International Journal of Humanoid Robotics

Visual Primitives: Local, Condensed, Semantically Rich Visual Descriptors and their Applications in Robotics

NICOLAS PUGEAULT*, FLORENTIN WÖRGÖTTER† and NORBERT KRÜGER‡

*Centre for Vision, Speech and Signal Processing,
University of Surrey, GU2 7XH Surrey (UK)
n.pugeault@surrey.ac.uk

†Computational Neurosciences Group
Georg-August-Universität Göttingen, 37083 Göttingen (Germany)
worgott@chaos.gwdg.de

‡The Maersk Mc-Kinney Moller Institute,
University of Southern Denmark, DK-5230 Odense M (Denmark)
norbert@mmmi.sdu.dk

We present a novel representation of visual information, based on local symbolic descriptors that we call visual primitives. These primitives: (1) combine different visual modalities, (2) associate semantic to local scene information, and (3) reduce the bandwidth while increasing the predictability of the information exchanged across the system. This representation leads to the concept of Early Cognitive Vision, that we define as an intermediate level between dense, signal based Early Vision and high level Cognitive Vision. The framework's potential is demonstrated in several applications, in particular in the area of robotics and humanoid robotics, which are briefly outlined.

Keywords: Early Cognitive Vision, Scene Representation, Vision based Robotics

1. Introduction

Visual perception aims at gathering information about an agent's surrounding, allowing the agent to plan, navigate, and interact with its environment. Three of the difficulties faced by visual perception are the large amount of noise in images, the endemic ambiguity of local information, and the weak semantic content carried by pixels. The extraction of more meaningful local features such as edges, surfaces, corners, and textures are processes subject to errors due to noise. Moreover, pixel information is only remotely related to geometric and other semantic properties of the scene.

Electronic version of an article published as *International Journal of Humanoid Robotics, 7(3),* 2010, pp. 379–405, DOI:10.1142/S0219843610002209 2010 World Scientific Publishing Company.
http://www.worldscinet.com/ijhr/07/0703/S0219843610002209.html

This article describes a visual scene representation based on a set of visual descriptors, hereafter called *visual primitives*. These primitives describe edge structures by means of a number of properties that are relevant for edges only. As a consequence, their extraction requires a prior step to distinguish edge structures from junctions and texture. Besides giving such description, they have been used to formalize different contexts in visual scenes—in particular, 6D motion and 3D spatial context, as outlined in section 6. Meanwhile, this descriptor has been used in a number of applications such as the learning of object representations [1], pose estimation [2], motion estimation [3], and vision based grasping [4]. In these applications, we observed the importance of three properties the edge descriptor fulfills: explicitness, orthogonality, and condensation:

Explicitness: The primitives express explicitly important structural properties of edges such as local orientation, phase, color, and motion; this information is encoded in a multi-dimensional feature vector, where geometric and appearance cues are separated (see discussion under 'orthogonality' below). The feature vector adapts structural parameters according to the presence of line or step edges and its position according to both local structure and neighboring descriptors. Stereopsis can be used to reconstruct the 3D equivalent of image primitives, encoded as a new feature vector that preserves the explicitness of its components. Using classical stereopsis and projective relations [5], it is possible to transfer the primitives from a 2D image plane representation to the 3D space, and conversely ; this allows the formalization of relations between primitives either in 2D or 3D, indifferently. In contrast, most popular feature descriptors have feature vectors that cannot be transferred so easily to 3D—apart from the local position. Note that although previous works reconstructed 3D curves [6,7] and curved surfaces [8,9], the present work additionally encodes appearance information (color and phase) in the reconstructed 3D entity. This explicitness of the primitives' encoding, the duality between 2D and 3D primitives and the relations defined between them endows the primitives with their rich semantic content. In this sense, when we mention 'semantics' in this article, we refer to the fact that the feature's properties can be directly related to meaningful image (or scene) events (e.g., local 2D and 3D orientation).

Orthogonality: The explicitness property allows primitives to separate visual information gathered from their support region by splitting *geometric* and *appearance* information. Geometric information covers position and orientation; appearance information covers phase and color—see also *split of identity* [10]. Geometric information varies with viewpoint change, but in a manner that is fully described by the knowledge of 3D motion. For example, in the simplified case of rigid objects the knowledge of an object's motion (that can be computed from correspondences of primitives) allows for fully predicting the change in geometric information of the primitives that describe this object. Appearance information, on the other hand, is robust to viewpoint changes, although still affected by lighting and shadows. Therefore, appearance information is mainly used for matching purposes, while geometric

information is used to estimate the actual motion underlying the viewpoint change.

Condensation: Because of the aperture problem, edges cannot—unlike points—be located uniquely: image contours form ridges in potential, every point on which is a valid location for the edge descriptor. This could potentially lead to a very large number of features, with a very high level of redundancy. In order to obtain a sparse yet complete representation of image contours, we extract primitives sparsely along contours, using a local competition mechanism to inhibit candidates that are too close to each other. The radius of this inhibition is chosen to be commensurate with the extent of the filtering operation used to extract the edge pixels, therefore ensuring the completeness of the representation.

The proposed approach is designed to fulfill the three properties of explicitness, orthogonality and condensation discussed above, leading to several differences to previous works. First, contours are modeled in 3D from the early stages on, allowing an intrinsic robustness to viewpoint changes. This is advantageous as most rigid motions in space lead to contour deformation when projected in images. Considering edges in both image and space domains allows to formulate assertions in the most convenient domain. For example, perceptual grouping is better addressed in images, whereas motion is better handled in space. A second difference is the handling of additional appearance properties, besides geometric information. The property of orthogonality ensures that all geometric information is encoded in position and orientation, whereas appearance cues are invariant to geometric transforms such as viewpoint changes.

A large amount of evidence suggests that the human visual system processes a number of aspects of visual data in its first cortical stages [11,12]; these aspects, in the following called visual modalities[a], cover, e.g., local orientation [14], color [14], junction structures [15], stereopsis [16] and optic flow [14]. In our representation, we bundle the different modalities in one visual descriptor which can be interpreted as a functional abstraction of a specific repetitive pattern at the first stage of cortical visual called *hyper-column* [11]—this biological analogy is discussed in an article by Krueger et al [17].

Information about these different modalities can be extracted from images by applying a variety of linear and non-linear local filtering operations [18]. In a stage that we will call *Early Vision* [19], computer vision has dealt to a large extent with these modalities separately and in many computer vision systems, one or more of the above-mentioned aspects are processed [19,20,21]. Aloimonos and Shulman [22] argued that such modules should be integrated and stressed the importance of such inter-module integration and feedback mechanisms for mutual disambiguation.

[a] We would like to stress that, in this work, the term 'modality' refers to different *visual* modalities such as motion, orientation, color etc. We are aware that in the literature the term modality has been used with two different meanings: it has been applied to distinguish between different visual modalities [13] as well as different sensory modalities. In this article, the term 'multi-modal' is meant to indicate different visual modalities and not *sensory* modalities.

In contrast to the local, pixel-wise information that suffers from a large amount of noise and ambiguity, higher cognitive functions (such as reasoning and planning) require a sparse, symbolic, temporally integrated and robust representation of knowledge; this stage is called *Cognitive Vision*[23]. Extracting such a representation directly from images or local filter responses is difficult, and is prone to failure due to the ambiguity and noise that is endemic to early vision. It has been discussed that such a representation can be obtained by a hierarchical structure, the layers of which are representations of increasing abstraction[24,22]. Recent work has shown that such a feature hierarchy can be learned directly from visual data[25,26,27,28]. In this article we therefore bypass a long and data intensive learning stage by devising the intermediate features directly from properties of the local signal. In a similar line of thoughts, this paper presents a concrete realization of such an intermediate layer of representation, fitting between the ambiguities of early vision and the demanding requirements of cognitive vision, which we call *Early Cognitive Vision*[29]. This level of representation makes use of an early-symbolic representation to disambiguate visual information and provides a suitable substrate for cognitive vision tasks as exemplified in section 6.

A large amount of evidence suggests that the human visual system processes a number of aspects of visual data in its first cortical stages[11,12]; these aspects, in the following called visual modalities[b], cover, e.g., local orientation[14], color[14], junction structures[15], stereopsis[16] and optic flow[14]. In our representation, we bundle the different modalities in one visual descriptor which can be interpreted as a functional abstraction of a specific repetitive pattern at the first stage of cortical visual called *hyper-column*[11]—this biological analogy is discussed in an article by Krueger et al[17].

Information about these different modalities can be extracted from images by applying a variety of linear and non-linear local filtering operations[18]. In a stage that we will call *Early Vision*[19], computer vision has dealt to a large extent with these modalities separately and in many computer vision systems, one or more of the above-mentioned aspects are processed[19,20,21]. Aloimonos and Shulman[22] argued that such modules should be integrated and stressed the importance of such inter-module integration and feedback mechanisms for mutual disambiguation.

