COMPACT (AND ACCURATE) EARLY VISION PROCESSING IN THE HARMONIC SPACE

First Author Name, Second Author Name
Institute of Problem Solving, XYZ University, My Street, MyTown, MyCountry
f_author@ips.xyz.edu, s_author@ips.xyz.edu

Third Author Name
Department of Computing, Main University, MySecondTown, MyCountry
t_author@xy.mu.edu

Keywords: Phase-based techniques, multidimensional signal processing, filter design, optic flow, stereo vision.

Abstract: The efficacy of anisotropic versus isotropic filtering is anayzed with respect to general phase-based metrics for early vision attributes. We verified that the spectral information content gathered through oriented frequency channels is characterized by high compactness and flexibility, since a wide range of visual attributes emerge from different hierarchical combinations of the same channels. We observed that it is preferable to construct a multichannel, multiorientation representation, rather than using a more compact representation based on an isotropic generalization of the analytic signal. The complete harmonic content is then combined in the phase-orientation space at the final stage, only, to come up with the ultimate perceptual decisions, thus avoiding an "early condensation" of basic features. The resulting algorithmic solutions reach high performance in real-world situations at an affordable computational cost.

1 INTRODUCTION

Although the basic ideas underlying early vision appear deceptively simple and their computational paradigms are known for a long time, early vision problems are difficult to quantify and solve. Moreover, in order to have high algorithmic performance in real-world situations, a large number of channels should be integrated with high efficiency. From a computational point of view, the visual signal should be processed in a "unifying" perspective that will allow us to share the maximum number of resources. In addition, from an implementation point of view, the resulting algorithms and architectures could fall short of their expectations when the high demand of computational resources for multichannel spatio-temporal filtering of high resolution images conflicts with real-time requirements. Several approaches and solutions have been proposed in the literature to accelerate the computation by means of dedicated hardware (e.g., see (Diaz et al., 2006; Kehtarnavaz and Gamadia, 2005)). Yet, the large number of products that must be computed to calculate each single pixel of each single frame for a couple of stereo images and at each time step still represents the main bottleneck. This is particularly true for stereo and motion problems to construct 3D representations of the world, for which establishing image correspondences in space and space-time is a prerequisite, but also their most challenging part.

In this paper, we propose (1) to define a systematic approach to obtain a "complete" harmonic analysis of the visual signal and (2) to integrate efficient multichannel algorithmic solutions to obtain high performance in real-world situations, and at the same time, an affordable computational load.

2 MULTICHANNEL BANDBASS REPRESENTATION

An efficient (internal) representation is necessary to guarantee all potential visual information can be made available for higher level analysis. At an early level, feature detection occurs through initial local *quantitative* measurements of basic image properties (e.g., edge, bar, orientation, movement, binocular disparity, colour) referable to spatial differential structure of the image luminance and its temporal evolution (cf. lin-

ear cortical cell responses). Later stages in vision can make use of these initial measurements by combining them in various ways, to come up with categorical *qualitative* descriptors, in which information is used in a non-local way to formulate more global spatial and temporal predictions (e.g., see (Krüger et al., 2004)).

The receptive fields of the cells in the primary visual cortex have been interpreted as fuzzy differential operators (or local *jets* (Koenderink and van Doorn, 1987)) that provide regularized partial derivatives of the image luminance in the neighborhood of a given point $\mathbf{x} = (x,y)$, along different directions and at several levels of resolution, simultaneously. Given the 2D nature of the visual signal, the spatial direction of the derivative (i.e., the orientation of the corresponding local filter) is an important "parameter". Within a local jet, the directionally biased receptive fields are represented by a set of similar filter profiles that merely differ in orientation.

Alternatively, considering the space/spatial-frequency duality (Gabor, 1946; Daugman, 1985), the local jets can be described through a set of independent spatial-frequency channels, which are selectively sensitive to a different limited range of spatial frequencies. These spatial-frequency channels are equally apt as the spatial ones. From this perspective, it is formally possible to derive, on a local basis, a complete harmonic representation (phase, energy/amplitude, and orientation) of any visual stimulus, by defining the associated analytic signal in a combined space-frequency domain through filtering operations with complex-valued band-pass kernels.

Formally, due to the impossibility of a direct definition of the analytic signal in two dimensions, a 2D spatial frequency filtering would require an association between spatial frequency and orientation channels. Basically, this association can be handled either (1) 'separately', for each orientation channel, by using Hilbert pairs of band-pass filters that display symmetry and antisymmetry about a steerable axis of orientation, or (2) 'as-a-whole', by introducing a 2D isotropic generalization of the analytic signal: the monogenic signal (Felsberg and Sommer, 2001), which allows us to build isotropic harmonic representations that are independent of the orientation (i.e., omnidirectional). By definition, the monogenic signal is a 3D phasor in spherical coordinates and provides a framework to obtain the harmonic representation of a signal respect to the dominant orientation of the image that becomes part of the representation itself.

