Processing Multi-modal Primitives from Image Sequences

Norbert Krüger[1], Michael Felsberg[2], and Florentin Wörgötter[3]

[1] Dep. of Computer Science,
Aalborg University Esbjerg,
6700 Esbjerg,
Denmark,
nk@aue.auc.dk

[2]Computer Vision Laboratory,
Dep. of Electrical Engineering,
Linköping University,
SE-58183 Linköping, Sweden,
mfe@isy.liu.se

[3]Computational Neuroscience,
Department of Psychology,
Stirling FK9 4LA,
University of Stirling,
worgott@cn.stir.ac.uk

Abstract

In this paper, we describe a new kind of image representation in terms of local multi–modal Primitives. Our local Primitives can be characterized by three properties: (1) They represent different aspects of the image in terms of multiple visual modalities. (2) They are adaptable according to context. (3) They provide a condensed representation of local image structure.

These three properties make them especially useful as a first stage of visual scene analysis. Our Primitives initialize a process of contextual integration that disambiguated locally ambiguous information in the artificial visual system [1, 2, 3].

1 Introduction

Vision faces the problem of an extremely high degree of vagueness and uncertainty in its low level processes such as edge detection, optic flow analysis and stereo estimation. This arises from a number of factors. Some of them are associated with image acquisition and interpretation: owing to noise in the acquisition process along with the limited resolution of cameras, only rough estimates of semantic information (e.g., orientation) are possible. The severeness of these problems increases for higher semantic information, such as curvature. Furthermore, illumination variation heavily influences the measured grey level values and is hard to model analytically (see, e.g., [4]). Extracting information across image frames, e.g., in stereo and optic flow estimation, faces (in addition to the above mentioned problems) the correspondence and aperture problem which interfere in a fundamental and especially awkward way (see, e.g., [5]).

Despite the above mentioned problem, the aim of this work is to compute reliable feature maps from natural scenes. Since local feature extraction is necessarily ambigious as explained above such reliable feature maps can only be computed through *integration across the spatial and temporal context and across visual modalities* (see, e.g., [6]). The European Project ECOVISION [7] focusses exactly on this issue. In this paper, we describe a new kind of image representation in terms of local multi–modal Primitives (see fig. 1). These Primitives are a central pillar of this ongoing EU-project. Our Primitives are local image descriptors that can be characterized by three properties:

Multi-modality: Different domains that describe different kinds of structures in visual data are well established in human vision and computer vision. For example, an edge in an image patch can be analyzed by local feature attributes such as orientation or energy in certain frequency bands. In addition, we can distinguish between line and step–edge like structures (constrast transition). Furthermore, color can be associated to the edge. This image patch also changes in time due to ego-motion or object motion. Therefore time specific features such as a 2D velocity vector (optic flow) can be associated to it. In this work we define local Primitives that realize all these multi-modal relations. The used modalities, in addition to the semantic parameters position and orientation, are contrast transition, color and optic flow (see fig. 1 and 7II). In addition, the Primitives carry also information about the 'edge–ness', 'junction–ness' or 'homogeneous–ness' of the local image patch.

Adaptability: Since the interpretation of local image patches in terms of the above mention attributes as well as classifications such as 'edgeness' or 'junctionness' are necessarilly ambigious when based on local processing, stable interpretations can only be achieved *through integration* by making use of contextual information [6]. Therefore, all attributes of our Primitives are equipped with confidence values that are *adaptable according to contextual information* expressing the reliability of this attribute. Furthermore, the feature attributes itself adapts according to the context (see section 3).

