Computer Vision and Image Understanding xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

A compact harmonic code for early vision based on anisotropic frequency channels ☆

Silvio P. Sabatini [a,*], Giulia Gastaldi [a], Fabio Solari [a], Karl Pauwels [b], Marc M. Van Hulle [b], Javier Diaz [c], Eduardo Ros [c], Nicolas Pugeault [d,1], Norbert Krüger [e]

[a] Dipartimento di Ingegneria Biofisica ed Elettronica, University of Genoa, Italy
[b] Laboratorium voor Neuro- en Psychofysiologie, K.U.Leuven, Belgium
[c] Departamento de Arquitectura y Tecnologia de Computadores, University of Granada, Spain
[d] School of Informatics, University of Edinburgh, UK
[e] The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark

ARTICLE INFO

Article history:
Received 1 September 2008
Accepted 24 March 2010
Available online xxxx

Keywords:
Early vision
Phase-based image analysis
Multichannel filtering
Image representations
Stereo
Motion
Bio-inspired vision processing

ABSTRACT

The problem of representing the visual signal in the harmonic space guaranteeing a complete characterization of its 2D local structure is investigated. Specifically, the efficacy of anisotropic versus isotropic filtering is analyzed with respect to general phase-based metrics for early vision attributes. We verified that the spectral information content gathered through channeled oriented frequency bands 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 constructing a multichannel, multiorientation representation is preferable than using a more compact one based on an isotropic generalization of the analytic signal. Maintaining a channeled (i.e., distributed) representation of the harmonic content results in a more complete structural analysis of the visual signal, and allows us to enable a set of "constraints" that are often essential to disambiguate the perception of the different features. 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. Such a difficulty is often related on the representation we adopt for the visual signal, which must be capable of capturing, through proper "channels", *what* is *where* in the visual signal, that is the structural ("what") and the positional ("where") information from the images impinging the retinas. Ever since the initial formulation of the channel concept, the problem arises of jointly handling the existence of spatial frequency channels on the one hand, and of orientation channels on the other. At a local operator level, the two-dimensional (2D) Gabor filter (proposed by J. Daugman [1] and S. Marcelja [2], as an extension of its one-dimensional (1D) counterpart [3]) retains the optimal joint information resolution in both the domains and meets thoroughly this demand, by underlining the 2D nature of the frequency representation and thus being isomorphic to the 2D character of the spatial manifold of the visual/retinal image. In this way, 2D Gabor filters reconciled the "atomistic" description of early vision, based on local feature detection in the space domain with the "undulatory" interpretation, based on a Fourier-like decomposition into spatial-frequency components. Yet, the picture remained still incomplete, if one considers the representation problem as a whole. Indeed, a hybrid approach asserted itself, consisting in an energy-based multichannel feature extraction (i.e., a parametric analysis in the image domain at different frequency bandwidths), which is still reminiscent of the more intuitive atomistic description, rather than a coarse, local, frequency analysis embedded within the global space-domain mapping. For years, the phase was the missing concept, and, even when it recovered its computational significance, its role within a unifying perspective of the optimal representation of the visual signal has never been fully explored.

In this paper, we propose a general and fully conciliatory position between the spatial and spectral (i.e., frequency) approaches to early vision. Specifically, if we include the local phase as a key

☆ This work has matured from a preliminary and shorter communication presented at the 2nd International Conference on Computer Vision Theory and Applications (VISAPP'07), 8-11 March, 2007 Barcelona, Spain.

* Corresponding author. Fax: +39 010 3532289.
E-mail address: silvio.sabatini@unige.it (S.P. Sabatini).
[1] Present address: University of Surrey, UK.

1077-3142/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2010.03.008

descriptive element of the visual signal, we gain a complete representation and the orientation becomes an integral part of the harmonic representation. Indeed, the extraction of local phase information in two dimensions is an intrinsically *anisotropic* problem that refers to a *selected* orientation, unless one introduces an isotropic extension of the analytic signal (Hilbert transform) and thereby the concept of *dominant* orientation. We will remark the conceptual complementarity of the two (spatial and spatial-frequency) descriptions, yet pointing out the optimal economy and the major richness of the information conveyed in the harmonic domain by anisotropic filters, which allow us to derive, with relative ease, more complex (higher-order) visual descriptors without resorting to complicated relational and/or symbolic constructors, but still operating at the signal level.

The rest of the paper is organized as follows: in Section 2, we introduce the problem of the representation of the visual signal in the harmonic space, according to different sets of band-pass filters. On that ground, we define/qualify early vision features in terms of specific phase properties and phase relationships. In Section 3, channel interaction is formalized. Section 4 presents a comparative analysis of the experimental results on both synthetic and natural image sequences. Concluding remarks in Section 6 are preceded by a general discussion in Section 5.

2. Visual features as measures in the harmonic space

The goal of early vision is to extract as much information as possible about the structural properties of the visual signal. As pointed out by [4,5], an efficient internal representation is necessary to guarantee all potential visual information can be made available for higher level analysis. The measurement of specific, significant visual "elements" in a local neighborhood of the visual signal has led to the concept of "feature" and of "feature extraction". An image feature can be defined in terms of attributes related to the visual data. Though, in practice, many features are also defined in terms of the particular procedure used to extract information about that feature, and, in more general terms, on the specific scheme adopted for the representation of the visual signal. At an early level, feature detection occurs through initial local *quantitative* measurements of basic image properties (e.g., edge, bar, orientation, movement, binocular disparity, color) referable to spatial differential structure of the image luminance and its temporal evolution (cf. linear visual cortical cell responses, see e.g. [6–8]). 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 [9]).

The receptive fields of the cells in the primary visual cortex have been interpreted as fuzzy differential operators (or local *jets* [4]) that provide regularized partial derivatives of the image luminance along different directions and at several levels of resolution, simultaneously. The jets characterize the local geometry in the neighborhood of a given point $\mathbf{x} = (x,y)$. The order of the jet determines the amount of geometry represented. 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 [3,1], the local jets can be described through a set of 2D spatial-frequency channels, which, for each spatial orientation, are selectively sensitive to a different limited range of spatial frequencies. These oriented spatial-frequency channels are equally apt as the spatial ones [10]. From this perspective, it is formally possible to derive, on a local basis, a complete harmonic representation (amplitude, phase, and orientation) of any visual stimulus, by defining the associated analytic signal in a combined space-frequency domain through filtering operations with complex-valued 2D band-pass kernels. Since spatial information is being linearly transformed from the space domain at the level of pixels, into a combined space-frequency domain at a cortical-like representation level, no actual analysis is taking place at this level, and the information is merely being put into another (presumably more useful) form [11].

Formally, due to the impossibility of a direct definition of the analytic signal in two dimensions, a full harmonic characterization of the 2D spatial vision channels in the Fourier domain requires explicit (1D) reference axes, and their association with the orientation of the spatial frequency channel must be discussed. Basically, this association can be handled either (1) 'separately', for each orientation, 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 [12], 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 θ, an image $I(\mathbf{x})$ is filtered with a complex-valued filter:

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

where $f_{\mathscr{H}}^\theta(\mathbf{x})$ is the Hilbert transform of $f^\theta(\mathbf{x})$ with respect to the axis orthogonal to the filter's orientation:

$$f_{\mathscr{H}}^\theta(\mathbf{x}) = f_{\mathscr{H}}(x_\theta, y_\theta) = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{f(\xi, y_\theta)}{\xi - x_\theta}\mathrm{d}\xi$$

with x_θ and y_θ the principal axes of the energy distribution of the filter in the spatial domain.

This results in a complex-valued *analytic image*:

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

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 $\phi_\theta = \mathrm{atan2}(S_\theta, C_\theta)$ envelopes measure the harmonic information content in a limited range of frequencies and orientations to which the channel is tuned (see Fig. 1a).

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

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

defined by a radial bandpass filter $f(\mathbf{x})$ (i.e., rotation invariant even filter) and a vector-valued isotropic odd filter $\mathbf{f}_{\mathscr{R}}(\mathbf{x}) = (f_{\mathscr{R},1}(\mathbf{x}), f_{\mathscr{R},2}(\mathbf{x}))^T$, obtained by the Riesz transform of $f(\mathbf{x})$ [12]:

$$\mathbf{f}_{\mathscr{R}}(\mathbf{x}) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{\boldsymbol{\xi}}{|\boldsymbol{\xi}|^3} f(\mathbf{x}-\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} \tag{4}$$

This results in a *monogenic image*:

$$Q_M(\mathbf{x}) = I * f_M(\mathbf{x}) = C(\mathbf{x}) + (i,j)\mathbf{S}(\mathbf{x}) = C(\mathbf{x}) + iS_1(\mathbf{x}) + jS_2(\mathbf{x}) \tag{5}$$

where using the standard spherical coordinates,

$$C(\mathbf{x}) = \rho(\mathbf{x})\cos\varphi(\mathbf{x})$$
$$S_1(\mathbf{x}) = \rho(\mathbf{x})\sin\varphi(\mathbf{x})\cos\vartheta(\mathbf{x})$$
$$S_2(\mathbf{x}) = \rho(\mathbf{x})\sin\varphi(\mathbf{x})\sin\vartheta(\mathbf{x}).$$

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,

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

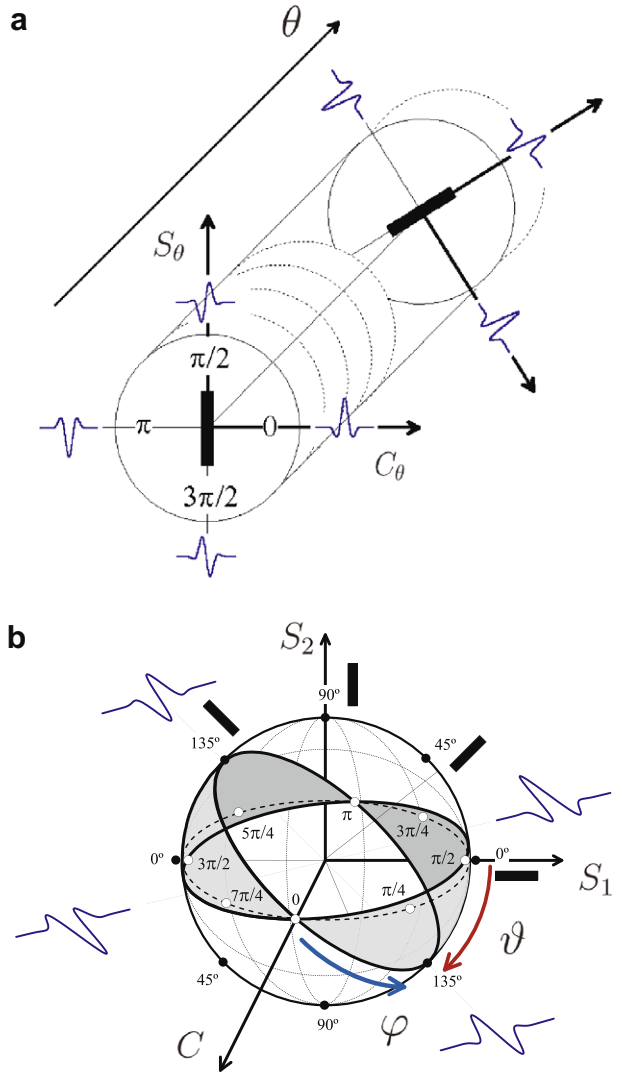

Fig. 1. (a) Multichannel harmonic representation through anisotropic frequency channels. (b) Isotropic harmonic representation through the monogenic signal.

for an intrinsically one-dimensional signal, φ and ϑ are the dominant phase and the dominant orientation, respectively (see Fig. 1b).

2.1. Compact band-pass filtering

In the harmonic space (e.g., phase-based approaches), it is in general an important requirement to have both the spatial width of the filters and the spatial frequency bandwidth small, so that good localization and good approximation of the harmonic information is realized simultaneously. Gabor functions reaching the maximal joint resolution in space and spatial frequency domains are specifically suitable for this purpose and are extensively used in computational vision [1]. Different band-pass filters have been proposed as an alternative to Gabor functions, on the basis of specific properties of the basis functions [13–20], or according to theoretical and practical considerations of the whole space-frequency transform [21–26]. A detailed comparison of the different filters evades the scope of this paper and numerous comparative reviews can be already found in the literature (e.g., see [27] [28,29]).

Since the main goal of this study is to analyze the efficacy of the two approaches (isotropic *vs.* anisotropic) in obtaining a complete and efficient representation of the visual signal, we consider 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, respectively. Moreover, as a choice in the middle between the two approaches, we will also take into consideration the classical steerable filter approach [23] that allows a continuous steerability of the filter with respect to any orientation. Hence, it is possible in principle to steer the filter with respect to the dominant orientation of the signal, which, yet, has to be known in advance and cannot be gained from the representation itself.

In order to guarantee a fair comparison among the different filters, we have considered Gabor functions with radially symmetric Gaussian envelopes. 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 [30] as an economic and efficient way to achieve a multi-resolution analysis (see also Section 3). Accordingly, we fixed the maximum radial peak frequency (ω_0) by considering the Nyquist condition, and a constant relative bandwidth β around one octave, that allows us to cover the frequency domain without loss of information. The result was an 11×11 filter mask capable of resolving sub-pixel phase differences. For Gabor and steerable filters, we should also consider the minimum number of oriented filters to guarantee a uniform orientation coverage. This number depends on the filter bandwidth and it is related to the desired orientation sensitivity of the filter (e.g., see [1,31]); we verified that, under our assumptions, the use of at least eight orientations is necessary. 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. Where possible, the filters have been expressed as sums of x–y separable functions to implement separate 1D convolutions instead of 2D convolutions in a similar way that [32], with a consequent further drop of the computational burden. For a detailed description of the filters used, see the Appendix.