In the first case, for each orientation channel θ, an image $I(\mathbf{x})$ is filtered with a complex-valued filter:

$$f_A^\theta(\mathbf{x}) = f^\theta(\mathbf{x}) - i f_{\mathcal{H}}^\theta(\mathbf{x}) \qquad (1)$$

where $f_{\mathcal{H}}^\theta(\mathbf{x})$ is the Hilbert transform of $f^\theta(\mathbf{x})$ with respect to the axis orthogonal to the filter's orientation. This results in a complex-valued *analytic image*:

$$Q_A^\theta(\mathbf{x}) = I * f_A^\theta(\mathbf{x}) = C_\theta(\mathbf{x}) + iS_\theta(\mathbf{x}) \,, \qquad (2)$$

where $C_\theta(\mathbf{x})$ and $S_\theta(\mathbf{x})$ denote the responses of the quadrature filter pair. For each spatial location, the amplitude $\rho_\theta = \sqrt{C_\theta^2 + S_\theta^2}$ and the phase $\phi_\theta = \arctan(S_\theta/C_\theta)$ envelopes measure the harmonic information content in a limited range of frequencies and orientations to which the channel is tuned.

In the second case, the image $I(\mathbf{x})$ is filtered with a *spherical quadrature filter* (SQF):

$$f_M(\mathbf{x}) = f(\mathbf{x}) - (i,j) \cdot \mathbf{f}_{\mathcal{R}}(\mathbf{x}) \qquad (3)$$

defined by a radial bandpass filter $f(\mathbf{x})$ (i.e., rotation invariant even filter) and a vector-valued isotropic odd filter $\mathbf{f}_{\mathcal{R}}(\mathbf{x}) = (f_{\mathcal{R},1}(\mathbf{x}), f_{\mathcal{R},2}(\mathbf{x}))^T$, obtained by the Riesz transform of $f(\mathbf{x})$ (Felsberg and Sommer, 2001). This results in a *monogenic image*:

$$\begin{aligned} Q_M(\mathbf{x}) &= I * f_M(\mathbf{x}) = C(\mathbf{x}) + (i,j)\mathbf{S}(\mathbf{x}) \qquad (4) \\ &= C(\mathbf{x}) + iS_1(\mathbf{x}) + jS_2(\mathbf{x}) \end{aligned}$$

where, using the standard spherical coordinates,

$$\begin{aligned} C(\mathbf{x}) &= \rho(\mathbf{x}) \cos\varphi(\mathbf{x}) \\ S_1(\mathbf{x}) &= \rho(\mathbf{x}) \sin\varphi(\mathbf{x}) \cos\vartheta(\mathbf{x}) \\ S_2(\mathbf{x}) &= \rho(\mathbf{x}) \sin\varphi(\mathbf{x}) \sin\vartheta(\mathbf{x}) \,. \end{aligned}$$

The amplitude of the monogenic signal is the vector norm of f_M: $\rho = \sqrt{C^2 + S_1^2 + S_2^2}$, as in the case of the analytic signal, and, for an intrinsically one-dimensional signal, φ and ϑ are the dominant phase and the dominant orientation, respectively.

In this work, we want to analyze the efficacy of the two approaches in obtaining a complete and efficient representation of the visual signal. To this end, we consider, respectively, a discrete set of oriented (i.e., anisotropic) Gabor filters and a triplet of isotropic spherical quadrature filters defined on the basis of the monogenic signal. Moreover, as a choice in the middle between the two approaches, we will also take into consideration the classical steerable filter approach (Freeman and Adelson, 1991) that allows a continuous steerability of the filter respect to any orientation. In this case, the number of basis kernels to compute the oriented outputs of the filters depends on the derivative order (n) of a Gaussian function. The basis filters corresponding to $n = 2$ or $n = 4$ turned out as an acceptable compromise between the representation efficacy and the computational efficiency.

