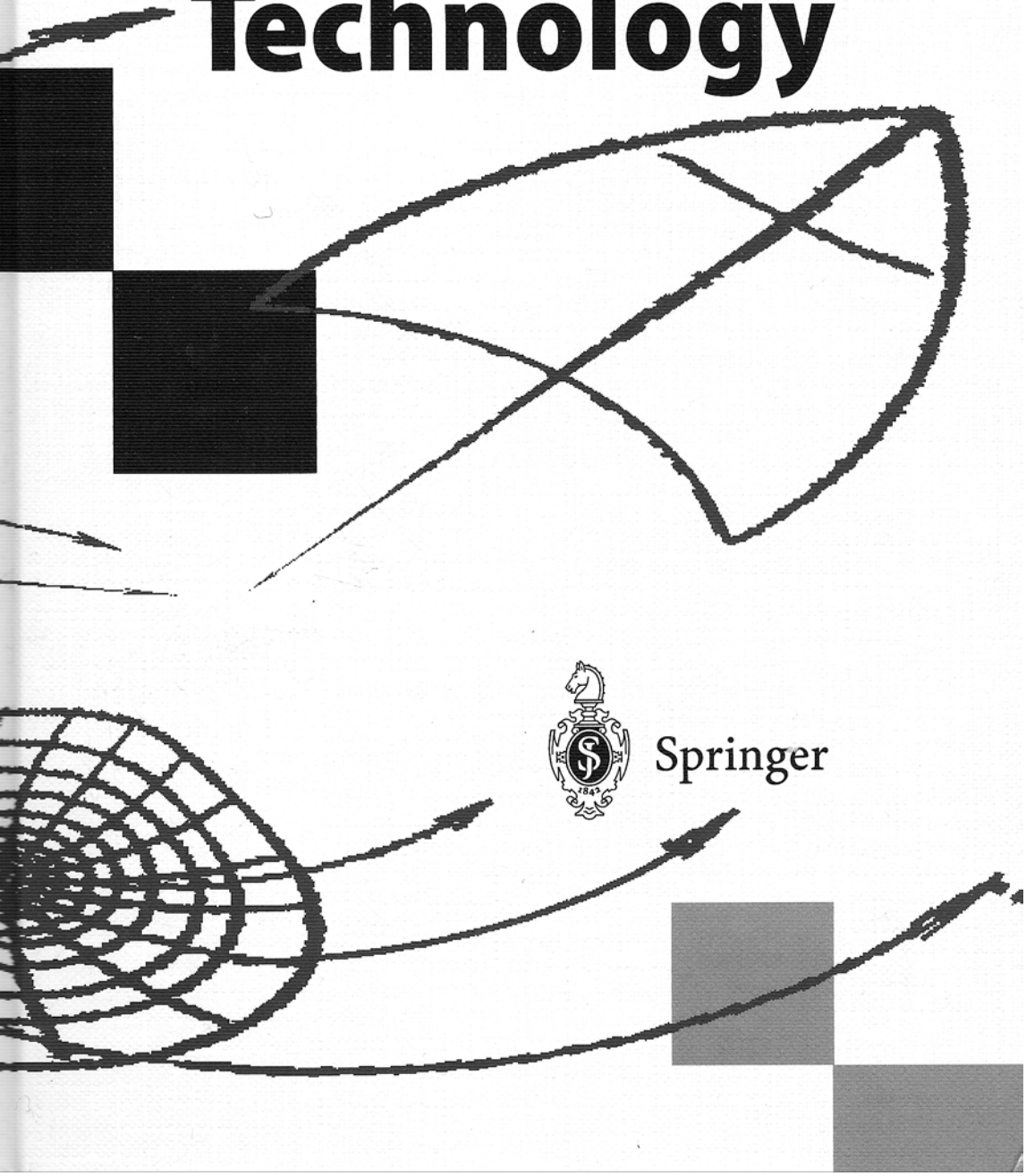
Jorge L. C. Sanz (Ed.)

Image Technology



Springer

Integrating Selective Attention and Space-Variant Sensing in Machine Vision

C. Colombo, M. Rucci, and P. Dario

Abstract. Studies on visual perception have demonstrated that selective attention mechanisms and space-variant sensing are powerful tools for focusing available computing resources to the process of relevant data. In this paper an overall architecture for an active, anthropomorphic robot vision system which integrates retina-like sensing and attention mechanisms is proposed. Gaze direction is shifted both on the basis of sensory and semantic characteristics of the visual input, which are extracted separately by means of a parallel and serial analysis. An implementation of the system by means of optical flow and neural network techniques is described, and the results of its application are discussed.

1 Introduction

During the last few decades, machine vision applications have often been hampered by the need of processing huge amounts of data. This led to the common belief that the major bottleneck in solving problems in vision was the computing power and image acquisition facilities [17]. Yet, only a small fraction of the raw image data may be relevant to the task at hand. That is, vision systems usually do not need to understand the scenes with which they deal, but they only need to extract the information required to accomplish specific tasks [7]. The idea of a system that purposively selects among visual data the relevant information and ignores irrelevant details is common to several recently proposed machine vision paradigms (e.g., [2, 4]), and it is crucial when a real time performance is required.

Specific mechanisms for data reduction and selection are also present at different levels of the human visual system, and play a major role to dramatically simplify visual computations. At the earliest stage of vision, the particular layout of the receptors of the retina – which are organized into a *space-variant* sampling structure including a high-resolution, small central *fovea*, and a *periphery* whose resolution linearly decreases with eccentricity – yields a good compromise between spatial acuity and field of view [34]. The mechanisms of *selective attention*, which allow for selectively processing simultaneous sources of visual information, act at later stages of human vision. By means of attention, computational resources are allocated to the processing of data included into a limited extent *spotlight*; the spotlight can be shifted through the visual field and can vary in size [19, 28]. Due to the space-variant structure of the retina, visual exploration in humans occurs by

actively shifting the fixation point, so as to exploit the high-resolution capabilities of the fovea [38]. Gaze control is mainly provided by attentional mechanisms, even if it has been shown that the spotlight can be voluntarily shifted independently of eye fixations [14]. As already hypothesized in 1890 by William James, attention can be drawn both by the *sensory* and *semantic* characteristics of a visual stimulus [18].

Due to the fact that the processing power of current computers is still by far lower than that of the brain, implementing similar mechanisms in machine vision is extremely important towards the development of effective, real time systems.

Recently, the space-variant structure of the human retina has received the interest of the research community, and a number of applications based on its simulations have been described [37, 16]. Furthermore, a hardware sensor mimicking the geometry of the human eye has been designed and developed, and it has been applied to bidimensional pattern recognition and motion estimation [32, 33, 35]. A few architectures have also been proposed, which attempt to replicate selective attention in machine vision systems, by the use of multi-resolution image data structures, or *pyramids* [29, 6, 10]. Data selection with such hierarchical structures is the result of a coarse-to-fine search through certain paths of the pyramids [7, 11, 27]. Also the concept of *saliency map* – a topographical map which combines individual sensory outputs into a global measure of attentional conspicuity, as first defined in 1987 by Koch and Ullman [21] – has been found useful for the implementation of attentive systems in robotics [9].