In contrast to the local, pixel-wise information that suffers from a large amount of noise and ambiguity, higher cognitive functions (such as reasoning and planning) require a sparse, symbolic, temporally integrated and robust representation of knowledge; this stage is called *Cognitive Vision*[23]. Extracting such a representation directly from images or local filter responses is difficult, and is prone to failure

[b] We would like to stress that, in this work, the term 'modality' refers to different *visual* modalities such as motion, orientation, color etc. We are aware that in the literature the term modality has been used with two different meanings: it has been applied to distinguish between different visual modalities[13] as well as different sensory modalities. In this article, the term 'multi-modal' is meant to indicate different visual modalities and not *sensory* modalities.

due to the ambiguity and noise that is endemic to early vision. It has been discussed that such a representation can be obtained by a hierarchical structure, the layers of which are representations of increasing abstraction[24,22]. In a similar line of thoughts, this paper presents a concrete realization of such an intermediate layer of representation, fitting between the ambiguities of early vision and the demanding requirements of cognitive vision, which we call *Early Cognitive Vision*[29]. This level of representation makes use of an early-symbolic representation to disambiguate visual information and provides a suitable substrate for cognitive vision tasks as exemplified in section 6.

Recent research has focused towards features that could be matched with better reliability and exhibit more general invariance properties, culminating in affine invariant descriptors—some examples of which are[30,31,32,33]. One needs to make the distinction between the two mechanisms that when combined, form a feature extraction process: First, an interest point detector is designed for finding a list of locations in the image that satisfy certain properties. Because of the need for accuracy and reliability in matching, modern detectors generally define interest points at locations that can be consistently extracted and matched. Common examples are Harris corners[34] and its affine version[35]. Second, a region descriptor is charged with encoding local information in a vector, for comparison and retrieval. Several reviews of interest point detectors and feature descriptors are available, notably Schmid et al[36] compared interest points detectors while Mikolajczyk & Schmid[37] and Moreels & Perona[38] compared region descriptors.

Although very successful for match-intensive tasks, such as the ones discussed above, this paradigm on feature descriptors can be complemented in certain respects, that prove to be important for the analysis of visual scenes. First, most visual scenes are dominated by edges which, in addition to being a candidate for finding correspondences, also express important structural properties which have been utilized for a variety of visual tasks. Further, local features can also be seen as initiators of the symbolic description of objects' shape, useful for generating generic pattern of interaction with objects (e.g., filling recipients, grasping planes, etc.); edges appear to be especially well suited for such a description[39]. For example, we can easily recognize objects and their potential use from line drawings[40]. In addition, edges also provide important information for actions such as grasping[41]. Finally, our visual descriptors are embedded in contextual relations that are used to disambiguate the extracted information being subject to noise as well as to link the visual information to actions (see section 6); these relations make use of the explicit semantics expressed in the descriptors.

Fig. 1 gives an overview of the framework presented in this paper: At the bottom of the figure, the system receives a stereo-pair of images obtained from a precalibrated stereo rig. The left and right images are then processed independently in the Early Vision layer (Fig. 1A): First, linear filtering operations are applied to the images, then combined (Fig. 1A-i) to extract the visual modalities: magnitude, orientation, phase, color, and optical flow. Each pixel represents the local flow at this

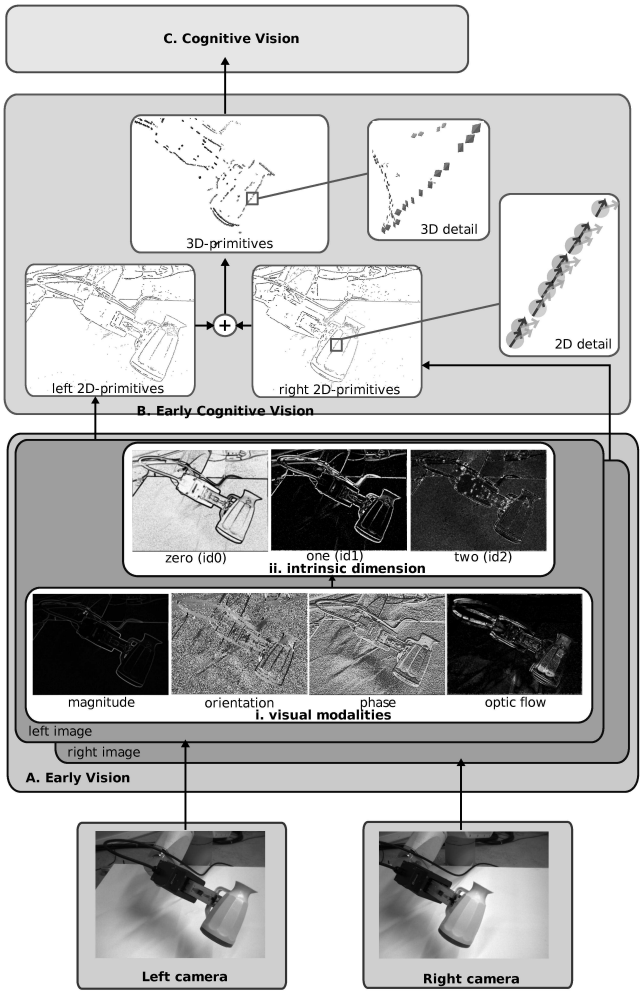
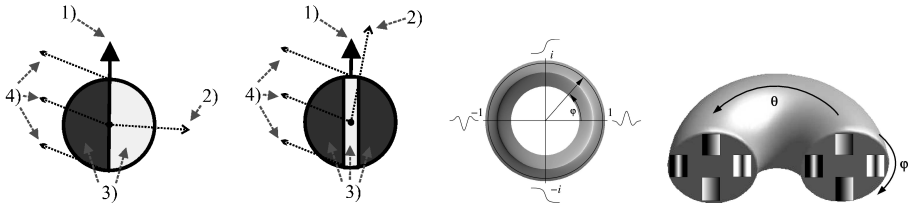


Fig. 1. Overview of the primitive extraction scheme.

location by its color: the hue indicates the orientation of the flow vector and the intensity, the magnitude of the flow (where black stands for zero flow). Then, the local signal is classified using a measure called *intrinsic dimension* (see 2.2), computed for each pixel in the image the confidence that it is an homogeneous surface (id0), an edge (id1), or a junction (id2). This is shown in Fig. 1A-ii, where white represents stronger confidence. In the next layer of the framework (Fig. 1B), coined Early Cognitive Vision in this paper, pixel-wise information provided by early vision is combined in a sparse, condensed set of feature vectors called *2D-primitives*—see section 3 and Fig. 2a,b. These primitives are then matched across the two stereo-views and correspondences allow for the reconstruction of *3D-primitives* that extend the primitive representation into 3D space—see section 4. This representation is then



(a) An edge primitive (b) A line primitive (c) phase (d) ori./phase space

Fig. 2. Illustration of the symbolic representation of (a) an edge primitive and (b) a line primitive, where 1) represents the orientation of the primitive, 2) the phase, 3) the color and 4) the optic flow. (c) The phase continuum (Figure courtesy of M. Felsberg[42]). (d) The torus topology of the orientation/phase space.

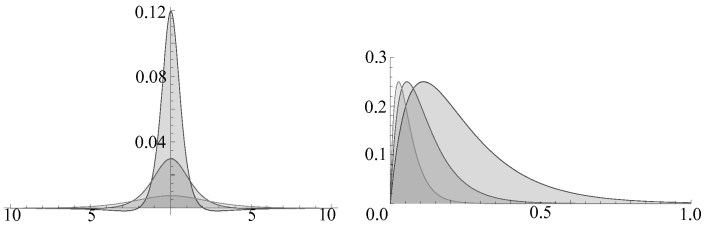
provided to higher level, Cognitive Vision processes (Fig. 1C).

An early version of the visual primitives was first introduced by Krüger & al[17], where 2D primitives were discussed as an analogy to cortical hypercolumns. This article did not dwell on the actual computation of the primitives, however, and only described a very early version of the primitives. Considerable work has since been invested in realizing these primitives towards computer vision and robotics applications, involving a parameter analysis that allowed to derive all parameters from the filters' bandwidth, and towards an extension of the primitives' semantic into 3D space. This lead to profound changes in the design of the primitives descriptors and the primitives extraction process. Therefore, the present article is the first detailed description of the visual primitives and the practical and theoretical issues involved in their computation.

In section 5, different relations between reconstructed 3D-primitives are discussed. Finally, section 6 presents applications of this framework to different vision and robotic problems and discusses primitives in the wider scope of a full fledged cognitive vision system.

2. Analysis of the Local Signal Structure

In a first stage, referred to as Early Vision, the image is processed using a collection of local, both linear and non-linear filtering operations. It has been shown that such local filtering operations can extract relevant structural information from the image[19,43]. Different filtering operations provide information on distinct aspects of the image structure, aspects that we will call *visual modalities*. In this work, we chose local edge structure as it encodes most relevant image information[39] and behaves well under scale changes[44] (section 2.1). Edge information can be locally described by the signal's magnitude, phase, and orientation, plus the local color information and the optical flow. Section 2.1 describes the low level filtering operations we use; section 2.2 describes how different kinds of local image structures are detected.



(a) Spatial response (b) Frequency response

Fig. 3. A 1D slice of the difference of Poissons (DOP) bandpass filter impulse responses for the three scales considered in this article—i.e., for $s = 1$, $s = 2$, and $s = 3$. Note that the filter is isotropic.