2.2. Phase-based metrics

By exploiting, on a local basis the spectral information content of the image signal (amplitude and phase), we can derive perceptual entities, useful to gain interpretative elements of the observed scene, such as edges/contours, motion, and binocular disparity. Although the major part of the classical algorithms available in the literature rely upon the amplitude information, in the last two decades alternative techniques based on phase measures have been asserted themselves. The importance of global (Fourier) phase has been first demonstrated with respect to image coding and representation, by comparing modulus-only and phase-only image reconstructions [33,34], and has been confirmed also in case of the local phase spectrum [35]. On that ground, the popularity of the phase information, as a robust feature descriptor, has risen in relation with the numerous important properties that have been reported and analyzed [36,31,37–44], such as: (1) the capacity of measuring changes much smaller than the spatial quantization (giving sub-pixel accuracy without a sub-pixel representation of the image, due to its continuous character); (2) the stability with respect to small geometric deformations of the input; and (3) – perhaps the most desirable property – the invariance with both mean luminance and contrast (e.g., with respect to smooth shading and lighting variations), which makes phase, in principle, robust against typical variations in image formation. For these reasons, during the recent past, 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

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

feature detection algorithms in different application domains [45,31,46,37,47,48,12,49–51,49,52–60,44,61–64]. Yet, to the best of our knowledge, a systematic analysis of the basic descriptive properties of the phase, which explicitly includes a discussion of the role of the oriented spatial frequency channels (e.g., see Section 5), 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, integrated across orientations. 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 analyzed in the spatio-temporal domain.

Let us summarize some basic principles.

2.2.1. Contrast feature detection

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 independent way [48]. 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 [48], by revealing different directions of contrast. Phase information is used as disambiguating feature whose values can be used to interpret the kind of contrast transition at its maximum, 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 ([48], see also [65]) .

Specifically, [66] show that phase congruency $PC(x)$, for an one-dimensional luminance profile, is equal to local energy $E(x)$ scaled by the sum of the amplitudes A_k of the Fourier series expansion of the visual signal at location x:

$$PC(x) = \frac{E(x)}{\sum_k A_k}. \tag{6}$$

Local energy and Fourier amplitudes A_k can be approximated by bandpass filter responses at different scales (k):

$$E(x) \simeq \sqrt{\left(\sum_k C_k(x)\right)^2 + \left(\sum_k S_k(x)\right)^2} \tag{7}$$

$$A_k(x) \simeq \sqrt{C_k^2(x) + S_k^2(x)} = \rho_k(x) \tag{8}$$

from which:

$$PC(x) = \frac{E(x)}{\sum_k \rho_k(x) + \epsilon}. \tag{9}$$

The appropriate value of ϵ depends on the precision with which we are able to perform convolutions and other operations on our signal.

We point out that in the following we will use a phase congruency formulation that takes into account the effect of noise, a more sensitive measure for localization, and the extension to 2D spatial domain [48].

2.2.2. 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 $I^L(x) = I^R[x + \delta(x)]$ [67,45]. Formally,

$$\delta(x) = \frac{\lfloor \phi^L(x) - \phi^R(x) \rfloor_{2\pi}}{\omega(x)} = \frac{\lfloor \Delta\phi(x) \rfloor_{2\pi}}{\omega(x)} \tag{10}$$

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 [46]. 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 [68] (see also [69,70]):

$$\lfloor \Delta\phi \rfloor_{2\pi} = \arg(Q^L Q^{*R}) == \operatorname{atan2}(C^R S^L - C^L S^R, C^L C^R + S^L S^R) \tag{11}$$

where Q^* denotes complex conjugate of Q.

2.2.3. Normal flow

Considering the conservation property of local phase measurements (phase constancy) [46], image velocities can be computed from the temporal evolution of equi-phase contours $\phi(\mathbf{x}, t) = c$. Differentiation with respect to t yields:

$$\nabla\phi \cdot \mathbf{v} + \phi_t = 0, \tag{12}$$

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 $\boldsymbol{\omega} = (\omega_x, \omega_y)$. In this way, the component velocity $\mathbf{v}_c$ can be estimated directly from the temporal phase gradient:

$$\mathbf{v}_c = -\frac{\phi_t}{\omega_0} \cdot \frac{\boldsymbol{\omega}}{|\boldsymbol{\omega}|}. \tag{13}$$

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) [51]:

$$(\phi_t, p) = \underset{\phi_t, p}{\operatorname{argmin}} \sum_t ((\phi_t \cdot t + p) - \phi(t))^2, \tag{14}$$

where p is the intercept. Note that this is different from the approach by [31], which involves a bank of *spatiotemporal* filters. Their method requires tiling the spatiotemporal frequency space with velocity-tuned filter pairs. The approach from [51] on the other hand, allows estimation of the temporal phase gradients irrespective of the velocity at that spatial location. Using spatial filters also renders the temporal span over which the optical flow field is estimated independent of the filter bank. The mean squared error (MSE) of the linear fit in Eq. (14) measures the *phase nonlinearity* and serves as a (strictly local) reliability criterion for the component velocity. Fleet and Jepson [46], instead, detect neighborhoods around *phase singularities*, requiring spatial derivatives, which cannot be computed purely locally.

3. Channel interactions

On the basis of the conceptual description of early vision features in terms of the local phase properties of the visual signal, illustrated in the previous section, we can formulate complete

solutions by combining the information gathered by a set of filters that properly tile the 2D spatial frequency domain. In other words, 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 localize the different features in the visual signal. If a multi-scale approach, that refines features' values through different levels of resolutions, is usually sufficient for combining information at different spatial frequency bandwidths, the blending of information across different orientation bandwidths, requires a major attention. In particular, we can decide to adopt a harmonic analysis through several oriented filters parameterized by θ, or prefer a measure of the dominant orientation through the monogenic approach, and refer the harmonic content to that (dominant) orientation axis. It is worth noting that the former approach maintains a channeled (i.e., distributed) representation of the harmonic content, to which corresponds a more complete structural analysis of the visual signal (see Discussion). This allows us to enable a set of "constraints" that are often essential to disambiguate the perception of the different features. On the other hand, the determination of the dominant orientation implies an early making of assumption on the structural properties of the image, which might result in less reliable feature estimation, or, sometimes, might restrict the solution of the perceptual problem.

3.1. Multi-scale analysis

In general, for what concerns the scale, a multi-resolution analysis can be efficiently implemented through a coarse-to-fine strategy that helps us to deal with large features values [71], which are otherwise unmeasurable by the small filters we have to use in order to achieve real-time performance. Specifically, a coarse-to-fine Gaussian pyramid [30] is constructed, where each layer is separated by an octave scale. Accordingly, the image is increasingly blurred with a Gaussian kernel $g(\mathbf{x})$ and sub-sampled:

$$I_k(\mathbf{x}) = (\mathscr{S}(g * I_{k-1}))(\mathbf{x}). \tag{15}$$

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

$$Q_k(\mathbf{x}) = (f * I_k)(\mathbf{x}). \tag{16}$$

3.2. Multi-orientation analysis

Using anisotropic filters such as Gabor functions or Gaussian derivatives it is likely that the local orientation of some features does not fit the discrete number (K) of orientations used ($K = 8$ in our implementation). Hence, for what concerns the interactions across the oriented spatial frequency channels, 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:

$$E_q(\mathbf{x}) = C_q^2(\mathbf{x}) + S_q^2(\mathbf{x}) = \rho_q^2(\mathbf{x})$$
$$\theta_q(\mathbf{x}) = \frac{q\pi}{K}$$
$$\phi_q(\mathbf{x}) = \operatorname{atan2}\big(S_q(\mathbf{x}), C_q(\mathbf{x})\big).$$

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 [72]) of the specific early vision problem considered, as it will be detailed in the following.

3.2.1. Contrast direction and orientation

Through a (modified) tensor-based method [73,74], by combining the magnitude responses from basis channels with different orientations $\theta_q, q = 0, \ldots, K-1$, but a common frequency, we can derive information about the local energy, the local phase and the dominant local orientation around each pixel location of the image:

$$\mathscr{E}(\mathbf{x}) = \sum_{q=0}^{K-1} E_q(\mathbf{x}) = \sum_{q=0}^{K-1} \left[C_q(\mathbf{x})^2 + S_q(\mathbf{x})^2\right] \tag{17}$$

$$\vartheta(\mathbf{x}) = \frac{1}{2}\arg\left[\sum_{q=0}^{K-1} \rho_q(\mathbf{x}) e^{2j\theta_q}\right] \tag{18}$$

$$\varphi(\mathbf{x}) = \operatorname{atan2}(\widehat{S}(\mathbf{x}), \widehat{C}(\mathbf{x})) \tag{19}$$

with

$$\widehat{C}(\mathbf{x}) = \sum_{q=0}^{K-1} C_q(\mathbf{x}) E_q(\mathbf{x}) |\cos[\theta_q - \vartheta(\mathbf{x})]|, \text{ and}$$
$$\widehat{S}(\mathbf{x}) = \sum_{q=0}^{K-1} S_q(\mathbf{x}) E_q(\mathbf{x}) \cos[\theta_q - \vartheta(\mathbf{x})]$$

where $\theta_q - \vartheta$ is the difference between the preferred orientation of the filter and the local dominant orientation. These values are comparable to what can be directly obtained by the monogenic signal approach (see Section 2). Alternative methods can be used, such as winner-take-all, weighted average or maximal steerable energy [23], though, on the basis of comparative analysis, we verified that (provided a uniform coverage of the orientation space) the tensor-based technique leads to the smaller error in the filter's frequency bandwidth.

3.2.2. Binocular disparity

The disparity computation from Eq. (10) can be extended to two-dimensional filters at different orientations θ_q by projection on the (horizontal) epipolar line in the following way:

$$\delta_q(x) = \frac{\lfloor \phi_q^L(\mathbf{x}) - \phi_q^R(\mathbf{x}) \rfloor_{2\pi}}{\omega_0 \cos\theta_q}. \tag{20}$$

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

$$\delta(x) = \underset{q \in V(x)}{\operatorname{median}}\, \delta_q(x), \tag{21}$$

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 [75]. A disparity map $\delta^k(x)$ is first computed at the coarsest level k. To be compatible with the next level, it must be up-sampled, using an expansion operator $\mathscr{X}$, and multiplied by two:

$$d^k(x) = 2 \cdot \mathscr{X}\left(\delta^k(x)\right). \tag{22}$$

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:

$$\delta_q^{k-1}(x) = \frac{\lfloor \phi_q^L(x) - \phi_q^R\left(x + d^k(x)\right) \rfloor_{2\pi}}{\omega_0 \cos\theta_q} + d^k(x). \tag{23}$$

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

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

Equivalently, disparity can be obtained from the monogenic phase difference [50] (cf. Eq. (11)):

$$\lfloor \Delta\varphi \rfloor_{2\pi} = \operatorname{atan2}\left(C^R\mathbb{S}^L - C^L\mathbb{S}^R, C^LC^R + \mathbb{S}^L\mathbb{S}^R\right) \tag{24}$$

where $\mathbb{S}^L = |\mathbf{S}^L|\text{sign}(S_1^L)$ and $\mathbb{S}^R = |\mathbf{S}^R|\text{sign}(S_1^R)$, defined as the phase angle between the two monogenic signals in the plane formed by the real signal and its Riesz transform, when equally oriented 1D dominant local structures are assumed in the stereo pair. The phase difference is associated to the "normal" displacement with respect to the direction of the dominant orientation signal component, which does not necessarily correspond to the direction along the horizontal epipolar line. In order to turn such displacement into a disparity measure the former must be still projected on the horizontal epipolar line:

$$\delta_M(\mathbf{x}) = \frac{\lfloor \Delta\varphi(\mathbf{x}) \rfloor_{2\pi}}{\omega_0 \cos\vartheta}. \tag{25}$$

3.2.3. Optic flow

Starting from the normal velocity components extracted for every spatial orientation, the estimation of the full velocity requires the combination of the information to solve the aperture problem. The reliability of each component velocity can be measured by the mean squared error (MSE) of the linear fit in Eq. (14) [51]. 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 [51]:

$$\mathbf{v}(\mathbf{x}) = \underset{\mathbf{v}(\mathbf{x})}{\operatorname{argmin}} \sum_{q\in O(\mathbf{x})} \left(|\mathbf{v}_{c,q}(\mathbf{x})| - \mathbf{v}(\mathbf{x})^{\mathrm{T}} \frac{\mathbf{v}_{c,q}(\mathbf{x})}{|\mathbf{v}_{c,q}(\mathbf{x})|} \right)^2, \tag{26}$$

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 used for disparity is adopted 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 [76].

The monogenic counterpart of this approach is not straightforward, since, by construction, it can provide for every pixel at each scale the velocity component normal to the dominant orientation $\vartheta(\mathbf{x})$, only: $\mathbf{v}_{c,\vartheta}(\mathbf{x})$ (cf. [77]). Such normal component is used to warp the filter outputs in the coarse-to-fine processing scheme. Actually, in analogy to the optimization strategy of Eq. (26), one can minimize over a set of spatial region-channels instead of oriented spatial frequency channels to obtain the full velocity estimation in the case of the monogenic signal (cf. [60]). Though, in this work, to focus on the local properties of the bandpass channels and on the direct comparison between anisotropic *vs.* isotropic frequency channels, we exclude any spatial contextual integration of the filter outputs in a spatial neighborhood, since such an integration is not necessary for anisotropic bandpass channels. It is also worth mentioning that avoiding the spatial integration has the advantage that the warping can be performed independently of neighboring estimates, thus being not necessary that the neighboring estimates are reliable, too. Hence, without resorting to additional spatial averaging over a larger area (cf. [60]), the local approximation of the visual stimuli by simple oriented dominant components, relentlessly prevents the solution, at each scale, of the aperture problem. Therefore, a direct comparison between the anisotropic and isotropic filters will not be possible.