So far, research on attention in machines has been based on traditional high-resolution cameras, thus not exploiting the further data reduction advantages achievable by space-variant sensing. Furthermore, apart from some speculative considerations, the work has been focused either on the sensory (*bottom-up*) or the semantic (*top-down*) characteristics of the input data, while their combined use into a single architecture has been neglected.

In this paper, the integration of selective attention mechanisms and space-variant sensing towards the development of a real-time anthropomorphic vision system is proposed. The focus here is on merging both semantic and sensory cues in order to achieve active gaze control with a robotic head. The architecture includes a simplified saliency map, which is activated simultaneously by bottom-up and top-down cues, as resulting from a parallel and a serial analysis of the visual input, respectively.

The system is designed so as to attend to different visual tasks, and to change the way it interacts with the environment depending on them. The system implementation is based on a hybrid architecture, including neural and conventional image analysis techniques, such as optical flow and edge extraction. Thanks to its intrinsic modularity, the architecture is suitable to be expanded so as to include different tasks and to be adapted to other requirements. In the present implementation, particular interest has been devoted to motion cues at the sensory level and to object recognition at the semantic level, so that the head is able to foveate on and explore moving objects, which can be eventually recognized as far as they are already known to the system. As the system's goal is to control the motion of the

sensors according to a predefined task, the architecture proposed in this paper can be included in the *active vision* paradigm [1, 3].

The paper is organized as follows: in Sect. 2 the system architecture is presented and theoretical issues are discussed, then in Sect. 3 implementation details are provided and the results of a simulation of system behavior are described. Finally, in Sect. 4 conclusions are drawn.

2 The System Architecture

A general scheme of system architecture is shown in Fig. 1. This includes a robotic head including retina-like sensors, two modules for visual analysis and an attention controller. A basic assumption of this scheme is that shifts of attention are always accompanied by corresponding movements of the sensors, i.e. that *the attentional spotlight is kept centered on the fovea.* Due to this constraint, controlling attention also implies gaze control.

The system is organized as a loop. The raw image data coming from the whole visual field of the space-variant sensors are analyzed by a *parallel processing module*, and a set of salient locations of the image space (*sensory cues*) which are candidates for drawing attention are produced. At the same time, the visual region covered by the attentional spotlight is analyzed by a *serial processing module*, which produces *semantic cues* as the result of knowledge-based expectations. The role of the attention controller is to combine current proprioceptive data and

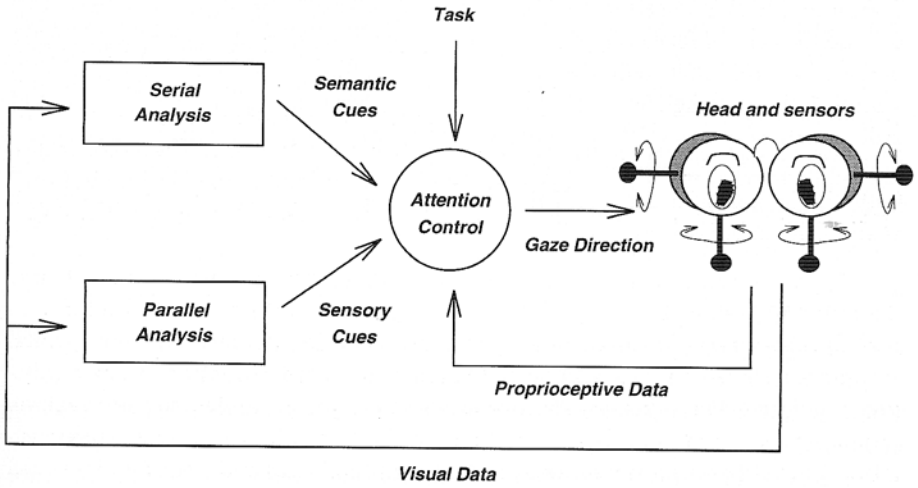


Fig. 1. The system architecture. Sensory and semantic cues are the result of a parallel and serial analysis of raw image data. The new direction of gaze is selected by the attention controller on the basis of the task at hand

salient cues so as to activate, also on the basis of the current task, a simplified version of the saliency map. Gaze direction is selected as the map location with maximum activation.

It is worth nothing that, as far as attention control is concerned, no formal distinction is made between sensory and semantic cues. These typically alternate in drawing attention. In face, after the detection and localization of a sensory relevant region of the visual field, movements of the eyes play an important role for its analysis and classification.

A number of theories of visual recognition have been proposed, based on the sequence of gaze directions [26, 25, 38]. According to these theories, the system performs recognition by serially looking at different parts and features of the examined object. The identification of a feature which characterizes a known object can stimulate the system to look in different directions searching for other features of that object, so as to better assess object identity. To this aim, the system incorporates a fragmentary representation of the objects to recognize.

The way the system responds to visual input changes according to the task at hand. That is, on the basis of the task, a priority can be assigned to different visual features.

In the following paragraphs, a detailed description of the elements of the architecture is given, in the case of a monocular implementation.

2.1 Head and Retina-Like Sensing

The sampling structure for retina-like sensing is shown in Fig. 2. The periphery around the central fovea is partitioned into M annuli $\times$ N angular sectors. The ratio of the outer and inner radii of each annulus is equal to a constant $a > 1$. By suitably scanning the sensing elements, the retinal periphery space (x, y) can be mapped onto a *cortical plane* (ξ, γ) by means of a *log-polar* (or *retino-cortical*) transformation [33]. Figure 3 gives an example of a space-variant image representation in both the retinal and the cortical planes. The log-polar transformation can be analytically expressed as:

$$\begin{cases} \xi = \log_a(\sqrt{x^2 + y^2}) - p \\ \gamma = q \arctan(y/x) \end{cases} \tag{1}$$

where p and q are constants determined by the size and layout of the sensing elements.

Figure 4 illustrates the geometry of a monocular head-eye system. In the case of a space-variant sensor characterized by equation (1), the field of view is 2δ, where

