Issues in Robot Vision

Goesta H. Granlund
Computer Vision Laboratory,
Linköping University,
581 83 Linköping, Sweden

Abstract

In this paper we will discuss certain issues regarding robot vision. We will deal with aspects of pre-attentive versus attentive vision, control mechanisms for low level focus of attention, representation of motion as the orientation of hyperplanes in multdimensional time-space. Issues of scale will be touched upon.

1 Introduction

In this paper we will discuss certain issues regarding robot vision. Due to the limited format, it will by no means be possible to give any comprehensive overview of the field. Most aspects will in fact have to be omitted, and it is unfortunately not possible to give due references to all significant contributions within the area.

Machine vision has developed over the years, and different methodologies have been emhasized as crucial at different times. Classically, the methodology of image analysis contains many procedures to perform various tasks [16, 3, 23]. A common problem is that these procedures are often not suitable as components of a larger system, where procedures interact. One reason is that information is represented in different ways for different types of features. It is difficult to have such descriptors cooperate and to control each other in a structured, parametrical way.

An important feature of current state of the art is the view that sufficiently efficient interpretation of complex scenes can only be implemented using an adaptive model structure. In the infancy of computer vision, it was believed that objects of interest could unequivocally be separated from the background using a few standard operations applied over the entire image. It turns out, however, that this simple methodology only works on simple images having a good separation between object and background. In the case of more difficult problems with noise, ambiguities and disturbances of different types, more sophisticated algorithms are required with provisions to adapt themselves to the image content. A further extension of this adaptivity is the current development of *Active Vision* [46].

It consequently turns out to be necessary to use what we may call different sub-algorithms on different parts of an image. The selection of a particular sub-algorithm is often based upon a tentative analysis of the image content. The reason for using different sub-algorithms is the simple fact that all possible events can not be expected in a particular context. In order for this handling of sub-algorithms to be manageable, it has to be implemented as a parameterization of more general algorithms.

In some of the work cited, as well as in our own work, there has been taken a great deal of impression from what is known about biological visual systems [24, 25, 35]. This is not to say that we assume that the structures presented are indeed models used in biological visual systems. Too little is so far known to

BMVC 1993 doi:10.5244/C.7.1

form any firm opinions on the structures used. The ultimate criterion is simply performance from a technical point of view.

2 Pre-attentive Versus Attentive Vision

Classically, most vision procedures have been applied uniformly over an image or a scene. Such an indiscriminate application of computation power is very costly, and as complexity in the desired processing is increasing, it becomes necessary to find methods to restrict the attention to regions of maximal importance.

Humans can shift the attention either by moving the fixation point or by concentrating on a part of the field of view. The two types are called *overt* and *covert* attention respectively. The covert attention shifts are about four times as fast as the overt shifts. This speed difference can be used to check a potential fixation point to see if it is worthwhile moving the gaze to that position.

A number of paradigms describing human focus of attention have been developed over the years [36]. We will here mainly discuss the *search light* metaphor [26]. A basic assumption of this metaphor is the division between *preattentive* and *attentive* perception. The idea is that the preattentive part of the system makes a crude analysis of the field of view. The attentive part then analyzes areas of particular interest more closely. The two systems should not be seen as taking turns in a time multiplex manner, but rather as a pipeline where the attentive part uses the continuous stream of results from the preattentive part as clues. The reason for having to focus the attention in this metaphor is that certain tasks are of inherently sequential nature, rather than amenable to a parallel processing.

What features or properties are important for positioning the fixation point? Yarbus pioneered the work on studying how humans move the fixation point in images depending on the wanted information [51]. For preattentional shifts, gradients in space and time, i.e high contrast areas or motion, are considered to be the important features. Abbott and Ahuja present a list of criteria for the choice of the next fixation point [1]. Many of the items in the list relate to computational considerations. A few clues from human visual behavior were also included, of which the following is a sample:

Absolute distance and direction If multiple candidates for fixation points are present, the ones closer to the center of the viewing field are more likely to be chosen. Upward movement is generally preferred over downward movement.

2D image characteristics If polygonal objects are presented, points close to corners are likely to be chosen as fixation points. When symmetries are present, the fixation point tends to be chosen along symmetry lines.

Temporal changes When a peripheral stimulus suddenly appears, a strong temporal cue often leads to a movement of the fixation point towards the stimulus.

Since fixation point control is a highly task dependent action, it is probably easy to construct situations that contradict the list above. The reader is urged to go back to the appropriate references in order to get a full description of how the results where obtained.

2.1 Focus of attention in machine vision

A number of research groups are currently working on incorporating focus of attention mechanisms in computer vision algorithms. This section is by no means a comprehensive overview, but rather a few interesting examples.