2.1. *Local edge structure: orientation and phase*

The edge structure of an image is characterized by local contrast and can be efficiently extracted using local filtering, as in [45,46]. There is a large amount of evidence that the human visual system, in its early stages, processes visual stimuli in a similar manner [47]. This work makes use of a local filtering operation called monogenic signal [10]. This operation is a rotationally invariant quadrature filter composed of a triplet of filters: a radial bandpass filter constructed from a difference of Poisson (DOP), and its two Riesz transforms. The impulse response of the bandpass filter, in spatial (S_e in Eq. (1) and Fig. 3a) and frequency (F_e in Eq. (2) and Fig. 3b) domains is given by:

$$S_\mathrm{e}(\boldsymbol{x}, s) = \frac{s}{2\pi(|\boldsymbol{x}|^2 + s^2)^{\frac{3}{2}}} - \frac{2s}{2\pi(|\boldsymbol{x}|^2 + 4s^2)^{\frac{3}{2}}} \tag{1}$$

$$F_\mathrm{e}(\boldsymbol{u}, s) = \exp(-2\pi|\boldsymbol{u}|s) - \exp(-4\pi|\boldsymbol{u}|s) \ . \tag{2}$$

where $\boldsymbol{x} = (x_1, x_2)^\top$ is the position, $\boldsymbol{u} = (u_1, u_2)^\top$ is the frequency, and s is the scale.

The split of identity (i.e., the separation of the signal into local amplitude, orientation and phase) is obtained by switching to appropriate polar coordinates—we refer to Felsberg and Sommer [10] for a discussion. This filter provides a qualitatively comparable response (for our purpose) as, e.g., Gabor wavelets, with fewer filtering operations: the filter is rotationally invariant thus there is no need to sample over all orientations. For a comparative study with other harmonic filtering alternatives, we refer the interested reader to the review by Sabatini et al [48].

The orientation θ encodes a local estimate of the edge orientation at this point. Because there is no unambiguous way to orient edges at this stage, local orientation is between $\theta \in [0, \pi)$. The phase φ encodes the local intensity transition across the edge into a continuum between $\varphi \in [-\pi, +\pi)$ in a compact way, using one parameter only (see Fig. 2c) [49]. For example, a pixel positioned on a bright line over a dark background has a phase of 0; a dark line on a bright background, a phase of π; a dark/bright edge, a phase of $\pi/2$; and a bright/dark edge, a phase of $-\pi/2$. Note that phase is 2π-periodic and continuous such that a phase of $-\pi$ represents the

same contrast transition as a phase of π. Also, the sign of the phase depends on the orientation interpretation; therefore local phase and orientation form a continuous space within $[0, \pi) \times [-\pi, +\pi)$, with a half-torus topology connected as illustrated in Fig. 2d. We compute filter responses for three different scales (for $s = 1$, $s = 2$ and $s = 4$).[c]

2.2. *Intrinsic dimension and symbolic interpretation*

Different kinds of image structures coexist in natural images: homogeneous image patches, edges, corners, and textures. Furthermore, certain concepts are only meaningful for specific classes of image structures. For example, the concept of orientation is well defined for edges or lines but not for junctions, homogeneous image patches, or most textures. Therefore, before associating concepts such as orientation or position, we need to classify image patches according to their junction-ness, edge-ness or homogeneous-ness. The intrinsic dimension[50] is a suitable classifier in this context. Ideal homogeneous image patches have an intrinsic dimension of zero (id0), ideal edges are intrinsically one-dimensional (id1) while junctions and most textures have an intrinsic dimension of two (id2). Going beyond common discrete classification[50], we use a *continuous* formulation[51] that, based in a compact triangular representation, provides three confidences that express the likelihood of an image patch being either id0, id1, or id2. For a detailed discussion of the relation of the concept of intrinsic dimension to other approaches, we refer to Felsberg et al[51].

3. Semantic Representation of Local Information

The abstraction of the pixel-wise information (as described in section 2) into a sparse set of primitives is done in three stages: First, in a *sampling* stage, interest points in the image are extracted with sub-pixel accuracy according to the signal's magnitude (section 3.2). Second, in an *elimination* stage, interest points that are too close to each other (and therefore would lead to redundant descriptors) compete and the weaker one is disregarded (section 3.3). Finally, in the *abstraction* stage, the image's local structure at selected interest points become interpreted and multiple visual modalities become associated (section 3.4). Before coming to the sampling stage (section 3.2), we need to define the scaled dependent parameters that it uses (section 3.1).

3.1. *Scale dependent parameters*

An important aspect of the condensation scheme is that all parameters are derived from the monogenic signal's filter equations. In particular, two quantities are of importance for the sampling scheme: the *line/edge bifurcation distance* and the

[c] Note that step edges have high amplitudes across all scales, whilst line structures are represented as a line at coarse scales, and as two step-edges at fine scales[44]—see also section 3.

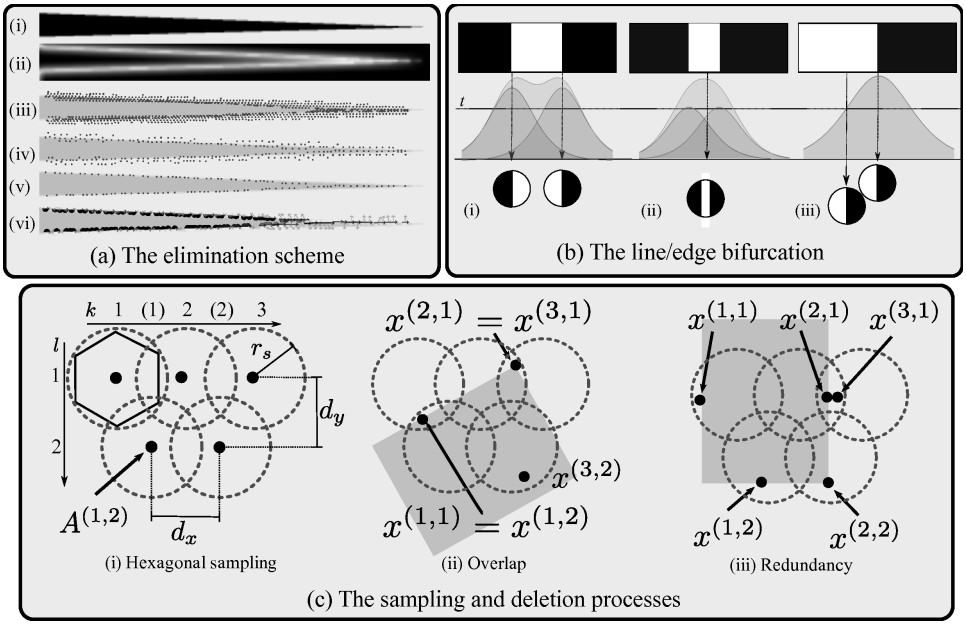


Fig. 4. (a) Illustration of the elimination scheme: (i) image; (ii) magnitude; interest points before elimination (iii) and after first (iv) then second (v) elimination; (vi) primitives. (b) Extraction of redundant primitives: (i) a double edge; (ii) a line primitive; (iii) redundant primitive. (c) Illustration of the sampling process: (i) hexagonal sampling; (ii) ambiguous cell association; (iii) redundant interest points.

filter extent. Both can be derived from the formulation of the DOP bandpass filter, the impulse response of which is given by Eq. (1) in the spatial domain, and by Eq. (2) in the frequency domain.

Three scales of processing are considered in this work, for $s = 1$, $s = 2$, and $s = 4$. The filter at the three scales have a peak response in the frequency domain, or *peak frequency*, that are computed using the root of the impulse response's derivative, to values of 0.110, 0.055, and 0.027, respectively—see Fig. 3(b).

Line/edge bifurcation distance (d_{leb}): The first filter property that is of relevance for sparse sampling is called *line/edge bifurcation distance*. It is defined for a given scale as the minimal distance between two edges for them to produce two distinct amplitude maxima in the filter response magnitude. Hence, a double edge will be represented by a pair of edge primitives if its width is larger than d_{leb}, and by only one line primitive otherwise. Fig. 4a-i shows a narrow triangle for which two edges get closer until they meet. Vertical sections of the local amplitude (Fig. 4a-ii) close to the vertex have only one maximum that splits into two distinct maxima further away from the vertex, where the distance between the two edges is larger—see also Fig. 4b.

The bandpass filter's total response for two edges separated by a distance a is

Table 1. The scale-dependent parameters of our representation.

Scale	s	1	2	4
Peak frequency	f_p	0.1103	0.0551	0.0275
Wavelength	λ	9.06	18.12	36.25
Line/edge bifurcation	d_{leb}	2(0.96)*	2(1.92)*	3.83
Influence radius	d_k	2.02663	4.05327	8.10653
Hexagonal spacing in x	d_x	2(1)*	2	4
Hexagonal spacing in y	d_y	2(1)*	2	3
Compression ratio	r_{co}	62% / **1.2%**	62% / **0.9%**	20% / **0.3%**

given by:

$$T_{\mathrm{e}}(x, a, s) = S_{\mathrm{e}}(x, s) + S_{\mathrm{e}}(x + a, s), \tag{3}$$

where s is the current scale. Each maximum of this function elicits a distinct candidate position. When the distance a is large, this function bears two distinct maxima, that become closer with diminishing distance a and finally become one when $a = d_{leb}$—see Fig. 4b. It follows that d_{leb} is equal to the distance a, for which the roots of the derivative $\frac{\partial T(x,a,s)}{\partial x}$ of equation (3) merge into a single one. This was numerically computed for all three scales and d_{leb} was found to grow linearly with scale $d_{leb} = \kappa s$, where $\kappa \backsimeq 0.95825$. Note that, because the magnitude is encoded in a discrete array (an image), the minimum distance between two maxima is always 2 pixels (or even $2\sqrt{2}$ along the diagonal). Therefore, d_{leb} is set to 2 for scales $s = 1$ and $s = 2$.