Condensation: Integration of information requires *communication between Primitives* expressing cross–modal [3], spatial (see, e.g., [2, 3]) and spatial–temporal dependencies (see, e.g., [1]). This communication has necessarily to be paid for with a certain cost. This cost can be reduced by limiting the amount of information transferred from one place to the other, i.e.,

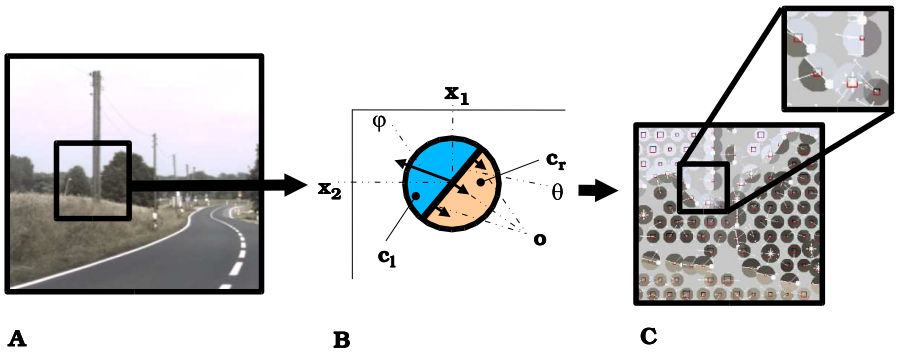


Figure 1: **A:** Image sequence and frame. **B:** Schematic representation of the multi–modal Primitives. **C:** Extracted Primitives at positions of high energy.

by reducing the bandwidth. Therefore, we are seeking after a *condensed* representation. Also for other tasks, e.g., to store objects, it is important to represent information in a *condensed way* to reduce memory requirements.

These three attributes make our Primitives especially useful as a first stage of an artificial visual system. Multi-modality ensures a rich description of visual structures. Especially geometric as well as appearance based information is coded which both have shown to be useful in different contexts and for different tasks (see, e.g., [8]). Adaptibility of Primitives prevents hard decisions at an early stage of processing that are necessarily ambiguous (as discussed above). Finally, the condensed representation of our Primitives allows for an efficient communication between visual events in the disambiguation process compared to a pixelwise representation.

In section 2, we describe the Primitive attributes and their extraction. In section 3, we refer to work that makes use of our Primitives for contextual integration.

2 Extraction of Multi–Modal Primitives

In subsection 2.1, we describe a recently developped filter approach, called the monogenic signal and derive spatial filters for this approach. Making use of the spatial realization of filters, in sub–section 2.2 we determine the position of the Primitives. In the following subsections, we then describe the extraction of the other Primitive attributes orientation, contrast transition, color, and optic flow.

2.1 Basic Filtering: The monogenic Signal

The basic pre–processing stage of our Primitives is based on a rotation invariant quadrature filter, which is derived from the concept of the *monogenic signal* [9]. Considered in polar coordinates, the monogenic signal performs a *split of identity*: it decomposes an intrinsically one-dimensional signal into intensity information (amplitude), orientation information, and phase information (contrast transition). These features are pointwise[1] mutually orthogonal. The intensity information can be used as an indicator for the likelihood of the presence of a certain structure while orientation and a contrast transition will be used as attributes of our Primitives.

Quadrature filters based on the monogenic signal are rotation invariant, i.e., they commute with the rotation operator. Hence, for an appropriate choice of polar coordinates, two coordinates do not change under rotations (amplitude and phase), whereas the third coordinate directly reflects the rotation angle. This kind of quadrature filter, which is called *spherical quadrature filter* [11], is formed by triplet of filters: a radial bandpass filter and its two Riesz transforms. We construct the bandpass filter from *difference of Poisson* (DOP) filters, in order to get analytic formulations of all filter components in the spatial domain and in the frequency domain. The DOP filter is an even filter (w.r.t. point reflections in the origin) and its impulse response (convolution kernel) and frequency response (Fourier transform of the kernel) are respectively given by:

$$h_{\mathrm{e}}(\mathbf{x}) = \frac{s_1}{2\pi(|\mathbf{x}|^2 + s_1^2)^{\frac{3}{2}}} - \frac{s_2}{2\pi(|\mathbf{x}|^2 + s_2^2)^{\frac{3}{2}}} \quad (1)$$

$$H_{\mathrm{e}}(\mathbf{u}) = \exp(-2\pi|\mathbf{u}|s_1) - \exp(-2\pi|\mathbf{u}|s_2) \ . \quad (2)$$

[1] In a global context, these features are related by the Riesz transform [10].

