

Early Cognitive Vision as a Front-end for Cognitive Systems

N. Krüger¹, N. Pugeault², E. Başeski¹, L.B.W. Jensen¹, S. Kalkan³, D. Kraft¹, J.B. Jessen¹, F. Pilz^{1,4}, A. Kjaer-Nielsen¹, M. Popovic¹, T. Asfour⁵, J. Piater⁶, D. Kragic⁷, and F. Wörgötter⁸

¹ Cognitive Vision Lab, Maersk Institute, University of Southern Denmark

² CVSSP, University of Surrey, United Kingdom

³ Dept. of Computer Eng., Middle East Technical University, Turkey

⁴ Medialogy, Aalborg University, Denmark

⁵ Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT), Germany

⁶ INTELSIG Laboratory, EECS Department, University of Lige, Belgium

⁷ Centre for Autonomous Systems, KTH, Sweden

⁸ University of Göttingen, Germany

Abstract. We discuss the need of an elaborated in-between stage bridging early vision and cognitive vision which we call ‘Early Cognitive Vision’ (ECV). This stage provides semantically rich, disambiguated and largely task independent scene representations which can be used in many contexts. In addition, the ECV stage is important for generalization processes across objects and actions. We exemplify this at a concrete realisation of an ECV system that has already been used in variety of application domains.

1 Introduction

The ability of human beings (and highly developed animals) to use vision as a versatile, precise and reliable source of information for tasks as diverse as pattern recognition, navigation or object manipulation is, to this day, still unparalleled and has not been achieved by any artificial system. For example, the simple action of opening a door involves a combined use of visual and motor abilities such as object recognition, navigation, grasping and precise manipulation. Considerable progress during the last decades has led to a mature understanding of the processing of different aspects of visual information (e.g., edge detection, optic flow, stereo) and to very successful solutions for specific problems such as object recognition with constrained numbers of classes and in constrained environments. Yet, there is still little understanding of the mechanisms that are required for designing a multi-purpose vision system such as the one found in humans.

In many computer vision approaches, visual processing is split into two complementary stages. At the first stage (often called *Early Vision* (EV)), a collection of image processing algorithms extract features from image sequences. This has been extensively studied, and these studies led to the design of a variety

of features that present different invariance qualities (see, e.g., [1, 2]). A second stage, often called *Cognitive Vision* (CV), is concerned with processing high-level visual entities (such as objects and trajectories thereof) to solve complex tasks, such as planning, navigation and surveillance. One fundamental difficulty encountered by the vision community is the semantic gap between the visual features produced by early vision, and the high-level concepts required by cognitive vision.

Today’s mainstream computer vision tends to circumvent this gap by designing systems for specific tasks—with very good results in some cases. One problem of this approach is that what is learned in the process of solving one task can generally not be used for solving another—for example, bag-of-features or bag-of-words approaches (see, e.g., [3]) are very successful representations for object detection, but not useful for pose estimation. In contrast, complex cognitive systems need to solve multiple tasks conjointly, such as object recognition, pose estimation, navigation, object manipulation, visual servoing, etc. Moreover, recent studies [4–6] suggest that combining several processing tasks in one system can lead to improved performance and that a shared hierarchical representation of visual features allows for an increase in performance compared to plain feature-based methods.

In this article, we argue for the design of an elaborated in-between layer, fitted between Early Vision (where different aspects of visual information are represented on a pixel level¹) and Cognitive Vision; we call this layer *Early Cognitive Vision* (ECV), and will define it in the following section. The role of the ECV layer is to provide a rich, symbolic and generic representation of visual information that can serve to mediate information and feedback between multiple visual processes. The complexity of the visual entities at the ECV level is lower than concepts such as objects and their trajectories, but higher than pixel-level information on local amplitude, optic flow and disparity. In particular, the ECV level allows for efficient learning for a number of tasks by relating the visual entities on the ECV level across objects, actions and tasks. This approach is more general than the usual task solving strategy in mainstream computer vision, where a minimal representation tailored to the specific task is used. By definition, the ECV level will process and mediate a lot more information than is strictly required for solving a single task, as this additional information enables inter-process feedback which is in particular important for learning and generalisation.

We argue that such a level is an essential component of a versatile cognitive vision system. Our research on the development of a concrete ECV system (which meanwhile has been used in a number of contexts ranging from cognitive robotics [10–12] to driver assistance systems [13]) is described in other articles (see, e.g., [14–16]). In this article, we will use this system as an example for an ECV system but we will not describe the system’s implementation in any detail, but rather discuss the theoretical implications of such an approach.

¹ This can, for example, be achieved in a harmonic representation based on a Gabor-wavelet like filter (see, e.g., [7, 8] as in the human visual system [9]).







		Homog. patches 	Edges 	Junctions 	Texture 
2D	geometry	none	2D orientation	2D intersection multiple orientations	none
	appearance	mean colour shading gradient	2 or 3 colours phase	probably too unstable	not understood
3D	geometry	surface patch 	3D point and 3D orientation	3D intersection multiple orientations (exception: T-junct.)	surface patch 
	appearance	as in 2D	as in 2D	probably too unstable	not understood

Fig. 1. Four different image structures and their symbolic representations. We distinguish between four different kinds of local image structures: Homogeneous patches, edges, junctions and textures. All four structures need to be represented by local symbolic descriptors covering appearance as well as geometric information, both in 2D and 3D.