Ballard and Brown have produced a series of experiments with ocular reflexes and visual skills [4, 9, 11, 10, 5]. The basic idea is to use simple and fast image processing algorithms in combination with a flexible, active perception system.

A focus of attention system based on salient features has been developed by Milanese [37]. A number of features are extracted from the input image and are represented in a set of feature maps. Features differing from their surroundings are moved to a corresponding set of conspicuity maps. These maps consist of interesting regions of each feature. The conspicuity maps are then merged into a central saliency map where the attention system generates a sequence of attention shifts based on the activity in the map.

Brunnström, Eklund and Lindeberg have presented an active vision approach to classifying corner points in order to examine the structure of the scene. Interesting areas are detected and potential corner points scrutinized by zooming in on them [12]. The possibility of actively choosing the imaging parameters, e.g. point of view and focal length, allows the classification algorithm to be much simpler than for static images or pre–recorded sequences.

A variation of the search light metaphor, called the attentional beam has been developed by Tsotsos and Culhane [14, 43, 44]. It is based on a hierarchical information representation where a search light on the top is passed downwards in the hierarchy to all processing units that contribute to the attended unit. Neighboring units are inhibited. The information in the 'beamed' part of the hierarchy is reprocessed, without the interference from the neighbors, the beam is then used to inhibit the processing elements and a new beam is chosen.

The ESPRIT Basic Research Action project 3038, Vision as Process [46], is designed to study the scientific hypothesis that vision should be handled as a continuous process. The project is aimed at bringing together knowhow from a wide variety of research fields ranging from low level feature extraction and ocular reflexes through object recognition and task planning.

Westelius, Knutsson and Granlund have developed a hierarchical gaze control structure for use with multi-resolution image sensors [47, 48].

2.2 Variable resolution sensors

The human eye has its highest resolution at the center of the optical axis, and it decays towards the periphery. There are a number of advantages in such an arrangement. To mention a few:

- Data reduction compared to having the whole field of view in full resolution.
- High resolution is combined with a broad field of view.
- The fovea marks the area of interest, and disturbing details in the surround are blurred.

These advantages can be utilized in a robot vision system as well. There are a number of research projects developing both hardware and algorithms for heterogeneous sampled image arrays, implementing the fovea concept in one form or another, e.g. [42].

2.3 Control mechanism components

Having full resolution only in the center part of the visual field makes it obvious that a good algorithm for positioning the fixation point is necessary. A number of focus-of-attention control mechanisms must be active simultaneously to be able to both handle unexpected events and perform an effective search. The different components can roughly be divided into the following groups:

1. Preattentive, data driven control. Non-predicted structured image information and events attract the focus-of-attention in order to get the information analyzed.
2. Attentive, model driven control. The focus-of-attention is directed toward an interesting region according to predictions using already acquired image information and knowledge from models.
3. Habituation. As image structures are analyzed and modeled their impact on preattentive gaze control is reduced.

The distinction between the preattentive and attentive parts is floating. It is more of a spectrum from pure reflexes to pure attentional movements of the fixation point.

2.4 Gaze control

We will discuss an example of a simple control system with three levels:

Camera Vergence. Cameras are verged towards the same fixation point using the disparity estimates from a stereo algorithm.

Edge tracker. Magnitude and phase from quadrature filters form a vector field drawing the attention towards and along lines and edges in the image [31, 47].

Object finder. Symmetry properties in the orientation estimates are used to indicate potential objects [8, 21].

The refinement of the positioning of the fixation point is handled with potential fields in the robots parameter space. It can be visualized as an 'energy landscape' where the trajectory is the path a little ball freely rolling around would take. The fixation point can be moved to a certain position by forming a potential well around the position in the parameter space corresponding to the robot looking in that direction. The potential fields from the different controlling modules are weighted together to get the total behavior.

2.4.1 Model acquisition and memory

The system marks the states in its parameter space that corresponds to the direction in which it has been tracking edges. This is the first step towards a memory of where it has looked before, and components of a model of its environment. In a general system, where many points in the parameter space might correspond to looking at the same thing, a more sophisticated handling of model properties is required. It is then important to remember and build up a model of not only WHERE but also WHAT the system has seen. For non-static scenes, WHEN becomes important. This leads to a procedure for model acquisition which is an ultimate goal for this process.

3 Image Measurements and Representation

In order for a system modeling a high structural complexity to be manageable and extendable, it is necessary that it exhibits modularity in various respects. This implies for example standardized information representations for interaction between operator modules. Otherwise, the complexity will be overwhelming and functional mechanisms will be completely obscure. One way to satisfy these requirements is to implement the model structure in a hierarchical, fragmented fashion. In order for such a structure to work efficiently, however, certain requirements have to be fulfilled for information representation and for operations.