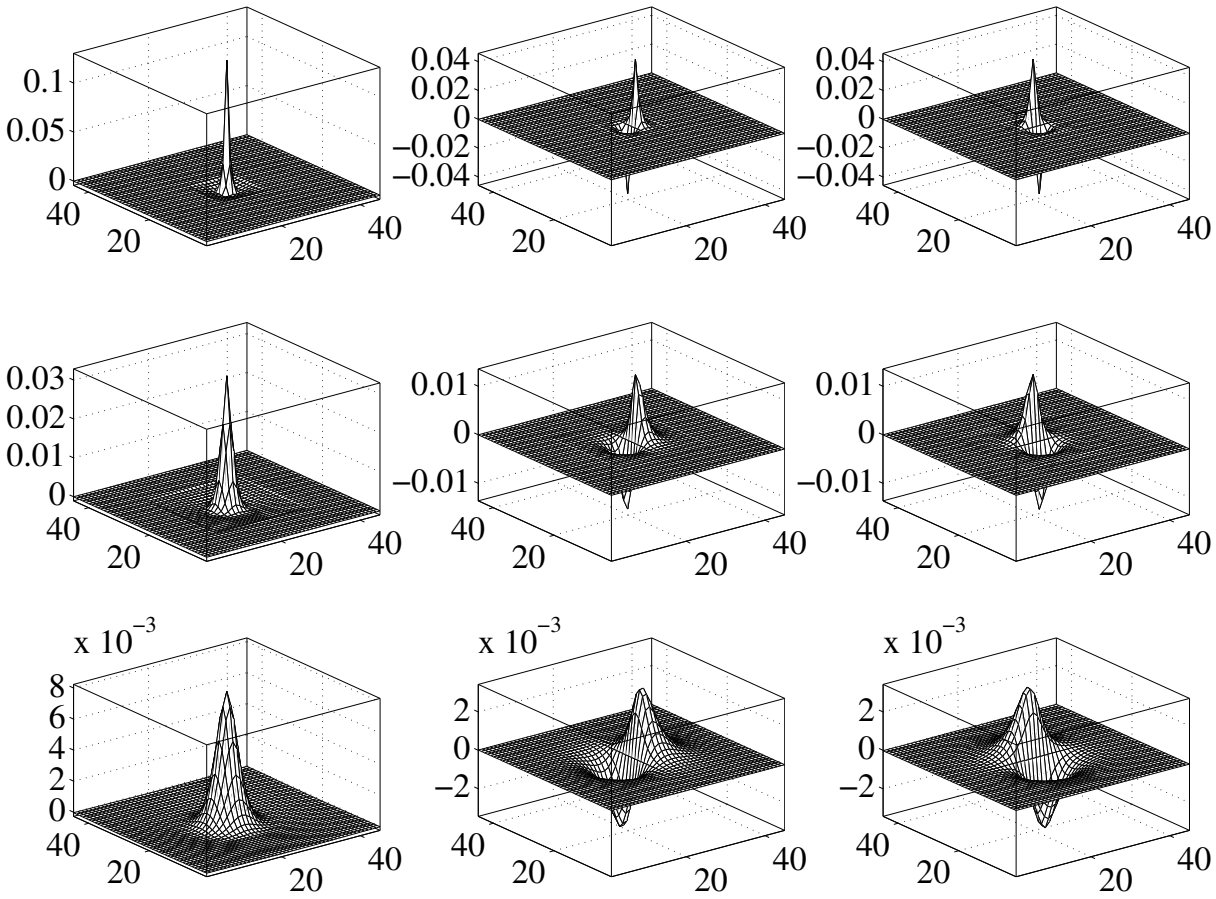


Figure 2: Impulse responses of the DOP filter and its Riesz transforms. From left to right: DOP filter, first Riesz transform, second Riesz transform. From top to bottom: scales (1,2), (2,4), (4,8).