Influence radius d_k: The second quantity that is of relevance for sparse sampling is the spatial extent of the filter's impulse response. This value is estimated as the distance x for which the bandpass filter's spatial response $S(x, s)$ reaches zero. This value was computed for all three scales considered, and was found to also grow linearly with scale $d_k = \eta s$, where $\eta \backsimeq 2.02663$.

3.2. *Sampling*

The concept of position can vary depending on what kind of features we are trying to locate. For example, in homogeneous areas there is no single location that can be identified; on edges, the aperture problem prevents us from identifying a specific location along the edge. Only corners, junctions, and other structures that are intrinsically two dimensional can be located unambiguously in the image. The present work focuses on locating edge features.

To get sparse candidates for our primitives, we first perform a hexagonal sampling (see Fig. 4c-i) of the image into overlapping areas $A^{(k,l)}$ of radius r_s, with (k, l) coding the hexagonal grid points. Hexagonal sampling has a number of advantages[52,53]; one amongst them is that the distance between the centers of neighbor tiles is uniform in an hexagonal grid while in a rectangular grid diagonal spacing is $\sqrt{2}$ times longer than for horizontal or vertical spacing. Since we want to extract

symbolic descriptors for each tile, hexagonal sampling allows for a more evenly distributed symbolic description and reflects more closely the isotropic structure of the original image filters. The parameters d_x and $d_y = \frac{\sqrt{3}}{2} d_x$ determine the spatial distance in x and y between the center $A_c^{(k,l)}$ of the tile $A^{(k,l)}$ and the centers of the neighbor tiles.[d] The optimal sampling distance d_x is set to be equal to the line/edge bifurcation distance d_{leb} for this scale, and therefore $d_x = d_{leb}$. All scale dependent parameters are shown in Table 1. Note that, due to sampling, the line/edge bifurcation distance has a minimum value of 2 pixels; in cases where the theoretical value is lower, it is recorded in parentheses after the effective value, and denoted as $(x)^*$.

For each cell center $A_c^{(k,l)}$, a circular neighborhood of radius r_s is searched for interest points. The radius r_s is chosen so that the image is fully covered by all cells' neighborhoods. In a hexagonal grid, the maximum distance to a tile's border is $\frac{1}{\sqrt{3}} d_x$, hence we set $r_s = \frac{d_x}{\sqrt{3}}$.

For a line or edge, the position $\boldsymbol{x}_{id1}^{(k,l)}$ can be defined through energy maxima that are organized as a one-dimensional manifold; therefore, an equidistant sampling along these energy maxima is appropriate. For this, we look within the area $A^{(k,l)}$ for the energy maximum along a line orthogonal to the orientation at $A_c^{(k,l)}$:

$$\boldsymbol{x}_{id1}^{(k,l)} = (\hat{x}, \hat{y}) = \arg \max_{(x,y) \in L^{(k,l)}} m(x, y), \tag{4}$$

where $L^{(k,l)}$ is a local line going through $A_c^{(k,l)}$ with orientation perpendicular to $\theta(A_c^{(k,l)})$. Then, candidate positions are computed with sub-pixel accuracy. Fig. 4a-iii shows the interest points for the test image in Fig 4a-i.

3.3. *Elimination of redundant descriptors*

Since areas $A^{(k,l)}$ are overlapping, the process described above can lead to identical positions found in neighboring areas: in Fig. 4c-ii, the putative positions $\boldsymbol{x}^{(2,1)}$ and $\boldsymbol{x}^{(3,1)}$, elicited by two distinct hexagonal cells, represent the same image location. Moreover, the filter's spatial extension can lead to proximate positions describing essentially the same image structure: in Fig. 4c-iii, the putative position $\boldsymbol{x}^{(3,1)}$ is redundant because it describes the same structure as $\boldsymbol{x}^{(2,1)}$, and less accurately.

In order to eliminate these redundant descriptors, an additional selection process is needed. This process faces the following challenges:

- Proximate, yet distinct, putative positions should be preserved as, for example, in Fig. 4b-i where two edges converge.
- Distant, yet redundant, putative positions should be discarded. Due to the kernel's spatial extent, a given image structure will generate significant response within a radius d_k that is larger than d_{leb} (as in Fig. 4b-iii).

[d] Note that the odd rows have an onset of $d_x/2$

These problems are addressed in a two stages elimination process, described in sections 3.3.1 and 3.3.2.

3.3.1. *Elimination based on the line/edge bifurcation distance d_{leb}*

In the first elimination step, stronger interest points suppress weaker candidates within a radius of d_{leb}; because d_{leb} is the minimal distance between two distinct edges, any closer candidate is known to be redundant—as in Fig. 4b-ii. In practice, all candidates $\boldsymbol{x}^{(k,l)}$ become ordered according to the associated amplitude $m(\boldsymbol{x}^{(k,l)})$. Starting with the candidates with highest local amplitude, all other candidates $\boldsymbol{x}^{(k',l')}$ within a radius d_{leb} are discarded.[e] Since candidates are ordered according to the local amplitude, a candidate corresponding to a stronger structure suppresses candidates from a weaker structure. Thereby, all non-distinct edges (according to the line/edge bifurcation distance) become deleted while redundant edges are preserved. The result of the first elimination stage on the test image in Fig. 4a-i is shown in Fig. 4a-iv.

3.3.2. *Elimination based on the influence radius d_k*

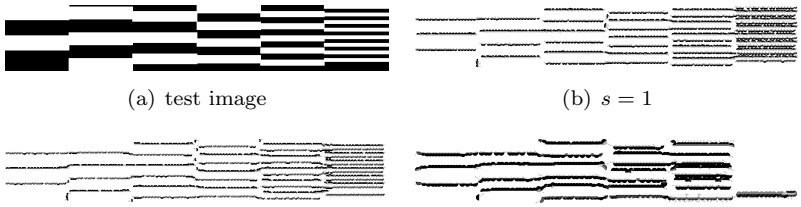
The local magnitude can be significantly affected by image structures within a radius d_k—as in Fig. 4b-iii. In the second elimination step, interest points that are not an amplitude maximum on a line orthogonal to the local orientation are suppressed. Each candidate position in a pair with distance smaller d_k, is tested whether it is an amplitude maximum, along a line orthogonal to the local orientation (see Fig. 4). This is achieved by comparing each candidate's amplitude to its direct neighbors, on both sides of the edge, as indicated by the local orientation.[f] Then, redundant structures, i.e., candidates that are not a local maximum, are discarded. The result of the second elimination stage on the test image in Fig. 4a-i is shown in Fig. 4a-v.

The effect of the double elimination process at different scales can be seen in Fig. 4a (for $s = 1$): Fig. 4a-i features a narrow triangle; from a signal perspective, this represents a double-edge narrowing until it becomes a line, and finally vanishes. Fig. 4a-ii illustrates the definition of the quantities d_{leb} and d_k. Fig. 4a-iii shows all candidate interest points. Figs. 4a-iv and 4a-v show the remaining interest points after the first and second elimination steps, respectively. Finally, Fig. 4a-vi show the primitives extracted (larger scales lead to an interpretation of the triangle as a line earlier towards the left of the triangle).

Fig. 5 shows the primitives extracted from an artificial test image, for different scales: The image in Fig. 5a shows vertically alternating black/white step-edges, getting narrower towards the right of the image; the primitives extracted at the three

[e] Note that for the quality of the process it is important that all positions are computed with sub-pixel accuracy already at this stage.

[f] Note that the criterion 'local maximum' that is applicable for id2 structures cannot be applied to edges, because edge-like structures form a ridge in the local amplitude surface (see Fig. 4a-ii).



(a) test image (b) $s = 1$

(c) $s = 2$ (d) $s = 4$

Fig. 5. Illustration of the primitives' sampling density: (a) shows an image with gradually (from left to right) narrower white and black bars; (b, c, and d) show the primitives extracted for different scales.

scales are shown in Fig. 5b, c, and d, respectively. Note that all of the narrower step edges (to the right of the image) are distinctly extracted at the finer scale ($s = 1$) in Fig. 5b; one is lost at a coarser scale ($s = 2$) in Fig. 5c; and for the coarsest scale ($s = 4$), in Fig. 5d, edges are not extracted anymore since their structure is not preserved at this scale.

3.4. *Association of visual attributes*

Once the redundant interest points have been discarded, other visual modalities are computed at the remaining positions $\boldsymbol{x}^i$, to form a feature vector. Orientation θ and phase φ are produced by the monogenic signal, as described in section 2.1. The local optic flow at pixel $\boldsymbol{x}$ is denoted by $\boldsymbol{f}(\boldsymbol{x})$ and computed using the well established Nagel–Enkelmann algorithm[54].