4. Experimental results

4.1. Oriented Gaborian channels

The phase-based methodologies described in Sections 2 and 3 allow us to perform a complete early vision analysis of the observed scene by combining (in different and specific ways) the output of the convolutions of the images with a complete set of filters that properly tile the two-dimensional frequency domain. For static early vision attributes, such as orientation and direction of contrast, and binocular disparity, an algebraic combination of the filters' outputs is sufficient, whereas for dynamic (time-varying) attributes, such as optic flow, an interpolation of the spatial phase values over time (typically five frames are sufficient) is required to derive an estimate of the temporal phase derivative and thence compute the local velocity component. Though, in general, the filtering stage that provides the harmonic representation of the visual signal is common for all the features. The resulting feature maps are shown in Fig. 2. The disparity map and the optic flow are obtained by the algorithms presented in Section 3. The contrast transition maps is obtained by performing non-maximal suppression on the raw phase congruency map followed by hysteresis thresholding proposed in [48].

As expected, the phase-based information provides dense and robust results; the use of an adequate number of oriented spatial frequency channels assures a good accuracy, too.

4.2. Comparative analysis

To compare the accuracy in feature extraction of the different band-pass representations, we have applied the same algorithmic procedure to the outputs of the spatial filtering stage, using as convolution kernels the three classes of filters defined in Section 2.1: Gabor-like kernels, spherical quadrature filters (related to monogenic signal), and steerable filters. At first, in order to obtain a quantitative measure of the accuracy, we used for benchmarking standard synthetic sequences with well-known ground-truth feature. Then, qualitative comparisons are obtained with real-world sequences.

It is worth noting that our goal is not comparing our algorithms with the state-of-the-art, but comparing the performances of the different filters. In their basic formulation, indeed, the algorithms are not competitive as such on those benchmarks, mainly due to the lack of global optimization strategies [78–81].

4.2.1. Synthetic sequences

4.2.1.1. Contrast direction and orientation. We have utilized a synthetic image (see Fig. 3a) where the feature type changes from a step edge to a line feature in a circular manifold [54] with continuously varying intensities in the background. Fig. 3b–d show the results obtained with Gabor filters for contour localization, by using the phase congruency [48], and for orientation and phase estimation, by using a modified tensor-based method [73,74].

The contour localization is based on the raw phase congruency map (only the values higher than a given threshold are shown). The unreliable values of the orientation estimation are discarded by using a reliability measure (the magnitude of the complex argument of the Eq. (18), see [73]). The reliability of the phase estimation is based on the local energy. It is worth noting that we choose to use and to show the phase estimation as a continuum of values, since the approaches for early vision feature extraction (e.g., binocular disparity and optic flow) exploit the whole range of phase values, it is not based only on specific phase values. As an alternative approach, [82] proposed the characteristic phase concept: the phase values that are consistent over a range of scales, named

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

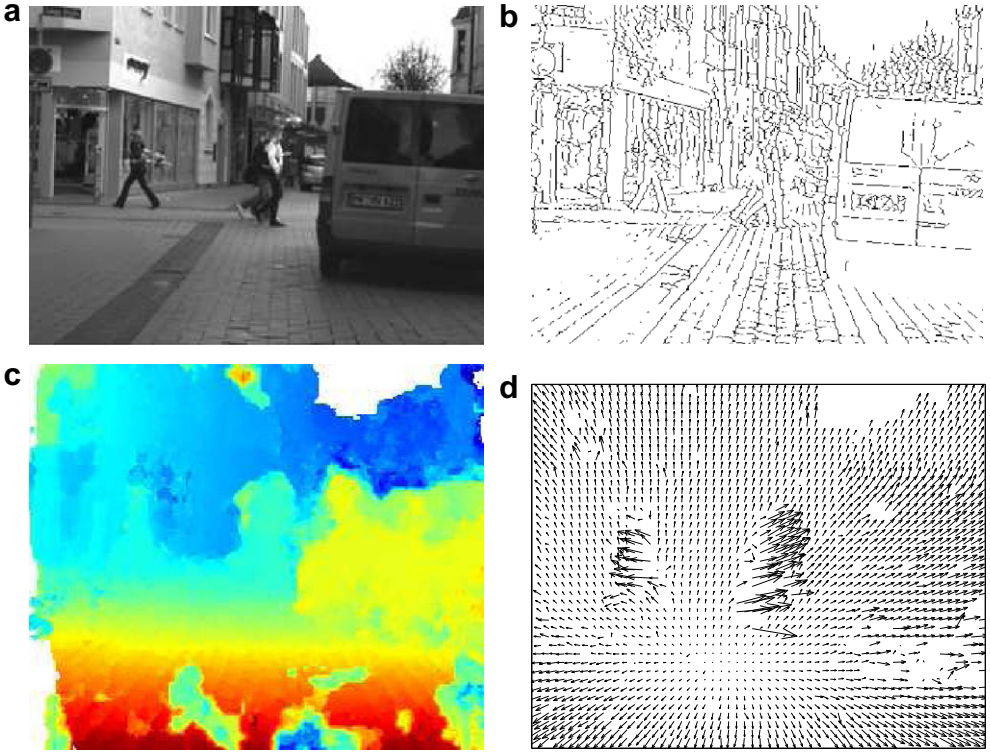

Fig. 2. Example of phase-based feature maps obtained with oriented Gabor channels for a real-world stereo sequence. (a) A frame of the test sequence for the left camera. The sequence is acquired by a stereo camera system rigidly installed behind the front shield of a car moving forward slowly. (b) Contrast transitions revealed by phase congruency. (c) Disparity map, coded from red (objects closer to the viewer) to blue (objects in the background). (d) Optic flow, subsampled and scaled five times. The motion of pedestrians crossing the street is superimposed to the translational ego-motion of the car. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

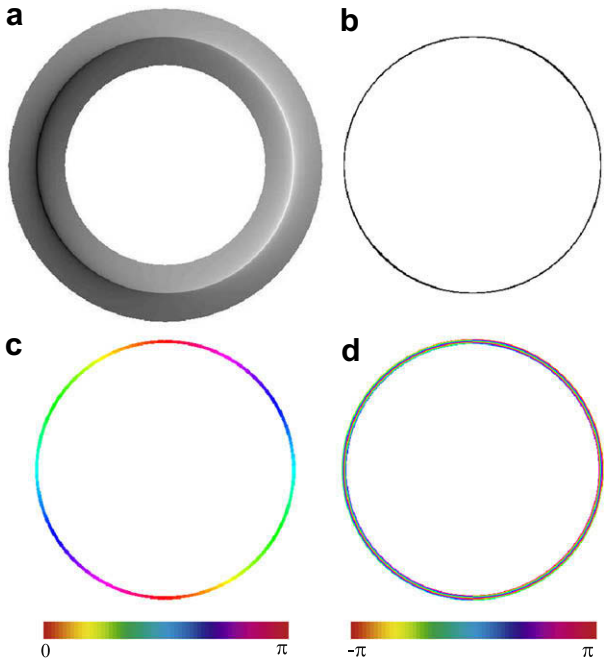

Fig. 3. (a) Test image representing a continuum of phases taking values between $-\pi$ and π corresponding to a continuum of oriented grey-level structures as expressed in the changing circular manifold (cf. [48,54]). The feature type changes circularly from a step edge to a line feature, while retaining perfect phase congruency. (b) Phase-based localization of contrast transitions. (c) Orientation estimation. (d) Local phase estimation. Note the linear variation of the phase across the contrast transition. Quantitative measures about correct localization, orientation and phase are reported in Tables 1–3.

characteristic phases, describe the local image structure; in particular, the consistent phase values through scales are 0 and π for lines, and $\pm\pi/2$ for edges.

We measured, for the different filters, the accuracy in the localization of contrast transitions, and in the phase and orientation estimation, comparing the results with the ground-truth. Different levels of white Gaussian noise, expressed in terms of the signal-to-noise ratio (SNR), have been added to the input image to validate the robustness of the approach. Tables 1–3 report the mean errors in localization, orientation and phase and their standard deviations.

To take into account the circular manifold of the changing feature in Fig. Appendix A, we gauge the accuracy in the localization by computing the position of the maximum value of phase congruency on a 1D segment orthogonal and symmetric with respect to every point of the circle of the ground truth. The accuracy in the orientation estimation is performed at the spatial position of the ground truth with respect to the value of the ground truth. Considering the small spatial period of the band-pass filters, to overcome the discretization problems of sub-pixel phase measures, the ground truth for the computation of the accuracy in the phase estimation is the phase difference between two circular manifolds with a constant offset; the phase estimation is performed at the spatial position of the ground truth independently, for the two circular manifolds.

We can see that, in the absence of noise, Gabor, fourth-order steerable filters (s4), and SQF yield to similar results. Second order steerable filters (s2) seems more noisy in its estimates. Though, it is worth noting that the deterioration of the results is more severe for the SQF than for the other filters. This may be justified by the fact that, as the noise increases, the synthetic image of contrast transitions looses its intrinsically 1D character, on which the performance of the monogenic approach is based. Hence, a higher

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

Table 1
Accuracy evaluation for localization of contrast transitions in the synthetic image of Fig. 3. The localization error is expressed in pixels. The SNR values are expressed in decibel.

	No noise		SNR = 40		SNR = 30		SNR = 25		SNR = 20	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
Gabor	0.033	0.175	0.032	0.173	0.043	0.198	0.060	0.238	0.090	0.288
s4	0.034	0.178	0.033	0.174	0.040	0.193	0.058	0.229	0.098	0.295
s2	0.057	0.228	0.057	0.227	0.060	0.237	0.088	0.285	0.156	0.410
SQF	0.035	0.181	0.041	0.195	0.067	0.253	0.125	0.353	0.268	0.492

Table 2
Accuracy evaluation for orientation in the synthetic image of Fig. 3. The orientation error is expressed in radians. The SNR values are expressed in decibel.

	No noise		SNR = 40		SNR = 30		SNR = 25		SNR = 20	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
Gabor	0.009	0.006	0.019	0.142	0.041	0.213	0.075	0.277	0.209	0.555
s4	0.009	0.007	0.018	0.116	0.063	0.311	0.095	0.335	0.221	0.554
s2	0.019	0.163	0.021	0.141	0.077	0.356	0.122	0.421	0.222	0.549
SQF	0.004	0.006	0.084	0.267	0.250	0.515	0.380	0.615	0.567	0.707

Table 3
Accuracy evaluation for phase in the synthetic image of Fig. 3. The phase error is expressed in radians. The SNR values are expressed in decibel.

	No noise		SNR = 40		SNR = 30		SNR = 25		SNR = 20	
	Avg	Std	Avg	Std	Avg	Std	Avg	Std	Avg	Std
Gabor	0.006	0.007	0.016	0.071	0.039	0.079	0.071	0.130	0.114	0.137
s4	0.021	0.016	0.027	0.060	0.053	0.109	0.091	0.166	0.155	0.205
s2	0.096	0.061	0.097	0.075	0.103	0.104	0.123	0.157	0.169	0.219
SQF	0.009	0.005	0.025	0.058	0.079	0.117	0.145	0.181	0.235	0.246

Table 4
Disparity results with consistency check.

	Tsukuba			Sawtooth			Venus			Teddy			Cones		
	Avg	Std	Dens	Avg	Std	Dens	Avg	Std	Dens	Avg	Std	Dens	Avg	Std	Dens
Gabor	0.27	0.40	96.2	0.26	0.82	94.5	0.18	0.47	95.9	0.58	2.11	84.1	0.22	0.90	92.8
s4	0.28	0.38	92.7	0.33	1.37	91.4	0.24	0.90	90.6	0.92	3.94	76.9	0.26	1.18	86.3
s2	0.33	0.46	80.2	0.61	1.74	69.4	0.56	2.03	71.3	1.65	5.71	53.5	0.84	4.53	52.8
SQF	0.32	0.49	87.0	0.47	1.09	77.4	0.47	1.69	76.6	0.75	3.27	62.3	0.56	2.84	67.4

reliability in detecting contrast direction and orientation is expected for anisotropic multichannel, multiorientation representation in real-world situations, where intrinsic 1D structures are more the exception than the rule.

4.2.1.2. Binocular disparity. A subset of stereo-pairs from the Middlebury stereo vision web-page [83,84] 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. To this end, in the following, we use a left-right consistency check, which is often used to detect occlusions [85], to evaluate the reliability of the disparity estimates. It amounts to comparing the disparity computed for the left frame, δ_L^R, to the disparity computed for the right frame, δ_R^L, at the corresponding pixels. The left frame disparity is used to find corresponding pixels. This results in the following measure:

$$E_\delta(x) = \left|\delta_L^R(x) + \delta_R^L(x + \delta_L^R(x))\right|. \tag{27}$$

The disparities $\delta_L^R(x)$ and $\delta_R^L(x)$ are computed by keeping respectively the left and right frame as (fixed) reference frame, and warping the other frame. Note that the differences result from the warping, so for the single scale case, this error is always zero. A threshold of 0.5 pixels is used to determine reliability. The disparity estimates are thus rejected if $E_\delta(x) > 0.5$. The results are shown in Table 4 and Fig. 4. The best results are obtained with the Gabor filters. Slightly worse are the results with fourth-order steerable filters and the second-order filters yield results about twice as bad as the fourth-order filters. The results obtained with SQFs are comparable with those obtained by the second-order steerable filters. Fig. 4 contains (from top to bottom) the left images of the stereo-pairs, the ground truth depth maps, and the depth maps obtained with the Gabor filters before and after applying the left-right consistency check.