2 Properties of an Early Cognitive Vision System

In the following, we define eight properties which we find essential for an ECV system.

Property 1. ECV bridges the gap between EV and CV. The ECV is an intermediate stage that mediates information between the initial low-level, pixel based EV stage, and the high-level, symbolic CV stage. This intermediate stage is important because it provides a semantically rich, disambiguated and hierarchically organized representation of visual information that enables multiple vision-based cognitive processes to bootstrap each other—as in the human visual system. Moreover, the ECV layer allows for the modulation and correction of low-level processes by transferring assumptions from high-level reasoning down to early image interpretation stages, in what we called ‘signal-symbol loops’ (see below and [17, 18]).

Property 2. ECV is generic and task independent. A cognitive agent has to solve a variety of vision dependent processes, often simultaneously: object recognition or categorisation, pose estimation, localisation and map-building, grasping, manipulation, tool use, etc. It is therefore advantageous for a cognitive agent to compute a *generic scene representation* that provides information as required by all common tasks.² Interestingly, the early stages in the human visual

² Note that this is in stark contrast with, e.g., the visual system of frogs where task specific ‘fly-detectors’ are processed in the retina [19] at very early stages of the processing. This underlines the difference between a cognitive vision system where

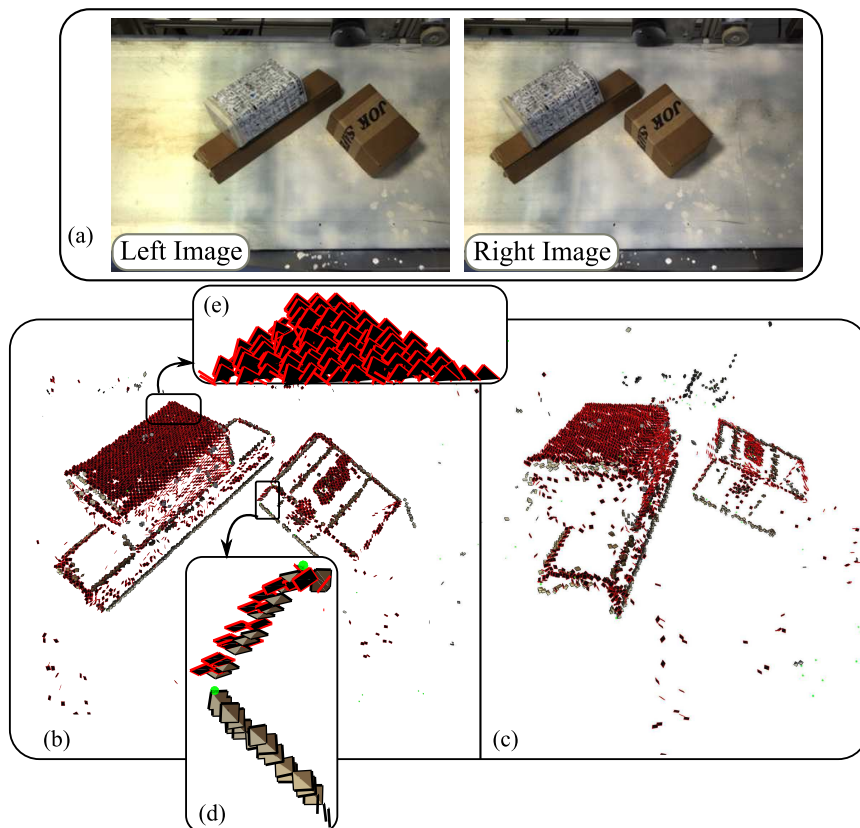


Fig. 2. Different image structures for a given scene. **(a)** Stereo image pair. **(b-c)** Different image structures in 3D from two different viewing angles. **(d)** 3D edge features denoted by 3D patches with two colors and junctions denoted by green spheres. **(e)** 3D texlets denoted by red 3D patches.

system devoted to feature extraction (realized in the areas V1 and V2) occupy much larger areas in the brain than all ‘higher’ stages of visual processing [20]; this shows how important is the extraction of a generic scene representation. Indeed, one might even argue that the extraction of an efficient visual representation *is* the key problem of visual perception, and that solving it makes ‘higher level processes’ solvable with fewer computational resources.

Property 3. ECV provides symbolic and contextually embedded representations. To provide visual information relevant for cognitive processes, the ECV system extracts condensed information from stereo sequences; this process realises two important properties (see [14]): First, the high-dimensional

versatility requests complex representations, and simpler, reactive vision systems, where the speed of response demands simpler connections

pixel space is decomposed into smaller units requiring fewer dimensions (*reduction of bandwidth* of information). Second, the *predictivity* of these dimensions (for example, for the task of contextual disambiguation, see below) becomes increased. In [21], it has been argued that these two requirements lead naturally to *symbolic local representations* where the semantic content of bits of visual information becomes increased. This, in particular, leads to a separation of relevant aspects of visual information into largely independent dimensions; for example, distinguishing appearance and geometric information as outlined in Fig. 1. As a consequence, the actual dimensions used in the representation relate to *explicit* and (e.g., geometrically) *interpretable* properties of visual information. Moreover, they become embedded into contextual relations (such as normal distance, coplanarity, cocolority and rigid body motion—see Fig. 3 for an illustration) which also exhibit a high degree of semantic content (e.g., the formalisation of the change of local descriptors under a rigid body motion or the angle between surfaces or contours). Note that this leads to fundamental differences to concepts such as SIFT features [1] (in which textural descriptors are formed by histograms) where different aspects of visual information stay scrambled and undistinguished.³