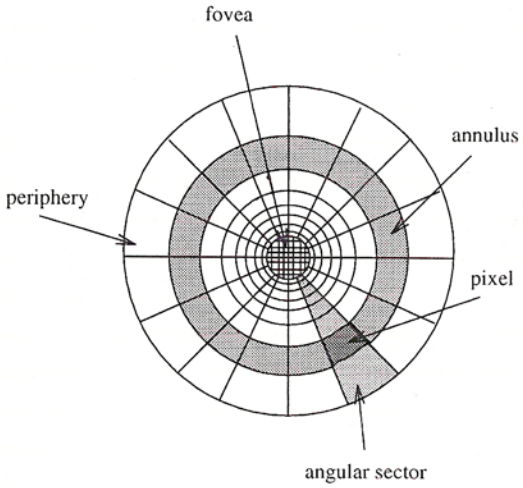


Fig. 2. The sampling structure of retina-like sensing

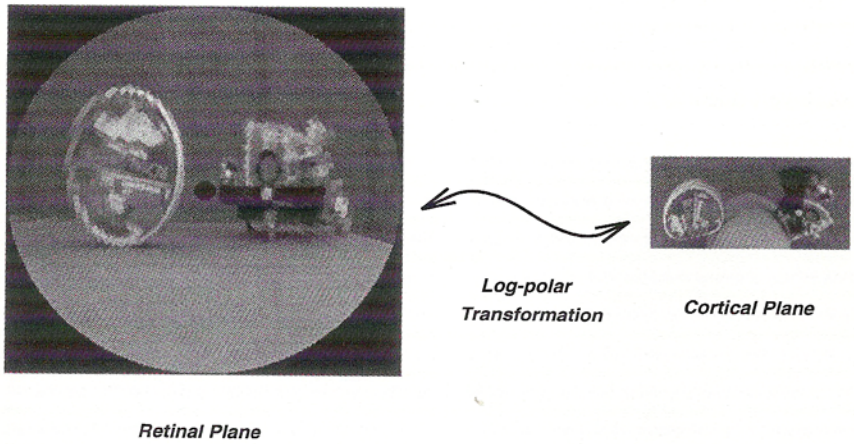


Fig. 3. Example of retina-like sensing. The objects imaged in the retinal plane (a cookie-box and a mini-robot) undergo a strong deformation after the log-polar transformation that maps them into the cortical plane. Notice also that image degradation is increasingly higher moving outwards from the fovea

$$\delta = \arctan(a^{M+p}/f), \tag{2}$$

and f denotes the focal length of perspective projection. The exponential term allows one to achieve either a wider field of view than with traditional cameras, if image resolution is kept fixed, or a reduction of the overall image data for a given field of view.

Figure 5 shows an hardware implementation of the retina-like sensor, which has been included in a binocular vision system recently developed at the ARTS Lab [12].

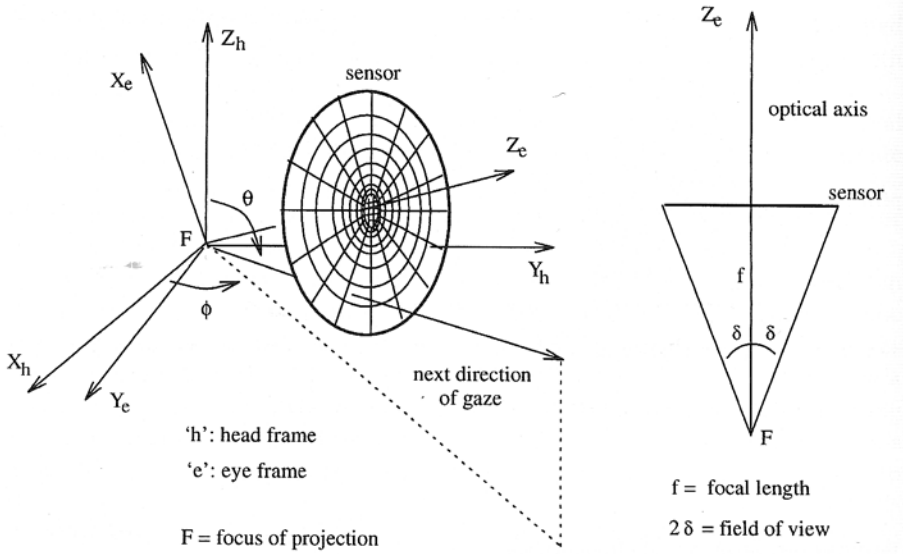


Fig. 4. The geometry of a monocular head-eye system. The two d.o.f. of the sensor are the *pan* (ϕ) and *tilt* (θ) angles. The head (fixed) and the eye (mobile) frame origins are assumed here to be coincident

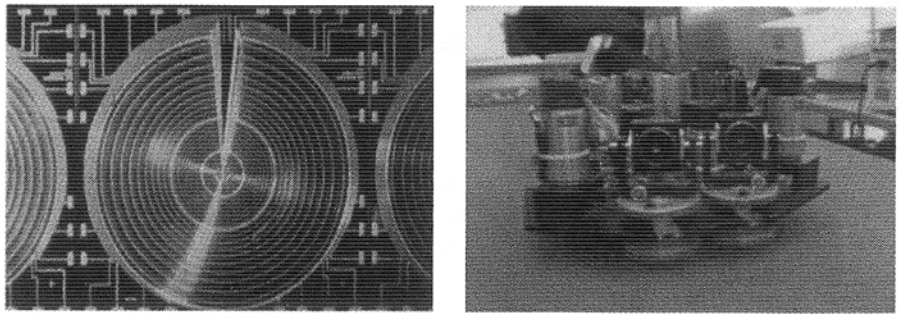


Fig. 5. *Left*: the CCD retina-like sensor built at IMEC, Leuven, Belgium. Several partners were involved in the design and development of the sensor: DIST – University of Genoa, Italy; University of Pennsylvania; ARTS Lab Pisa, Italy. *Right*: the robotic head developed at the ARTS Lab. The system is composed of a mechanical unit which includes two retina-like sensors, each actuated with two DOFs through DC servomotors, and a transputer-based control architecture

2.2 Parallel Analysis

Figure 6 presents the schematic organization of the parallel analysis module. The module includes as many pyramids as the number of different sensory features that are candidates for drawing system attention – e.g., image brightness, color, motion features, edge density, texture, etc.

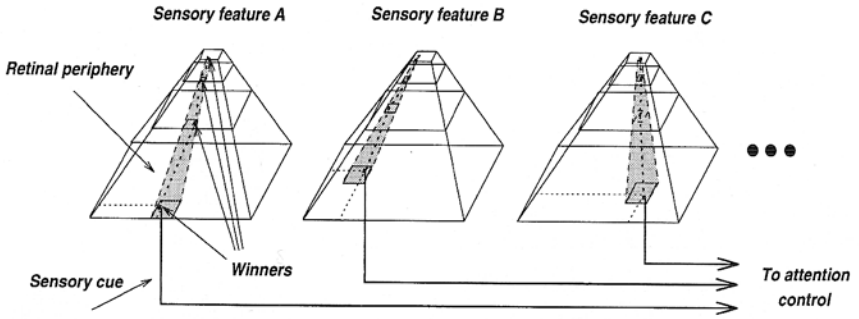


Fig. 6. Sensory pyramids and the Winner-Take-All mechanism at the parallel stage of analysis

A sensory cue is produced for each pyramid as the result of a coarse-to-fine search through all of its levels. All pyramids are scanned in parallel, and at a generic level L a *Winner-Take-All* mechanism acts so that only one "winner" is propagated to level $L - 1$, while "loosers" are inhibited [11, 36]. In such a way computations are greatly reduced, as only a small part of the visual data is processed at high resolution, while the rest is explored only at low resolution. The image coordinates of the 0th level winner of a pyramid are delivered to the attention control module as the sensory cue for that pyramid.