It is apparent that there are two issues related to hierarchies and pyramid structures. One has to do with level of abstraction, and the other with size or scale. Although they are conceptually different, there are certain relations. With increased level of abstraction generally follows an increase of the scale over which we relate phenomena [39].

Hierarchical structures is nothing new in information processing in general, or in computer vision in particular. A regular organization of algorithms has always been a desired goal for computer scientists.

Among the first structured approaches were those motivated by knowledge about biological visual systems. The perceptron approach by Rosenblatt [40], has attracted new attention as neural network theory has become a hot research topic [22]. The work on layered networks continued, where such networks would accept image data at their bottom level [45, 41, 19].

The Fourier transform has found considerable use in signal analysis. In image analysis, however, the global Fourier transform representation gives rise to problems due to the loss of spatial localization in the transform domain. The Short Time Fourier Transform, or windowed Fourier transform, is one way to modify the Fourier transform for better performance on non-stationary signals. The widely chosen windowing function is the Gabor function due to its simultaneous concentration in both domains [17]. Gabor and wavelet transforms have proved to be very useful.

Most of the work so far has dealt with hierarchies relating to size or scale, although they have indirectly given structural properties. Granlund introduced an explicit abstraction hierarchy [18], employing symmetry properties implemented by Gaussian wavelets in what today is commonly referred to as Gabor functions [17].

Burt introduced an approach to hierarchical image decomposition using the Laplacian or DOLP (Difference Of Low Pass) pyramid [13]. In this way an image is transformed into a set of descriptor elements. The image can then be reconstructed from its set of primitives.

The concept of scale or size as a dimension, was further extended in the so called *scale space* representation of images [50, 33, 34].

3.1 Representation of Motion as Orientation in 3-D

Motion of a point in 2-D can be viewed as a line in 3-D time-space. Correspondingly, the motion of a line in 2-D can be viewed as a plane in 3-D time-space. There are however some complications, not only due to the increased volume of data, but also from a more fundamental point of view. In two dimensions, the orientation of a line or an edge can unambiguously be represented by a vector in a "double angle" representation [18]. The mapping requirements of operations in multiple dimensions are more severe than for two dimensions [32]. With a hemisphere as the original space, an equvialent of the complication encountered in 2-D occurs: Surfaces that differ by a small angle can end up being represented by vectors that are very different, i e close to opposite sides of the rims of the hemispheres. This is of course unacceptable if the metric properties of the space are of any consequence, which will always be the case if there is a next stage where the information is to be further processed. Consider e g the case of differentiation when the vector passes in a step-like fashion from one side of the hemisphere to the other. It is necessary therefore, that a mapping is established that "closes" the space in the same manner as earlier discussed for the two-dimensional case.

It turns out that information can for this purpose advantageously be represented by *tensors* [27]. The tensor representation can be used for filtering in volumes and in time sequences, implementing spatio-temporal filters [49]. It can also be used for computation of higher level features such as curvature [6] or acceleration. The tensor mapping can be controlled by transforms to implement adaptive filtering of volume data or time sequences [29].

The tensor representation of the local orientation of a neighbourhood with one single orientation in 3-D, is given by

$$\mathbf{T} = \frac{1}{x}\begin{pmatrix} x_1^2 & x_1x_2 & x_1x_3 \\ x_1x_2 & x_2^2 & x_2x_3 \\ x_1x_3 & x_2x_3 & x_3^2 \end{pmatrix} \tag{1}$$

where $\mathbf{x} = (x_1, x_2, x_3)$ is a normal vector to the plane of the neighbourhood and $x = \sqrt{x_1^2 + x_2^2 + x_3^2}$. The magnitude of $\mathbf{x}$ is determined by the local energy distribution estimated by filters.