For convenience, we combine the two Riesz transforms of the DOP filter in a complex, odd filter, yielding the impulse response and the frequency response:

$$h_{\mathrm{o}}(\mathbf{x}) = \frac{x_1 + ix_2}{2\pi(|\mathbf{x}|^2 + s_1^2)^{\frac{3}{2}}} - \frac{x_1 + ix_2}{2\pi(|\mathbf{x}|^2 + s_2^2)^{\frac{3}{2}}} \tag{3}$$

$$H_{\mathrm{o}}(\mathbf{u}) = \frac{u_2 - iu_1}{|\mathbf{u}|}(\exp(-2\pi|\mathbf{u}|s_1) - \exp(-2\pi|\mathbf{u}|s_2)) \ , \tag{4}$$

respectively. The impulse responses of the filters for $(s_1, s_2) = (1, 2), (2, 4), (4, 8)$ are shown in figure 2. A representation of the filters in the spatial domain is needed for our feature generation since it allows for the definition of a 'spatial region of influence' where the sum of the amplitude of the convolution kernel is significantly above zero. In the sparsification process described in section 2.2 this defines naturally a sampling distance for the extraction of our Primitives and a minimal distance between Primitives since inside this region filter responses are necessarily redundant while outside these regions they describe different image patches (see section 2.2).

In order to illustrate the effect of the quadrature filter to an image I, we applied the three filters to the image in the top left of figure 3. The convolution results $I_{\mathrm{e}} = h_{\mathrm{e}} * I$ and $I_{\mathrm{o}} = h_{\mathrm{o}} * I$ (complex) can be found in the same row.

The split of identity is obtained by switching to appropriate polar coordinates. In particular, we transform the filter responses according to

$$m(\mathbf{x}) = \sqrt{I_{\mathrm{e}}(\mathbf{x})^2 + |I_{\mathrm{o}}(\mathbf{x})|^2} \tag{5}$$

$$\theta(\mathbf{x}) = \arg I_{\mathrm{o}}(\mathbf{x}) \pmod{\pi} \tag{6}$$

$$\varphi(\mathbf{x}) = sign(\Im\{I_{\mathrm{o}}(\mathbf{x})\}) \arg(I_{\mathrm{e}}(\mathbf{x}) + i|I_{\mathrm{o}}(\mathbf{x})|) \ , \tag{7}$$

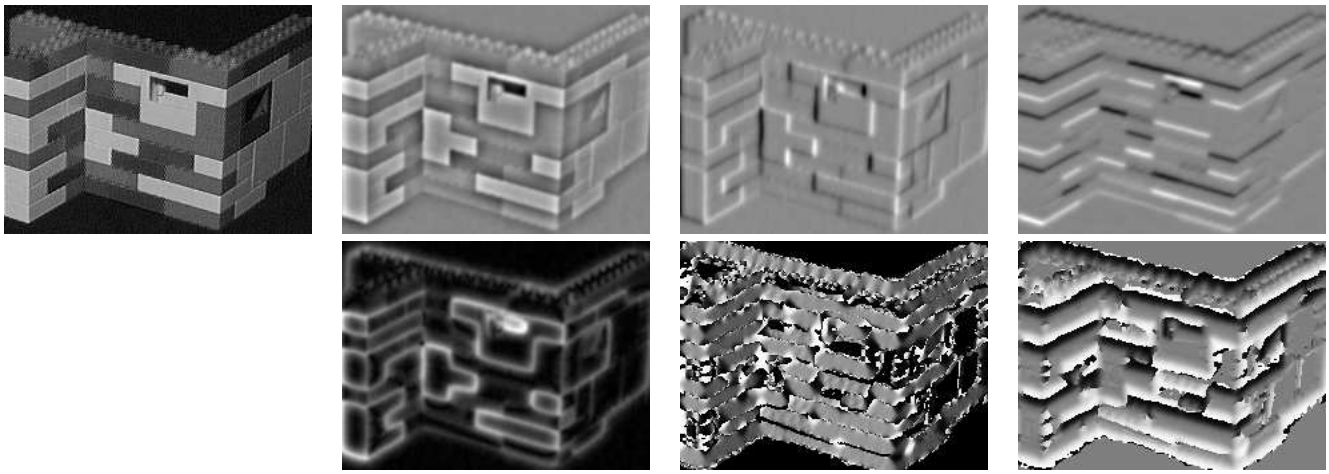


Figure 3: Upper row, from left to right: original image, image filtered with even kernel, image filtered with real part of odd kernel, and image filtered with imaginary part of odd kernel. Bottom row: interpretation in polar coordinates. From left to right: amplitude, masked orientation, and masked phase.