The different type of structures handled by the ECV layer are illustrated in Fig. 1 and Fig. 2. The semantic content is very different for the different kinds of structures: *Homogeneous patches* in 2D do not carry any geometric information since neither reliable point nor orientation information can be extracted. The appearance information only contains a mean color and eventually some weak gradient information (possibly useful for shape from shading computation). A reasonable 3D representation of a homogeneous patch is a 3D surface patch [22, 23]. For *edges*, an orientation can be reliably computed as well as a position on a one-dimensional manifold (aperture problem). The color information needs to be coded depending on the orientation as well as the local signal structure (e.g., line versus step-edge) which can be characterized by the local phase [24, 25]. An appropriate representation of the 3D geometry is a local 3D line segment with position and 3D orientation information. *Junctions* are intersections of edges and have a complex 2D geometry covering the intersection point as well as half-lines extending from it. Because of this complexity, a large degree of ambiguity can be expected in the computation of the junction parameters and it is unlikely that reliable enough appearance information for any practical use can be computed. The complex geometry extends to the 3D domain where an important distinction is whether the lines intersecting in 2D also intersect in 3D. *Texture* is characterized by an intrinsic complexity which is difficult to characterize in 2D. This complexity, however, allows for the computation of reliable correspondences for stereo and optic flow processing. A reasonable 3D interpretation is a 3D surface patch, which in contrast to homogeneous patches, can be computed

³ There is no doubt that SIFT-like descriptors are very useful for finding correspondences but they are more or less useless for analysing what is actually going on in a local image structure and to make this knowledge in some way explicit for higher level processes.

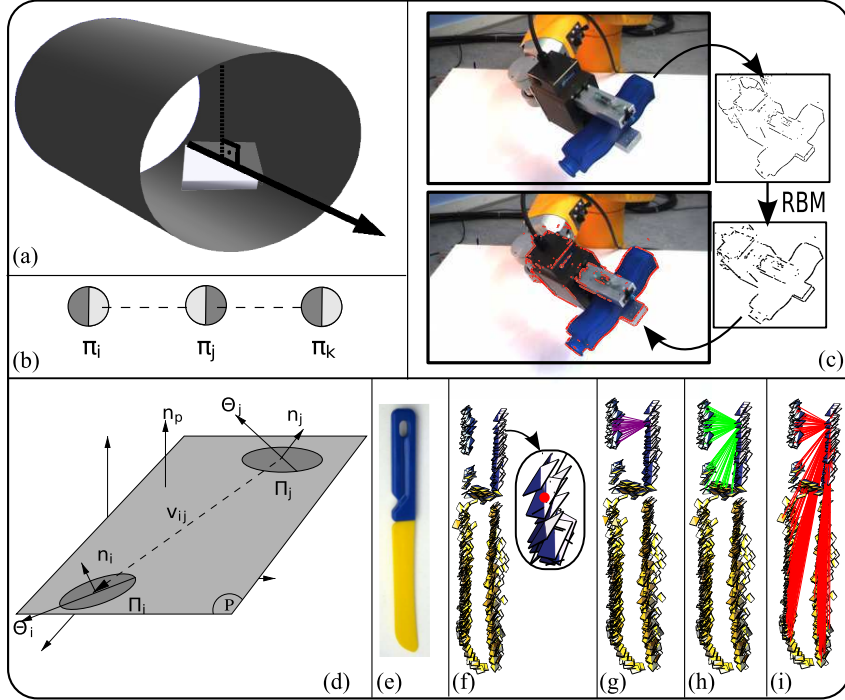


Fig. 3. Sample relations between edge primitives and their illustration on an example. (a) Normal distance. (b) Cocolarity. (c) Rigid body motion (d) Coplanarity. (e-f) Image and the 3D primitives of a sample object. A selected primitive is shown magnified. (g) All primitives that have a normal distance less than 1.5 cm to the selected primitive. (h) All primitives that are cocolor with the selected primitive. (i) All primitives that are coplanar with the selected primitive.

reliably by stereo correspondence. However, also irregular structures (e.g., trees) in 3D create 2D textures. Hence, a 3D representation of the geometric information probably also requires at least two different descriptors (surface patch and irregular structure). The descriptors used in our system are exemplified in Fig. 2.

Property 4. ECV disambiguates visual information. It is known that local processes extracting image information in all domains (edge and junction detection, stereo, optic flow, etc.) face a high degree of uncertainty due to various reasons ranging from the actual noise in the recording process (due to factors such as low contrast and motion blur) to fundamental problems such as the correspondence problem in stereo and optic flow computation. It is evident that by local processing, these ambiguities in general cannot be resolved. However, by means of the *contextual embedding*, the inherent ambiguity of locally extracted information (in particular dominant on the level of EV, see, e.g., [26]) becomes reduced in the ECV system. For example, in Fig. 6, 3D information for edge

and junction descriptors extracted from stereo (b) is compared to a temporally disambiguated representation (c).