4.2.1.3. Optic flow. We have evaluated the different filters with respect to optic flow estimation on the *diverging tree* and *Yosemite* sequences from [86], using the error measures presented there. The cloud region was excluded from the *Yosemite* sequence. The results are presented in Table 5 and similar conclusions can be drawn as in the previous paragraph. Gabor and fourth-order steerable filters yield comparable results whereas second-order steerable filters score about twice as bad.

Fig. 5 shows the center images, ground truth optic flow fields, and optic flow fields computed with the Gabor filters.

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

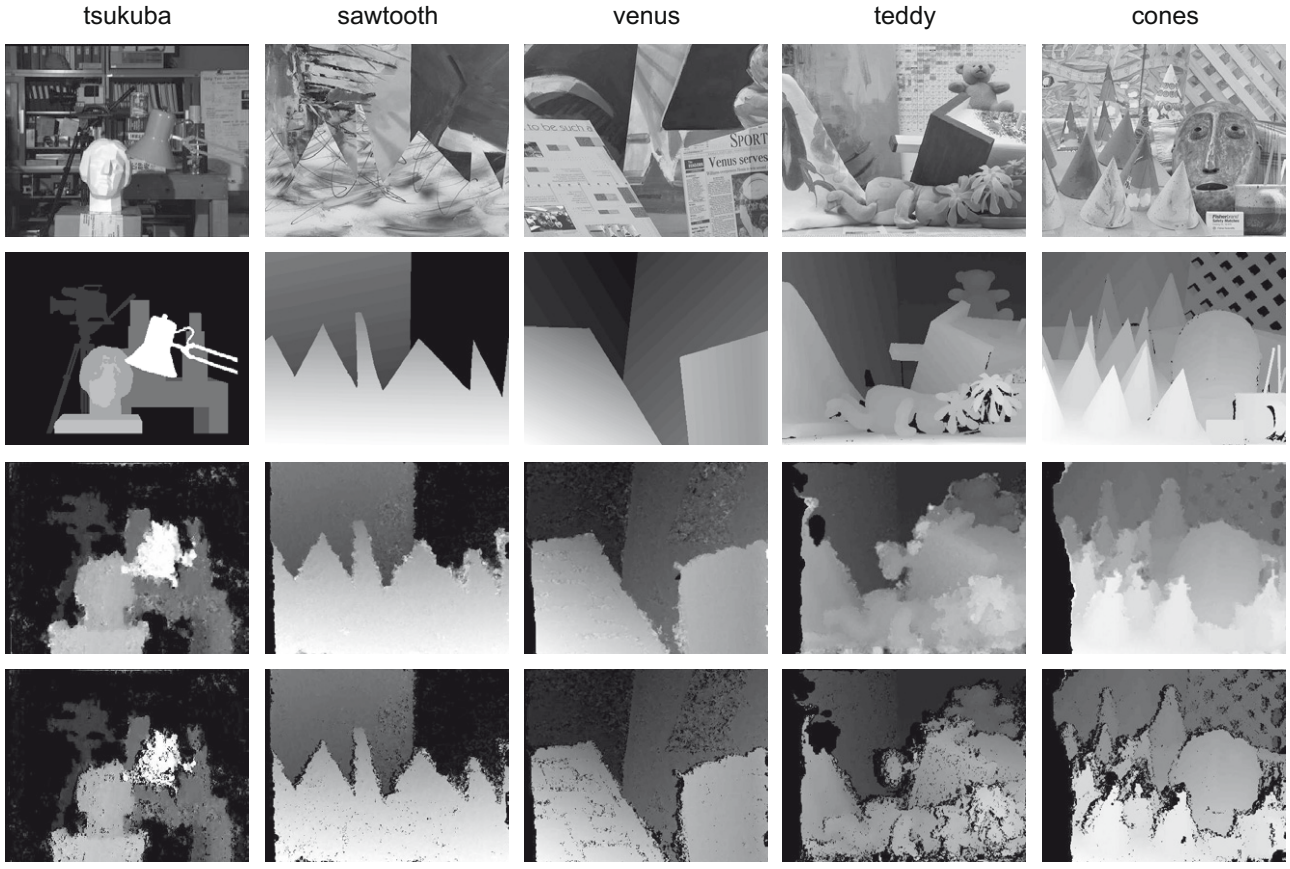

Fig. 4. From top to bottom: left image, ground-truth disparity, Gabor disparity, Gabor disparity after removing inconsistent estimates (left/right consistency check).

Table 5
Average and standard deviation of the optic flow errors (in degs) 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	9.79	8.37	47.1	15.75	16.24	31.4

As already mentioned, the SQF-based optic flow allows the computation of the velocity component orthogonal to the dominant orientation. To make a fairer comparison between the Gabor and the monogenic filters we can back-project the velocities obtained by the two filters on the motion directions given by the ground truth. We adopt this approach just to have a measure of the precision with which the monogenic filters (SQFs) compute the normal flow. Indeed, by definition, the normal velocity component is the projection of the full velocity along the dominant orientation. In this way, by back-projecting the velocity obtained by isotropic filters on the real velocity direction we can obtain the estimate of the full velocity from the monogenic filters. It is worth noting that, in this way, we do not solve the aperture problem for isotropic filters since we make an *a posteriori* evaluation of the full velocity that can be obtained when one knows the ground truth velocity map. The average optic flow errors (in pixel/frame), along the motion directions given by the ground truth, are shown in Table 6. Still, the anisotropic multichannel approach outperforms the isotropic one.

4.2.2. Real-world sequences

In this Section, we want to qualify comparatively the reliability of the disparity and optic flow estimates obtained in real-world scenes, for the different filters considered. Since we evaluate the algorithms on real-world data for which ground-truth is unavailable, some measure of the reliability of the estimates is required. For disparity, we use the left-right consistency check already introduced and adopted in Section 4.2.1. Concerning optic flow, by assuming a simplified version of our algorithm that uses the temporal phase difference between two subsequent frames as an approximation of the temporal phase gradient:

$$\mathbf{v}_{c,q}(\mathbf{x}) = \frac{\lfloor \phi_q(\mathbf{x},t) - \phi_q(\mathbf{x},t+1) \rfloor_{2\pi}}{\omega_0} \cdot \frac{\boldsymbol{\omega}}{|\boldsymbol{\omega}|}, \tag{28}$$

and integrating the component velocities as before (Eq. (26)), an analogous consistency measure can be used:

$$E_{\mathbf{v}}(\mathbf{x}) = \left\| \mathbf{v}_t^{t+1}(\mathbf{x}) + \mathbf{v}_{t+1}^{t}(\mathbf{x} + \mathbf{v}_t^{t+1}(\mathbf{x})) \right\|, \tag{29}$$

where the two-frame optic flow fields, $\mathbf{v}_t^{t+1}(\mathbf{x})$ and $\mathbf{v}_{t+1}^{t}(\mathbf{x})$, are obtained by fixing respectively frame t and frame $t+1$, and warping the other frame. Both for disparity and optic flow a threshold of 0.5 pixels is used to determine reliability. The disparity estimates are thus rejected if $E_\delta(x) > 0.5$, and the optic flow estimates if $E_{\mathbf{v}}(\mathbf{x}) > 0.5$.

We show results on two real-world sequences, recorded with a stereo camera system rigidly installed behind the front shield of a moving car (see Fig. 6). In the *town* sequence, the car is moving forward slowly, inducing a mostly translational motion field on the scene. There are also pedestrians crossing the street. In the *tour* sequence, the car is negotiating a curve at a much higher speed.

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

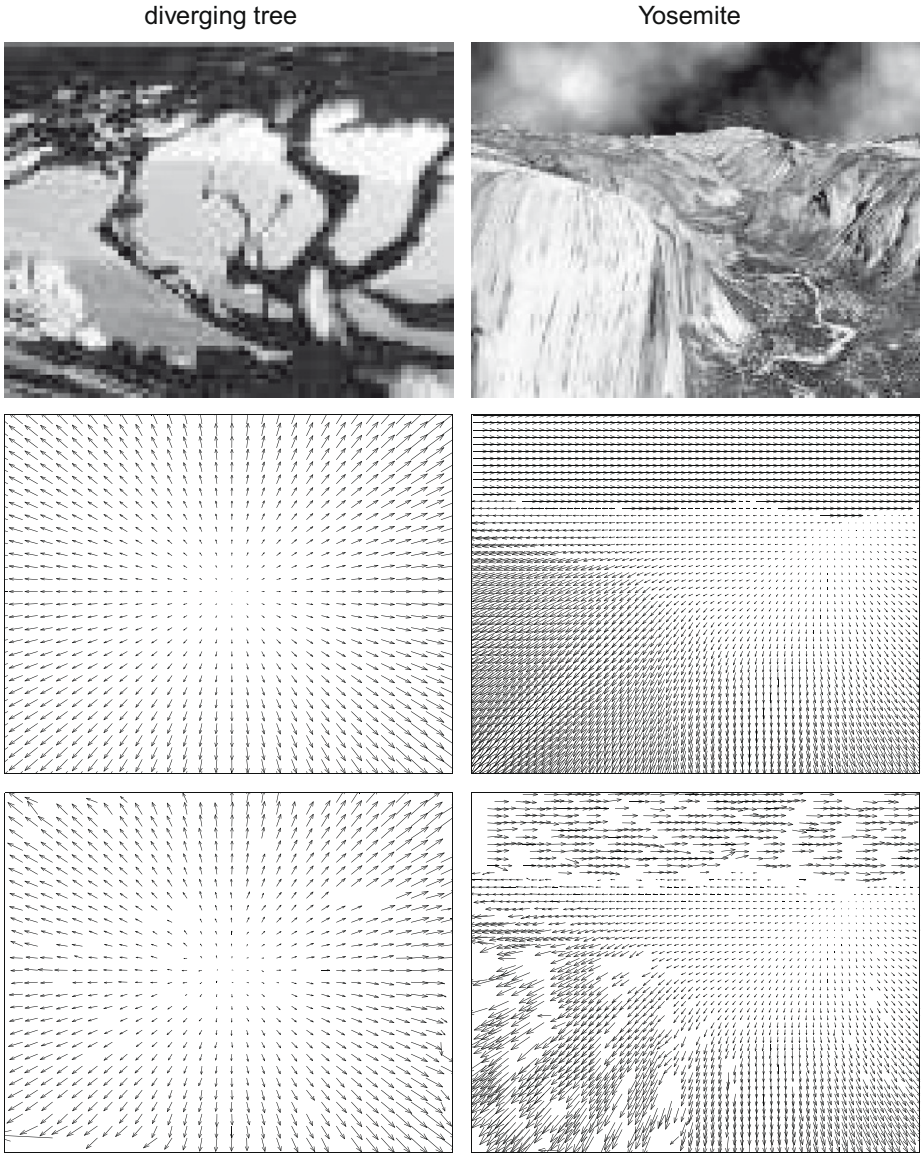

Fig. 5. Center frame (top row), ground truth optic flow (middle row) and estimated optic flow obtained with Gabor filters (bottom row). All optic flow fields have been scaled and sub-sampled five times.

Table 6
Average and standard deviation of the optic flow errors (in pixel/frame) and optic flow density (in percent).

	Diverging tree			Yosemite (no cloud)		
	Avg	Std	Dens	Avg	Std	Dens
Gabor	0.04	0.05	95.6	0.08	0.14	81.8
SQF	0.16	0.78	47.1	0.41	3.82	31.4

For both sequences, the disparity results are similar for Gabor and s4 (with Gabor slightly better). The disparity obtained with s2, on the other hand, contains more noise, and the left/right consistency correctly discards these estimates. The isotropic disparity maps, obtained by SQF, retain far less of the scene's fine structure and contain many noisy estimates that have been removed on the basis of $E_\delta(x)$. A similar result is obtained for optic flow: for both sequences, the optic flow results are similar for Gabor and s4; the flow fields obtained with s2, on the other hand, clearly contain more noise. In general, the anisotropic optic flow estimates, compared to those obtained by SQF, are better and denser flow fields remain after thresholding. Yet, it is worth recalling that the results obtained by the point-wise application of the SQF are plagued by the aperture problem, as discussed in Section 4.2.1. These two-frame flow fields are quite good, considering that no reliability measure is used during the coarse-to-fine processing. Estimates obtained with the five-frame algorithm contain much less noise, as it is noticeable for the *town* sequence, shown previously in Fig. 2d.