which gives the desired amplitude, orientation, and phase information. In the case of our example from figure 3, we obtain the results illustrated in the bottom row of figure 3. Using the amplitude and the variance of orientation in an image patch confidences for the existence of a feature will be derived in section 2.2. In section 2.3 orientation and phase are used as features attributes of our Primitives.

Figure 4 shows a radial slice cut through the DOP bandpass filters for a certain range of scales and their superposition, demonstrating a homogenous covering of the frequency domain. For infinitely many bandpass filters, the superpostion is one everywhere, except for the origin. In our system, we apply filters on three frequency levels (see figure 2). The applied bandpasses are indicated by the darker color in figure 4.

The application of spherical quadrature filters for the processing of our Primitives bears two main advantages:

1) Compared to the application of a Gabor wavelet transform (see, e.g., [12]) we do not need to sample different orientations but the orientation is a direct output of the computation. That means we only need to apply 3 filter operations compared to 16 in case of a sampling with 8 complex orientation filters (as done, e.g., in [13]). This gives us a speed up of this time–expensive processing stage of more than a factor of 5.

2) The phase is a direct output of the sperical quadrature filter processing and can directly be used as an attribute that describes the kind of contrast transition of the oriented image structure (see figure 5 and section 2.3).

2.2 Intrinsic Dimension and Position

There exist special sub–structures for local image patches that can be characterized by their intrinsic dimensionality (see, e.g., [14]). Homogeneous image patches have an intrinsic dimension of zero (i0D), edge–like structures are intrinsically 1–dimensional (i1D) while junctions and most textures have an intrinsic dimension of two (i2D). The concept of position is different for a junction compared to an edge or an homogeneous image patch (see figure 6). Our Primitives carry information about the intrinsic dimension by three values $c_{i0D}, c_{i1D}, c_{i2D}$. c_i represents the confidence that the image patch, which is represented by the Primitive, has intrinsic dimension i. Our coding of intrinsic dimension as well as its application in our image representation is described in detail in [15].

We want to express a local image patch corresponding to an intrinsically one–dimensional signal patch by a Primitive E. For this we first perform an hexagonal sampling of the image into areas $A^{(i,j)}$. The sampling distance depends on the frequency level and by this, on the size of the basic filter. We then look for optimal positions inside the patch. Naturally, we distinguish three cases:

Homogeneous image patches (i0D): At homogeneous image patches the position can not be defined by properties of the local signal since the signal is everywhere the same. Therefore, the position of a Primitive should be defined by an equidistant sampling (see figure 6a).

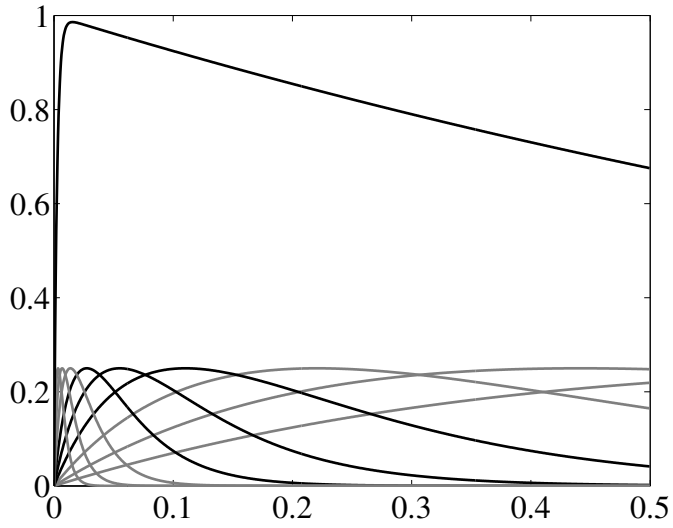


Figure 4: DOP bandpass filters and their superposition approaching the identity. The superposition and the filters applied in this paper are indicated by the darker lines.