More specifically, the contextual embedding is based on two kinds of to-be-distinguished regularities, namely *deterministic* and *probabilistic* regularities [27]. While deterministic regularities are relatively easy to formulate analytically (e.g., basic laws of Euclidean geometry), statistical regularities express probabilistic relations between visual events (corresponding to the well-known Gestalt laws, see, e.g., [28–30]). The increased predictivity of the local descriptors in combination with the contextual relations based on these two regularities allow for the utilisation of the high degree of redundancy in visual data to substitute unreliable locally extracted information by contextually verified information.

Property 5. ECV represents 2D and 3D aspects as well as their contexts in a hierarchical congruent way. The condensation of local information in 2D and 3D takes place for *four different kinds of image structures* that coexist in natural images: textured areas, junctions, edges and homogeneous image patches (see Fig. 1 and 2). This differentiation comes naturally from the different semantic contents in the 2D image domain as well as the underlying 3D spatial domain corresponding to these different structures (a first investigation of the dependency between 2D image structures and underlying depth information can be found at [23, 31–33]). As a consequence, for the representation of these image structures, different kinds of local symbolic descriptors are required (as outlined in Fig. 1) which are engaged in rather different contexts and embeddings to the same or the other kinds of structural descriptors (in Fig. 5, these contexts are represented by arrows, the circular one denoting the context between the same kind of symbolic descriptor).

For example, to describe the change of a junction under a rigid body motion, essentially the movement of a point needs to be represented while the rigid body motion of an edge segment needs to take the aperture problem [34] into account. In addition, edges and junctions can have a very different role for motion estimation (for a detailed discussion see, e.g., [13, 35, 36]) in terms of strength, frequency of occurrence and precision. Firstly, point correspondences give much stronger constraints than edge correspondences (see, e.g., [36]). Secondly, the frequency of occurrence of edges in natural scenes [37] is much higher than the frequency of occurrence of reliable point features. Thirdly, edges can be localized with high precision due to the redundancy in the intrinsically one dimensional signal structure. Moreover, the global embedding of these descriptors into spatially extended units (required for more global relational reasoning processes) is very different for different kinds of descriptors. For example, edge-primitives are embedded in contours while texlets are usually embedded in surfaces and junctions are natural end-points of contours. In Fig. 5, this is indicated by the second box around the different descriptors. As a consequence, in our ECV approach, the visual scene representation for the different image structures (see Fig. 1 and

Fig. 2) and their associated contexts (Fig. 4 exemplifies the edge context) differs depending on the local signal structure⁴.

Disambiguation is realized by feedforward processes, feedback processes across levels as well as processes that reach down to modulate EV processes. While feedforward processes lead to higher levels of abstraction (e.g., grouping of local edge primitives into 2D and 3D contours, see Fig. 4), feedback processes across levels of representation stabilise and disambiguate information on the same (symbolic) level (for example in Fig. 6, the accumulation of edge and junction primitives based on motion information is demonstrated [40]). Moreover, global reasoning on higher level symbolic entities can be used to disambiguate early visual processes by re-translating the disambiguated symbolic information into the signal domain and merging it with the EV information. Due to this transition of symbolic information to the signal-level, this process has been called ‘Signal-Symbol Loop’ [17, 18]. To realise these kinds of processes, the ECV system possesses hierarchical congruency, i.e., it allows for a controlled (and reversible) mapping between the different levels of abstraction (e.g., local versus global, 2D versus 3D, symbolic versus signal based).

Property 6. ECV reflects the bias/variance dilemma. It is known that complex learning tasks cannot be solved without including a certain degree of bias in the representations. This somehow frustrating insight was summarized in the so-called Bias/Variance Dilemma [41]: Although large degrees of freedom in the actual learning algorithm would lead to a principal ability to deal with any kind of learning problem, the actual data and amount of learning cycles required to allow for a suitable generalisation is beyond any reasonable limits. Hence, a careful selection of prior knowledge, guided by the genetically coded structures in the human visual system, is a crucial part of the ECV system. In our system, this concerns in particular the design of the local feature descriptors (as discussed above) as well as the contextual relations utilised. Interestingly, there are indications that competences based on *deterministic* regularities are likely to correspond to innate structures in the brain, whereas competences based on *statistical* regularities, significant learning during the first year of human development plays a crucial role (for a more detailed discussion, see [27, 42]).

3 Discussion

Fundamentally, computer vision is all about recovering information about a 3D world using 2D image information. In the early days, much effort concentrated on vision systems able to localise in 2D images objects characterized by known 3D contour models [43–45], to interpret 3D scenes in terms of domain knowledge [46], etc. While the geometric reasoning, in principle, is straightforward, these systems did not prevail, largely because it proved difficult to reliably extract

⁴ In our system, this is done by making use of the concept of the intrinsic dimensionality of the local image signal [38] which is an extension of the so-called ‘Harris operator’ [39].