The orientation estimation requires a number of precomputed quadrature filtersevenly spread in one half of the Fourier space [27, 28]. The minimum number of quadrature filters required for orientation estimation in 3-D is 6, where the filters are directed as the vertices of a semiicosahedron, see Figure 1:

$$\begin{array}{lclcrrrl} \hat{\mathbf{n}}_1 & = & c & (& a, & 0, & b &)^t \\ \hat{\mathbf{n}}_2 & = & c & (& -a, & 0, & b &)^t \\ \hat{\mathbf{n}}_3 & = & c & (& b, & a, & 0 &)^t \\ \hat{\mathbf{n}}_4 & = & c & (& b, & -a, & 0 &)^t \\ \hat{\mathbf{n}}_5 & = & c & (& 0, & b, & a &)^t \\ \hat{\mathbf{n}}_6 & = & c & (& 0, & b, & -a &)^t \end{array} \tag{2}$$

with

$$\begin{array}{lcl} a & = & 2 \\ b & = & 1+\sqrt{5} \\ c & = & (10+2\sqrt{5})^{-1/2} \end{array} \tag{3}$$

A quadrature filter designed with a lognormal function, is given in the frequency domain by:

$$\begin{cases} F_k(\boldsymbol{\omega}) \;=\; F_\omega(\omega)(\hat{\boldsymbol{\omega}} \cdot \hat{\mathbf{n}}_{\mathbf{k}})^2 & if \;\; \boldsymbol{\omega} \cdot \hat{\mathbf{n}}_{\mathbf{k}} > 0 \\ F_k(\boldsymbol{\omega}) \;=\; 0 & otherwise \end{cases} \tag{4}$$

The spatial filter coefficients are found by a straightforward 3-D-DFT or by use of an optimization technique. The resulting spatial filter is complex-valued. This procedure is used to obtain the six quadrature filters.

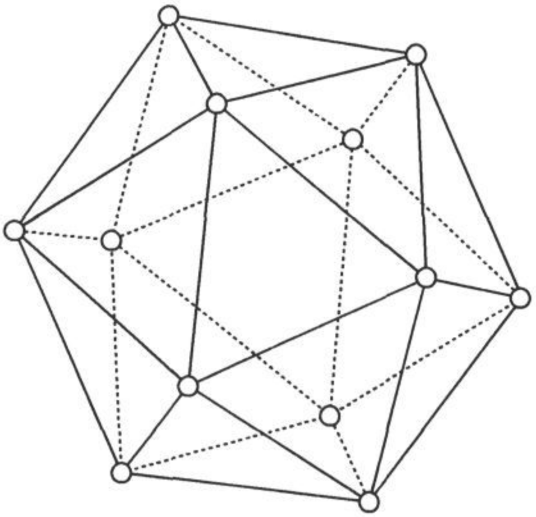

Figure 1: Orientation in 3-D space of filter symmetry axes, in the form of an icosahedron.

It is easy to implement the orientation algorithm with these precomputed filters [27]. The tensor describing the neighbourhood is given by:

$$\mathbf{T^e} = \sum_k q_k(\mathbf{N}_k - \frac{1}{5}\mathbf{I}) \tag{5}$$

where q_k again denotes the magnitude of the output from filter k. $\mathbf{N}_k = \hat{\mathbf{n}}_k\hat{\mathbf{n}}_k^t$ denotes the direction of the filter expressed in the tensor representation and $\mathbf{I}$ is the unity tensor. A less compact description of Eq. 5 is:

1. Convolve the input data with the six complex-valued filters, i.e. perform twelve scalar convolutions.
2. Compute the magnitude of each complex-valued filter by

   $$q_k = \sqrt{q_{ke}^2 + q_{ko}^2}$$

   where q_{ke} denotes the filter output of the real part of filter k and q_{ko} denotes the filter output of the imaginary part of filter k.
3. Compute the tensor $\mathbf{T^e}$ by Eq. 5, i.e.

   $$\mathbf{T^e} = \begin{pmatrix} T_{11} & T_{12} & T_{13} \\ T_{12} & T_{22} & T_{23} \\ T_{13} & T_{23} & T_{33} \end{pmatrix}$$

   where

   $$\begin{aligned} T_{11} &= A(q_1 + q_2) + B(q_3 + q_4) - S \\ T_{22} &= A(q_3 + q_4) + B(q_5 + q_6) - S \\ T_{33} &= A(q_5 + q_6) + B(q_1 + q_2) - S \\ T_{12} &= C(q_3 - q_4) \\ T_{13} &= C(q_1 - q_2) \\ T_{23} &= C(q_5 - q_6) \end{aligned}$$

for

$$S = \frac{1}{5}\sum_{k=1}^{6} q_k$$

$$A = \frac{4}{10+2\sqrt{5}}$$

$$B = \frac{6+2\sqrt{5}}{10+2\sqrt{5}}$$

$$C = \frac{2+2\sqrt{5}}{10+2\sqrt{5}}$$

3.1.1 Evaluation of the Representation Tensor

It is shown in [27] that the eigenvector corresponding to the largest eigenvalue of $\mathbf{T^e}$ is the normal vector of the plane best describing the neighbourhood. This implies that an eigenvalue analysis is appropriate for evaluating the tensor. Below the eigenvalue distribution and the corresponding tensor representation are given for three particular cases of $\mathbf{T^e}$, where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ are the eigenvalues in decreasing order, and $\hat{\mathbf{e}}_\mathbf{i}$ is the eigenvector corresponding to λ_i.