Lines and edges (i1D): For a line or edge, the position can be defined through energy maxima that are organized as a one–dimensional manifold. Therefore, an equidistant sampling along these enery maxima is appropriate (see figure 6b).

Junction–like structures (i2D): For a junction the position can be defined un–ambigiously as energy maxima in a local region (see figure 6c).

To avoid similar descriptors in our image representation we look in a circle around each $\mathbf{x}^{(i,j)}$ whether there exists an $\mathbf{x}^{(i',j')}$ close to $\mathbf{x}^{(i,j)}$. If this is the case we delete the one with smaller amplitude [2]. This gives us the position $\mathbf{x}$ of a Primitive.

2.3 Frequency, Orientation and Contrast Transition

We now describe the coding of modalities associated to our Primitives that are directly related to the the filter approach described in section 2.1.

Frequency: We describe the signal on 3 different frequency levels f independently. Often the decision in which frequency band the relevant information does occur is difficult, therefore we leave this decision open to be decided at later stages of processing.

Orientation: The local orientation associated to the image patch is described by θ. The orientation θ is computed by interpolating across the orientation information of the whole image patch to achieve a more reliable estimate. This holds also true for the feature attributes contrast transition, colour and optic flow.

Contrast transition: Contrast transition is coded in the phase φ of the applied filter [9]. The phase codes the local symmetry, for example a bright line on a dark background has phase 0 while a bright/dark edge has phase $-\pi/2$ (in fig. 7 the line that marks the border of the street in represented as a line or two edges depending on the distance). Of course, there is a continuum between these different grey level structure. The local phase as additional feature allows to take the grey level information into account (as one parameter in addition to orientation) in a very compact way (see figure 5 and, e.g., [16, 17, 9]). In case of boundaries of objects, phase represents a description of the transition between object and background.

2.4 Color and Optic Flow

Color $(\mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r)$ is processed by integrating colour pixel values over image patches in accordance to the local orientation and edge/line structure (i.e., integrating separately over the left and right side of the edge as well as a middle strip in case of a line structure). In case of an homogenous image patch $\mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r$ are similar and only one vector needs to be stored. In contrast to phase, for a boundary edge of a moving object at least the color at one side of the edge is expected to be stable since it represents a description of the object.

Finally, local displacements $\mathbf{o}$ are computed by Nagel's optic flow algorithm [18].

[2] The radius of the circle depends of the frequency of the applied filter.

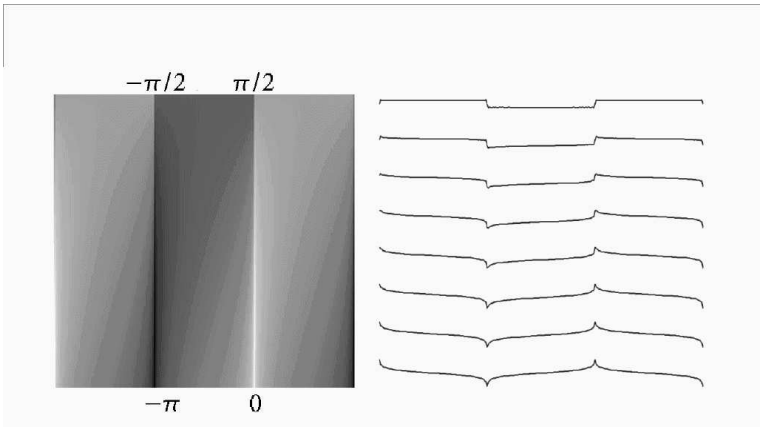


Figure 5: **Left:** Variation of contrast transition according to phase variation. Note that a phase of π codes a dark line on bright background, a phase of$-\pi/2$ coded a bright/dark edge, a phsae of 0 codes a bright line on a dark background while a phase of $\pi/2$ codes a dark/bright edge. As can be seen the continuum between these case is also be coded by the phase. **Right:** Luminance profiles corresponding to left image.