4.2.3. Computational load

Since we are interested in computing different image features with the maximum accuracy and the lower processor requirements, the computational cost of the different filters adopted must be considered, too. The utilization of the different filtering approaches leads to different computing load requirements. Focusing on the convolution operations on which the filters are based, we have analyzed each approach to evaluate their complexity. Spherical filters require three non-separable 2D convolutions operations, which makes this approach quite expensive in terms of the re-

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

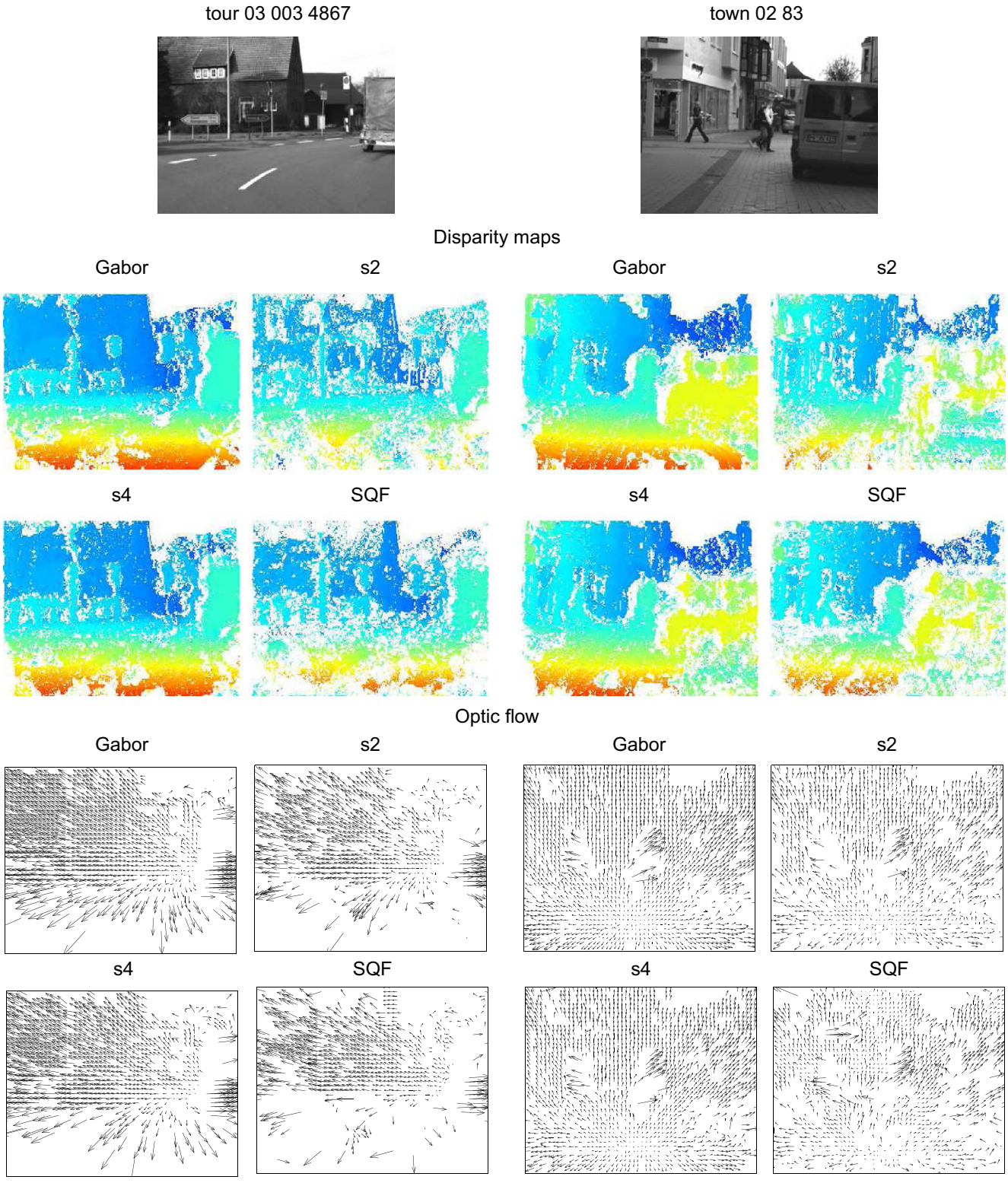

Fig. 6. Comparative results for disparity and motion estimates on two real-world sequences (*tour*) and *town*), for the different filters used: Gabor filters, steerable filters (s2 and s4) and spherical quadrature filters (SQF). For optic flow comparison, a simplified version of our algorithm is adopted, which approximates the temporal phase gradient from two subsequent frame, only. Though the results contain much more noise (with respect to the five-frame algorithm), especially for the *town* sequence, where the motion is very small, this simplification allows us to apply a simple consistency validation, as a qualitative measure of reliability (see text). Both for disparity and optic flow, the results obtained with s2 and SQF retain far less of the scene's fine structure and contain much more noisy (i.e., unreliable) estimates, that have been removed after the consistency check. The estimates for Gabor and s4 are much better and reliable, since much denser feature maps remain after validation thresholding.

quired computational resources (this might be true also for modern platforms like GPUs e.g., see [87], where fast convolution algorithms still exploit filter separability). The eight oriented Gabor filters require eight 2D non separable convolution, but they can be efficiently computed through a linear combination of separable kernels as it is indicated in [32], thus significantly reducing the

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

Table 7
Computational complexity of the filtering stage with the different kernels.

	# 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

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 1D convolutions is given by $4L + 6$, where L is the differentiation order.

The complexity of computing the harmonic representation with the different set of filters is summarized in Table 7. It is worth noting that the anisotropic solutions require more than 8/3 time memory resources than the isotropic approach, which has to be taken into account for embedded systems; nevertheless, it is not an issue for current computing machine.

5. Discussion

5.1. Complete harmonic analysis of the visual signal

The understanding of a visual signal involves carefully defining which feature to extract, or, from a different perspective, which kind of representation to adopt for the visual signal itself. Although the two issues cross relate each other, there is a substantial distinction between them, which set the "feature detection hypothesis" against the "signal analysis hypothesis". The former relies on matched operators that extract the most informative (symbolic) elements of an image, such as points and lines (but that inevitably discard part of the signal), the latter performs a mapping to a quasi holographic description, meaning that the visual signal is described in terms of more general structural properties of a local portion of the "plenoptic" space [5]. Such structural properties comprise (first-order) spatial relationships in the light intensity functions (e.g., oriented similarities along contours and transitions across them, as well as basic symmetries), but also their comparisons/relationships in time and between different viewpoints. For many image processing tasks, it is commonly used to represent an image by oriented spatial-frequency (scale-space) channels (cf. the wavelet transform) in which some properties of the image are better represented than in image space. The spatial behavior in each channel, and the relationships between the channels are critically important for extracting primary early vision information.

In this paper, we have revised the harmonic description of the visual signal based on oriented spatial frequency channels to account for a complete characterization of the 2D local structure of the visual signal in terms of the phase properties/relationships from all the available channels. The orientation of the channels is instrumental to measure the complete harmonic information content, since it provides reference axes with respect to which one can evaluate the spatial symmetries of the signal. Provided that a sufficient number of oriented channels (and scales, to account for the different granularity of the image structures) are used, the description allows us to achieve optimal perceptual performances with a minimal computational cost. This confirms the richness of the representation, which is capable, as a whole, to fully describe the structural properties of the original signal (cf. the split of identity concept [74]). Accordingly, the information content of the original signal is preserved, with the advantage that comparative structural analysis can be performed in a more efficient way. Indeed, in general, as evidenced in several studies (e.g., see [37,42,44]), by using harmonic patterns for matching instead of image luminance measures, the resulting perception is more reliable (i.e., stable), denser and immune to lighting conditions.

5.2. Anisotropic vs. isotropic channels

In general, the phase information of a multidimensional signal provides information about the symmetries of the signal with respect to different hyperplanes. Given the 2D character of the spatial manifold of the image signal, the phase information of a visual signal provides information about the spatial (a-)symmetries with respect to different oriented axes. For each symmetry axis, we can measure a phase and, if the signal has a rich structure, several symmetry axes should be considered. The number of relevant axes for characterizing the local structure of a visual signal can be related to the notion of intrinsic dimensionality, introduced by [88], as a measure of the degree of redundancy of a signal in a local neighborhood (i.e., image patch), on the basis of the spatial distribution of its energy spectrum.

Accordingly, an image point $\mathbf{x}$ can be classified as (i) intrinsically 0D (i0D) (ii) intrinsically 1D (i1D), or (iii) intrinsically 2D (i2D) depending on the two-dimensional image-intensity function $I(\mathbf{x})$ in the neighborhood of that point, which can (i) be constant in all directions, (ii) be constant in one direction, or (iii) vary in all directions. Edges, lines, and gratings characterized by iso-curves of $I(\mathbf{x})$ with a common direction [89], correspond to i1D patches. All other structures like corners, junctions, complex textures, and noise correspond to i2D patches.

Each oriented channel is capable of measuring the phase of an i1D signal with respect to its linear (characteristic) symmetry axis. However, it is worth noting that, although in principle the value of the phase is correct, its confidence decreases as far as the linear symmetry axis deviates from the orientation axis of the filter. Hence, we can state that the energy value of the associated Hilbert transform is not isotropic, since it is not invariant under rotations of the signal. The isotropy of the representation is recovered when one considers the whole set of oriented channels, only (see Fig. 7).

Furthermore, if we do not pursue feature detection, but the image signal analysis, the different responses of the simultaneously active channels increase the number of dimensions in the harmonic representation of the image patch, thus favoring an holistic *vs.* reductionist approach for visual perception. An important source of difficulties that arise in an attempt to make structural comparison (cf. also grouping pixels together) is the distributed nature of the information that can be potentially relevant to perceptual decisions. Indeed, such information may be inherently holistic: the ultimate interpretation of an image fragment usually depends on its context, if not on the entire image.

More precisely, for i1D signals (and low values of noise), the phase measurements obtained through oriented frequency channels can be interpreted either with respect to the orientation of the signal or with respect to the orientation of the filter. Indeed, each filter measures the phase of the signal with respect to its characteristic symmetry axis (i.e., across its dominant orientation). From a different perspective, each filter gathers information about the signal's phase with reference to its oriented bandwidth (i.e., across the orientation of the filter) and a vector averaging operation (cf. Haglund [73]) must be used to decode the local phase. For i2D signals, such as corners and most textures, there does not exist a single (characteristic) symmetry axis, and, even if we arbitrarily select one, the phase measure would be influenced (and thus corrupted) by the signal's energy distribution along other symmetry axes that characterize its complex structure. Therefore, in this case phase measures with respect to the orientations of the filters are the only practicable. In other words, when there is not a characteristic (i.e., dominant) orientation of the signal it is

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

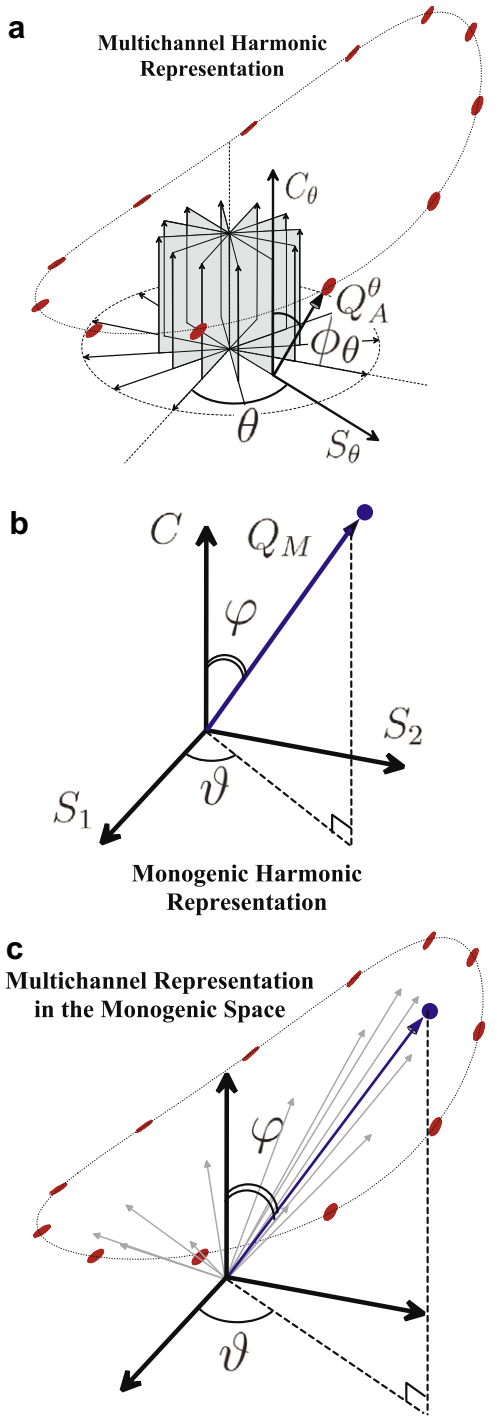

Fig. 7. Comparison between anisotropic *vs.* isotropic harmonic representations. (a) Multi-channel harmonic representation: the analytic image Q_A^θ is defined in a complex plane parameterized by a discrete number of orientations θ. (b) Monogenic harmonic representation: the monogenic image Q^M provides a synthetic description of the i1D dominant structure (local phase, orientation, and amplitude) of the image signal as a single 3D phasor in spherical coordinates. (c) Multi-channel representation in the monogenic space: the integration of information over different oriented spatial frequency channels allows us to obtain single estimates of the dominant orientation and phase, while retaining the distributed/channeled representation of the image signal.

not possible to determine a single value of phase. Yet, we can consider the different orientations of the filters, with respect to which one can extract the different phase values that characterize the local structure of the signal (e.g., most textures are characterized by multiple orientations at different frequencies). As a consequence, the use of anisotropic *vs.* isotropic frequency channels allows us to describe not only i1D, but also i2D image signals, without paying the price of postulating a dimension reduction of the intrinsic structural properties of the signals.

In the computational strategies adopted and illustrated in the previous Sections, aimed at the interpretation of the scene from stereo and motion information, we evidenced that the phase relationships among oriented spatial features (e.g., edge and line correspondences) are not really important, but the harmonic content gathered by a full set of oriented channels.