A characteristic of the pyramids used in this system is that they are built *on cortical images rather than on retinal images.* Such hierarchical structures will be referred to as "cortical pyramids." Note that, since cortical images encode only information from the periphery of the retina, data from the fovea are not considered at the parallel stage, but their analysis is left to the serial analysis module. Pixels in the retinal plane have nonuniform layout and dimensions – they are in fact distributed according to a polar-exponential geometry. Nonetheless, the corresponding pixels in the cortical plane are distibuted according to the same Cartesian geometry as the pixels of traditional raster cameras, thus allowing one to build up the cortical pyramids according to a classical algorithm [6, 29]. That is, the generic level $L + 1$, $L \geq 0$ of the pyramid is constructed by first smoothing and then subsampling by a factor of two in both the ξ and γ directions the cortical image at level L. As a result, as shown in Fig. 7, the new level of the hierarchy can be interpreted in the retinal plane as a novel retina with characteristic parameters $M_{L+1} = M_L/2$, $N_{L+1} = N_L/2$, $a_{L+1} = a_L^2$, $p_{L+1} = p_L/2$, $q_{L+1} = q_L/2$. Figure 8 shows a three-level cortical pyramid, and the corresponding retinal pyramid.

Note that, although constructing cortical pyramids has of course a smoothing effect, which contributes to reduce the noise in the raw image data, sensory cues are usually very noisy, and thus inadequate for a quantitative (absolute) analysis of

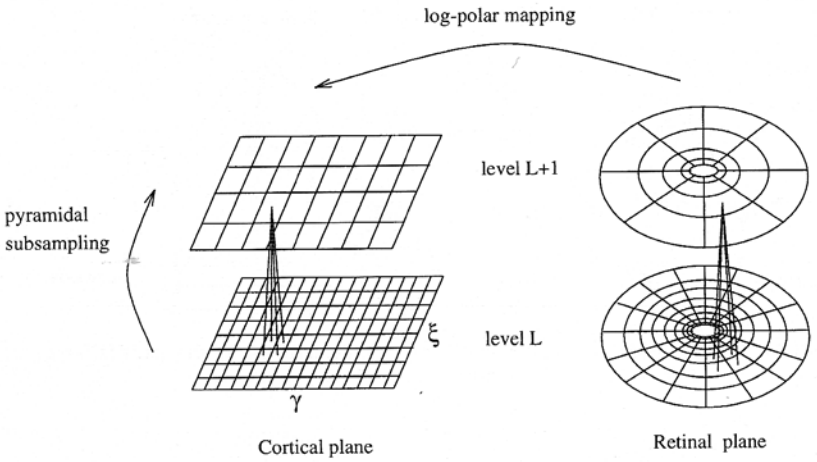


Fig. 7. *Left*: building binary cortical pyramids. *Right*: backtransforming a cortical pyramid yields a retinal pyramid: notice the presence of a hollow inner region correnspondingto the fovea

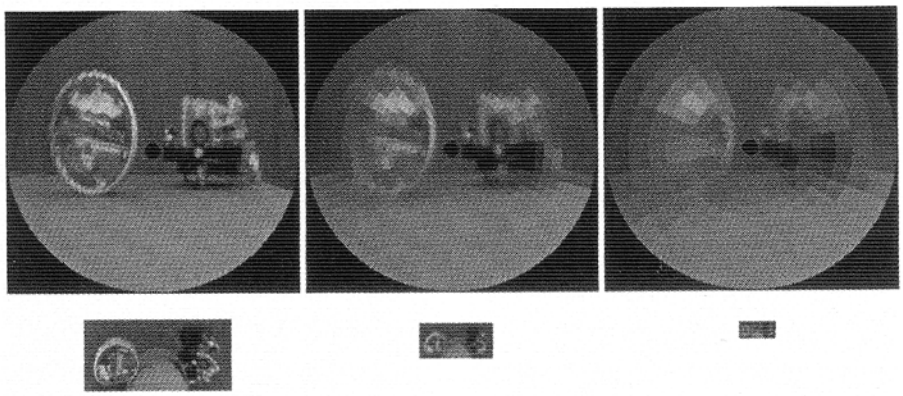


Fig. 8. An example of cortical pyramid. *Top*: retinal plane. *Bottom*: cortical plane

the visual scene[1]. The system is though robust enough to exploit the information present in the *relative* feature magnitude, and to draw attention to specific objects that have to be recognized.

Note also that, as far as a b/w sensor is considered, sensory pyramids related to features of increasing complexity can all be built starting from an image brightness pyramid. This leads to a further simplification, in that pyramid level L for a

[1] A related problem is the generation of spurious cues produced by image noise, which could produce gaze shifts to irrelevant visual directions. To avoid that, a simple thresholding mechanism is used at the lowest level of each sensory pyramid, in order to discriminate the signal from noise.

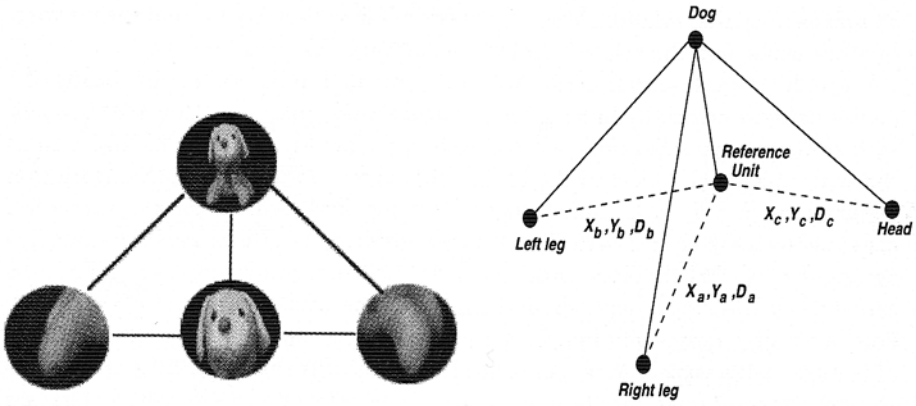


Fig. 9. *Left*: the fragmentary representation used for object recognition is composed of a set of icons having different levels of resolution, and linked by spatial relationships. *Right*: the representation is implemented with a neural network which includes a set of feature units (dashed circles) and an object unit (solid circle)

given sensory feature is actually constructed only when – and if – required by the top-down search mechanism. In the Appendix, the case of motion analysis is discussed, and the recursive construction of sensory pyramids for the *optical flow* and related features – such as its first-order *differential invariants* and the *immediacy of collision* – starting directly from cortical data is described. Such motion features provide powerful cues for the attentive control of a robot head. They can be effectively used to implement several different system tasks – such as searching for specific kinds of motions (e.g., pure translations or rotations), shifting gaze on the image feature with largest immediacy of collision, or pointing out the object in the scene with largest speed. In Sect. 3, some experiments are described in which the parallel analysis is based on the optical flow magnitude and the immediacy of collision.