1. $\lambda_1 > 0\,;\ \lambda_2 = \lambda_3 = 0;$
   $\mathbf{T^e} = \lambda_1 \hat{\mathbf{e}}_\mathbf{1} \hat{\mathbf{e}}_\mathbf{1}^t$

   This case corresponds to a neighbourhood that is perfectly *planar*, i.e. is constant on planes in a given orientation. The orientation of the normal vectors to the planes is given by $\hat{\mathbf{e}}_1$.

2. $\lambda_1 = \lambda_2 > 0\,;\ \lambda_3 = 0;$
   $\mathbf{T^e} = \lambda_1 \left(\mathbf{I} - \hat{\mathbf{e}}_\mathbf{3} \hat{\mathbf{e}}_\mathbf{3}^t\right)$ This case corresponds to a neighbourhood that is constant on *lines*. The orientation of the lines is given by the eigenvector corresponding to the least eigenvalue, $\hat{\mathbf{e}}_3$.

3. $\lambda_1 = \lambda_2 = \lambda_3 > 0;$
   $\mathbf{T^e} = \lambda_1 \mathbf{I}$

   This case corresponds to an *isotropic* neighbourhood, meaning that there exists energy in the neighbourhood but no orientation, e.g. in the case of noise.

The eigenvalues and eigenvectors are easily computed with standard methods such as the Jacobi method, e.g. [38]. Note that the spectral decomposition theorem states that all neighborhoods can be expressed as a linear combination of these three cases.

3.1.2 Velocity Estimation

If the signal to analyze is a time sequence, a plane implies a moving line and a line implies a moving point. The optical flow will be obtained by an eigenvalue analysis of the estimated representation tensor. The projection of the eigenvector corresponding to the largest eigenvalue onto the image plane will give the flow field. However, the so-called aperture problem will give rise to an unspecified velocity component, the component moving along the line. The aperture problem is a problem in all optical flow algorithms which rely on local operators. On the other hand, the aperture problem does not exist for moving points in the sequence. In this case of velocity estimation the correspondence between the energy in the spatial dimensions and the time dimension is established to get correct velocity estimation.

By examining the relations between the eigenvalues in the orientation tensor it is possible to divide the optical flow estimation into different categories, [7, 20]. Depending on the category, different strategies can be chosen, see Section 3.1.1. Case number two in Section 3.1.1, i.e. the line case, gives a correct estimation of the velocity in the image plane and is thus very important in the understanding of the motion.

To do this division of different shapes of the tensor the following functions are chosen:

$$p_{plane} = \frac{\lambda_1 - \lambda_2}{\lambda_1} \tag{6}$$

$$p_{line} = \frac{\lambda_2 - \lambda_3}{\lambda_1} \tag{7}$$

$$p_{iso} = \frac{\lambda_3}{\lambda_1} \tag{8}$$

These expressions can be seen as the probability for each case. The discrimination is made by selecting the case having the highest probability.

The calculation of the optical flow is done using Eq. 9 for the plane case and Eq. 10 for the line case. In neighborhoods classified as 'isotropic' no optical flow is computed. The 'true' optical flow in neighborhoods of the 'plane' type, such as moving lines, cannot be computed by optical flow algorithms using only local neighbourhood operations as mentioned earlier. The optical flow is computed by

$$\begin{aligned} \mathbf{x} &= \hat{\mathbf{e}}_1 \\ \mathbf{v}_{line} &= (-x_1 x_3 \hat{\mathbf{x}}_1 - x_2 x_3 \hat{\mathbf{x}}_2)/(x_1^2 + x_2^2) \end{aligned} \tag{9}$$

where $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ are the orthogonal unit vectors defining the image plane.

The aperture problem does not exist for neighborhoods of the 'line' type, such as moving points. This makes them, as mentioned, very important for motion analysis. The optical flow is computed by

$$\begin{aligned} \mathbf{x} &= \hat{\mathbf{e}}_3 \\ \mathbf{v}_{point} &= (x_1 \hat{\mathbf{x}}_1 + x_2 \hat{\mathbf{x}}_2)/x_3 \end{aligned} \tag{10}$$

The use of certainty measures is one of the central mechanisms in the hierarchical framework and the optical flow is not used directly. Separate estimates of the direction of movement and velocity are accompanied with certainty measures computed by combining the tensor norm and the appropriate discriminant function. It is possible to use the confidence statement to process incomplete or uncertain data, as well as data emerging from spatial or temporal transients [30].