For what concerns the compactness of channels in tiling the frequency space, it is worth noting that a certain trade-off emerges between two opposite design constraints. For i1D signals a "population coding" design strategy, based on a reduced set of overlapping channels, can prevail (provided that a reasonable size of the spatial neighborhood is maintained). In general, for i2D signals, an "interval coding" approach might be preferable, based on a large set of narrow-band channels with minimal overlap, such that they do not compromise too much the spatial localization (cf. local structural neighborhood), but that, at the same time, allow a proper identification of the orientation axis for extracting the phase information (cf. orientation bandwidth). The choice made in this paper (eight oriented Gabor filters, within one octave bandwidth) meets an optimal trade-off between the two approaches. The aspect ratio of the Gabor filters can be an additional degree of freedom, which has not been exploited in order to allow a better fairness among the other filters used in the comparison.

5.3. Towards mid-level descriptors

A transition from a pixel based representation to a more condensed symbolic representation, based on the harmonic code has been realized in [9] (see Fig. 8).

In this representation, local image patches covering edge and line-like structures (i.e., i1D structures) are coded by a symbolic descriptors covering position, orientation and phase computed from the harmonic code. This corresponds to a decomposition of the local signal into amplitude information, orientation information, and phase information (split of identity [12]). The amplitude information can be used as an indicator for the likelihood of the presence of a certain structure while the orientation and contrast transition will be used as attributes of the symbolic descriptors. Furthermore, color is coded according to the local structure either as two color vectors representing the left and right side of a step edge or as three color vectors (in addition a middle color in case of a line structure (see Fig. 8f). The local phase is used to distinguish between edge- and line-like structures (e.g., see [74]) and Fig. 3).

The transition to a local symbolic descriptor has certain advantages in the context of an early cognitive architecture [90,91]. First, since it reduces the number of bits representing a local area, comparison between or relations to other image areas can be performed more efficiently. Moreover, the memorizing of information at higher levels of vision becomes facilitated. Another aspect is the higher predictability of the condensed information, for example in the context of predicting the change of a local patch under motions (for details see [92]). Concerning the transition to a symbolic representation, an isotropic harmonic code leads to difficulties. In [93,94], a triangular representation of different image structures has been developed, based on the concept of the intrinsic dimensionality of the local signal. A triangle with its three corners representing homogeneous image patches (i0D), edge-like structures (i1D) and junction/noise like structures (i2D) can be established and hence these structures can be distinguished efficiently (see [94]). However, the isotropic representation might be adequate

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

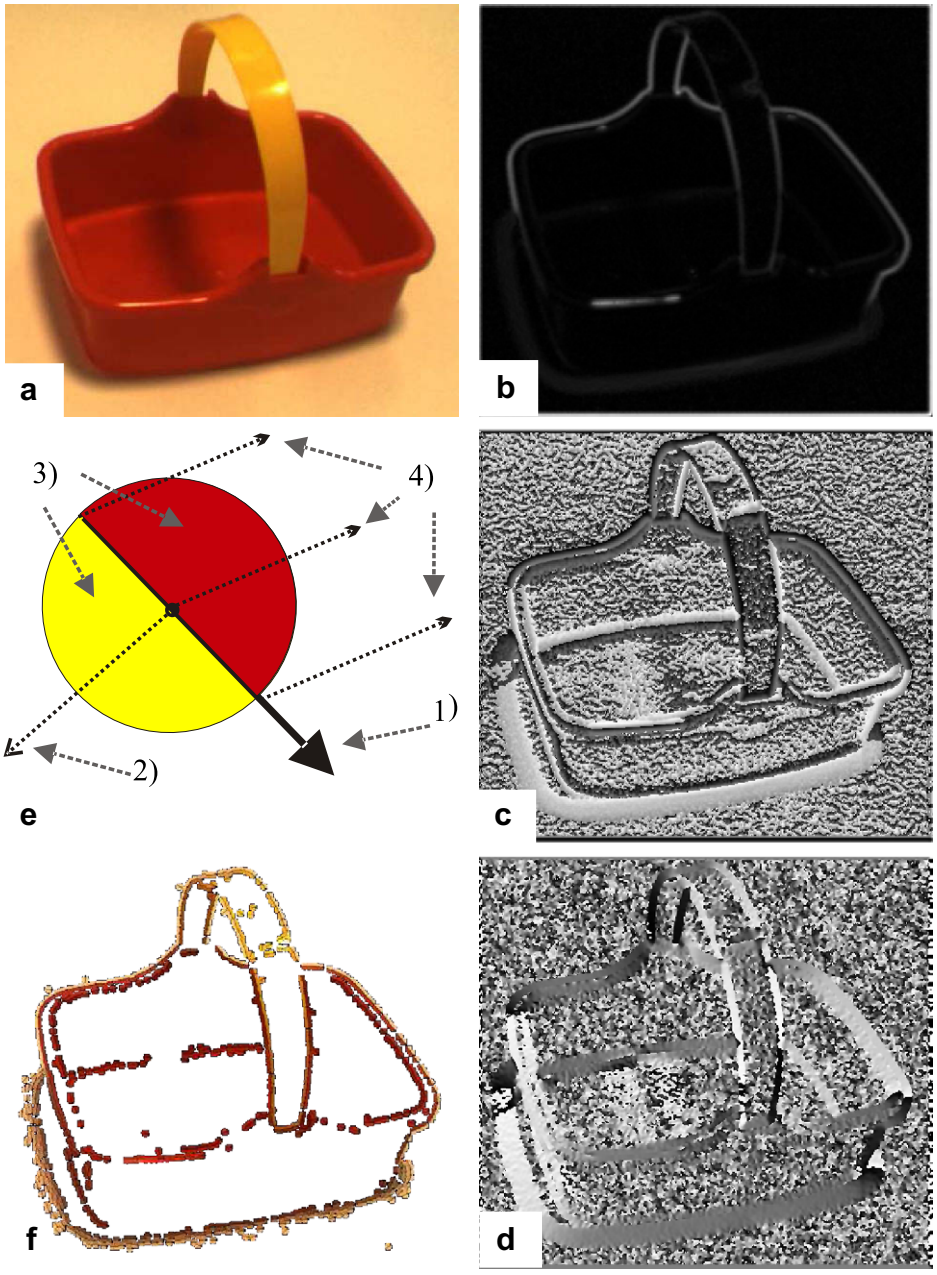

Fig. 8. Illustration of the primitive extraction process. (a) An image of an object. The signal can be decomposed in the magnitude (b), phase (c), and orientation (d) information. This information is encoded into sparse symbolic vectors called primitives. (e) A pictorial representation of a primitive, with: the orientation (1), the phase (2), the color (3) and the optic flow (4). (f) The primitives extracted from the image.

for i1D signals (although still questions can be raised about for example more elongated filter that increase orientation and position specificity [6]). However, there is a strong point in favor of an explicit sampling of orientations for junction- and in particular texture-like structures since these kinds of image structures are characterized by high local amplitudes in multiple orientations in parallel. Here an adequate description of the local signal (as well as any transition to a symbolic level) requires a non-isotropic harmonic code. In addition, it worth mentioning that a distributed representation (over the oriented spatial frequency channels) might favor adaptation mechanisms, by providing a higher number of potential "entry points" for top-down contributions from mid-level back to low-level. Such feedback or re-entrant mechanisms can mimik the contextual or attentional modulation of the sensorial input that occur through dense intra- and inter-area feedback interconnections in visual cortical areas, adapting visual cells tuning and refining their selectivity (e.g., [95,96]).

5.4. Phase-based second-order motion features

On the basis of the early phase-based features extracted as described in Section 3, it is possible to build more complex visual descriptors based on second-order motion properties. By example, the perception of motion in the 3D space relates to second-order measures, which can be gained either by inter-ocular velocity differences or temporal variations of binocular disparity [97]. In [52], 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:

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

$$\frac{d\delta}{dt} \simeq \frac{\partial\delta}{\partial t} = \frac{\phi_t^L - \phi_t^R}{\omega_0} \simeq v^R - v^L, \tag{30}$$

where v^R and v^L are the velocities along the (horizontal) epipolar lines.

The approximations depend on the robustness of phase information, and the error made is the same as the one that affects the measurement of phase components around singularities [37]. 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:

$$\begin{aligned}\frac{\partial\delta}{\partial t} &= \left[\frac{\mathrm{Im}[Q_t^L Q^{*L}]}{|Q^L|^2} - \frac{\mathrm{Im}[Q_t^R Q^{*R}]}{|Q^R|^2}\right]\frac{1}{\omega_0} \\ &= \left[\frac{S_t^L C^L - S^L C_t^L}{(S^L)^2 + (C^L)^2} - \frac{S_t^R C^R - S^R C_t^R}{(S^R)^2 + (C^R)^2}\right]\frac{1}{\omega_0}\end{aligned} \tag{31}$$

thus avoiding explicit calculation and differentiation of phase, and the attendant problem of phase unwrapping. The terms S_t and C_t refer to the temporal variations of S and C, respectively, which can still be obtained through a linear fitting procedure (cf. Eq. (14)) over five consecutive frames. The direct determination of temporal variations of the disparity, through filtering operations, better tolerates the problem of the limit on maximum disparities due to "wrap-around" [45], yielding correct estimates even for disparities greater than one-half the wavelength of the central frequency of the Gabor filter. The monocular terms in Eq. (31) can be interpreted as the binocular velocities along the epipolar lines v_q^L and v_q^R, for any given orientation q.

Although the motion-in-depth is a second-order measure, by exploiting the direct determination of the temporal derivative of the disparity and by combining information over the oriented spatial frequency channels, the inter-ocular velocity difference, and thence the motion-in-depth can be directly calculated from filters' outputs by:

$$\mathrm{V_Z}(x) = \underset{q\in W_L(x)}{\mathrm{median}}\, v_q^L(x) - \underset{q\in W_R(x)}{\mathrm{median}}\, v_q^R(x), \tag{32}$$

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.

Since binocular test sequences with the ground truth and a sufficiently high frame rate are not available, making quantitative comparisons among the different filters has not been possible. However, considering that 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 (not shown) obtained in real-world sequences preliminarily confirmed this conclusion.

6. Concluding remarks

The first stages of a vision system ("early vision") consists of a set of parallel pathways each analyzing 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 adopting thresholding procedures, which depend on a specific and restricted environmental context, limits the possibility of building 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 importance of this for vision in the brain was highlighted in [98] as being a natural way to reduce the computational complexity of visual processing. 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 band-pass channel); (2) coverage of the frequency domain; and (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 constructing a multichannel, multi-orientation representation is preferable in order to avoid 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. It is worth noting that phase-based signatures are texture-based (or "correspondenceless") measures, hence, they do not suffer from the problem of false matches and provide dense feature maps, as opposite to edge matching algorithms, provided that we have sufficient texture information over local image patches. On the other hand, it is likewise true that phase information measurements appear to be more stable in the neighborhood of localized salient image features, such as edges, bars, and ramps. In the vicinity of these points the different harmonics sum themselves coherently (i.e., in phase), thereby improving the signal to noise ratio. This observation establishes an interesting bridge between sparse (feature-based) and pure dense techniques (such as area correlation), which are both embraced within a phase-based approach.

Acknowledgments

We wish to thank Michael Felsberg for stimulating discussion and his critical review of a previous version of the manuscript, and Alexander Rotter of Hella KGaA Hueck & Co. for having provided the driving sequences. This work was mostly developed in the "DrivSco" project (16276-2) funded in the 6th Framework Programme of the European Union IST-FET "Future and Emergent Technologies".

Appendix A. Filter design specification

Gabor filters – a Gabor oriented filter along an angle θ with respect to the horizontal axis is defined by:

$$f_{\mathrm{Gabor}}^{\theta}(x,y) = \mathrm{e}^{-\frac{x^2+y^2}{2\sigma^2}}\mathrm{e}^{\mathrm{j}\omega_0(x\cos\theta + y\sin\theta)} \tag{A.1}$$

where ω_0 is the peak frequency of the filter and σ determines its spatial extension. The spatial window has been chosen as four times σ. At the highest scale $\omega_0 = \pi/2$ and $\sigma = 2.67$. Following [32], we implemented the oriented filters as sums of separable filters. By exploiting symmetry considerations, all eight even and odd filters (see Fig. A.1) can be constructed on the basis of twenty four 1D convolutions. The 1D filters are modified by enforcing zero DC sensitivity on the resulting 2D filters in which they take part, and by minimizing the difference with the theoretical 2D Gabor filters. Specific care have been paid to adjust the coefficients of each filter function so that the even and odd symmetry is respected. To this purpose, a constrained non-linear multivariable minimization is adopted.

Steerable filters – following [23], an approximation of a complex-valued Gabor filter of arbitrary orientation θ can be synthesized by taking a linear combination of steerable quadrature

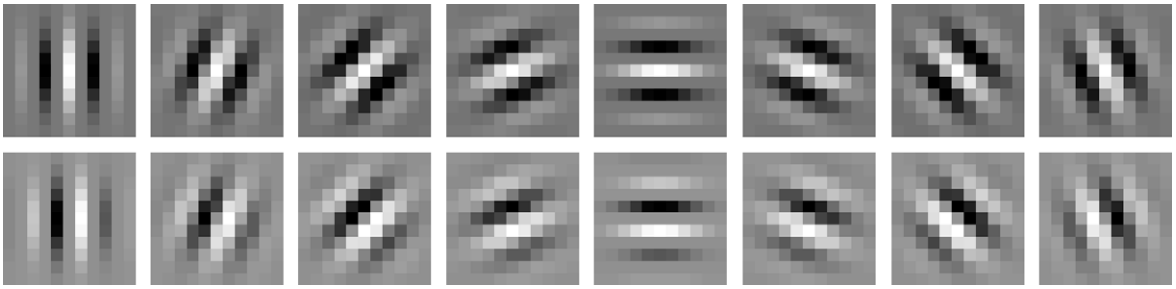

Fig. A.1. The resulting 11×11 quadrature pair of Gabor filters for $\omega_0 = \pi/2$ and eight orientations.