2.3 Serial Analysis

As emphasized at the beginning of this section, semantic cues correspond to visual field locations where characteristics of the hypothesized object are expected to be found. The system includes a fragmentary representation of each object to recognize. As shown in the left part of Fig. 9, an *object representation* is basically a graph composed of fixed-dimension images (*icons*) reproducing different parts (or *features*[2]) and views of the object at varying level of resolution. The icons are linked

[2] The term "feature" indicates in this paragraph a characteristic part of an object and should not be confused with the sensory feature of the previous paragraph.

by means of spatial relationships, which specify how they are located with respect to other icons, and the relative dimensions of their features.

An object representation has been implemented in the system by means of a neural network which includes a set of *feature units* linked to a single *object unit* with excitatory connections (see the right part of Fig. 9). Each feature unit is sensitive to a specific part of the object, that is to all the icons reproducing that feature, and it acts as a cumulator, by storing and cumulating the activation provided by a matching network. As shown in the figure, the spatial relationships among the features are given with respect to the feature unit sensitive to the icons centered on the object centroid and enclosing the whole object (*reference unit*). For each object representation, the parameters (x_i, y_i) and D_i indicate the position of feature i with respect to object centroid and the dimension that the attentional spotlight should have for its examination, respectively. All the parameters are given in the retinal space, and they are normalized with respect to the object size. In this way, if the reference unit is activated when the width of the attentional spotlight is equal to D_s, so that gaze is shifted, with respect to feature i by an amount proportional to $(x_i\, D_s, y_i\, D_s)$ in the image plane and the spotlight size is correspondingly set to $D_i\, D_s$.

Starting from each feature, the parameters for the examination of each other part of the object are determined by following the graph and passing through the reference unit. As will be explained later on, when required the spatial relationships are estimated by a neural network so as to adapt to the actual orientation of the examined object. The use of a scale-independent object representation allows object recognition capabilities even with different focal lengths and object distances.

All object units inhibit each other in a Winner-Take-All fashion, so that only one of them has a positive value of activation at a given time, while all the others are inhibited (output equal to zero). This is equivalent to the formulation of an hypothesis on the identity of the observed object. In this terms, recognition is intended as the eventual determination of a winning object unit.

The structure of the serial analysis module is illustrated in Fig. 10 – a preliminary of its application to a simple case of object recognition using conventional sensors is described in detail in [30]. Image data are organized into a multiresolution edge pyramid, which represents image edges at different levels of resolution. The pyramid is built on both data coming from the fovea and the periphery of the sensor. The edges are extracted at the lowest pyramid level by a gradient operator and data are then propagated at successive stages by means of Gaussian smoothing. As shown in the figure, the pyramid is scanned by an attentional spotlight centered on the fovea which, by moving through a "fovea-centered cylinder," samples a fixed amount of information at different levels of resolution. As a result, a trade-off is built between the resolution level and the spatial extension of the considered area, and an increment of the width of the area implies a corresponding decrement of the level of resolution at which data are examined. The spotlight performs an expansion of the gray-level dynamic and produces a fixed dimension *attentional icon*. By means of the expansion, parts of the examined area

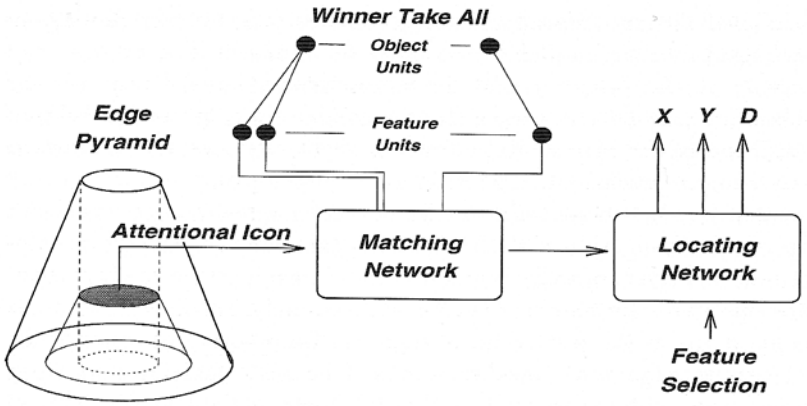


Fig. 10. The architecture for visual recognition (serial processing module)

with stronger edges and/or higher edge concentration are emphasized with respect the others. When it is not involved in a recognition sequence, the spotlight scans the cylinder from the bottom to the top of the pyramid. This is equivalent to performing a radial expansion of the considered region, starting from the fovea and moving towards the periphery. The resulting attentional icon is compared at each time with all the icons included in the object representation. If a "semantic match" is found, the pyramid scanning process stops and the spotlight dimension is stored into a short-term memory (not shown in the figure). The size of the attentional spotlight is later used for calculating the spatial parameters of subsequent fixations.

The attentional icon is matched in parallel with all the icons of the memory, so as to activate hypotheses on object identity. Hypotheses are ordered according to their plausibility and their are sequentially verified. Each verification involves the serial analysis of all the parts and features of the hypothesized object so as to test if other matches are achieved.

The matching process is carried out by a *matching net*[3], which is implemented with the counterpropagation paradigm [13]. The matching network has as many output units as are the features to identify, and each output is linked with the corresponding feature unit in the object representations. As proposed by Kohonen [23], the topological self-organizing map at the second layer of the net is trained by means of an unsupervised learning process which requires the recursive presentation of patterns of the training set to the net. During the training, the weight vectors of the maximally responding unit and those of its immediate neighborhood are modified towards the input pattern, whereas the size of the neighborhood and the parameters regulating weight changes decrease. The train-

[3] A "Weak perspective" projection model [24] is assumed here, thus avoiding significant deformations of the peripheral parts of the foveated objects; such deformations could seriously complicate the iconic representation.

ing set includes all the icons reproducing the features selected for representing the objects, extracted in a large number of images. By means of learning, the map selects actually which icons to use for the representations on the basis of their statistical relevance, and stores generalized versions of them in the first layer of weights. In this way, the map self-organizes so as produce several disconnected areas where units are sensitive to similar icons.

The spatial parameters are estimated by a *locating net* which is trained with the back-propagation algorithm [31]. The network, illustrated in Fig. 11, includes four full-connected layers of weights and it is split in two parts in the first layer, where different input information are processed separately. The first set of inputs has as many inputs as the feature units' number. The number of input of the second set is equal to the sum of the dimensions of the self-organizing map. Three sets of outputs code, with a sparse coding, the coordinates of the spatial location of the feature and the spotlight dimension required for its examination, respectively. The locating net is trained so as to produce the spatial parameters of a feature when the feature is selected with the first set of inputs, while the position of the winning unit on the self-organizing map which activated the reference unit (stored in the short-term memory) is specified with the second set. If a feature not corresponding to the reference unit is examined, the location of the centroid and the object size – i.e., the parameters needed for activating the reference unit in the subsequent fixation – are found by coding the current winning position and the desired feature. In this way it is always possible to reach a global analysis starting from a detail, and to search then again for other details.

