Using 3D Contours and Their Relations for Cognitive Vision and Robotics

Emre Başeski*, Leon Bodenhagen*, Nicolas Pugeault*, Sinan Kalkan†, Justus Piater‡ and Norbert Krüger*
*The Mærsk Mc-Kinney Møller Institute, University of Southern Denmark, Odense, Denmark
Email: {emre,leon,nicolas,norbert}@mmmi.sdu.dk
†Bernstein Center for Computational Neuroscience, University of Göttingen, Göttingen, Germany
Email: sinan@bccn-goettingen.de
‡Department of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium
Email: Justus.Piater@ULg.ac.be

***Abstract*—In this work, we make use of 3D visual contours carrying geometric as well as appearance information. Between these contours, we define 3D relations that encode structural information relevant to object-level operations such as similarity assessment and grasping. We show that this relational space can also be used as input features for learning which we exemplify for the grasping of unknown objects. Our representation is motivated by the human visual system in two respects. First, we make use of a visual descriptor that is motivated by hyper-columns in V1. Secondly, the contours can be seen as one stage in a visual hierarchy bridging between local symbolic descriptors to higher level stages of processing such as object coding and grasping.**

I. Introduction

The human visual system is characterized by a hierarchy in which entities of increasing complexity are processed (see, e.g., [1]). At the first cortical stage, different aspects of visual information in terms of a variety of visual modalities (e.g., orientation, color and local motion) are computed in so-called hypercolumns [2]. This local information is embedded into the spatio-temporal context at intermediate stages through mechanisms of perceptual organization, resulting in spatially extended visual descriptors that code semi-global shapes such as surfaces and contours. This coding allows for the detection and utilization of even more global relations between visual entities, leading to a semantic understanding of the visual scenes. In this work, we want to mimic this hierarchical scheme in the edge domain. Starting with functional abstractions of hypercolumns in terms of multi-modal[1] local 2D and 3D descriptors [3]. We group these descriptors into multi-modal 2D and 3D contours which we then apply in different tasks in the context of vision and vision-based robotics.

Recent progress in the research of salient 2D local features (see, e.g., [4], [5], [6]) has provided successful and robust results in certain contexts such as object identification and classification. On the other hand, they heavily rely on texture and do not typically give information about the shape of the object (see, e.g., [7]). Their usage is also limited when objects need to be manipulated. These drawbacks can be compensated by using 3D global entities such as visual contours and their relations [8], [9], [10], [11].

As discussed in, e.g., [10], global entities complement the local approaches by providing a more global overview of the scene, which makes reasoning about the geometry and shape easier. Also, once the global reasoning processes take place in 3D, they become independent from viewpoint and transformations. There is substantial evidence from human vision showing the importance of contours and relations between them. For example, in [12] and [13] it has been shown that scrambled objects are difficult to recognize. Moreover, certain relations between groups of features are found to be salient [14] and shown to be non-accidental (see, e.g., [10]).

Visual contours and their relations have been used in computer vision and robotics in various contexts. For example, in [8] and [15], visual contours have been used for object recognition and classification (see, [9] for a recent review). Their relations also have been utilized as features for object recognition (see, e.g., [16]). Similarly, Henricsson [17] makes use of geometric relations such as proximity, curvilinearity and symmetry between contours to describe objects by combinations of these relations. These non-accidental contour relations have also been utilized by Dickinson et al. [18] to create *aspect hierarchies* which are then used for recovery and recognition of 3D objects from a single 2D image.

A popular approach for dealing with the relational space is the relational histogram (see, e.g., [19], [20]) where the histogram bins count certain combinations of relations. Once the objects are encoded as relational histograms, their similarity can be measured by simple histogram intersection [21]. Such histograms then can be used for comparing objects.

The aim of this article is to discuss the importance of 3D contours and second-order 3D contour relations. We introduce four relations, namely, co-planarity, co-colority, normal distance and angle that are important in the context of robotics and vision. While more relations can be defined for different contexts, the aim of this article is not to give a complete set of relations but to show the importance of 3D contours and their relations in different domains. We show that the relational space can provide a basis for discriminating different object structures and can be used for the definition of visual feature-grasp associations. In this context, we also discuss the robustness of computation when using 3D contours instead of local descriptors. Moreover, we show that through learning relevant aspects in the space of contour relations the success rate of grasping can be increased, indicating that contours and their relations span a relevant space for learning.