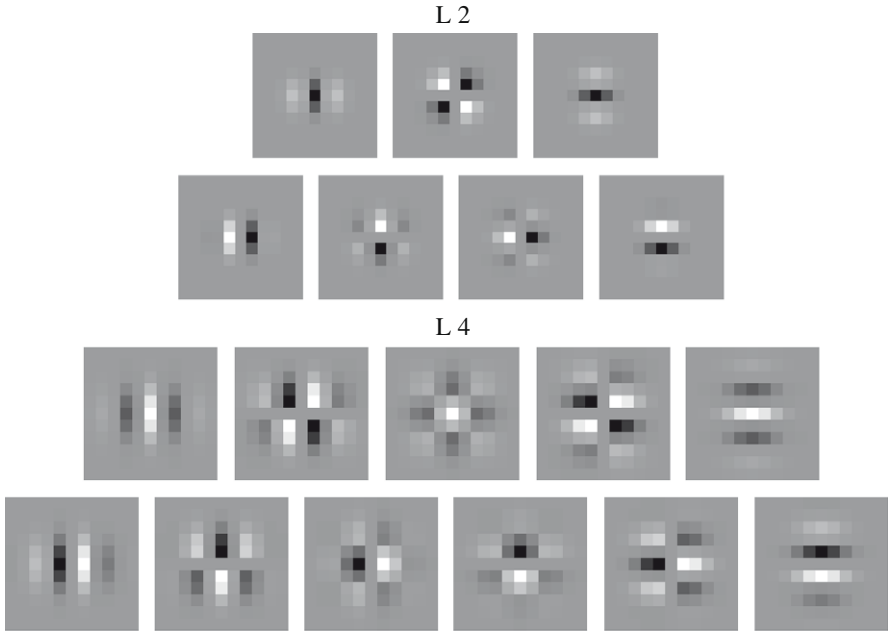

Fig. A.2. The 11×11 x–y separable, steerable quadrature pair basis filters for two different orders of differentiation. The width of the Gaussian function has been adjusted to have, for both cases, a resulting $\omega_0 = \pi/2 : \sigma = 0.90$ for $L = 2$ and $\sigma = 1.27$ for $L = 4$.

pairs of 2D Gaussian directional derivatives, along the cardinal axes:

$$g_0(x,y) = \mathrm{e}^{-\frac{x^2+y^2}{2\sigma^2}} \tag{A.2}$$

$$g_L(x,y) = \frac{\partial^{L-l}}{\partial x^{L-l}} \frac{\partial^l}{\partial y^l} g_0(x,y) \quad l = 0, 1, \ldots, L-1 \tag{A.3}$$

$$f^{\theta}_{\mathrm{Steer}}(x,y) = g_0(x) \sum_{l=1}^{L} b_l(\theta) P_{l,\sigma}(x) Q_{l,\sigma}(y) \tag{A.4}$$

where $b_l(\theta)$ are the interpolation functions:

$$b_l(\theta) = (-1)^l \binom{L}{l} \cos^{L-l}\theta \sin^l\theta \tag{A.5}$$

L is the order of differentiation, and P_l and Q_l are polynomial functions defined as:

$$P_{l,\sigma}(x) Q_{l,\sigma}(y) = \left(\frac{x^{L-l}}{\sigma^{2(L-l)}} + \cdots\right)\left(\frac{y^l}{\sigma^{2l}} + \cdots\right). \tag{A.6}$$

Gaussian derivatives asymptotically coincide to a Gabor function with a radial peak frequency $\omega_0 = \sigma^{-1}\sqrt{L+1}$ and an absolute bandwidth $\Delta\omega = \sigma^{-1}\sqrt{2}/2$ [4]. Since the peak frequency and the bandwidth are jointly defined by σ, it is not possible to design banks of steerable filters with an arbitrarily constant relative bandwidth. We therefore adjusted the spatial extension of the Gaussian function (σ) in order to have the same peak frequency of the Gabor filters, and deduced as a consequence the relative bandwidth. The number of basis kernels to compute the oriented outputs of the filters depends on their derivative order. The quadrature pair of these filters has been obtained by approximating their Hilbert transform as a the least square fit to a polynomial times a Gaussian described in [23]. The basis filters corresponding to Gaussian derivatives of second- or fourth-order (see Fig. A.2) turned out as an acceptable compromise between the representation efficacy (i.e., optimality in terms of the Heisenberg–Weyl uncertainty principle) and the computational efficiency.

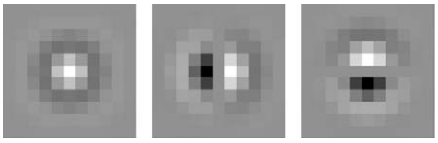

Fig. A.3. The 11×11 spherical quadrature filter (SQF) triplet.

Spherical Quadrature filters – the 2D SQF (see Fig. A.3) is constructed as

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

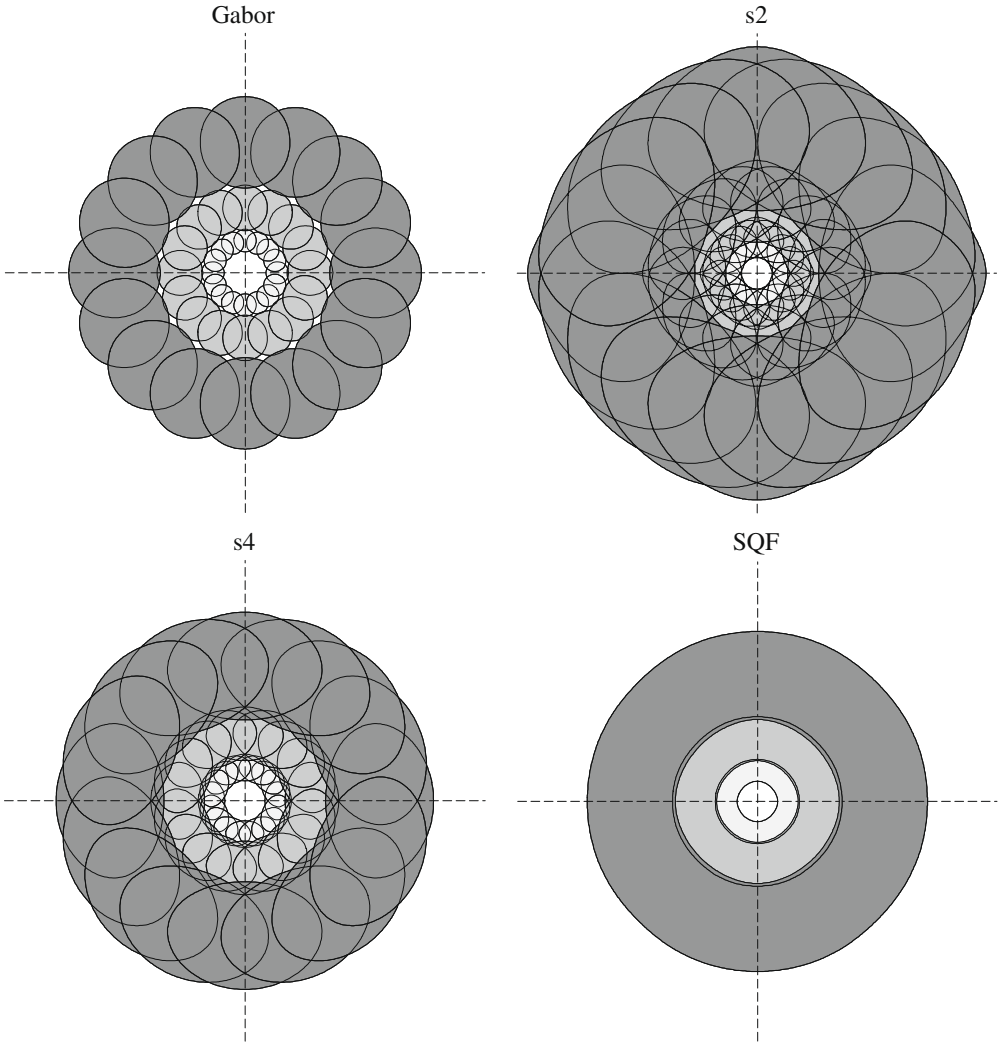

Fig. A.4. (a–c) Rosette-like diagrams of the multichannel frequency representation for the Gabor, and the steerable filters s2 ($L = 2$) and s4 ($L = 4$), respectively. It is worth noting that the orientation bandwidth of the steerable filters is larger than that obtained with Gabor filters. (d) Bandpass isotropic channels obtained by the monogenic signal. Contours correspond to half-width cut-off frequencies, and each corona is separated by an octave scale.

1. A rotation invariant bandpass filter obtained as the sum all the real parts of the oriented Gabor filter bank:

$$f(x,y) = \sum_{\theta} \mathrm{Re}[f^{\theta}_{\mathrm{Gabor}}(x,y)] \tag{A.7}$$

2. A vector-valued filter with the desired isotropic oddness, obtained by filtering $f(x,y)$ with the convolution kernel of the Riesz transform:

$$\begin{bmatrix} h_1(x,y) \\ h_2(x,y) \end{bmatrix} = \left[\frac{-x}{2\pi(x^2+y^2)^{\frac{3}{2}}}, \frac{-y}{2\pi(x^2+y^2)^{\frac{3}{2}}} \right]^T \tag{A.8}$$

$$f_{\mathrm{SQF}}(x,y) = [f(x,y), (h_1 * f)(x,y), (h_2 * f)(x,y)] \tag{A.9}$$

All the filters have been normalized prior to their use in order to have constant energy. The corresponding rosette-like frequency representation of the filters used is shown in Fig. A.4, for three different scales (octaves).

From the frequency representation of the Gabor-based spherical filter we observe slight deviations from isotropy due to numerical approximation errors, which though does not affect the results presented in this paper. We can observe that a more isotropic SQF filter (but not comparable with Gabor's) can be obtained starting with first-order Gaussian derivatives as Riesz components and by numerically compute the rotation invariant bandpass filter [99].

References

[1] J. Daugman, Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, J. Opt. Soc. Am. A 2 (1985) 1160–1169.
[2] S. Marčelja, Mathematical description of the responses of simple cortical cells, J. Opt. Soc. Am. A 70 (1980) 1297–1300.
[3] D. Gabor, Theory of communication, J. Inst. Elect. Eng. 93 (1946) 429–459.
[4] J. Koenderink, A. van Doorn, Representation of local geometry in the visual system, Biol. Cybern. 55 (1987) 367–375.
[5] E. Adelson, J. Bergen, The plenoptic and the elements of early vision, in: M. Landy, J. Movshon (Eds.), Computational Models of Visual Processing, MIT Press, 1991, pp. 3–20.
[6] J. Jones, L. Palmer, The two-dimensional spatial structure of simple receptive fields in cat striate cortex, J. Neurosci. 58 (1987) 1187–1211.
[7] G. De Angelis, I. Ohzawa, R. Freeman, Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation, J. Neurophysiol. 69 (1993) 1118–1135.
[8] M. Carandini, J. Demb, V. Mante, D. Tolhurst, Y. Dan, A. Olshausen, J. Gallant, N. Rust, Do we know what the early visual system does?, J Neurosci. 25 (2005) 10577–10597.
[9] N. Krüger, M. Lappe, F. Wörgötter, Biologically motivated multi-modal processing of visual primitives, Interdiscipl. J. Artif. Intell. Simulat. Behav. 1 (5) (2004) 417–428.
[10] J. Daugman, Spatial visual channels in the Fourier plane, Vision Res. 24 (1984) 891–910.