For all the filters considered, we chose the design parameters to have a good coverage of the space-frequency domain and to keep the spatial support (i.e.,

the number of taps) to a minimum, in order to cut down the computational cost. Therefore, we determined the smallest filter on the basis of the highest allowable frequency without aliasing, and we adopted a pyramidal technique (Adelson et al., 1984) as an economic and efficient way to achieve a multiresolution analysis (see also Section 3.2). Accordingly, we fixed the maximum radial peak frequency (ω_0) by considering the Nyquist condition and a constant relative bandwidth of one octave ($\beta = 1$), that allows us to cover the frequency domain without loss of information. For Gabor and steerable filters, we should also consider the minimum number of oriented filters to guarantee a uniform orientation coverage. This number still depends on the filter bandwidth and it is related to the desired orientation sensitivity of the filter (e.g., see (Daugman, 1985; Fleet and Jepson, 1990)); we verified that, under our assumptions, it is necessary to use at least eight orientations. To satisfy the quadrature requirement all the even symmetric filters have been "corrected" to cancel the DC sensitivity. The monogenic signal has been constructed from a radial bandpass filter obtained by summing the corrected bank of oriented even Gabor filters. All the filters have been normalized prior to their use in order to have constant unitary energy. A detailed description of the filters used can be found at `http://130.251.51.86/VISAPP07/`.

3 PHASE-BASED EARLY VISION ATTRIBUTES

3.1 Basic principles

During the last two decades, the phase from local bandpass filtering has gained increasing interest in the computer vision community and has led to the development of a wide number of phase-based feature detection algorithms in different application domains (Sanger, 1988; Fleet et al., 1991; Fleet and Jepson, 1990; Fleet and Jepson, 1993; Kovesi, 1999; Gautama and Van Hulle, 2002). Yet, to the best of our knowledge, a systematic analysis of the basic descriptive properties of the phase has never been done. One of the key contributions of this paper is to formulate a *single* unified representation framework for early vision grounded on a proper phase-based metrics. We verified that the resulting representation is characterized by high compactness and flexibility, since a wide range of visual attributes emerge from different hierarchical combinations of the same channels (i.e., the same computational resources).

The harmonic representation will be the base for a systematic phase-based interpretation of early vision processing, by defining perceptual features on measures of phase properties. From this perspective, edge and contour information can come from *phase-congruency*, motion information can be derived from the *phase-constancy* assumption, while matching operations, such as those used for disparity estimation, can be reduced to *phase-difference* measures. In this way, simple local relational operations capture signal features, which would be more "complex" and less stable if directly analysed in the spatio-temporal domain.

Contrast direction and orientation. Traditional gradient-based operators are used to detect sharp changes in image luminance (such as step edges), and hence are unable to properly detect and localize other feature types. As an alternative, phase information can be used to discriminate different features in a contrast independently way (Kovesi, 1999). Abrupt luminance transitions, as in correspondence of step edges and line features are, indeed, points where the Fourier components are maximally in phase. Therefore, both they are then signaled by peaks in the local energy, and the phase information (i.e, the 'phase-variance') can be used to discriminate among them (see (Kovesi, 1999)). Phase information is used as disambiguating feature whose values can be used to interpret the kind of contrast transition at its maximum (Kovesi, 1999), e.g., a phase of $\pi/2$ corresponds to a dark-bright edge, whereas a phase of 0 corresponds to a bright line on dark background (see also (Krüger and Felsberg, 2003)) .

Binocular disparity. In a first approximation, the phase-based stereopsis defines the disparity $\delta(x)$ as the one-dimensional shift necessary to align, along the direction of the (horizontal) epipolar lines, the phase values of bandpass filtered versions of the stereo image pair $I^R(x)$ and $I^L(x) = I^R[x + \delta(x)]$ (Sanger, 1988). Formally,

$$\delta(x) = \frac{\lfloor \phi^L(x) - \phi^R(x) \rfloor_{2\pi}}{\omega(x)} = \frac{\lfloor \Delta\phi(x) \rfloor_{2\pi}}{\omega(x)} \qquad (5)$$

where $\omega(x)$ is the average instantaneous frequency of the bandpass signal, at point x, that only under a linear phase model can be approximated by ω_0 (Fleet et al., 1991). Equivalently, the disparity can be obtained by direct calculation of the principal part of phase difference, without explicit manipulation of the left and right phase and thereby without incurring the 'wrapping' effects on the resulting disparity map (So-

lari et al., 2001):

$$\lfloor \Delta\phi \rfloor_{2\pi} = \lfloor \arg(Q^L Q^{*R}) \rfloor_{2\pi} \quad (6)$$

where Q^* denotes complex conjugate of Q.

Normal Flow. Considering the conservation property of local phase measurements (phase constancy), image velocities can be computed from the temporal evolution of equi-phase contours $\phi(\mathbf{x},t) = c$ (Fleet et al., 1991). Differentiation with respect to t yields:

$$\nabla\phi \cdot \mathbf{v} + \phi_t = 0 \,, \quad (7)$$

where $\nabla\phi = (\phi_x, \phi_y)$ is the spatial and ϕ_t is the temporal phase gradient. Note that, due to the aperture problem, only the velocity component along the spatial gradient of phase can be computed (normal flow). Under a linear phase model, the spatial phase gradient can be substituted by the radial frequency vector $\omega = (\omega_x, \omega_y)$. In this way, the component velocity $\mathbf{v}_c$ can be estimated directly from the temporal phase gradient:

$$\mathbf{v}_c = -\frac{\phi_t}{\omega_0} \cdot \frac{\omega}{|\omega|} \,. \quad (8)$$

The temporal phase gradient can be obtained by fitting a linear model to the temporal sequence of spatial phases (using *e.g.* five subsequent frames) (Gautama and Van Hulle, 2002):

$$(\phi_t, p) = \underset{\phi_t, p}{\operatorname{argmin}} \sum_t \left((\phi_t \cdot t + p) - \phi(t) \right)^2 \,, \quad (9)$$

where p is the intercept.