The distinguishing feature of our approach can be summa-

[1]Note that here we refer to different visual modalities and not different sensory modalities

rized as follows:

- We make use of 3D relations while keeping the possibility of accessing the 2D information from which the 3D information is calculated. Although a lot of non-accidental relations can be defined in 2D (see, e.g., [10]), relations like co-planarity or distance heavily depend on 3D information.
- The contour representation we use covers geometric as well as appearance information (as a straightforward extension of the local entities from which they are computed).
- The geometric and appearance information encoded in the 3D contours leads to semantically rich dependencies between scene areas that can be expressed in 3D relations between these contours.

The rest of the article is organized as follows. In Section II, the visual representation used in this work is briefly explained. After the discussion about the object structure encoding in Section III-A, the application of the relational space on grasping unknown objects and learning how to refine grasping is presented in Section III-B. We conclude with a discussion on the potential of 3D contours and their relations in Section IV.

II. Visual Representation

A. Local Multi-modal Edge Descriptors

In this work we make use of a visual representation based on an Early Cognitive Vision (ECV) system which creates local edge descriptors called primitives [3], [22], [23]. They are extracted sparsely along image contours and form a feature vector that contains visual modalities such as position, orientation, phase, color and optical flow ($\pi = (\mathbf{x}, \theta, \phi, (\mathbf{c_l}, \mathbf{c_m}, \mathbf{c_r}), \mathbf{f})$) where the color of a patch is defined by left, right and middle color (see Fig. 1(c)). An important property of the primitives is that they encode geometric information (position and orientation) as well as appearance information (color and phase) as separate modalities. 2D-primitives are matched across two stereo views. Such pairs of corresponding primitives afford the reconstruction of a 3-dimensional equivalent called 3D-primitive which is encoded by the vector $\Pi = (\mathbf{X}, \Theta, \Phi, (\mathbf{C_l}, \mathbf{C_m}, \mathbf{C_r}))$. The extraction process of 2D and 3D primitives for a sample stereo image pair is illustrated in Fig. 1.

During the reconstruction of 3D entities from 2D stereo data, the uncertainty of 2D data propagates via the operations that are used for reconstruction. The statistical uncertainty of the reconstructed entity can be modeled, if these operations are linear and the error model for 2D data is known. For the visual features discussed above, the uncertainty calculation was shown by Pugeault et al. [24].

B. Multi-modal 3D Contours and Their Relations

The local 2D and 3D primitives are grouped together by using the perceptual organization scheme described in [25] to create semi-global contour structures (Fig. 1(d)). Since contours are based on good continuation of local primitives in terms of geometry and appearance, they also contain