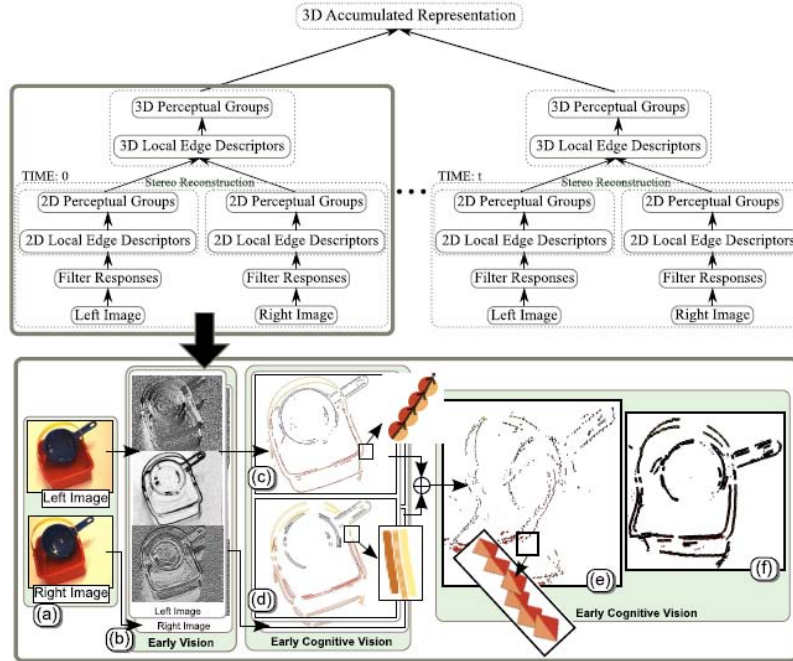


Fig. 4. Hierarchy of the ECV system in the edge domain and its temporal embedding: The hierarchy in the visual representation. (a) Stereo image pair, (b) Filter responses, (c) 2D primitives, (d) 2D contours, (e) 3D primitives, (f) 3D contours.

object contours from natural images. Arguably, the representations they used were too far removed from the available visual stimulus.

With the introduction of appearance-based methods, this gap was effectively closed by bringing the model description all the way down to the image level. This resulted in breakthroughs in a number of important computer vision applications, such as object detection, recognition and tracking. The problem of identifying specific object instances, even under clutter and illumination variations, is widely considered a solved problem. Therefore, the community has moved to object categories and the detection and segmentation of arbitrary instances of these categories. Progress in this area is traced in recent years by the Visual Object Classes (VOC) challenges organised by the European PASCAL network of excellence and its successor [47]. Current computer vision research appears to be dominated by modern evolutions of appearance-based methods that rely on local features and their statistics together with powerful techniques for statistical machine learning (see any contributions to contemporary PASCAL VOC Challenges). Some methods use various degrees of spatial (part-based models, constellation models) [48, 49] or conceptual structure (probabilistic latent semantic analysis, latent Dirichlet allocation) [50, 51].

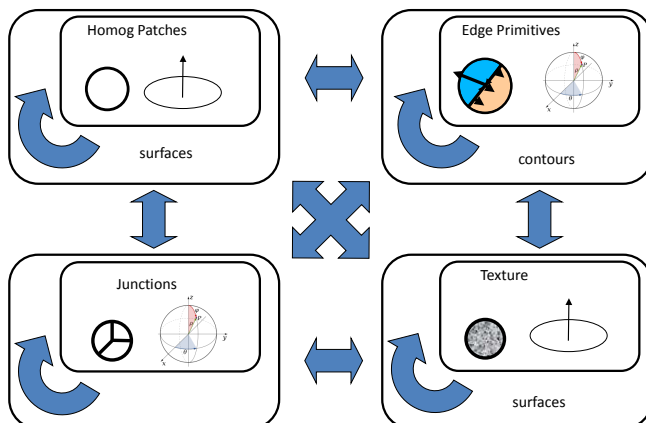


Fig. 5. The four different kinds of image structures are represented by four different symbolic descriptors which themselves are embedded in different kinds of contexts as well as of more global structures. For example, *edge primitives* are embedded in *contours* (embedding into more global structure). Moreover, junctions are natural endings of contours (denoted by the arrow between junctions and contours) and contours are the boundaries for surfaces (embedding into different kinds of contexts).

While the VOC Challenges showcase remarkable progress, they also highlight what we consider a fundamental shortcoming of current know-how in computer vision with respect to real-world visual tasks: Vision problems—involving 2D projections of a structured 3D world—are addressed using methods that remain entirely in 2D and do not capture the 3D nature of the scene. Moreover, most methods capture very little structure at all, which is unsurprising since the original 3D structure is generally much simpler than its 2D projection—a single 3D structure can give rise to a variety of distinct 2D structures under different projections. In the VOC Challenges, this results in consistently low performance on objects that are better described by structure than by appearance such as furniture, bottles, and potted plants [52]. Thus, modern methods have lost their structural, representational power.

To build more general vision systems, the representational gap between the 2D image array and 3D structure must be bridged, a problem that has been mostly overlooked by modern computer vision research [53]. The ECV system we propose represents one step in this direction by providing a richly structured, 3D representation at an intermediate level of abstraction, situated between the pixel and the semantic levels.