Motion-in-depth. The perception of motion in the 3D space relates to 2nd-order measures, which can be gained either by interocular velocity differences or temporal variations of binocular disparity (Harris and Watamaniuk, 1995). Recently (Sabatini et al., 2003), it has been demonstrated that both cues provide the same information about motion-in-depth, when the rate of change of retinal disparity is evaluated as a total temporal derivative of the disparity:

$$\frac{d\delta}{dt} \simeq \frac{\partial\delta}{\partial t} = \frac{\phi_t^L - \phi_t^R}{\omega_0} \simeq v^R - v^L \,, \quad (10)$$

where v^R and v^L are the velocities along the epipolar lines. By exploiting the chain rule in the evaluation of the temporal derivative of phases, one can obtain information about motion-in-depth directly from convolutions Q of stereo image pairs and by their temporal derivatives Q_t:

$$\frac{\partial\delta}{\partial t} = \left[\frac{\mathrm{Im}[Q_t^L Q^{*L}]}{|Q^L|^2} - \frac{\mathrm{Im}[Q_t^R Q^{*R}]}{|Q^R|^2} \right] \frac{1}{\omega_0} \quad (11)$$

thus avoiding explicit calculation and differentiation of phase, and the attendant problem of phase unwrapping.

3.2 Channel interactions

The harmonic information made available by the different basis channels must be properly integrated across both multiple scales and multiple orientations to optimally detect and localise different features at different levels of resolution in the visual signal.

In general, for what concerns the scale, a multiresolution analysis can be efficiently implemented through a coarse-to-fine strategy that helps us to deal with large features values, which are otherwise unmeasurables by the small filters we have to use in order to achieve real-time performance. Specifically, a coarse-to-fine Gaussian pyramid (Adelson et al., 1984) is constructed, where each layer is separate by an octave scale. Accordingly, the image is increasingly blurred with a Gaussian kernel $g(\mathbf{x})$ and subsampled:

$$I_k(\mathbf{x}) = (\mathcal{S}(g * I_{k-1}))(\mathbf{x}) \,. \quad (12)$$

At each pyramid level k the subsampling operator $\mathcal{S}$ reduces to a half the image resolution respect to the previous level $k-1$. The filter response image Q_k at level k is computed by filtering the image I_k with the fixed kernel $f(\mathbf{x})$:

$$Q_k(\mathbf{x}) = (f * I_k)(\mathbf{x}) \,. \quad (13)$$

For what concerns the interactions across the orientation channels a first important distinction must be done according that one uses isotropic or anisotropic filtering.

Isotropic filtering. The monogenic signal directly provides a *single* harmonic content with respect to the dominant orientation:

$$\rho(\mathbf{x}) \stackrel{\text{def}}{=} \sqrt{C^2(\mathbf{x}) + |\mathbf{S}(\mathbf{x})|^2} = \mathcal{E}(\mathbf{x})$$

$$\theta(\mathbf{x}) \stackrel{\text{def}}{=} \text{atan2}\,(S_2(\mathbf{x}), S_1(\mathbf{x})) = \vartheta(\mathbf{x})$$

$$\phi(\mathbf{x}) \stackrel{\text{def}}{=} \text{sign}[\mathbf{S}(\mathbf{x}) \cdot \mathbf{n}_\vartheta(\mathbf{x})] \text{atan2}\,(|\mathbf{S}(\mathbf{x})|, C(\mathbf{x})) = \varphi(\mathbf{x}),$$

with $\mathbf{n}_\vartheta(\mathbf{x}) = (\cos\vartheta(\mathbf{x}), \sin\vartheta(\mathbf{x}))$.