4 Spatio-Temporal Channels

The human visual system has difficulties handling high spatial frequencies simultaneously with high temporal frequencies [2, 15]. This means that objects with high

	Spatial subsampling					
	1/8	**1/4**	**1/2**	**1**		
	1/64	1/16	1/4	1	**1**	
Relative	1/128	1/32	1/8	1/2	**1/2**	Temporal
data content	1/256	1/64	1/16	1/4	**1/4**	subsampling
	1/512	1/128	1/32	1/8	**1/8**	

	Spatial subsampling					
	1/8	**1/4**	**1/2**	**1**		
	ch30	ch20	ch10	ch00	**1**	
Notation for	ch31	ch21	ch11	ch01	**1/2**	Temporal
sequence	ch32	ch22	ch12	ch02	**1/4**	subsampling
	ch33	ch23	ch13	ch03	**1/8**	

Table 1: Data content and name convention for the different spatio-temporal channels.

velocity cannot be seen sharply without tracking. One aspect of this is that the visual system performs an effective data reduction. The data reduction is made in such a way that high spatial frequencies can be handled if the temporal frequency is low, and vice versa. This strategy is possible to use in a computer vision model for time sequences.

An input image sequence is subsampled both spatially and temporally into different channels. In Table 1 the data content in the different channels relatively to a reference sequence, *ch00*, is shown. for a typical example. The reference sequence has maximum resolution in all dimensions; typically this means a video signal of 50 Hz, height 576 and width 720 pixels. The frequency difference between adjacent channels is one octave, i.e. a subsampling factor of 2 is used.

The numbers in Table 1 indicate that a large data reduction can be made by not using the channels with high resolution in both spatial and temporal domains. For instance, the channels on the diagonal together contain approximately 1/4 of the data in the reference sequence (*ch00*). There is a signal theoretical reason to use a pyramid representation of the image. A single filter has a particular limited pass band, both temporally and spatially, which may or may not be tuned to the different features to describe. In Figure 2a the upper cut-off frequency for a spatio-temporal quadrature filter set is indicated. The lower cut-off frequency is not plotted for the sake of clarity. Only the first quadrant in the ω_s, ω_t plane is plotted. The use of this filter set on different subsampled channels corresponds to using filters with different center frequencies and constant relative bandwidth. Figure 2b indicates the upper cut-off frequency when convolving the channels on the diagonal in Table 1 with this filter set.

To avoid aliasing in the subsampling, the sequence must be prefiltered with a lowpass filter. As the resulting channel shall be further processed, the design of the lowpass filter is critical. The estimation of optical flow from Eq. 9 and Eq. 10 utilizes the relationship of energies originating from spatial variations and from temporal variations. The lowpass filter used for anti-aliasing should then *not* influence this relationship.

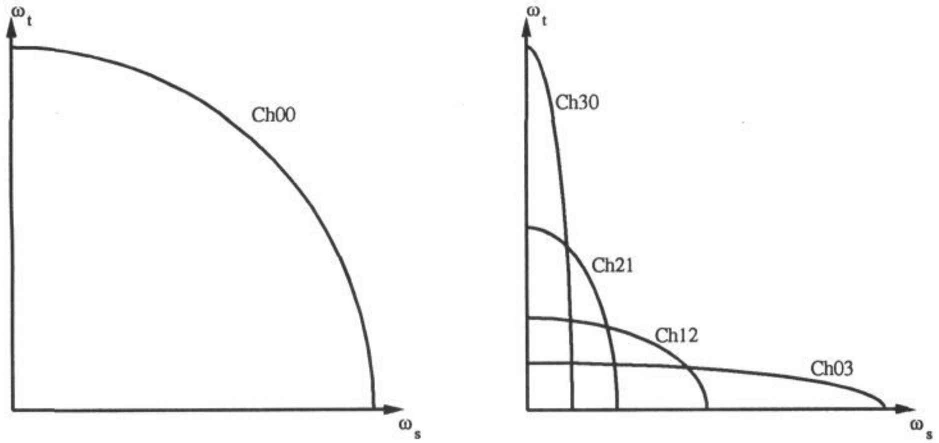

Figure 2: Cut-off frequency for a spatio-temporal filter.

5 Acknowledgements

The author wants to acknowledge the financial support of the Swedish National Board of Technical Development, which supported a great deal of the local research and documentation work mentioned in this overview. Considerable credit should be given to the staff of the Computer Vision Laboratory of Linkoeping University, for discussion of the contents as well as for text and figure contributions to different parts of the manuscript.