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

[11] R. De Valois, K. De Valois, Spatial Vision, Oxford University Press, New York, 1990.
[12] M. Felsberg, G. Sommer, The monogenic signal, IEEE Trans. Signal Process. 48 (12) (2001) 3136–3144.
[13] R. Young, The Gaussian derivative theory of spatial vision: analysis of cortical cell receptive field line-weighting profiles, Tech. Rep. GMR-4920, General Motors Research, 1985.
[14] A. Watson, The cortex transform: rapid computation of simulated neural images, Comput. Vison Graph. Image Process. 39 (1987) 311–327.
[15] M. Hawken, A. Parker, Spatial properties of neurons in the monkey striate cortex, Proc. Roy. Soc. Lond. B 231 (1987) 251–288.
[16] D. Field, Relations between the statistics of natural images and the response properties of cortical cells, J. Opt. Soc. Am. A 4 (1987) 2379–2394.
[17] J. Martens, The Hermite transform – theory, IEEE Trans. Acoust., Speech, Signal Process. 38 (1990) 1595–1606.
[18] D. Stork, H. Wilson, Do Gabor functions provide appropriate decriptions of visual cortical receptive fields?, J Opt. Soc. Am. A 7 (8) (1990) 1362–1373.
[19] J. Yang, Do Gabor functions provide appropriate descriptions of visual cortical receptive fields?: comment, J. Opt. Soc. Am. A 9 (2) (1992) 334–336.
[20] S. Klein, B. Beutter, Minimizing and maximizing the joint space-spatial frequency uncertainty of Gabor-like functions: comment, J. Opt. Soc. Am. A 9 (2) (1992) 337–340.
[21] S. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (7) (1989) 674–693.
[22] T. Reed, H. Wechsler, Segmentation of textured images and Gestalt organization using spatial/spatial frequency representations, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1990) 1–12.
[23] W. Freeman, E. Adelson, The design and use of steerable filters, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 891–906.
[24] P. Perona, Steerable-scalable kernels for edge detection and junction analysis, Image Vis. Comput. 10 (1992) 663–672.
[25] E. Simoncelli, W. Freeman, E. Adelson, D. Heeger, Shiftable multiscale transforms, IEEE Trans. Inform. Theory 38 (2) (1992) 587–607.
[26] M. Felsberg, G. Sommer, The monogenic scale-space: a unifying approach to phase-based image processing in scale-space, J. Math. Imag. Vision 21 (2004) 5–26.
[27] L. Jacobson, H. Wechsler, Joint spatial/spatial-frequency representation, Signal Process. 14 (1988) 37–68.
[28] H. Wechsler, Computational Vision, Academic Press, 1990.
[29] R. Navarro, A. Tabernero, G. Cristobal, Image representation with Gabor wavelets and its applications, in: P.W. Hawkes (Ed.), Advances in Imaging and Electron Physics, Academic Press, San Diego CA, 1996, pp. 1–84.
[30] E. Adelson, C. Anderson, J. Bergen, P. Burt, J. Ogden, Pyramid methods in image processing, RCA Eng. 29 (6) (1984) 33–41.
[31] D.J. Fleet, A.D. Jepson, Computation of component image velocity from local phase information, Int. J. Comput. Vision 1 (1990) 77–104.
[32] O. Nestares, R. Navarro, J. Portilla, A. Tabernero, Efficient spatial-domain implementation of a multiscale image representation based on Gabor functions, J. Elect. Imag. 7 (1) (1998) 166–173.
[33] T. Huang, W. Bumett, G. Deczky, The importance of phase in image processing, IEEE Trans. Acoust., Speech, Signal Process. 23 (6) (1975) 529–542.
[34] A. Oppenheim, J. Lim, The importance of phase in signals, Proc. IEEE 69 (1981) 529–541.
[35] J. Behar, M. Porat, Y. Zeevi, The importance of localized phase in vision and image representation, in: Proceedings of SPIE 1001, Visual Communications and Image Processing, 1988, pp. 61–68.
[36] M. Morrone, D. Burr, Feature detection in human vision: a phase-dependent energy model, Proc. Roy. Soc. Lond. B 235 (1988) 221–245.
[37] D. Fleet, A. Jepson, Stability of phase information, IEEE Trans. Pattern Anal. Mach. Intell. 15 (12) (1993) 1253–1268.
[38] L. Haglund, D. Fleet, Stable estimation of image orientation, in: Proceedings of IEEE-ICIP'94, Austin, Texas, 1994, pp. 68–72.
[39] D. Fleet, Disparity from local weighted phase-correlation, in: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, vol. 1, 1994, pp. 48–54.
[40] P. Kovesi, Invariant measures of image features from phase information, Ph.D. thesis, The University of Western Australia, 1996.
[41] A. Cozzi, B. Crespi, F. Valentinotti, F. Woergoetter, Performance of phase-based algorithms for disparity estimation, Mach. Vision Appl. 9 (5–6) (1997) 334–340.
[42] P. Kovesi, Phase congruency: a low-level image invariant, Psychol. Res. 64 (2000) 136–148.
[43] G. Carneiro, A. Jepson, Multi-scale phase-based local features, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03), vol. 1, 2003, pp. 736–743.
[44] A. Ogale, Y. Aloimonos, A roadmap to the integration of early visual modules, Int. J. Comput. Vision 72 (2007) 9–25.
[45] T. Sanger, Stereo disparity computation using Gabor filters, Biol. Cybern. 59 (1988) 405–418.
[46] D. Fleet, A. Jepson, M. Jenkin, Phase-based disparity measurement, CVGIP: Image Understan. 53 (2) (1991) 198–210.
[47] W.M. Theimer, H.A. Mallot, Phase-based binocular vergence control and depth reconstruction using active vision, CVGIP: Image Understan. 60 (3) (1994) 343–358.
[48] P. Kovesi, Image features from phase congruency, Videre: J. Comput. Vision Res., vol. 1, MIT Press, 1999. pp. 1–26.
[49] M. Felsberg, G. Sommer, Image features based on a new approach to 2d rotation invariant quadrature filters, in: A. Heyden, G. Sparr, M. Nielsen, P. Johansen (Eds.), Computer Vision – ECCV 2002, Lecture Notes in Computer Science, vol. 2350, Springer, 2002, pp. 369–383.
[50] M. Felsberg, Disparity from monogenic phase,, in: L.V. Gool (Ed.), 24. DAGM Symposium Mustererkennung, Zürich, Lecture Notes in Computer Science, vol. 2449, Springer, Heidelberg, 2002, pp. 248–256.
[51] T. Gautama, M. Van Hulle, A phase-based approach to the estimation of the optical flow field using spatial filtering, IEEE Trans. Neural Networks 13 (5) (2002) 1127–1136.
[52] S. Sabatini, F. Solari, P. Cavalleri, G. Bisio, Phase-based binocular perception of motion in depth: Cortical-like operators and analog VLSI architectures, EURASIP J. Appl. Signal Process. 7 (2003) 690–702.
[53] H. Foroosh, W. Hoge, Motion information in the phase domain, in: M. Shah, R. Kumar (Eds.), Video Registration, Kluwer Academic Publishers, 2003, pp. 36–71.
[54] M. Felsberg, Optical flow estimation from monogenic phase, in: Proceedings of 1st International Workshop on Complex Motion, Günzburg, October 12–14, 2004.
[55] D. Boukerroui, J.A. Noble, M. Brady, On the choice of band-pass quadrature filters, J. Math. Imag. Vison 21 (1) (2004) 53–80.
[56] Z. Xiao, Z. Hou, C. Miao, J. Wang, Using phase information for symmetry detection, Pattern Recogn. Lett. 26 (13) (2005) 1985–1994.
[57] K. Pauwels, M. Van Hulle, Optic flow from unstable sequences containing unconstrained scenes through local velocity constancy maximization, in: Proceedings of British Machine Vision Conference, Edinburgh, 4–7 September, 2006.
[58] A. Darabiha, W. MacLean, J. Rose, Reconfigurable hardware implementation of a phase-correlation stereoalgorithm, Mach. Vision Appl. 17 (2) (2006) 116–132.
[59] I. Ulusoy, E. Hancock, A statistical approach to sparse multi-scale phase-based stereo, Pattern Recogn. 40 (2007) 2504–2520.
[60] L. Zang, D. Wietzke, C. Schmaltz, G. Sommer, Dense optical flow estimation from the monogenic curvature tensor, in: Scale Space and Variational Methods in Computer Vision, Lecture Notes in Computer Science, vol. 4485, Springer, 2007, pp. 239–250.
[61] R. Li, S. Sclarof, Multi-scale 3D scene flow from binocular stereo sequences, Comput. Vision Image Understan. 110 (2008) 75–90.
[62] J. Monaco, A. Bovik, L. Cormack, Stereoscopic phase-differencing: multiscale synthesis, in: Proceedings of IEEE Southwest Symposium on Image Analysis and Interpretation, 2008, pp. 33–36.
[63] X. Yang, Y. Zhou, T. Zhang, E. Zheng, J. Yang, Gabor phase based gait recognition, Electron. Lett. 44 (10) (2008) 620–621.
[64] J. Diaz, E. Ros, R. Carrillo, A. Prieto, Real-time system for high-image-resolution disparity, IEEE Trans. Image Process. 16 (1) (2007) 280–285.
[65] N. Krüger, M. Felsberg, A continuous formulation of intrinsic dimension, in: Proceedings of British Machine Vision Conference, 2003.
[66] S. Venkatesh, R. Owens, An energy feature detection scheme, in: Proceedings of International Conference on Image Processing, 1989, pp. 553–557.
[67] A. Jenkin, M. Jepson, The measurement of binocular disparity, in: Z. Pylyshyn (Ed.), Computational Processes in Human Vision, Ablex Publ., New Jersey, 1988.
[68] F. Solari, S. Sabatini, G. Bisio, Fast technique for phase-based disparity estimation with no explicit calculation of phase, Elect. Lett. 37 (23) (2001) 1382–1383.
[69] A. Jepson, M. Jenkin, The fast computation of disparity from phase differences, in: Procedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'89), 1989, pp. 398–403.
[70] C.-J. Westelius, Preattentive gaze control for robot vision, Lic. Thesis LiU-Tek-Lic-1992:14, ISY, Linköping University, SE-581 83 Linköping, Sweden, Thesis No. 322, ISBN 91-7870-961-X (June 1992).
[71] T. Lindeberg, Scale-Space Theory in Computer Vision, Kluwer Academic Publishers., 1994.
[72] D. Marr, Vision, Freeman, New York, 1982.
[73] L. Haglund, Adaptive multidimensional filtering, Tech. rep., Linköing University, Sweden, PhD Dissertation No. 284 1992.
[74] G. Granlund, H. Knutsson, Signal Processing for Computer Vision, Kluwer Academic Publishers, 1995.
[75] J. Bergen, P. Anandan, K. Hanna, R. Hingorani, Hierarchical model-based motion estimation, in: Proceedings of ECCV'92, 1992, pp. 237–252.
[76] K. Pauwels, M. Van Hulle, Optic flow from unstable sequences through local velocity constancy maximization, Image Vision Comput. 27 (5) (2009) 579–587.
[77] M. Felsberg, Optical flow estimation from monogenic phase, in: B. Jähne, R. Mester, E. Barth, H. Scharr (Eds.), 1st International Workshop on Complex Motion (ICM04), LNCS, vol. 3417, 2007, pp. 1–13.
[78] B. Horn, B. Schunck, Determining optical flow, Artif. Intell. 17 (1981) 185–203.
[79] M. Tistarelli, Multiple constraints for optical flow, in: J.-O. Eklundh (Ed.), Lecture Notes in Computer Science, vol. 801, Springer, 1994, pp. 61–70.
[80] A. Bruhn, Variational optic flow computation – accurate modelling and efficient numerics, Ph.D. thesis, Fakultät für Mathematik und Informatik, Universität des Saarlandes, Saarbrücken, 2006.
[81] T. Nir, A.M. Bruckstein, R. Kimmel, Over-parameterized variational optical flow, Int. J. Comput. Vision 76 (2) (2008) 205–216.

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008

[82] P.-E. Forssén, G. Granlund, Sparse feature maps in a scale hierarchy, in: AFPAC, Algebraic Frames for the Perception Action Cycle, Springer Verlag, 2000, pp. 186–196.
[83] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, Int. J. Comput. Vision 47 (1–3) (2002) 7–42.
[84] D. Scharstein, R. Szeliski, High-accuracy stereo depth maps using structured light, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03), vol. 1, Madison, WI, 2003, pp. 195–202.
[85] G. Egnal, R. Wildes, Detecting binocular half-occlusions: empirical comparisons of five approaches, IEEE Trans. Pattern Anal. Mach. Intell. 24 (8) (2002) 1127–1133.
[86] J. Barron, D. Fleet, S. Beauchemin, Performance of optical flow techniques, Int. J. Comput. Vision 12 (1994) 43–77.
[87] E. Lindholm, J. Nickolls, J. Montrym, NVDIA Tesla: a unified graphics and computing architecture, IEEE Micro 28 (2008) 39–55.
[88] C. Zetzsche, E. Barth, Fundamental limits of linear filter in the visual processing of two-dimensional signals, Vision Res. 30 (7) (1990) 1111–1117.
[89] J. Bigun, Vision with Directions: A Systematic Introduction to Image Processing and Computer Vision, Springer-Verlag, Berlin Heidelberg, 2006.
[90] F. Wörgötter, N. Krüger, N. Pugeault, D. Calow, M. Lappe, K. Pauwels, M.V. Hulle, S. Tan, A. Johnston, Early cognitive vision: using Gestalt-laws for task-dependent, active image-processing, Nat. Comput. 3 (3) (2004) 293–321.
[91] G. Granlund, The complexity of vision, Signal Process. 74 (1999) 101–126.
[92] P. König, N. Krüger, Perspectives: symbols as self-emergent entities in an optimization process of feature extraction and predictions, Biol. Cybern. 94 (4) (2006) 325–334.
[93] M. Felsberg, N. Krüger, A probabilistic definition of intrinsic dimensionality for images, in: Proceedings of DAGM Symposium Mustererkennung, 2003.
[94] A. Felsberg, S. Kalkan, N. Krüger, Continuous dimensionality characterization of image structures, Image Vision Comput. 27 (6) (2009) 628–636.
[95] J. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, F. Nuflo, Modeling visual attention via selective tuning, Artif. Intell. 78 (1995) 507–547.
[96] R. Crist, W. Li, C. Gilbert, Learning to see: experience and attention in primary visual cortex, Nat. Neurosci. 4 (2001) 519–525.
[97] J. Harris, S.N. Watamaniuk, Speed discrimination of motion-in-depth using binocular cues, Vision Res. 35 (7) (1995) 885–896.
[98] J. Tsotsos, A 'complexity level' analysis of vision, in: Proceedings of ICCV'87 – Human and Machine Vision Workshop, London, 8–11 June, 1987, pp. 346–355.
[99] M. Felsberg, Personal Communication, 2008.

Please cite this article in press as: S.P. Sabatini et al., A compact harmonic code for early vision based on anisotropic frequency channels, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.03.008