The basic strategy followed by the system for recognition is to examine sequentially all the features of an hypothesized object. When a feature to be analyzed is chosen, its spatial location and the spotlight size required for its

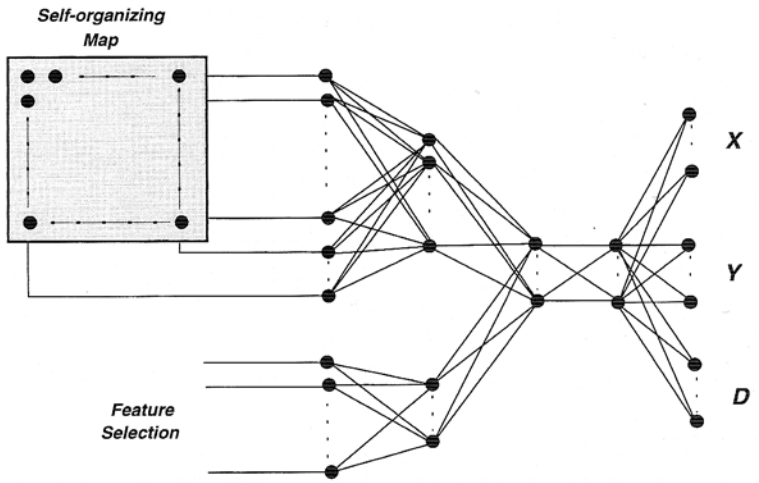


Fig. 11. The locating network is a five-layer full-connected net, which has been trained with the backpropagation algorithm to provide the spatial parameters of the representation

examination are found by converting the numerical values provided by the locating net by means of the parameters stored in the short term memory. Recognition is achieved if the global input to the object unit is larger than a predetermined threshold. If an hypothesis fails to be confirmed by successive attentional fixations, all the feature units in the representation of the rejected object are reset. In this way, the second most probable hypothesis wins the competition and its features are then analyzed. The cycle is repeated until recognition is achieved or all the formulated hypotheses are sequentially examined.

2.4 Attention Control

The sensory cues generated by parallel analysis and the semantic cues produced by the serial analysis activate spatial locations in the saliency map. The saliency map can be seen as a not retinotopic plane including all visual directions. Each gaze direction is represented in the map by the point identified by the spherical coordinates θ (co-latitude) and ϕ (longitude) relative to a head frame centered on the focus of perspective projection – see Fig. 4.

Salient cues are expressed by the processing stages in retinotopic coordinates, and they are transformed later into visual directions on the basis of the focal length – sensor plane-eye coordinate transformation – and of "proprioceptive" data – the relative rotation between the eye and the head coordinate systems – obtained from the position sensors of the actuation subsystem. Specifically, if g_s is a vector that represents the next direction of gaze as expressed in the sensor coordinate frame, its expression g_h in the head reference frame, which has to be provided to the motor subsystem, can be simply evaluated as

$$g_h = {}^{e}_{h}\mathrm{T}\,{}^{s}_{e}\mathrm{T}\,g_s, \tag{3}$$

where the matrices ${}^{e}_{h}\mathrm{T}$ and ${}^{s}_{e}\mathrm{T}$ encode the homogeneous transformations between the eye/head and between the sensor/eye frames, respectively – refer again to Fig. 4.

It should be noted that the set of directions of gaze explored by the sensor at a given time is a subset of the map, whose size depends on the field of view. At everytime, the activations provided to the saliency map (salient cues) are gated by the considered task. That is, the task selects the level of priority for a generical cue c by assigning it a weight $w_c \in [0, 1]$. By properly arranging the weight values, it is possible to select a sensory feature with respect to others and/or to inhibit irrelevant cues. In this way, the task currently accomplished by the system changes the way it responds to visual stimuli and, eventually, the way it interacts with the external world.

3 Simulation Results

In this section, experimental results obtained with a simulation of the system behavior are described and discussed. The simulation has been carried out on a

SUN SPARCstation, with a program written in C. Space-variant images have been obtained by resampling high-resolution 256 × 256 images acquired with a b/w raster camera.

Three toy-objects, which are shown in the sequence of Fig. 12, were included into the system memory for recognition: a dog, a train, and a cat. The icons were circles of 156 pixels – approximately the size of the fovea – and they were classified by a topological self-organized map composed of 400 units (20 units along each axis). Each object representation included 5 feature units. The locating network was composed of 55 (40 + 15) input units, 15 units in both the hidden layers and 15 (5 + 5 + 5) units in the output layer. A number of images for each object were used to train the nets of the system: windows located on the representative features were extracted from the images and their positions and dimensions were stored. The topological feature map was trained by means of the icons corresponding to the selected windows and the locating net with their spatial relationships. To achieve a further computational gain, only the considered part – i.e., the attentional icon – of the edge pyramid was produced.

The retina was composed of $M = 64$ annuli and $N = 128$ angular sectors, the radii-ratio was $a = 1.04621$, and p was set to 43.32067. The field of view was approximately 45 deg. Three pyramid levels were built in the cortical plane, using a Gaussian-shaped smoothing filter with a mask size of 3 pixels. At the top pyramid level, which was the only one to be completely processed, a data reduction of approximately two orders of magnitude was achieved with respect to raster images. Two motion features were used as sensory features for the system: the magnitude of the optical flow, and the upper bound on the immediacy of collision,

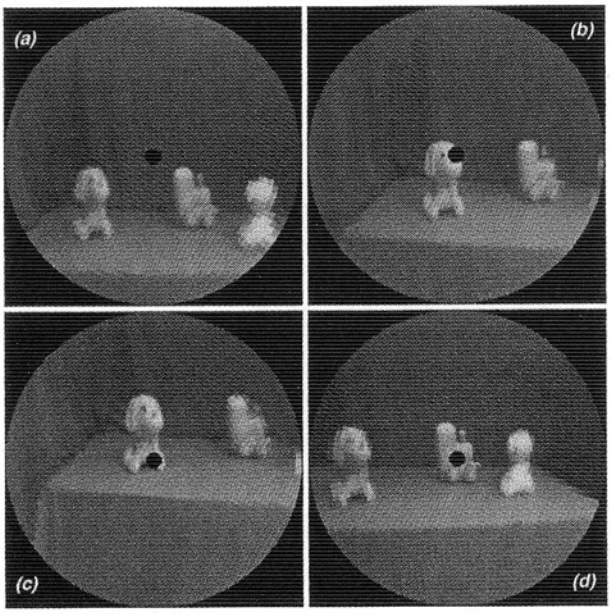


Fig. 12. Some frames taken by the system at work. Black dots indicate the foveal region, that surrounds the fixation point

obtained from the cortical flow and its derivatives using equations (5) and (9), respectively – see Appendix. The cortical flow was computed on a three-frame basis, by means of a motion-boundary preserving algorithm based on a multiwindow least squares technique [5]. The cortical flow was regularized by a vector median filter, having a filter mask size of 3 pixels.

In the following, an experiment illustrating the system behavior is described. The three objects represented into the system were located in the scene, and different motions were assigned to each of them. In a first phase, the dog approaches the camera, the cat goes away, and the train is still. After a while, the dog stops and the train departs moving parallel to the camera. Task weights are assigned so that recognition has the highest priority ($w_r = 1.0$), followed by immediacy detection ($w_i = 0.6$) and speed detection ($w_s = 0.2$).