Sub-pixel interpolation of visual modalities: Since descriptors' positions are computed with sub-pixel accuracy, we can also interpolate sub-pixel values for orientation, phase, and optic flow using bi-linear interpolation. Let $\tilde{x}$ and $\tilde{y}$ be the positions computed with sub-pixel accuracy (see section 3.2); let $x_l = \lfloor \tilde{x} \rfloor$, $y_l = \lfloor \tilde{y} \rfloor$, $x_h = \lceil \tilde{x} \rceil$, and $y_h = \lceil \tilde{y} \rceil$ (where $\lfloor \tilde{x} \rfloor$ is the integral part of the real number x and $\lceil \tilde{x} \rceil = \lfloor \tilde{x} \rfloor + 1$); then the bi-linear interpolation computation leads to the formula:

$$\tilde{M}(\tilde{x}, \tilde{y}) = \sum_{\hat{x}} \sum_{\hat{y}} \sigma(\tilde{x}, \hat{x})\ \sigma(\tilde{y}, \hat{y})\ \hat{M}(\hat{x}, \hat{y}), \tag{5}$$

for $\hat{x} \in \{x_l, x_h\}$ and $\hat{y} \in \{y_l, y_h\}$, $\sigma(x, y) = 1 - |x - y|$, and $M \in \{\theta, \varphi, \boldsymbol{f}\}$ a visual modality for which we have discrete measurements at every pixel $\hat{M}(\hat{x}, \hat{y})$, and $\tilde{M}(\tilde{x}, \tilde{y})$ is then interpolation at sub-pixel location $(\tilde{x}, \tilde{y})$ for this modality. Note that, for the interpolation of orientation and phase, the specific topology of the orientation/phase space needs also to be taken into account. Hence, $\hat{\theta}$ is transformed such that the distance between all pairs of the set $\hat{\theta}(x_l, y_l)$, $\hat{\theta}(x_l, y_h)$, $\hat{\theta}(x_h, y_l)$, $\hat{\theta}(x_h, y_h)$ is smaller than $\frac{\pi}{2}$ and $\hat{\theta}(\tilde{x})$ is in $[0, \pi)$.

Color sampling: Although color information is available at each pixel position, it is heavily redundant, especially for edge and line structures. Moreover, an edge

indicates a separation between two areas with distinct properties; therefore, special care is required when associating color information to edge-like structures. First, because the primitives are the result of a local filtering operation, it is appropriate to sample color over an area of the image that is commensurate with the filter's extent. Second, in the case of a double edge, it is important that the color on the inside of the double-edge is sampled separately from the color on the outside (see Fig. 2). Because two edges have a minimal distance of d_{leb}, the color is sampled within a radius $r = d_{leb}$. Therefore, we sample the pixels within a neighborhood $N(\boldsymbol{x}, r)$, that contains all image pixels $\boldsymbol{y}$ such that $\boldsymbol{y} \in N(\boldsymbol{x}, r) \Leftrightarrow ||\boldsymbol{y} - \boldsymbol{x}|| < r$.

The color modality is encoded in two different ways depending on the phase value: for a step-edge structure ($\frac{\pi}{4} < |\varphi| < \frac{3\pi}{4}$), it is natural to distinguish between the color on each side of the edge $(\boldsymbol{c}_l, \boldsymbol{c}_r)$; for a line structure ($|\varphi| \leq \frac{\pi}{4}$ or $|\varphi| \geq \frac{3\pi}{4}$), the color of a middle strip $\boldsymbol{c}_m$ (i.e.on the actual line) is also encoded (see Figs. 2 and 4a-vi). It follows that pixels $\boldsymbol{y} \in N(\boldsymbol{x}, r)$ are binned into two (if the phase indicates an edge) or three (if the phase indicates a line) areas. If we consider a vector $\boldsymbol{n} = (\cos\theta, \sin\theta)^\top$, normal to the local orientation, the three binning areas are defined by:

$$\boldsymbol{y} \in \begin{cases} B_l(\boldsymbol{x}, r, \theta) & \text{if } (\boldsymbol{y} - \boldsymbol{x}) \cdot \boldsymbol{n} > +w, \\ B_r(\boldsymbol{x}, r, \theta) & \text{if } (\boldsymbol{y} - \boldsymbol{x}) \cdot \boldsymbol{n} < -w, \\ B_m(\boldsymbol{x}, r, \theta) & \text{else.} \end{cases} \tag{6}$$

where w is the width of the middle strip where the line color is sampled. For edges, color is encoded in only two vectors, one for each side, and $w = 0$; for lines, color is encoded in three vectors. Because the maximal width of a line to be encoded by a single primitive is d_{leb}, the line color is sampled on a middle strip of this width: $w = d_{leb}/2$. The red, green, and blue components are then averaged in each bin.

Feature vector: From all this, we obtain a parametric description of local image patches that we call a primitive $\boldsymbol{\pi}_i$. For a step-edge this representation is

$$\boldsymbol{\pi}_i = (\boldsymbol{x}_i, \theta(x_i), \varphi(x_i), (\boldsymbol{c}_l(x_i), \boldsymbol{c}_r(x_i)), \boldsymbol{f}(x_i)) \tag{7}$$

and for line structures

$$\boldsymbol{\pi}_i = (\boldsymbol{x}_i, \theta(x_i), \varphi(x_i), (\boldsymbol{c}_l(x_i), \boldsymbol{c}_m(x_i), \boldsymbol{c}_r(x_i)), \boldsymbol{f}(x_i))\,. \tag{8}$$

The primitives' parameters are explicit and the set of all primitives provides a condensed representation of the image. The condensation factor can be computed by the ratio of the number of bits required to encode the list of primitives compared to the number of bits taken by two RGB color images (to account for the temporal information recorded by the optical flow modality). The condensation ratio r_{co} is recorded in Table 1, with first the worst case condensation (assuming one primitive is extracted in each cell of the grid), and the effective rate on a real image in bold face. For the finest scale ($s = 1$) we obtain an effective condensation ratio of $r_{co} \simeq 1.2\%$ (and $r_{co} \simeq 62\%$ in the worst case).

4. Computation of 3D-Primitives

So far we have presented multi-modal image descriptors that code 2D information. However, these descriptors represent visual events occurring at a certain 3D position in space. This depth information is essential for higher level processes because of two reasons: First, humans and robots act in a 3D world where depth information is valuable for, e.g., navigation or grasping. Second, since many structural dependencies of visual events (e.g., rigid body motion) take place in 3D, depth information is essential for their formalization and for the disambiguation processes they underlie[55]. In the following, we describe an extension of the image primitives (2D-primitives) to spatial primitives (3D-primitives); thus, the semantic information coded in the image primitives is transferred into the 3D domain. It is well known that pairs of matching features between two stereo views can be used to reconstruct depth[5,56]. Given a pair of corresponding points between the left and right images, a meaningful 3D interpretation of this stereo-pair is a 3D point. The 2D-primitives presented herein, however, encode multi-modal information, and therefore a stereo-pair of matching 2D-primitives ought to reconstruct this multi-modal information in space: the resulting entity is called a *3D-primitive*.

4.1. *Stereo matching of 2D-primitives*

There exists a wealth of studies on stereo matching of image edge features: e.g., using the sign of the zero-crossings and their orientation[57,58], length and orientation of line segments[59], normalized cross-correlation between the pixels' surrounding lines (or curves)[6]. The present work uses of the multi-modal information carried by the 2D-primitives as a matching criterion and assumes a calibrated stereo set-up and known epipolar geometry to reduce the correspondence search to a line[5,56].

The similarity distance between two primitives $\boldsymbol{\pi}^l$ and $\boldsymbol{\pi}^r$ is given by:

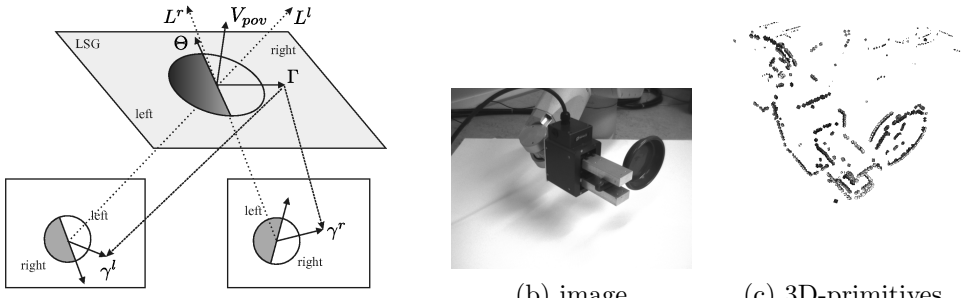
$$d_{\boldsymbol{\pi}}(\boldsymbol{\pi}^l, \boldsymbol{\pi}^r) = \boldsymbol{w} \cdot (d_\theta(\theta^l, \theta^r), d_\varphi(\varphi^l, \varphi^r), d_{\boldsymbol{c}}(\boldsymbol{c}^l, \boldsymbol{c}^r), d_{\boldsymbol{f}}(\boldsymbol{f}^l, \boldsymbol{f}^r))^\top, \tag{9}$$

where $\boldsymbol{w}$ is a vector containing the relative weight of each modality, with $\|\boldsymbol{w}\| = 1$. The actual distance measures $d_m(m^l, m^r)$ can be chosen for each visual modality $m \in \{\theta, \varphi, \boldsymbol{c}, \boldsymbol{f}\}$. For example, the color distance can be computed in different color spaces.