Vision research has been influenced substantially by David Marr’s paradigm [54]. One of Marr’s main contributions was to combine the findings and the theories of his time from neurophysiology, psychology and artificial intelligence into a coherent and complete vision theory. As Marr’s approach, the design of our ECV system is motivated and guided by knowledge about the human visual system in several aspects. Firstly, the visual modalities represented in the

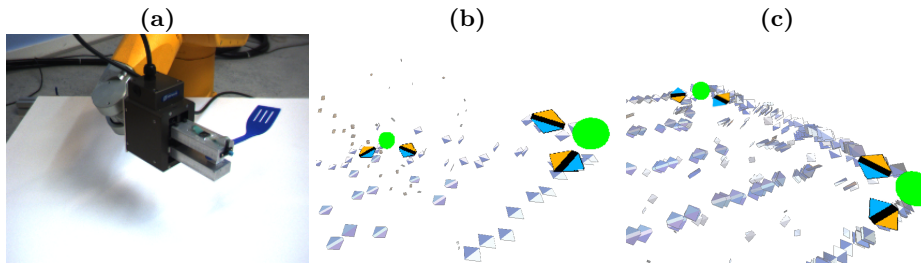


Fig. 6. Disambiguation by temporal accumulation of 3D edge and junction descriptors. (a) left input image (b) 3D reconstruction from stereo (c) accumulated representation after 72 frames.

different dimensions of the local descriptors (orientation, phase, stereo and optic flow in a harmonic representation [8] and a loosely coupled color representation) are motivated by the visual features processed in the areas of V1 as well as their organization in terms of hyper-columns [55]. Secondly, this feature space has a large degree of hard-coded structure in terms of local feature extraction [56] as well as deterministic regularities [27] corresponding to well-documented innate structures in the human visual system (for a more in-depth discussion see [27, 42]). Thirdly, the EV and ECV layers constitute a hierarchy in analogy to the human visual system where abstraction of information and extent of supporting region are increased, whereas ambiguity is decreased upwards the visual pathway [20], this culminates into the integrated, robust representation of knowledge required by higher cognitive functions.

A major difference to Marr's approach is that ECV explicitly account for the ambiguity of visual information at early stages and uses of disambiguation processes to arrive at more reliable information at higher levels. The complexity of the processes involved is huge. Today, we can make use of a large body of work of the computer vision community addressing sub-aspects of such a representational hierarchy (e.g., processing of visual modalities such as stereo and optic flow, reliable parametrisation of extraction of 3D entities of different complexity, etc.). Forty years ago, Marr's paradigm was simply not realisable. However, new insights in computational theory and efficient algorithms that have been developed for early vision processes as well as the increased computational power (also supported by special hardware such as GPUs [57]) allow for the realisation of processes with a complexity as required in the ECV system; hence, we suggest that Marr's ideas should be revisited under the light of developments in vision research during the last 40 years.

Another important aspect, ECV provides a natural interface for 'Vision for Action'. Above, we emphasized that the ECV representation is task independent. As it is the general visual front-end, it also needs to be rich enough to support the full set of CV tasks to be addressed by humans. When performing a 'task', the cognitive agent will receive continuous visual feedback about its own doing.

Thus, the ECV system not only has to support different tasks but it also has to be able to dynamically react within such a feedback loop. At the cognitive top level, tasks can often be defined and represented in a very abstract way (like phrasing a task as a sentence). The actual execution of any such task will, however, usually require to visually analyse multiple aspects of the scene, like the relations of visual entities in 3D (pointing to object relations), or the patterns in which certain entities move, etc. Most of the time, such a multi-factorial analysis has to be performed in parallel at the same time. Thus, the agent needs to have access to a rich representation, which allows access to such different aspects, while at the same time, the representation needs to be condensed to a degree which makes processing efficient. Both—richness and condensation—are fundamental attributes of the ECV representation as described above. Hence, ECV can form an ideal interface to allow for vision for action. Not surprisingly, many of the applications of our ECV system are linked to robot actions (see, e.g., [10, 11, 58, 59])

Our concrete implementation of an ECV vision system [14–16] is one attempt at the design of an ECV system that has demonstrated its usefulness in a number of applications (see, e.g., [10–13]). We are currently in the process of making this system accessible to the community⁵. However, there are still significant gaps to be filled to arrive at a complete ECV system such as a full integration of all descriptors and their relations. Also, although, for some steps, real-time processing has already been achieved [60], significant work has to be done until the full complexity of the system is realised in a sufficient frame-rate.

Acknowledgement

This research has been funded by the the Interreg project IRFO (Intelligent Robots for handling of flexible Objects).