References

[1] A. L. Abbott and N. Ahuja. Surface reconstruction by dynamic integration of focus, camera vergence and stereo. In *Proceedings IEEE Conf. on Computer Vision*, pages 523–543, 1989.

[2] M. A. Arbib and A. Hanson, editors. *Vision, Brain and Cooperative Computation*, pages 187–207. MIT Press, 1987.

[3] D. H. Ballard and C. M. Brown. *Computer Vision.* Prentice-Hall, 1982.

[4] Dana H. Ballard. Animate vision. Technical Report 329, Computer Science Department, University of Rochester, Feb. 1990.

[5] D.H. Ballard and Altan Ozcandarli. Eye fixation and early vision: kinetic depyh. In *Proceedings 2nd IEEE Int. Conf. on computer vision*, pages 524–531, december 1988.

[6] H. Bårman, G. H. Granlund, and H. Knutsson. Tensor field filtering and curvature estimation. In *Proceedings of the SSAB Symposium on Image Analysis*, pages 175–178, Linköping, Sweden, March 1990. SSAB. Report LiTH–ISY–I–1088, Linköping University, Sweden, 1990.

[7] H. Bårman, L. Haglund, H. Knutsson, and G. H. Granlund. Estimation of velocity, acceleration and disparity in time sequences. In *Proceedings of IEEE Workshop on Visual Motion*, pages 44–51, Princeton, NJ, USA, October 1991.

[8] J. Bigün. *Local Symmetry Features in Image Processing.* PhD thesis, Linköping University, Sweden, 1988. Dissertation No 179, ISBN 91–7870–334–4.

[9] C. M. Brown. The Rochester robot. Technical Report 257, Computer Science Department, University of Rochester, Aug. 1988.

[10] C. M. Brown. Gaze control with interactions and delays. *IEEE systems, man and cybernetics*, 20(1):518–527, march 1990.

[11] C. M. Brown. Prediction and cooperation in gaze control. *Biological cybernetics*, 63:61–70, 1990.

[12] K. Brunnström, J. O. Eklundh, and T. Lindeberg. Active detection and classification of junctions by foveating with a head–eye system guided by the scale–space primal sketch. Technical Report TRITIA-NA-P9131, CVAP, NADA, Royal Institute of Technology, Stockholm, Sweden, 1990.

[13] P. J. Burt and E. H. Adelson. Merging images through pattern decomposition. In *Applications of digital image procesing VIII.* SPIE, 1985. vol. 575.

[14] S. Culhane and J. Tsotsos. An attentinal prototype for early vision. In *Proccedings of the 2:nd European Conf. on computer vision*, Santa Margharita Ligure, Italy, May 1992.

[15] Hugh Davson, editor. *The Eye*, volume 2A. Academic Press, New York, 2nd edition, 1976.

[16] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis.* Wiley-Interscience, New York, 1973.

[17] D. Gabor. Theory of communication. *Proc. Inst. Elec. Eng.*, 93(26):429–441, 1946.

[18] G. H. Granlund. In search of a general picture processing operator. *Computer Graphics and Image Processing*, 8(2):155–178, 1978.

[19] A. R. Hansen and E. M. Riseman. Constructing semantic models in the visual analysis of scenes. In *Proceedings Milwaukee Symp. Auto. & Contr. 4*, pages 97–102, 1976.

[20] O. Hansen. *On the use of Local Symmetries in Image Analysis and Computer Vision.* PhD thesis, Aalborg University, March 1992.

[21] O. Hansen and J. Bigun. Local symmetry modeling in multidimensional images. In *Pattern Recognition Letters, Volume 13, Nr 4*, 1992.

[22] J. J. Hopfield. Neural networks and physical systems with emergent collective computational capabilities. *Proceedings of the National Academy of Sciences*, 79:2554–2558, 1982.

[23] B. K. P. Horn. *Robot vision.* The MIT Press, 1986.

[24] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, 160:106–154, 1962.

[25] David H. Hubel. *Eye, Brain and Vision*, volume 22 of *Scientific American Library*. W. H. Freeman and Company, 1988.

[26] B. Julesz. Early vision and focal attention. *Review of Modern physics*, 63(3):735–772, 1991.

[27] H. Knutsson. Representing local structure using tensors. In *The 6th Scandinavian Conference on Image Analysis*, pages 244–251, Oulu, Finland, June 1989. Report LiTH–ISY–I–1019, Computer Vision Laboratory, Linköping University, Sweden, 1989.

[28] H. Knutsson, H. Bårman, and L. Haglund. Robust orientation estimation in 2d, 3d and 4d using tensors. In *Proceedings of International Conference on Automation, Robotics and Computer Vision*, September 1992.