Anisotropic filters. Basic feature interpolation mechanisms must be introduced. More specifically, if we name E_q and ϕ_q the "oriented" energy and the "oriented" phase extracted by the filter f_q steered to the angle $\theta_q = q\pi/K$, the harmonic features computed with this filter orientation are:

$$\begin{aligned} \rho_q(\mathbf{x}) &= \sqrt{C_q^2(\mathbf{x}) + S_q^2(\mathbf{x})} = E_q(\mathbf{x}) \\ \theta_q(\mathbf{x}) &= \frac{q\pi}{K} \\ \phi_q(\mathbf{x}) &= \text{atan2}\,(S_q(\mathbf{x}), C_q(\mathbf{x})) \,. \end{aligned}$$

Under this circumstance, we require to interpolate the feature values computed by the filter banks in order

to estimate the filter's output at the proper signal orientation. The strategies adopted for this interpolation are very different, and strictly depend on the 'computational theory' (in the Marr's sense (Marr, 1982)) of the specific early vision problem considered, as it will be detailed in the following.

Contrast direction and orientation. According to (Krüger and Felsberg, 2004) the phase is used to describe the local structure of i1D signals in an image (see Figure 1). Therefore, we determine maxima of the local amplitude orthogonal to the main orientation with sub–pixel accuracy and compute orientation and phase information at this sub-pixel position using bi-linear interpolation in the phase–orientation space. Sub–pixel accuracy is achieved by computing the center of gravity in a window with size depending on the frequency level. For the bilinear interpolation we need to take care of the topology of the orientation–phase space that has the form of a half–torus. The precision of sub–pixel accuracy calculation as well as the precision of the phase estimate depending on the different harmonic representations is discussed in Section 4.

Binocular disparity. The disparity computation from Eq. (5) can be extended to two-dimensional filters at different orientations θ_q by projection on the epipolar line in the following way:

$$\delta_q(x) = \frac{\lfloor \phi_q^L(x) - \phi_q^R(x) \rfloor_{2\pi}}{\omega_0 \cos\theta_q} \ . \tag{14}$$

In this way, multiple disparity estimates are obtained at each location. These estimates can be combined by taking their median:

$$\delta(x) = \underset{q \in V(x)}{\text{median}}\, \delta_q(x) \ , \tag{15}$$

where $V(x)$ is the set of orientations where valid component disparities have been obtained for pixel x. Validity can be measured by the filter energy.

A coarse-to-fine control scheme is used to integrate the estimates over the different pyramid levels (Bergen et al., 1992). A disparity map $\delta^k(x)$ is first computed at the coarsest level k. To be compatible with the next level, it must be upsampled, using an expansion operator $\mathcal{X}$, and multiplied by two:

$$d^k(x) = 2 \cdot \mathcal{X}\left(\delta^k(x)\right) \ . \tag{16}$$

This map is then used to reduce the disparity at level $k+1$, by warping the phase or filter outputs before computing the phase difference:

$$\delta_q^{k+1}(x) = \frac{\lfloor \phi^L(x) - \phi^R\left(x - d^k(x)\right) \rfloor_{2\pi}}{\omega_0 \cos\theta_q} + d^k(x) \ . \tag{17}$$

In this way, the remaining disparity is guaranteed to lie within the filter range. This procedure is repeated until the finest level is reached.

Optic flow. The reliability of each component velocity can be measured by the mean squared error (MSE) of the linear fit in Eq. (8) (Gautama and Van Hulle, 2002). Provided a minimal number of reliable component velocities are obtained (threshold on the MSE), an estimate of the full velocity can be computed for each pixel by integrating the valid component velocities (Gautama and Van Hulle, 2002):

$$\mathbf{v}(\mathbf{x}) = \underset{\mathbf{v}(\mathbf{x})}{\text{argmin}} \sum_{q \in O(\mathbf{x})} \left(|\mathbf{v}_{c,q}(\mathbf{x})| - \mathbf{v}(\mathbf{x})^{\mathrm{T}} \frac{\mathbf{v}_{c,q}(\mathbf{x})}{|\mathbf{v}_{c,q}(\mathbf{x})|} \right)^2 , \tag{18}$$

where $O(\mathbf{x})$ is the set of orientations where valid component velocities have been obtained for pixel $\mathbf{x}$. A coarse-to-fine control scheme, similar to that of Section 3.2 is used to integrate the estimates over the different pyramid levels. Starting from the coarsest level k, the optic flow field $\mathbf{v}^k(\mathbf{x})$ is computed, expanded, and used to warp the phases or filter outputs at level $k+1$. For more details on this procedure we refer to (Pauwels and Van Hulle, 2006).

Motion-in-depth. Although the motion-in-depth is a 2nd-order measure, by exploiting the direct determination of the temporal derivative of the disparity (see Eq.11), the binocular velocity along the epipolar lines can be directly calculated for each orientation channel, and thence the motion-in-depth

$$\mathrm{V_Z} = \underset{q \in W_L(x)}{\text{median}}\, v_q^L(x) - \underset{q \in W_R(x)}{\text{median}}\, v_q^R(x) \ , \tag{19}$$

where for each monocular sequence, $W(x)$ is the set of orientations for which valid components of velocities have been obtained for pixel x. As in the previous cases, a coarse-to-fine strategy is adopted to guarantee that the horizontal spatial shift between two consecutive frames lie within the filter range.