References

1. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* **2** (2004) 91–110
2. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27** (2005) 1615–1630
3. Marszaek, M., Schmid, C.: Spatial weighting for bag-of-features. In: *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Volume 2.* (2006) 2118–2125
4. Torralba, A., Murphy, K., Freeman, W.: Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29** (2007) 854–869

⁵ The ECV system is realised in the C++ library CoViS (Cognitive Vision Software) which will be put under a BSD license. We are currently in the process of testing relevant demo programs and extending the documentation. We expect a first release in autumn 2010. Scientist who are interested in using the CoViS already now please contact info@covig.org

5. Fidler, S., Leonardis, A.: Towards scalable representations of object categories: Learning a hierarchy of parts. In: IEEE CVPR. (2007)
6. Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision* **77** (2008) 259–289
7. Granlund, G.: In search of a general picture processing operator. *Computer Graphics and Image Processing* **8** (1978) 155–173
8. Sabatini, S.P., Gastaldi, G., Solari, F., Diaz, J., Ros, E., Pauwels, K., Hulle, K.M.M.V., Pugeault, N., Krüger, N.: A compact harmonic code for early vision based on anisotropic frequency channels. *Computer Vision and Image Understanding* **114** (2010) 681–699
9. Jones, J., Palmer, L.: An evaluation of the two dimensional Gabor filter model of simple receptive fields in striate cortex. *Journal of Neurophysiology* **58** (1987) 1223–1258
10. Popović, M., Kraft, D., Bodenhagen, L., Bašeski, E., Pugeault, N., Kragic, D., Asfour, T., Krüger, N.: A strategy for grasping unknown objects based on co-planarity and colour information. (*Robotics and Autonomous Systems*) (Accepted).
11. Kraft, D., Pugeault, N., Bašeski, E., Popović, M., Kragic, D., Kalkan, S., Wörgötter, F., Krüger, N.: Birth of the Object: Detection of Objectness and Extraction of Object Shape through Object Action Complexes. Special Issue on “Cognitive Humanoid Robots” of the *International Journal of Humanoid Robotics* **5** (2009) 247–265
12. Detry, R., Pugeault, N., Piater, J.: A probabilistic framework for 3d visual object representation. *IEEE transactions on Pattern Analysis and Machine Intelligence* **31** (2009) 1790–1803
13. Pilz, F., Pugeault, N., Krüger, N.: Comparison of point and line features and their combination for rigid body motion estimation. *Statistical and Geometrical Approaches to Visual Motion Analysis*, Springer LNCS 5604 (2009)
14. Krüger, N., Lappe, M., Wörgötter, F.: Biologically Motivated Multi-modal Processing of Visual Primitives. *Interdisciplinary Journal of Artificial Intelligence & the Simulation of Behaviour*, AISB Journal **1** (2004) 417–427
15. Bašeski, E., Pugeault, N., Kalkan, S., Bodenhagen, L., Piater, J.H., Krüger, N.: Using Multi-Modal 3D Contours and Their Relations for Vision and Robotics. *Journal of Visual Communication and Image Representation* (**accepted**) (2009)
16. Pugeault, N., Wörgötter, F., Krüger, N.: Visual primitives: Local, condensed, and semantically rich visual descriptors and their applications in robotics. *International Journal of Humanoid Robotics* (Special Issue on Cognitive Humanoid Vision) (accepted)
17. Kalkan, S., Yan, S., Krger, V., Wörgötter, F., Krüger, N.: A signal-symbol loop mechanism for enhanced edge extraction. *International Conference on Computer Vision Theory and Applications (VISAPP)* (2008)
18. Ralli, J., Diaz, J., Kalkan, S., Kruger, N., Ros, E.: Disparity disambiguation by fusion of signal- and symbolic-level information. *Machine Vision and Applications* (in press)
19. Lettvin, J.Y., Maturana, H.R., McCulloch, W.S., Pitts, W.H.: What the frog’s eye tells the frog’s brain. *Proceedings of the Institute of Radio Engineers* **47** (1959) 1950 – 1961
20. Wurtz, R., Kandel, E.: Central visual pathways. In Kandell, E., Schwartz, J., Messel, T., eds.: *Principles of Neural Science* (4th edition). (2000) 523–547

21. König, P., Krüger, N.: Perspectives: Symbols as self-emergent entities in an optimization process of feature extraction and predictions. *Biological Cybernetics* **94** (2006) 325–334
22. Grimson, W.: Surface consistency constraints in vision. *CVGIP* **24** (1983) 28–51
23. Kalkan, S., Wörgötter, F., Krüger, N.: Statistical analysis of local 3d structure in 2d images. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2006) 1114–1121
24. Granlund, G.H., Knutsson, H.: *Signal Processing for Computer Vision*. Kluwer Academic Publishers, Dordrecht (1995)
25. Kovese, P.: Image features from phase congruency. *Videre: Journal of Computer Vision Research* **1** (1999) 1–26
26. Aloimonos, Y., Shulman, D.: *Integration of Visual Modules — An extension of the Marr Paradigm*. Academic Press, London (1989)
27. Krüger, N., Wörgötter, F.: Statistical and deterministic regularities: Utilisation of motion and grouping in biological and artificial visual systems. *Advances in Imaging and Electron Physics* **131** (2004) 82–147
28. Ellis, W., ed.: *Gestalt Theory, A source book for Gestalt Psychology*. (1938)
29. Koffka, K.: *Principles of Gestalt Psychology*. Lund Humphries, London (1935)
30. Köhler, K.: *Gestalt Psychology: An introduction to new concepts in psychology*. New York: Liveright (1947)
31. Huang, J., Lee, A., Mumford, D.: Statistics of range images. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2000) 1324–1331
32. Yang, Z., Purves, D.: Image/source statistics of surfaces in natural scenes. *Network: Computation in Neural Systems* **14** (2003) 371–390
33. Potetz, B., Lee, T.S.: Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes. *Journal of the Optical Society of America* **20** (2003) 1292–1303
34. Stumpf, P.: Über die Abhängigkeit der visuellen Bewegungsrichtung und negativen Nachbildes von den Reizvorgängen auf der Netzhaut. *Zeitschrift für Psychologie* **59** (1911) 321–330
35. Ansar, A., Daniilidis, K.: Linear pose estimation from points or lines. In: *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV*, London, UK, Springer-Verlag (2002) 282–296
36. Rosenhahn, B., Sommer, G.: Adaptive pose estimation for different corresponding entities. In van Gool, L., ed.: *Pattern Recognition, 24th DAGM Symposium*. Springer Verlag (2002) 265–273
37. Zetsche, C., Barth, E.: Fundamental limits of linear filters in the visual processing of two dimensional signals. *Vision Research* **30** (1990) 1111–1117
38. Felsberg, M., Kalkan, S., Krüger, N.: Continuous dimensionality characterization of image structures. *Image and Vision Computing* **27** (2009) 628–636
39. Harris, C.G., Stephens, M.: A combined corner and edge detector. In: *4th Alvey Vision Conference*. (1988) 147–151
40. Pugeault, N., Wörgötter, F., Krüger, N.: Accumulated Visual Representation for Cognitive Vision. In *Proceedings of the British Machine Vision Conference (BMVC)* (2008)
41. Geman, S., Bienenstock, E., Doursat, R.: Neural networks and the bias/variance dilemma. *Neural Computation* **4** (1995) 1–58
42. Kellman, P., Arterberry, M.: *The Cradle of Knowledge*. MIT-Press (1998)
43. Huttenlocher, D.P., Ullman, S.: Object recognition using alignment. In: *First International Conference on Computer Vision*. (1987) 102–111