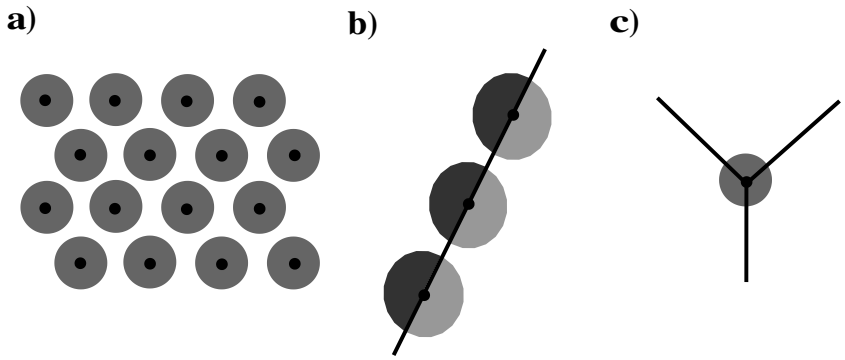


Figure 6: a) At homogeneous image patches an equidistant sampling is performed. b) For a line or edge an equidistant sampling along the one-dimensional manifold of magnitude maxima is appropriate. c) For a junction the position can be defined as the local magnitude maximum.

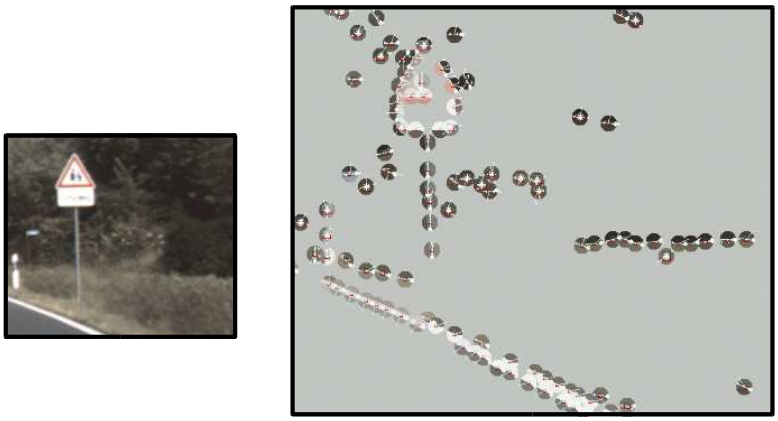
With our confidences for the different intrinsic dimensions $c_{i0D}, c_{i1D}, c_{i2D}$ (see section 2.2 and [15]) we end up with a parametric description of a Primitive as

$$E = (\mathbf{x}, f, \theta, \varphi, (\mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r), \mathbf{o}; c_{i0D}, c_{i1D}, c_{i2D}).$$

As mentioned above, in addition, to each of the parameters $\varphi, (\mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r), \mathbf{o}$ there exist confidences $c_i, i \in \{\varphi, \mathbf{c}^l, \mathbf{c}^m, \mathbf{c}^r, \mathbf{o}\}$ that code the reliabilty of the specific sub–aspects that are subject to contextual adaptation.

3 Integration of Contextual Information and Conclusion

The introduced image representation has been used in different contexts. Firstly, an image patch also describes a certain region of the 3D space and therefore 3D attributes can be associated such as a 3D-position and a 3D-direction. In [19] a stereo similarity function is defined that makes use of multiple-modalities to enhance matching performance. It could be shown that the use of multiple modalities enhances stereo performance significantly and that optimal performance is allways achieved by using all modalities represented in our Primitives. Secondly, the Primitives can be subject to spatial contextual modification. Groups of Primitives are defined based on a purely statistical criterion in [2]. Once these groups are defined, confidences of the Primitives become modulated: confidences become increased if the Primitives are part of a bigger group, otherwise the confidences are decreased. Again, it could be shown that the integrative power of visual events corresponding to the Gestalt law 'good continuation' increases by making use of the different modalities of the Primitives. Thirdly, orientation and position attributes habe been stabilized according to the temporal context. In [1] it is made use of the motion of an object to predict feature occurrences across frames and is has been shown that stereo processing can be stabilized by modifying the confidences



A B

Figure 7: A: Original Image. B: Extracted Primitives with high amplitude.

according to the temporal context. We are currently working on extending this approach to other modalities, such as, contrast transition and colour.