At the beginning, as shown in Fig. 12a, none of the objects is foveated, so that no semantic cues are produced by the serial analysis stage. The results of the parallel analysis are shown in Fig. 13a,b. As it can be noticed, the magnitude of the optic flow field is greater for the cat than for the dog. Nonetheless, system attention is drawn by the approaching dog, due to its larger value of immediacy of collision. The train is not considered at all in this phase, because it is not in motion.

After the first foveation on the dog (Fig. 12b), a semantic match occurs, so that a recognition sequence is activated. Figure 12c shows the last of the four foveations required for the recognition process.

While the system is involved in the recognition "scanpath," the train starts moving, but the corresponding sensory cue (Fig. 13c,d) does not gain the system

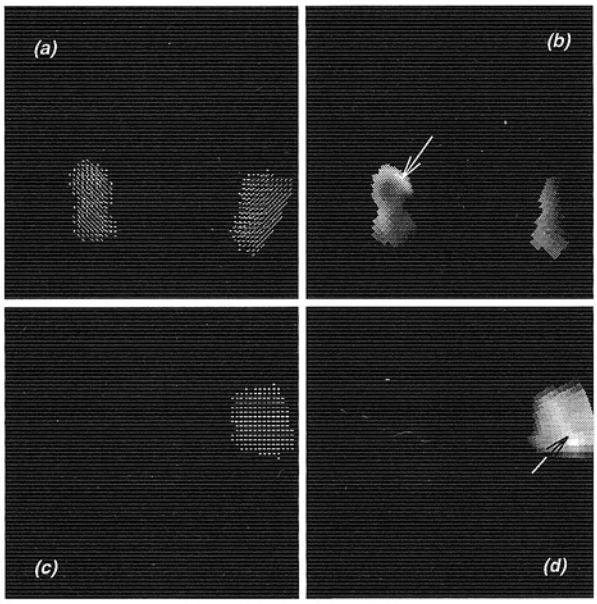


Fig. 13. Optical flow and sensory features. Brighter areas correspond to higher feature values. Features are as they would look if the whole 0th level of the pyramid was processed. Actually, only the cues indicated by the arrows were computed

attention. However, after that object identity is assessed, no more semantic cues are available and optical flow magnitude becomes the most prominent cue. As a result, gaze direction is shifted to the train (Fig. 12d).

Other experiments have shown that recognition performance is good, in that objects are correctly identified in more than 90 percent of the cases, all the errors being due to the failing of the first semantic match during the pyramid scanning.

4 Conclusions

Studies on visual perception have demonstrated that evolution has developed powerful tools – such as selective attention mechanisms and space-variant sensing – for focusing the available computing resources to the process of relevant data. Implementing similar mechanisms in machine vision is extremely important towards the development of effective systems. In fact, in spite of the great improvements of the recent years, the processing power of current computers is still by far lower than that of the brain.

The system described in this paper is an example of how biological theories and concepts can be effectively exploited in machine vision applications. The simultaneous use of selective attention mechanisms and retina-like sensing allows a dramatical computational gain, which encourages complex real-time applications. Besides, the integration of sensory and semantic characteristics of the visual data and the dependency on the task at hand are indicated as fundamental issues for the development of an "intelligent" behavior.

Basically, two directions of future research can be outlined. On the one side, other visual features and even other sensory modalities, such as touch and hearing, could be included in the system processing stages so as to produce new cues for the saliency map. On the other, the learning capabilities of the system could be improved. By means of on-line learning, achievable among other techniques also by different neural network paradigms, the system could be able to autonomously select relevant features and to create new object representations.

Acknowledgements. The work described in this paper has been supported by MURST and by the Special Project on Robotics of the National Research Council of Italy. We thank Dr. D.M. De Micheli for providing us the implementation details of the ARTS head.

References

[1] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. *Active vision.* International Journal of Computer Vision **1**(4), pp. 333–356, 1988.
[2] Y. Aloimonos. *Purposive and qualitative active vision.* In: "Artificial Intelligence and Computer Vision", Y.A. Feldman and A. Bruckstein eds., Elsevier, 1991.
[3] R. Bajcsy. *Active perception.* Proc. of the IEEE **76**(8), pp. 996–1005, 1988.
[4] D.H. Ballard. *Animate vision.* Artificial Intelligence **48**, pp. 57–86, 1991.

[5] F. Bartolini, V. Cappellini, C. Colombo, and A. Mecocci. *Multiwindow least squares approach to the estimation of optical flow with discontinuities.* Optical Engineering, **32**(6), pp. 1250–1256, 1993.
[6] P.J. Burt and E.H. Adelson. *The Laplacian pyramid as a compact image code.* IEEE Transactions on Communications **31**(4), pp. 532–540, 1983.
[7] P.J. Burt. *Smart sensing within a pyramid vision machine.* Proc, of the IEEE **76**(8), pp. 1006–1015, 1988.
[8] R. Cipolla and A. Blake. *Surface orientation and time to contact from image divergence and deformation.* Proc. 2nd European Conference on Computer Vision, pp. 187–202, S. Margherita Ligure (Italy) 1992.
[9] J.J. Clark and N.J. Ferrier. *Attentive visual servoing.* In: "Active vision", A. Blake and A. Yuille eds., MIT Press, 1992.
[10] J.L. Crowley. *A representation for visual information.* Tech. Rep. CMU-RI-TR-82-7, Carnegie-Mellon University, 1987.
[11] S.M. Culhane and J.K. Tsotsos. *An attentional prototype for early vision.* Proc. 2nd European conference on Computer Vision, pp. 551–560, S. Margherita Ligure (Italy) 1992.
[12] D.M. De Micheli, M. Bergamasco, and P. Dario. *An anthropomorphic active vision system based on a retina-like sensor.* Proc. 3rd International Symposium on Measurement and Control in Robotics, Torino (Italy) September 1993.
[13] R. Hect-Nielsen. *Applications of the counter-propagation networks.* Neural Networks **2**(1), 1988.
[14] H. von Helmholtz. "Psychological optics". J.P.C. Sothall ed., Dover, New York, 1866/1925.
[15] B.K.P. Horn and B.G. Schunck. *Determining optical flow.* Artificial Intelligence **17**, pp. 185–203, 1981.
[16] R. Jain, S.L. Bartlett, and N. O'Brien. *Motion stereo using ego-motion complex logarithmic mapping.* IEEE Transactions on Pattern Analysis and Machine Intelligence **9**(3), pp. 356–369, 1987.
[17] R.C. Jain and T.O. Binford. *Ignorance, myopia, and naiveté in computer vision systems.* Computer Vision, Graphics and Image Processing: Image Understanding **53**(1), pp. 112–117, 1991.
[18] W. James. "The principles of psychology", Harvard University Press, Cambridge, 1890/1983.
[19] W.A. Johnston and V.J. Dark. *Selective attention.* Annual Review of Psychology **37**, pp. 43–75, 1986.
[20] K. Kanatani. "Group-theoretical methods in image understanding", Springer, 1990.
[21] C. Koch and S. Ullman. *Shifts in selective visual attention: toward the underlying neural circuitry.* In: "Matters of intelligence". L.M. Vaina ed., D. Reidel Pub. Comp., 1987.
[22] J.J. Koenderink and A.J. van Doorn. *Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer.* Optica Acta **22**(9), pp. 773–791, 1975.
[23] T. Kohonen. *Self-organized formation of topologically correct feature maps.* Biological Cybernetics **43**, pp. 59–69, 1982.
[24] J.L. Mundy and A. Zisserman. *Projective geometry for machine vision.* In: "Geometric invariance in computer vision", J.L. Mundy and A. Zisserman eds., MIT Press, 1992.
[25] K. Nakayama. *The iconic bottleneck and the tenuous link between early visual processing and perception.* In: "Vision: coding and efficiency", C. Blakemore ed., University Press, 1991.
[26] D. Noton and L. Stark. *Eye movements and visual perception.* Scientific American, **224**(6), pp. 34–43, 1971.
[27] B. Olshausen. *A neural model of visual attention and invariant pattern recognition.* Tech. Rep. CalTech, CNS Memo 18, September 1992.