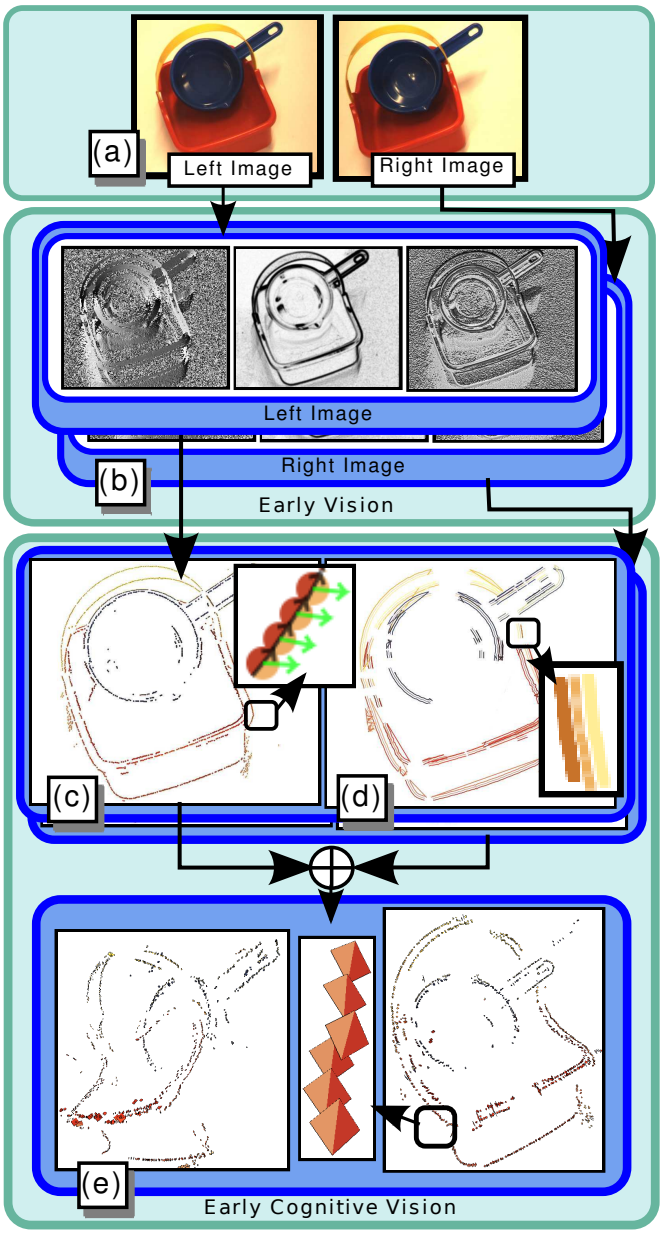

Fig. 1. An overview of the visual representation. **(a)** Stereo image pair, **(b)** Filter responses, **(c)** 2D primitives, **(d)** 2D contours, **(e)** 3D primitives.

modalities covering geometry and appearance. This sparse and symbolic nature of the primitives and the contours allows the definition of perceptual relations on them that express relevant spatial relations in 2D and 3D (e.g., co-planarity, co-colority) which can be applied in different contexts. In the rest of this section, a brief description of the relations used as visual cues in this work is given.

Co-planarity: Co-planarity of entities is measured by their distance to a common plane. An important point is that analytical operations may be misleading for 3D data produced by stereo, because of the uncertainty of the data. The co-planarity calculation benefits from a plane fitting algorithm based on *renormalization* [26] that takes the uncertainty of

Fig. 2. Illustration of 3D relations. **(a)** All 3D contours of the object in Fig. 1(a). **(b)** A selected 3D contour shown in red. **(c)** All contours that are co-planar to the selected contour are shown in green. **(d)** All contours that have a distance of less than or equal to 30 mm to the selected contour are shown in green. **(e)** All contours that have color difference of less than or equal to 5 to the selected contour are shown in green.

the data into account. The co-planarity between i^{th} and j^{th} contour in the scene is defined as:

$$R_C = \frac{1}{N}\sum_{k=0}^{N}(\mathbf{p}_i^k)^T(\Lambda_i^k)^{-1}(\mathbf{p}_i^k) + \frac{1}{M}\sum_{l=0}^{M}(\mathbf{p}_j^l)^T(\Lambda_j^l)^{-1}(\mathbf{p}_j^l), \tag{1}$$

where N and M are the numbers of primitives in the i^{th} and j^{th} contours respectively, $\mathbf{p}_i^k = (\mathbf{x}_i^k - \hat{\mathbf{x}}_i^k)$, Λ_i^k is the uncertainty matrix of the k^{th} feature in the i^{th} contour, $\mathbf{x}_i^k$ is the position vector of it and $\hat{\mathbf{x}}_i^k$ is the closest point to $\mathbf{x}_i^k$ on the plane that has been fit to feature locations of the i^{th} and j^{th} contours by renormalization. An example is shown in Fig. 2(c).

Normal Distance: The distance between two 3D contours is defined by the distance of the centroid of one contour to the line created by the orientation of the principal component[2] and the centroid of the other. Therefore, the distance between i^{th} and j^{th} contour in the scene is defined as:

$$R_D = \frac{|\mathbf{w}_i - (\mathbf{w}_i \cdot \mathbf{u}_i)\mathbf{u}_i| + |\mathbf{w}_j - (\mathbf{w}_j \cdot \mathbf{u}_j)\mathbf{u}_j|}{2}, \tag{2}$$

where $\mathbf{w}_i$ is $(\mathbf{k}_j - \mathbf{k}_i)$, $\mathbf{w}_j$ is $(\mathbf{k}_i - \mathbf{k}_j)$, $\mathbf{k}_i$ is the centroid and $\mathbf{u}_i$ is the principal component of the i^{th} contour. An example is presented in Fig. 2(d). Note that this is a parallel distance, and even if a contour is split into small parts it does not change, which makes the definition independent from contour length.

Angle: Similar to distance, angle between two contours is defined by using the principal components of the contours as:

$$R_A = \arccos\left(\frac{\mathbf{u}_i \cdot \mathbf{u}_j}{|\mathbf{u}_i||\mathbf{u}_j|}\right) \tag{3}$$

Co-colority: Since contours are created from primitives with good continuation in terms of geometry and color, similar color needs to be shared among those primitives that create the contour. Therefore, similar to primitives, every contour has mean left, right and middle colors as well. Co-colority of two contours (R_{Cl}) is defined as the color difference between the mean colors of the contours' sides that are facing each other. The color difference is calculated using the CIE 1994 color difference formula [27]. A mean color is calculated for the contours' sides that are facing each other and the color difference is then used to estimate how reliable the mean color is (see Fig. 2(e) for an example).