4 RESULTS

We are interested in computing different image features with the maximum accuracy and the lower processor requirements. The utilization of the different filtering approaches leads to different computing load requirements. Focusing on the convolutions operations on which the filters are based, we have analyzed each approach to evaluate their complexity. Spherical filters require three non-separable convolutions operations, which makes this approach quite

expensive in terms of the required computational resources. The eight oriented Gabor filters requires eight 2-D non separable convolution but they can be efficiently computed through a linear combination of separable kernels as it is indicated in (Nestares et al., 1998), thus significantly reducing the computational load. For steerable filters, quadrature oriented outputs are obtained from the filter bases composed of separable kernels. The higher is the Gaussian derivative order, the higher the number of basis filters. More specifically, the number of 1-D convolutions is given by $4n+6$ where n is the differentiation order.

Summarizing, the complexity of computing the harmonic representation with the different set of filters is summarized in Table 1.

Table 1:

	# filters	# taps	products	sums
Gabor	24	11	264	240
s4	22	11	242	220
s2	14	11	154	140
SQF	3	11×11	363	360

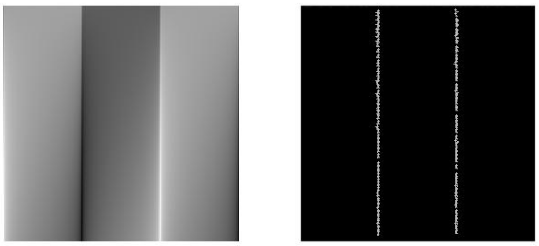
The accuracy of the different filters has been evaluated using synthetic images with well-known ground-truth feature.

Contrast direction and orientation. We have utilized a synthetic image (see Figure 1) where the feature type changes from a step edge to a line feature from top to bottom (Kovesi, 1999). By rotating the image by stepwise angles in $[0, 2\pi)$, we constructed a set of test images and measured the contour localization accuracy, phase and orientation with the different approaches, comparing the results with the ground-truth. In Table 2 the mean errors in localisation, orientation and phase and their standard deviations are reported. It is worth noting that the features were extracted with sub-pixel accuracy.

Table 2: Accuracy evaluation for localization, phase and orientation in the synthetic image of Figure 1. The localization error is expressed in pixels, whereas the orientation and phase errors are in radians.

	localization		orientation		phase	
	avg	std	avg	std	avg	std
Gabor	0.067	0.026	0.021	0.007	0.025	0.005
s4	0.072	0.027	0.022	0.008	0.032	0.006
s2	0.076	0.017	0.042	0.011	0.340	0.203
SQF	0.124	0.062	0.026	0.021	0.198	0.092

We can see that Gabor and 4th-order steerable filters (s4) produce the most accurate results for phase and edge localization, with low variance. Second order steerable filters (s2) and SQFs seem very noisy in their phase estimation.



(a) (b)

Figure 1: (a) Test image representing a continuum of phases taking values between $-\pi$ and π corresponding to a continuum of oriented grey-level structures as expressed in a changing "circular" manifold (cf. (Kovesi, 1999)). The feature type changes progressively from a step edge to a line feature, while retaining perfect phase congruency. (b) Phase-based localization of contours obtained with the Gabor filters.

Binocular disparity. The *tsukuba*, *sawtooth* and *venus* stereo-pairs from the Middlebury stereo vision page (Scharstein and Szeliski, 2002) are used in the evaluation. Since we are interested in the precision of the filters we do not use the integer-based measures proposed there but instead compute the mean and standard deviation of the absolute disparity error. So as not to distort the results with outliers, the error is evaluated only at regions that are textured, non-occluded and continuous. The results are shown in Table 3. The best results are obtained with the Gabor filters. Slightly worse are the results with 4th-order steerable filters and the 2nd-order filters yield results about twice as bad as the 4th-order filters. The results obtained with SQFs are comparable with those obtained by the 2nd-order steerable filters. Figure 2 contains the left images of the stereo-pairs, the ground truth depth maps, and the depth maps obtained with the Gabor filters.