[28] M. Posner. *Orienting of attention.* Quarterly Journal of Experimental Psychology **32**, pp. 3–25, 1980.
[29] A. Rosenfeld *et al.* "Multiresolution image processing and analysis", A. Rosenfeld ed., Springer, 1984.
[30] M. Rucci and P. Dario. *Selective attention mechanisms in a vision system based on neural networks.* Proc. International Conference on Intelligent Robots and Systems, Yokohama (Japan) July 1993.
[31] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. *Learning internal representations by error propagation.* In: "Parallel Distributed Processing", MIT Press, 1986.
[32] G. Sandini and V. Tagliasco. *An anthropomorphic retina-like structure for scene analysis.* Computer Graphic and Image Processing **14**(3), pp. 365–372, 1980.
[33] G. Sandini and P. Dario. *Active vision based on space-variant sensing.* Proc. 5th International Symposium of Roboticsw Research, pp. 408–417, Tokio 1989.
[34] E.L. Schwartz. *Spatial mapping in the primate sensory projection: analytic structure and relevance to perception.* Biological Cybernetics **25**, pp. 181–194, 1977.
[35] M. Tistarelli and G. Sandini. *Estimation of depth from motion using an anthropomorphic visual sensor.* Image and Vision Computing **8**(4), pp. 271–278, 1990.
[36] J.K. Tsotsos. *Analyzing vision at the complexity level.* The Behavioral and Brain Sciences **13**, pp. 423–469, 1990.
[37] C.F.R. Weiman. *Tracking algorithms using log-polar mapped image coordinates.* Proc. SPIE Intelligent Robots and Computer Vision VIII: Algorithms and Techniques, pp. 843–853, Philadelphia (Pennsylvania) 1989.
[38] A.L. Yarbus. "Eye movements and vision", Plenum Press, 1967.

A Space-Variant Motion Analysis

In this Appendix, the case of motion analysis, and specifically the construction of cortical pyramids for the optical flow and related features, is described.

Although defined as an approximation of the velocity field $(\dot{x}, \dot{y})$ – the projection onto the retinal plane of the tridimensional speed of the objects in the scene relative to the sensor –, the optical flow of points of the retinal periphery can be computed directly from cortical image data using the log-polar transformation (1). Let us define the "cortical flow" as the optical flow-like field $(\dot{\xi}, \dot{\gamma})$ arising due to brightness changes in the cortical plane. This field satisfies the equation [15]:

$$\frac{\partial E}{\partial \xi}\dot{\xi} + \frac{\partial E}{\partial \gamma}\dot{\gamma} + \frac{\partial E}{\partial t} = 0, \tag{4}$$

where $E(\xi, \gamma, t)$ is the cortical image brightness. Then, assuming that $(\dot{\xi}, \dot{\gamma})$ is the cortical flow at a generic pyramid level L (the subscript $_L$ from the retinal parameters is omitted for the sake of simplicity), the corresponding optical flow is evaluated as

$$\begin{cases} \dot{x} = a^{\xi+p}\left[(\ln a\ \dot{\xi})\cos(\gamma/q) - \left(\frac{1}{q}\dot{\gamma}\right)\sin(\gamma/q)\right] \\ \dot{y} = a^{\xi+p}\left[(\ln a\ \dot{\xi})\sin(\gamma/q) + \left(\frac{1}{q}\dot{\gamma}\right)\cos(\gamma/q)\right]. \end{cases} \tag{5}$$

In the same way – that is, by exploiting the retino-cortical transformation – , other interesting motion parameters can be computed from the cortical flow pyramid. The *differential invariants* of the optical flow, which are nothing but proper combinations of the first spatial derivatives of the flow [20]:

$$\begin{cases} \text{divergence} = \dfrac{\partial \dot{x}}{\partial x} + \dfrac{\partial \dot{y}}{\partial y} \\ \text{curl} = -\dfrac{\partial \dot{x}}{\partial y} + \dfrac{\partial \dot{y}}{\partial x} \\ \text{shear} = \left[\left(\dfrac{\partial \dot{x}}{\partial x} - \dfrac{\partial \dot{y}}{\partial y} \right)^2 + \left(\dfrac{\partial \dot{x}}{\partial y} + \dfrac{\partial \dot{y}}{\partial x} \right)^2 \right]^{1/2} \end{cases}, \tag{6}$$

are related to some simple geometrical and kinematic characteristics of the imaged scene [22]. Using again equation (1), it can be easily shown that the optical flow invariants can be computed from the cortical flow and its spatial derivatives as:

$$\begin{cases} \text{divergence} = 2 \ln a \, \dot{\xi} + \dfrac{\partial \dot{\xi}}{\partial \xi} + \dfrac{\partial \dot{\gamma}}{\partial \gamma} \\ \text{curl} = \dfrac{2}{q} \dot{\gamma} - q \ln a \dfrac{\partial \dot{\xi}}{\partial \gamma} + \dfrac{1}{q \ln a} \dfrac{\partial \dot{\gamma}}{\partial \xi} \\ \text{shear} = \left[\left(\dfrac{\partial \dot{\xi}}{\partial \xi} - \dfrac{\partial \dot{\gamma}}{\partial \gamma} \right)^2 + \left(q \ln a \dfrac{\partial \dot{\xi}}{\partial \gamma} + \dfrac{1}{q \ln a} \dfrac{\partial \dot{\gamma}}{\partial \xi} \right)^2 \right]^{1/2} \end{cases}. \tag{7}$$

Besides, simple linear combinations of the optical flow invariants can be directly related to the *immediacy of collision* (the reciprocal of the *time to collision*, or the time it takes before the object and sensor collide) and the *cyclotorsion* – the rotational component of the sensor speed along the optical axis. If the field of view is sufficiently small, these two quantities can be bounded by [8]:

$$\text{immediacy} \approx \frac{\text{divergence} \pm \text{shear}}{2} \tag{8}$$

$$\text{cyclotorsion} \approx \frac{\text{curl} \pm \text{shear}}{2}. \tag{9}$$