4.2. *Geometric information: pose reconstruction*

3D-primitives need to encode the reconstructed 3D orientation $\boldsymbol{\Theta}$ beside the 3D position $\boldsymbol{X}$. Considering a stereo pair of corresponding 2D-primitives $\boldsymbol{\pi}^l$ and $\boldsymbol{\pi}^r$, this orientation is computed as the intersection of two planes in space, each defined by the optical center of one camera and the line $L(\boldsymbol{\pi}^l, \boldsymbol{\pi}^r)$ in the image plane described by the 2D-primitives position ($\boldsymbol{x}^l$ and $\boldsymbol{x}^r$) and orientation (θ^l and θ^r) (see Fig. 6):

$$L(\boldsymbol{\pi}^l, \boldsymbol{\pi}^r) \equiv P^l(\boldsymbol{x}^l, \theta^l) \cap P^r(\boldsymbol{x}^r, \theta^r), \tag{10}$$



(a) stereo reconstruction (b) image (c) 3D-primitives

Fig. 6. Illustration of the reconstruction of a 3D-primitive from a stereo pair of 2D-primitives.

where $P^l(\boldsymbol{x}^l, \theta^l)$ is the 3D plane back-projected by the 2D line formed from a primitive with position $\boldsymbol{x}^l$ and orientation θ^l, in the left image. The intersection of these two planes in space is a 3D line $L(\boldsymbol{\pi}^l, \boldsymbol{\pi}^r)$ and the normalized orientation vector of this line defines the reconstructed 3D-primitive's orientation in space ($\boldsymbol{\Theta}$).

The 3D position $\boldsymbol{X}(\boldsymbol{\pi}^l, \boldsymbol{\pi}^r)$ is given by the intersection between both optical rays. Due to sampling error, those two rays will rarely intersect in space, and it is customary to reconstruct the point in space that minimizes the distance to the two back-projected rays[56]. In this case, this would generally lead to a point that does not lie on the reconstructed 3D line; therefore we will rather compute the intersection between the left back-projected ray and the right back-projected plane.

$$\boldsymbol{X}(\boldsymbol{\pi}^l, \boldsymbol{\pi}^r) \equiv L^l(\boldsymbol{x}^l) \cap P^r(\boldsymbol{x}^r, \theta^r) \tag{11}$$

where $L^l(\boldsymbol{x}^l)$ is the ray back-projected from the position $\boldsymbol{x}^l$ in the left image. This biases the reconstructed position towards the position in the left image but ensures consistency with the reconstructed line.

4.3. *Appearance information: reconstruction of phase and color*

Phase Φ and color $\boldsymbol{C}$ are reconstructed in space as the mean value between the two corresponding image primitives:

$$\Phi = \tfrac{1}{2}(\varphi^L + \varphi^R) \quad \text{and} \quad \boldsymbol{C} = \tfrac{1}{2}(\boldsymbol{c}^L + \boldsymbol{c}^R). \tag{12}$$

Note that, because 2D-primitives are matched across stereo according to multimodal similarity, these phase and color values will be very similar, justifying the use of the mean.

Because color and phase encode surface information (respectively contrast and color transition across an edge), we need to define a 3D surface patch onto which they apply. Unfortunately, it is not possible to reconstruct the exact surface from local information: for a pure edge, the surface on one side does not allow finding the additional correspondence that would be required for the reconstruction of a 3D surface. Moreover, in case of a depth discontinuity, the color information might

come from a 3D position that is completely independent from the 3D orientation information (e.g., an object in the background). Therefore, we define an *a priori* 3D surface (see Fig. 6a) using the 3D orientation of the primitive, and an additional *Local Surface Guess Vector* $\Gamma = \boldsymbol{\Theta} \times V_{pov}$ such that the surface is normal to a vector V_{pov} that encodes the observer's perspective on the 3D-primitive.

The vector V_{pov} is defined as $V_{pov} = \frac{1}{2}\left(L^l + L^r\right)$, where L^l and L^r are the two optical rays joining the location of the 3D-primitive $\boldsymbol{X}$ with the optical center of the left and right cameras. The vector Γ also identifies each side of the 3D line, which is critical for modalities like color and phase that describe the modality transition across the contour. Each side of the 3D-primitive is associated to the corresponding side of the 2D-primitives by back-projecting this vector on each image plane, as indicated by γ^l and γ^r in Fig. 6a. Note that this vector is merely used for display purposes; a more reliable estimate needs to be inferred at a later processing stage from global considerations such as an explicit description of the 3D contour, the surface or even the object it belongs to.

4.4. *3D-primitive feature vector*

In summary, the above sections have defined a scheme the reconstruction of spatial primitives $\boldsymbol{\Pi}^{(i,j)}$, with a parametric description:

$$\boldsymbol{\Pi}^{(i,j)} = (\boldsymbol{X}, \boldsymbol{\Theta}, \Phi, (\boldsymbol{C}_l, \boldsymbol{C}_m, \boldsymbol{C}_r)), \tag{13}$$

where the j index represents the alternative 3D entities generated from different stereo correspondences in the right image to the i^{th} primitive in the left image. Since a final decision about stereo match can usually not be made solely based on local information, multiple hypotheses can be kept at this stage; this allows disambiguation by later processes such as perceptual grouping[60,55], or motion estimation[1,55]. Fig. 6c shows the 3D primitives reconstructed from the image in Fig. 6b.

5. Relations and Operations Defined on Primitives

Because primitives are local *symbolic* descriptors, they carry additional semantic information, that defines the way a given primitive relates to its neighbors and what combinations of primitives are of relevance; such combinations can be expressed as a set of relations between pairs or groups of primitives. For example, a smooth planar surface will be described by several strings of collinear primitives that represent its contours; those primitives will generally be co-planar (see Fig. 7c,h) and isochromatic (see Fig. 7b,g). Hence, an important aspect of the primitive representation is that such second-order relations between primitives can be defined efficiently, while condensation keeps the relational space within a tractable size.

Good continuation (collinearity): We defined 2D-primitives as local edge descriptors, and therefore we expect them to lie on image contours; conversely, image contours are described by strings of primitives. In a study by Pugeault et

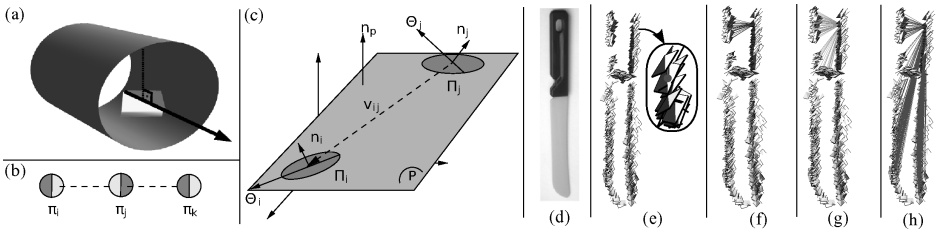


Fig. 7. Sample relations between primitives and their illustration on an example: (a) normal distance; (b) isochromacy; (c) coplanarity; (d-e) image and the 3D primitives of a sample object. Illustration of all primitives related to a selected one: (f) all primitives that has a normal distance of maximum 1.5 cm to the selected primitive; (g) all primitives that are isochromatic with the selected primitive; (h) all primitives that are coplanar with the selected primitive.

al[60], a measure of the likelihood that two 2D-primitives belong to the same image contour is defined using a mixture of Gestalt rules of proximity, good continuation and similarity.

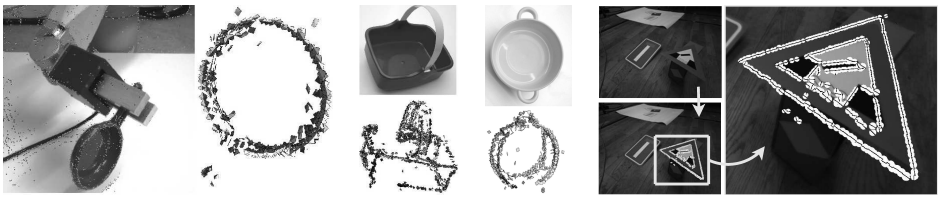
Co-planarity: 3D-primitives have a position and an orientation in space. Therefore, we can say that two primitives are coplanar if the second primitive's orientation $\boldsymbol{\Theta}_2$ lies in the plane formed by the two positions $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and the first orientation $\boldsymbol{\Theta}_1$ (see Fig. 7c,h).

Isochromacy: The relation of isochromacy expresses the similarity between two primitives' color. Assuming that the two primitives describe the same surface, we compare their color modality on the inner side, neglecting the color on the outer side (see Fig. 7b,g).

Normal distance: In addition to the Euclidean distance, the normal distance (i.e., the distance between the line that goes through one primitive to the other primitive) expresses a meaningful relation between parallel lines (see Fig. 7a,f).

Rigid body motion: One important type of motion, called *Rigid Body Motion* (RBM), describes the possible motions of rigid objects. For example, it can describe the motion of manipulated objects, or the camera itself (ego-motion). If the motion of an object is known, the future position and appearance of a 3D-primitive can be predicted from this motion: the change of position and orientation induced by a RBM ($\mathcal{M}(\Pi)$) can be computed analytically[5]; phase and color can be approximated to be constant.

The condensation of information allows to actually use such second-order relations between primitives which would be computationally intractable at the pixel level. Note that the relations defined above can be combined to form even stronger visual events, e.g., a combination of isochromacy with co-planarity allows us to infer probable surfaces in the image, and is used for generating grasping hypotheses in section 6.2.