Table 3: Average and standard deviation of the absolute errors in the disparity estimates (in pixels).

	tsukuba		sawtooth		venus	
	avg	std	avg	std	avg	std
Gabor	0.32	0.61	0.41	1.26	0.25	0.77
s4	0.36	0.68	0.50	1.86	0.40	1.30
s2	0.47	0.79	1.12	2.50	0.98	2.44
SQF	0.46	0.85	0.93	2.20	0.95	2.40

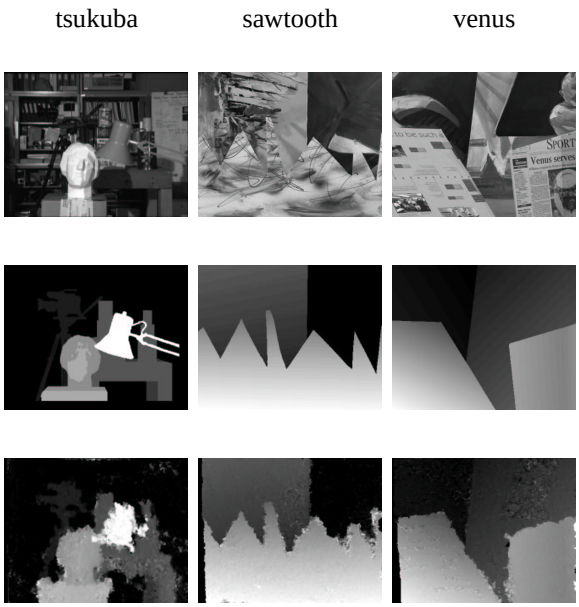


Figure 2: Left frame (top row), ground truth disparity (middle row), and estimated disparity using Gabor filters (bottom row).

Optic flow. We have evaluated the different filters with respect to optic flow estimation on the *diverging tree* and *yosemite* sequences from (Barron et al., 1994), using the error measures presented there. The cloud region was excluded from the *yosemite* sequence. The results are presented in Table 4 and similar conclusions can be drawn as in the previous Section. Gabor and 4th-order steerable filters yield comparable results whereas 2nd-order steerable filters score about twice as bad. The results obtained by SQFs are slightly worse, since the resulting optic flow have larger errors but a higher density. Figure 3 shows the center images, ground truth optic flow fields, and optic flow fields computed with the Gabor filters.

Table 4: Average and standard deviation of the optic flow errors (in pixels) and optic flow density (in percent).

	diverging tree			yosemite (no cloud)		
	avg	std	dens	avg	std	dens
Gabor	2.05	2.28	95.6	2.15	3.12	81.8
s4	2.39	2.62	93.2	2.96	4.46	85.0
s2	4.20	4.58	90.6	6.51	9.23	81.9
SQF	12.9	13.4	95.1	18.7	17.8	99.1

Motion-in-depth. Since binocular test sequences with the ground truth and a sufficiently high frame rate are not available, it has not been possible to make quantitative comparisons. However, considering that

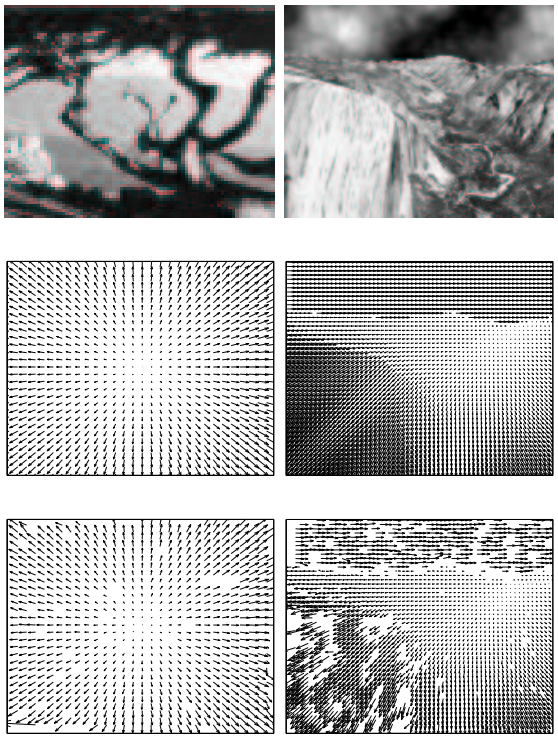


Figure 3: Center frame (top row), ground truth optic flow (middle row) and estimated optic flow obtained with Gabor filters (bottom row). All optic flow fields have been scaled and subsampled five times.

motion-in-depth is a 'derived' quantity, we expected, that the multichannel anisotropic filtering has the same advantages over isotropic filtering alike those observed for stereo and motion processing. Qualitative results obtained in real-world sequences preliminarily confirmed this conclusion.

5 CONCLUSIONS

The first stages of a vision system (early vision) consists of a set of parallel pathways each analysing some particular aspects of the visual stimulus, on the basis of proper local descriptors. Hence, early vision processing can be reconducted to measuring the amount of a particular type of local structure with respect to a specific representation space. The choice for an early selection of features by adopying thresholding procedures, which depend on a specific and restricted environmental context, limits the possibility to build on the ground of such representations an artificial vision system with complex functionalities.