Conclusion: We have introduced a novel kind of image representation in terms of visual Primitives. These Primitives are multi-modal and give a dense and meaningful description of a scene. Our Primitives are used as a first stage in the artificial visual system [1, 2, 3] where they initialize a disambiguation process. In this process confidences associated to our Primitive attributes adapt according to spatial and temporal context and in this way stabilize the locally unrealiable feature extraction.

References

[1] N. Krüger, M. Ackermann, and G. Sommer, "Accumulation of object representations utilizing interaction of robot action and perception," *Knowledge Based Systems*, vol. 15, pp. 111–118, 2002.

[2] N. Krüger and F. Wörgötter, "Multi modal estimation of collinearity and parallelism in natural image sequences," *Network: Computation in Neural Systems*, vol. 13, pp. 553–576, 2002.

[3] N. Krüger, M. Felsberg, C. Gebken, and M. Pörksen, "An explicit and compact coding of geometric and structural information applied to stereo processing," *Proceedings of the workshop 'Vision, Modeling and VISUALIZATION 2002'*, 2002.

[4] K. Ikeuchi and B.K.P. Horn, "Numerical shape from shading and occluding boundaries," *Artificial Intelligence*, vol. 17, pp. 141–184, 1981.

[5] N. Ayache, *Stereovision and Sensor Fusion*, MIT Press, 1990.

[6] Y. Aloimonos and D. Shulman, *Integration of Visual Modules — An extension of the Marr Paradigm*, Academic Press, London, 1989.

[7] ECOVISION, "Artificial visual systems based on early-cognitive cortical processing (EU–Project)," *http://www.pspc.dibe.unige.it/ecovision/project.html*, 2003.

[8] Joseph L. Mundy, A. Liu, Nic Pillow, Andrew Zisserman, S. Abdallah, Sven Utcke, S. Nayar, and Charlie Rothwell, "An experimental comparison of appearance and geometric model based recognition," in *Object Representation in Computer Vision*, 1996, pp. 247–269.

[9] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE Transactions on Signal Processing*, vol. 49, no. 12, pp. 3136–3144, December 2001.

[10] M. Felsberg and G. Sommer, "The monogenic scale-space: A unifying approach to phase-based image processing in scale-space," *Journal of Mathematical Imaging and Vision*, 2003, to appear.

[11] M. Felsberg, *Low-Level Image Processing with the Structure Multivector*, Ph.D. thesis, Institute of Computer Science and Applied Mathematics, Christian-Albrechts-University of Kiel, 2002.

[12] J.G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by 2d visual cortical filters," *Journal of the Optical Society of America*, vol. 2, no. 7, pp. 1160–1169, 1985.

[13] L. Wiskott, J.M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–780, 1997.

[14] C. Zetzsche and E. Barth, "Fundamental limits of linear filters in the visual processing of two dimensional signals," *Vision Research*, vol. 30, 1990.

[15] N. Krüger and M. Felsberg, "A continuous formulation of intrinsic dimension," *Proceedings of the British Machine Vision Conference*, 2003.

[16] G. H. Granlund and H. Knutsson, *Signal Processing for Computer Vision*, Kluwer Academic Publishers, Dordrecht, 1995.

[17] P. Kovesi, "Image features from phase congruency," *Videre: Journal of Computer Vision Research*, vol. 1, no. 3, pp. 1–26, 1999.

[18] H.-H. Nagel, "On the estimation of optic flow: Relations between different approaches and some new results.," *Artificial Intelligence*, vol. 33, pp. 299–324, 1987.

[19] N. Pugeault and N. Krüger, "Multi–modal matching applied to stereo," *Proceedings of the BMVC 2003*, 2003.