(a) pan (b) basket (c) pot (d) pose

Fig. 8. Learning object models from manipulation.

6. Applications

The primitives have been used in a number of applications that require general 3D scene representations, both in computer vision [2,3,1] and in vision-based robotics [61,62], including the humanoid system ARMAR [63] (see also Fig. 9e). The broad applicability of this representation stems from the goal to develop a generic vision machinery, in analogy to the human visual system, providing a disambiguated, rich and explicit interpretation of visual scenes.

The computation of the 3D-primitives, including the extraction of 2D-primitives in two stereo images (512×512 pixels) as well as the stereo matching and the reconstruction, is currently performed with $\simeq$5Hz on a hybrid hardware architecture consisting of one GPU, on which the pixel-wise filtering processes are performed, and an 8 core machine, on which the higher level computation is done. A real-time implementation of this system is discussed by Jensen et al [64].

The following briefly discusses the applications in the robotic domain, stressing the importance of the three properties explicitness, orthogonality, and condensation.

6.1. *Object model learning*

Structure from motion [56] allows for the reconstruction of an object's shape from the knowledge of its motion. The visual representation presented in this study presents some advantages for this purpose: encoding the object's contours allows us to represent the object's shape and appearance (via the color and phase modalities) in a compact yet descriptive manner. As discussed in section 5, motion knowledge allows for accurate prediction of a primitive's feature vector at a later stage.

For this task, an object is manipulated by a robotic arm in front of a pair of stereo cameras. Since the arm's motion is known and the stereo and robot systems are properly calibrated, the arm's motion can be used to track reliably the 3D-primitives describing the manipulated object. Conversely, it can be inferred that 3D-primitives that do not move according to the arm's motion are not part of the manipulated object. Furthermore, object features that were not initially visible (e.g., occluded) can be added later on to the object representation; in this way, a full 3D model of the object can be generated [1].

This is illustrated in Fig. 8. There, Fig. 8a shows on the left hand side the robotic

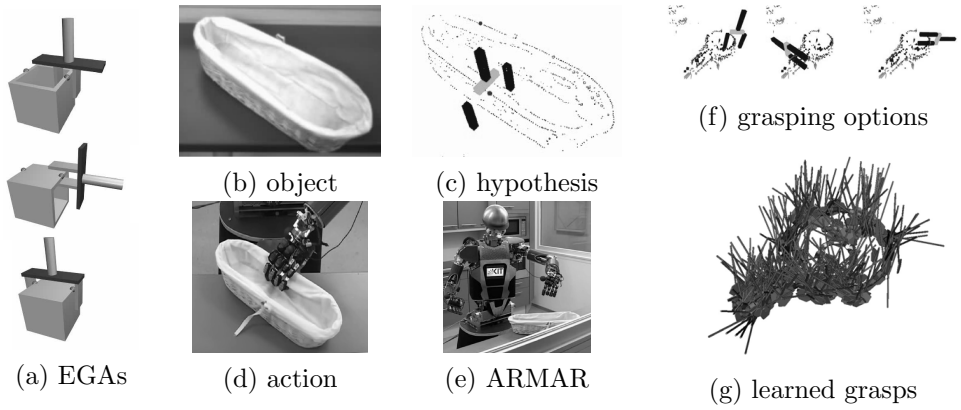


Fig. 9. The learned grasping options associated to the learned object

setup holding a pan-like object; on the right hand side, the learned object model is shown from a different viewpoint[g]. Also, Figs. 8b and c show the shape model obtained for two different objects.

The acquired representations have been used for pose estimation using a Markov network[2]; the primitives' property of explicitness allows for the definition of a strong local matching function, providing confidences that become propagated through the network to reach a global statement about the 6D pose of the object. In Figure 8d, the result of a pose estimation with the learned model is shown.

6.2. *Generating grasping hypotheses*

The proposed representation has also been used to define grasping hypotheses in a scene filled with unknown objects[4] (see Fig. 9). Essentially, pairs of co-planar and isochromatic (see section 5) primitives are used to define planes that are good candidates for an initial grasping hypothesis. Each plane's pose in space is fully defined by a pair of non-collinear 3D-primitives and it elicits several grasping hypotheses—see Fig. 9a. Fig. 9c shows one grasping hypothesis elicited by coplanar pairs of 3D-primitives in the visual representation extracted from the object in Fig. 9b. Fig. 9d shows the realization of the grasp by the robot ARMAR[63] (Fig. 9e). The fact that multiple grasping hypotheses become generated by the feature-action association indicated in Fig. 9a (see Fig. 9f for three out of usually hundreds of such associations connected to one object) allows the system to attempt grasping objects in its environment without object knowledge or prior segmentation. In practice, such a straightforward association of pairs of primitives to actions achieved success rates of around 50%. It is the explicitness and orthogonality properties of the primitives that allow for these associations: the appearance information is used to associate

[g] The gap in the representation on the pan's handle correspond to the part occluded by the gripper and could be filled by using at least one alternative grasp

primitives likely to belong to the same object, and the geometric information allows for the computation of the 6D pose of the plane the gripper can grasp.

6.3. *Birth of the object and the learning of grasp affordances*

If evaluated as successful (making use of haptic information), a grasping action such as the one proposed in section 6.2 endows the system with physical control over objects, as required by the object learning sketched in section 6.1. This provides a robot with a basic exploratory behavior: 1) try to grasp the (unknown) environment; 2) if successful, manipulate the object; and 3) learn a full 3D representation of the object. Using this scheme, a robot is able to learn about new objects in its environment, and to associate successful grasping actions to pairs of 3D-primitives that are part of the object model (see figure 9g). We coined the term *Birth of the object*[61] for the robot's ability to discover unknown objects from the combination of these three behaviors. In this way, an initially naive robot is able to progressively learn an internal representation of the world with only minimal prior world knowledge[61].

Making use of the ability to do pose estimation (as discussed in section 6.1) and grasp hypotheses computed as described in section 6.2, successful grasps can be associated to the object while the robot is playing with the object. In Fig. 9g we show the learned object with the associated successful grasping actions; black lines denote successful grasps, and their orientations reflects the corresponding pose of the robot's gripper's.

7. Conclusion

This paper presented a framework that initiates the transition from the signal-level representation of visual information towards symbolic representation which is motivated by processing in the human visual cortex. The resulting representation is condensed and carries rich and explicit semantic information where geometric and appearance information is coded in an orthogonal manner. This allowed for the definition of higher level relations between the primitives that were used for disambiguation in, e.g., object learning processes as well as for visual feature-action associations.

The three properties of this representation, together with the relations defined on them, allowed for their use in a number of applications, in particular as a vision interface of a cognitive robot system. The visual system provides the system with a powerful front-end giving it access to important structural properties of the visual scene that allows for efficient learning and bootstrapping processes. Future work will be on the enhancement of the early cognitive vision system in terms of the kind of image structures we represent in addition to edges (e.g., junctions and textures) as well as the relations between them.

Nicolas Pugeault obtained an M.Sc. from the University of Plymouth in 2002 and an Engineer degree from the Ecole Supérieure d'Informatique, Électronique, Automatique (Paris) in 2004. He obtained his Ph.D. from the University of Göttingen in 2008, and subsequently worked as a Research Associate at the University of Edinburgh and as an assistant professor at the university of Southern Denmark for two years. He is now working as a Reseach Fellow at the University of Surrey, United Kingdom.

Nicolas Pugeault is the author of over 30 technical publications, proceedings, editorials and books. His research interests include cognitive vision, machine learning and articifial intelligence.

Norbert Krüger is a Professor at the Mærsk McKinney Møller Institute, University of Southern Denmark. He holds a M.Sc. from the Ruhr-Universität Bochum, Germany and his Ph.D. from the University of Bielefeld. He is a partner in several EU and national projects: PACO-PLUS, Drivsco, NISA, Handyman. Norbert Krüger is leading the Cognitive Vision Lab which is focusing on computer vision and cognitive systems, in particular the learning of object representations in the context of grasping. He has also been working in the areas of computational neuroscience and machine learning.

Florentin Wörgötter has studied Biology and Mathematics in Düsseldorf. He received his PhD in 1988 in Essen working experimentally on the visual cortex before he turned to computational issues at the Caltech, USA (1988-1990). After 1990 he was researcher at the University of Bochum concerned with experimental and computational neuroscience of the visual system. Between 2000 and 2005 he had been Professor for Computational Neuroscience at the Psychology Department of the University of Stirling, Scotland where his interests strongly turned towards "Learning in Neurons". Since July 2005 he leads the Department for Computational Neuroscience at the Bernstein Center at the University of Göttingen. His main research interest is information processing in closed-loop perception-action systems, which includes aspects of sensory processing, motor control and learning/plasticity. These approaches are tested in walking as well as driving robotic implementations. His group has developed the RunBot a fast and adaptive biped walking robot.