Hence, it is more convenient to base further perceptual processes on a more general representation of the visual signal. The harmonic representation discussed in this paper is a reasonable representation of early vision process since it allows for an efficient and complete representation of (spatially and temporally) *localized* structures. It is characterized by: (1) compactness (i.e., minimal uncertainty of the bandpass channel); (2) coverage of the frequency domain; (3) robust correspondence between the harmonic descriptors and the perceptual 'substances' in the various modalities (edge, motion and stereo). Through a systematic analysis we investigated the advantages of anisotropic *vs* isotropic filtering approaches for a complete harmonic description of the visual signal. In particular, we observed that it is preferable to construct a multichannel, multiorientation representation, thus avoiding an "early condensation" of basic features. The harmonic content is then combined in the phase-orientation space at the final stage, only, to come up with the ultimate perceptual decisions.

REFERENCES

Adelson, E., Anderson, C., Bergen, J., Burt, P., and Ogden, J. (1984). Pyramid methods in image processing. *RCA Engineer*, 29(6):33–41.

Barron, J., Fleet, D., and Beauchemin, S. (1994). Performance of optical flow techniques. *International Journal of Computer Vision*, 12:43–77.

Bergen, J., Anandan, P., Hanna, K., and Hingorani, R. (1992). Hierarchical model-based motion estimation. In *Proc. ECCV'92*, pages 237–252.

Daugman, J. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Amer. A*, A/2:1160–1169.

Diaz, J., Ros, E., Pelayo, F., Ortigosa, E., and Mota, S. (2006). FPGA based real-time optical-flow system. *IEEE Trans. on Circuits and Systems for Video Technology*, 16(2):274–279.

Felsberg, M. and Sommer, G. (2001). The monogenic signal. *IEEE Transactions on Signal Processing*, 48(12):3136–3144.

Fleet, D. and Jepson, A. (1993). Stability of phase information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(12):1253–1268.

Fleet, D., Jepson, A., and Jenkin, M. (1991). Phase-based disparity measurement. *CVGIP: Image Understanding*, 53(2):198–210.

Fleet, D. J. and Jepson, A. D. (1990). Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 1:77–104.

Freeman, W. and Adelson, E. (1991). The design and use of steerable filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13:891–906.

Gabor, D. (1946). Theory of communication. *J. Inst. Elec. Eng.*, 93:429–459.

Gautama, T. and Van Hulle, M. (2002). A phase-based approach to the estimation of the optical flow field using spatial filtering. *IEEE Trans. Neural Networks*, 13(5):1127–1136.

Harris, J. and Watamaniuk, S. N. (1995). Speed discrimination of motion-in-depth using binocular cues. *Vision Research*, 35(7):885–896.

Kehtarnavaz, N. and Gamadia, M. (2005). *Real-Time Image and Video Processing: From Research to Reality*. Morgan & Claypool Publishers.

Koenderink, J. and van Doorn, A. (1987). Representation of local geometry in the visual system. *Biol. Cybern.*, 55:367–375.

Kovesi, P. (1999). Image features from phase congruency. *Videre: A Journal of Computer Vision Research, MIT Press*, 1(3):1–26.

Krüger, N. and Felsberg, M. (2003). A continuous formulation of intrinsic dimension. In *Proc. British Machine Vision Conference*.

Krüger, N. and Felsberg, M. (2004). An explicit and compact coding of geometric and structural information applied to stereo matching. *Pattern Recognition Letters*, 25(8):849–863.

Krüger, N., Lappe, M., and Wörgötter, F. (2004). Biologically motivated multi-modal processing of visual primitives. *The Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour*, 1(5):417–428.

Marr, D. (1982). *Vision*. New York: Freeman.

Nestares, O., Navarro, R., Portilla, J., and Tabernero, A. (1998). Efficient spatial-domain implementation of a multiscale image representation based on Gabor functions. *Journal of Electronic Imaging*, 7(1):166–173.

Pauwels, K. and Van Hulle, M. (2006). Optic flow from unstable sequences containing unconstrained scenes through local velocity constancy maximization. In *Proc. British Machine Vision Conference*, Edinburgh, 4-7 September.

Sabatini, S., Solari, F., Cavalleri, P., and Bisio, G. (2003). Phase-based binocular perception of motion in depth: Cortical-like operators and analog VLSI architectures. *EURASIP Journal on Applied Signal Processing*, 7:690–702.

Sanger, T. (1988). Stereo disparity computation using Gabor filters. *Biol. Cybern.*, 59:405–418.

Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1–3):7–42.

Solari, F., Sabatini, S., and Bisio, G. (2001). Fast technique for phase-based disparity estimation with no explicit calculation of phase. *Elect. Letters*, 37(23):1382–1383.