[29] H. Knutsson, L. Haglund, and G. H. Granlund. Tensor field controlled image sequence enhancement. In *Proceedings of the SSAB Symposium on Image Analysis*, pages 163–167, Linköping, Sweden, March 1990. SSAB. Report LiTH–ISY–I–1087, Linköping University, Sweden, 1990.

[30] H. Knutsson and C-F Westin. Normalized and differential convolution: Methods for interpolation and filtering of incomplete and uncertain data. In *Proceedings of CVPR*, New York City, USA, June 1993. IEEE.

[31] Hans Knutsson. *Filtering and Reconstruction in Image Processing*. PhD thesis, Linköping University, Sweden, 1982. Diss. No. 88.

[32] Hans Knutsson. Producing a continuous and distance preserving 5-D vector representation of 3-D orientation. In *IEEE Computer Society Workshop on Computer Architecture for Pattern Analysis and Image Database Management - CAPAIDM*, pages 175–182, Miami Beach, Florida, November 1985. IEEE. Report LiTH–ISY–I–0843, Linköping University, Sweden, 1986.

[33] J. J. Koenderink and A. J. van Doorn. The structure of images. *Biological Cybernetics*, 50:363–370, 1984.

[34] L. M. Lifshitz. Image segmentation via multiresolution extrema following. Tech. Report 87-012, University of North Carolina, 1987.

[35] Ralph Linsker. Development of feature-analyzing cells and their columnar organization in a layered self-adaptive network. In Rodney M. L. Cotteril, editor, *Computer Simulation in Brain Science*, chapter 27, pages 416–431. Cambridge University Press, 1988.

[36] R. Milanese. Focus of attention in human vision: a survey. Technical Report 90.03, Computing Science Center, University of Geneva, Geneva, August 1990.

[37] R. Milanese. Detection of salient features for focus of attention. In *Proc. of the 3rd Meeting of the Swiss Group for Artificial Intelligence and Cognitive Science*, Biel-Bienne, October 1991. World Scientific Publishing.

[38] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes.* Cambridge University Press, 1986.

[39] J. Princen, J. Illingworth, and J. Kittler. A hierarchical approach to line extraction based on the Hough transform. *Computer Vision, Graphics, and Image Processing*, 52, 1990.

[40] F. Rosenblatt. *Principles of Neurodynamics: Perceptrons and the theory of brain mechanisms.* Spartan Books, Washington, D.C., 1962.

[41] S. L. Tanimoto and T. Pavlidis. A hierarchical data structure for picture processing. *Computer Graphics and Image Processing*, 2:104–119, June 1975.

[42] M. Tistarelli and G. Sandini. Direct estimation of time–to–impact from optical flow. In *Proceedings of IEEE Workshop on Visual Motion*, pages 52–60, Princeton, USA, October 1991. IEEE, IEEE Society Press.

[43] J. K. Tsotsos. Localizing stimuli in a sensory field using an inhibitory attentinal beam. Technical Report RBCV–TR–91–37, Department of Computer Science, University of Toronto, October 1991.

[44] J. K. Tsotsos. On the relative complexity of active vs. passive visual search. *Int. Journal of Computer Vision*, 7(2):127–142, Januari 1992.

[45] L. Uhr. Layered 'recognition cone' networks that preprocess, classify and describe. In *Proceedings Conf. on Two-Dimensional Image Processing*, 1971.

[46] Esprit basic research action 3038, vision as process, final report. Project document, April 1992.

[47] C-J Westelius, H. Knutsson, and G. H. Granlund. Focus of attention control. In *Proceedings of the 7th Scandinavian Conference on Image Analysis*, pages 667–674, Aalborg, Denmark, August 1991. Pattern Recognition Society of Denmark.

[48] C-J Westelius, H. Knutsson, and G.H. Granlund. Hierarchical gaze control using a multi-resolution image sensor. In *Proceedings from Robotics Workshop*, Linköping, June 1993.

[49] J. Wiklund, L. Haglund, H. Knutsson, and G. H. Granlund. Time sequence analysis using multi-resolution spatio-temporal filters. In *The 3rd International Workshop on Time-Varying Image Processing and Moving Object Recognition*, pages 258–265, Florence, Italy, May 1989. Invited Paper. Report LiTH–ISY–I–1014, Computer Vision Laboratory, Linköping University, Sweden, 1989.

[50] A. Witkin. Scale-space filtering. In *8th Int. Joint Conf. Artificial Intelligence*, pages 1019–1022, Karlsruhe, 1983.

[51] A. L. Yarbus. *Eye movements and vision.* Plenum, New York, 1969.