References

1. N. Pugeault, F. Wörgötter, and N. Krüger. Accumulated visual representation for cognitive vision. In *Proc. of the BMVC*, 2008.
2. R. Detry, N. Pugeault, and J. Piater. A probabilistic framework for 3D visual object representation. *IEEE TPAMI*, 2009.
3. F. Pilz, N. Pugeault, and N. Krüger. Comparison of point and line features and their combination for rigid body motion estimation. In *Dagstuhl-Seminar 08291 LNCS-Postproceeding*, 2009. under revision.
4. Mila Popović, Dirk Kraft, Leon Bodenhagen, Emre Başeski, Nicolas Pugeault, Danica Kragic, Tamim Asfour, and Norbert Krüger. A strategy for grasping unknown objects based on co-planarity and colour information. *Robotic and Autonomous Systems*, 58(5):551–565, 2010.
5. O.D. Faugeras. *Three–Dimensional Computer Vision*. MIT Press, 1993.
6. C. Schmid and A. Zisserman. The geometry and matching of lines and curves. *IJCV*, 40(3):199–233, 2000.
7. G. Li and S. W. Zucker. Contextual inference in contour-based stereo correspondence. *IJCV*, 69(1):59–75, 2006.
8. G. Li and S. W. Zucker. Surface geometric constraints for stereo in belief propagation. In *Proc. of the IEEE CVPR*, 2006.
9. G. Li and S. W. Zucker. Differential geometric consistency extends stereo to curved surfaces. In *Proc. of the ECCV*, 2006.
10. M. Felsberg and G. Sommer. The monogenic signal. *IEEE Transactions on Signal Processing*, 49(12):3136–3144, December 2001.
11. D.H. Hubel and T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiology*, 160:106–154, 1962.
12. M.W. Oram and D.I. Perrett. Modeling visual recognition from neurobiological constraints. *Neural Networks*, 7:945–972, 1994.
13. M. Tistarelli, E. Grosso, and G. Sandini. Dynamic stereo in visual navigation. In *Proc. IEEE CVPR*, pages 186–193, 1991.
14. D.H. Hubel and T.N. Wiesel. Anatomical demonstration of columns in the monkey striate cortex. *Nature*, 221:747–750, 1969.
15. I.A. Shevelev, N.A. Lazareva, A.S. Tikhomirov, and G.A. Sharev. Sensitivity to cross–like figures in the cat striate neurons. *Neuroscience*, 61:965–973, 1995.
16. H. Barlow, C. Blakemore, and J.D. Pettigrew. The neural mechanisms of binocular depth discrimination. *J. of Physiology (London)*, 193:327–342, 1967.
17. N. Krüger, M. Lappe, and F. Wörgötter. Biologically motivated multi-modal processing of visual primitives. *AISB Journal*, 1(5):417–427, 2004.
18. G. H. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer Academic Publishers, 1995.
19. D. Marr. *Vision*. Freeman, 1982.
20. B. Schiele and J.L. Crowley. Probabilistic object recognition using multidimensional receptive field histograms. *NIPS*, 8:865–871, 1996.
21. M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R.P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamik link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.
22. Y. Aloimonos and D. Shulman. *Integration of Visual Modules — An Extension of the Marr Paradigm*. Academic Press, London, 1989.
23. D. Vernon. Cognitive vision: The case for embodied perception. *Image and Vision Computing*, 26(1):127–140, 2008.
24. G.H. Granlund. The complexity of vision. *Signal Processing*, 74:101–126, 1999.

25. S. Fidler, G. Berginc, and A. Leonardis. Hierarchical statistical learning of generic parts of object structure. In *CVPR06*, pages 182–189, 2006.
26. S. Fidler and A. Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *CVPR07*, 2007.
27. S. Fidler, M. Boben, and A. Leonardis. Similarity-based cross-layered hierarchical representation for object categorization. In *CVPR08*, pages 182–189, 2008.
28. L.L. Zhu, C. Lin, H. Huang, Y. Chen, and A. Yuille. Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion. In *ECCV02*, volume 2, pages 759–773, 2008.
29. F. Wörgötter, N. Krüger, N. Pugeault, D. Calow, M. Lappe, K. Pauwels, M. Van Hulle, S. Tan, and A. Johnston. Early cognitive vision: Using gestalt-laws for task-dependent, active image-processing. *Natural Computing*, 3(3):293–321, 2004.
30. D.G. Lowe. Distinctive Image Features from Scale–Invariant Keypoints. *IJCV*, 60(2):91–110, November 2004.
31. T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *Proc. of the ECCV*, pages 228–241. Springer–Verlag, 2004.
32. T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *IJCV*, 59(1):61–85, 2004.
33. H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In *Proc. of the ECCV*, 2006.
34. C.G. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conference*, pages 147–151, 1988.
35. K. Mykolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. of the ECCV*. Springer–Verlag, 2002.
36. C. Schmid, R. Mohr, and C. Baukhage. Evaluation of interest point detectors. *IJCV*, 37(2):151–172, 2000.
37. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE TPAMI*, 27(10):1615–1630, 2005.
38. P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3d objects. In *Proc. of the ICCV*, volume 1, pages 800–807, 2005.
39. J.H. Elder. Are edges incomplete ? *IJCV*, 34:97–122, 1999.
40. I. Biederman and G. Ju. Surface vs. edge-based determinants of visual recognition. *Cognitive Psychology*, 20:38–64, 1988.
41. G. Smith, Lee E., Goldberg K., K. Böhringer, and J. Craig. Computing parallel-jaw grips. In *Proc. IEEE ICRA*, 1999.
42. M. Felsberg. Optical flow estimation from monogenic phase. In B. Jähne, R. Mester, E. Barth, and H. Scharr, editors, *1st International Workshop on Complex Motion*, volume 3417 of *LNCS*, pages 1–13, 2006. In press.
43. J.J Koenderink and A.J. van Doorn. Generic neighbourhood operators. *IEEE TPAMI*, 14:597–605, 1992.
44. T. Lindeberg. Edge detection and ridge detection with automatic scale selection. *IJCV*, 30(2):117–156, 1998.
45. W.T. Freeman and E.H. Adelson. The design and use of steerable filters. *IEEE TPAMI*, 13(9):891–906, 1991.
46. J.G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by 2D visual cortical filters. *J. of the Optical Society of America*, 2(7):1160–1169, 1985.
47. J.P. Jones and L.A. Palmer. An evaluation of the two dimensional Gabor filter model of simple receptive fields in striate cortex. *J. of Neurophysiology*, 58(6):1223–1258, 1987.

48. S. P. Sabatini, G. Gastaldi, F. Solari, J. Diaz, E. Ros, K. Pauwels, K. M. M. Van Hulle, N. Pugeault, and N. Krüger. A compact harmonic code for early vision based on anisotropic frequency channels. *Computer Vision and Image Understanding*, 114:681–699, 2010.
49. P. Kovesi. Image features from phase congruency. *Videre: Journal of Computer Vision Research*, 1(3):1–26, 1999.
50. C. Zetzsche and E. Barth. Fundamental limits of linear filters in the visual processing of two dimensional signals. *Vision Research*, 30, 1990.
51. M. Felsberg, S. Kalkan, and N. Krüger. Continuous dimensionality characterization of image structures. *Image and Vision Computing*, 27:628–636, 2009.
52. L. Middleton and J. Sivaswamy. *Hexagonal Image Processing : A Practical Approach.* Springer Verlag, 2005.
53. R.C. Staunton and N. Storey. A comparison between square and hexagonal sampling methods for pipeline image processing. *Proc. SPIE*, 1194:142–151,, 1989.
54. H.-H. Nagel and W. Enkelmann. An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *IEEE TPAMI*, 8:565–593, 1986.
55. N. Pugeault. *Early Cognitive Vision: Feedback Mechanisms for the Disambiguation of Early Visual Representation.* PhD thesis, Georg-August-Universität Göttingen, 2008.
56. R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2000.
57. J.E.W. Mayhew and J.P. Frisby. Psychophysical and computational studies towards a theory of human stereopsis. *Artificial Intelligence*, 17:349–385, 1981.
58. W.E.L. Grimson. Computational experiments with a feature based stereo algorithm. *IEEE TPAMI*, 7:17–24, 1985.
59. N. Ayache and B. Faverjon. Efficient registration of stereo images by matching graph descriptions of edge segments. *IJCV*, pages 107–131, 1987.
60. N. Pugeault, F. Wörgötter, and N. Krüger. Disambiguating multi–modal scene representations using perceptual grouping constraints. *PLoS ONE*, 5(6), 2010.
61. D. Kraft, N. Pugeault, E. Başeski, M. Popović, D. Kragic, S. Kalkan, F. Wörgötter, and N. Krüger. Birth of the object: Detection of objectness and extraction of object shape through object action complexes. *IJHR Special Issue on "Cognitive Humanoid Robots"*, 5:247–265, 2009.
62. R. Detry, E. Baseski, M. Popovic, Y. Touati, N. Krüger, O. Krömer, J. Peters, and J. Piater. Learning object-specific grasp affordance densities. In *Int. Conf. on Development and Learning*, 2009.
63. T. Asfour, P. Azad, N. Vahrenkamp, K. Regenstein, A. Bierbaum, K. Welke, J. Schröder, and R. Dillmann. Toward humanoid manipulation in human-centred environments. *Robot. Auton. Syst.*, 56(1):54–65, 2008.
64. L.B.W. Jensen, A. Kjær-Nielsen, K. Pauwels, J.B. Jessen Jessen, M. Van Hulle, and N. Krüger. A two-level real-time vision machine combining coarse and fine grained parallelism. *J. of Real-time Image Proc.*, accepted.