44. Lamdan, Y., Wolfson, H.J.: Geometric hashing: A general and efficient model-based recognition scheme. In: Second International Conference on Computer Vision. (1988) 238–249
45. Brooks, R.: Model-based 3-D interpretations of 2-D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **5** (1983) 140–150
46. Firschein, O., ed.: *RADIUS: Image Understanding for Imagery Intelligence*. Morgan Kaufmann, San Francisco (1997)
47. Everingham, M., Zisserman, A., Williams, C.K.I., van Gool, L., Allan, M., Bishop, C.M., Chapelle, O., Dalal, N., Deselaers, T., Dorko, G., Duffner, S., Eichhorn, J., Farquhar, J.D.R., Fritz, M., Garcia, C., Griffiths, T., Jurie, F., Keysers, D., Koskela, M., Laaksonen, J., Larlus, D., Leibe, B., Meng, H., Ney, H., Schiele, B., Schmid, C., Seemann, E., Taylor, J.S., Storkey, A., Szedmak, S., Triggs, B., Ulusoy, I., Viitaniemi, V., Zhang, J.: The 2005 PASCAL Visual Object Classes Challenge. In: *Pascal Challenges Workshop*. Volume 3944 of *LNAI*. Springer (2006) 117–176
48. Felzenszwalb, P., McAllester, D., Ramaman, D.: A discriminatively trained, multiscale, deformable part model. In: *IEEE International Conference on Computer Vision and Pattern Recognition*. (2008)
49. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: *IEEE International Conference on Computer Vision and Pattern Recognition*. (2003)
50. Sivic, J., Russell, B., Efros, A., Zisserman, A., Freeman, W.: Discovering objects and their locations in images. In: *International Conference on Computer Vision*. (2005) 370–377
51. Fei-Fei, L., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: *IEEE Conference on Computer Vision and Pattern Recognition*. (2005)
52. Everingham, M., Gool, L.V., Williams, C., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2009 (VOC2009). Summary presentation at the 2009 PASCAL VOC workshop (2009)
53. Dickinson, S.: The evolution of object categorization and the challenge of image abstraction. In Dickinson, S., Schiele, B., Tarr, M., eds.: *Object Categorization: Computer and Human Vision Perspectives*. Cambridge University Press (2009) 1–37
54. Marr, D.: *Vision*. Freeman (1982)
55. Hubel, D., Wiesel, T.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiology* **160** (1962) 106–154
56. Wiesel, T., Hubel, D.: Ordered arrangement of orientation columns in monkeys lacking visual experience. *J. Comp. Neurol.* **158** (1974) 307–318
57. Che, S., Boyer, M., Meng, J., Tarjan, D., Sheaffer, J.W., Skadron, K.: A performance study of general-purpose applications on graphics processors using cuda. *J. Parallel Distrib. Comput.* **68** (2008) 1370–1380
58. Kraft, D., Detry, R., Pugeault, N., Başeski, E., Piater, J., Krüger, N.: Learning objects and grasp affordances through autonomous exploration. In: *International Conference on Computer Vision Systems (ICVS)*. (2009)
59. Detry, R., Kraft, D., Buch, A.G., Krüger, N., Piater, J.: Refining grasp affordance models by experience. *International Conference on Robotics and Automation* (2010)
60. Jensen, L.B.W., Kjær-Nielsen, A., Pauwels, K., Jessen, J.B., Hulle, M.V., Krüger, N.: A two-level real-time vision machine combining coarse and fine grained parallelism. *Journal of Real-Time Image Processing* (accepted)