[2]The eigen-vector of the highest eigen-value in PCA.

III. Using 3D contours in cognitive vision and grasping tasks

In this section, we want to exemplify the use of our multi-modal contours and their relations. We have chosen two very different domains – object encoding and grasp learning – to demonstrate the general usability of the representation.

A. Object Structure Encoding via Relations

In this section, we discuss encoding the structure of an object as a relation histogram to show the potential of the relational space in a visual task. The main idea behind relation histograms is to split objects into parts that are both geometrically and visually similar. Therefore, the histograms have axes that contain both geometrical and appearance based information. We selected co-planarity, normal distance, and (L,a,b) components of the mean color as the axes for the relation histograms. Note that, since the color is encoded with three components, the relation histograms are 5-dimensional. Also, to obtain visually similar parts, instead of calculating the histograms from every feature pair, only co-color pairs are used.

The similarity between two histograms is calculated using a histogram intersection technique introduced in [21]. The intersection between normalized histograms H_1 and H_2 can be formulated as

$$\mathcal{D}(H_1, H_2) = \sum_{i=1}^{n}\min(H_1(i), H_2(i)), \tag{4}$$

where n is the total number of bins and $H_j(i)$ is the value stored in the i^{th} bin of the j^{th} histogram. For normalized histograms, a perfect match results in 1 and a total mismatch results in 0. One may of course use more sophisticated methods for comparing two objects in relational space, but metrics are not the focus of this paper. Our aim in this section is to demonstrate the potential of 3D contour relations.

50 relation histograms were created based on a set of 10 objects (see Fig. 3), viewed from 5 poses each. Each histogram is compared with all the others using Equation 4, and the resulting confusion matrix is given in Fig. 4. High values within object-aligned blocks on the diagonal in Fig. 4 indicate good matches, whereas other high values indicate either false positives or structural similarity between objects. For example,

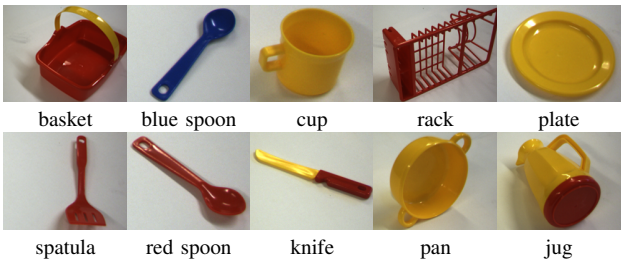

Fig. 3. Objects used in our experiments.

all poses of the object 'basket' were found to be similar to all poses of the object 'rack'. On the other hand, the relation histograms of object 'knife' responded weakly to every object in the experiment.

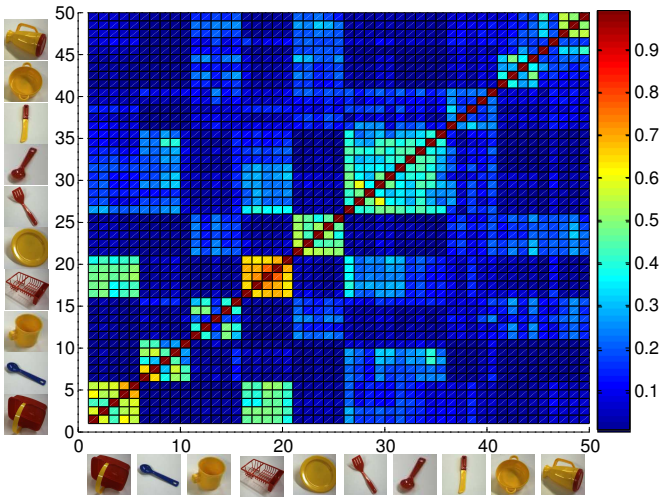

Fig. 4. Similarity measures between 5 different poses of the 10 objects in Fig. 3.

For further analysis, the 50x50 similarity matrix was reduced to a dimension of 50x2 using multi-dimensional scaling [28]. The results are shown in Fig. 5 where we observe that both appearance-based (e.g., small distance between 'spatula' and 'red spoon' clusters) as well as structural information (e.g., small distance between 'pan' and 'plate' clusters, 'red spoon' and 'blue spoon' clusters) is encoded successfully. Also, the high degree of similarity between objects in Fig. 4 (e.g., 'basket' and 'rack') appears in Fig. 5 as well as the low degree of similarity (e.g., very large spread of the 'knife' cluster).

B. Using 3D Contours for Grasping Unknown Objects

In the absence of prior knowledge about the 3D model of an object, sensory data must be used to calculate grasping hypotheses (see, e.g., [29]). Compared to grasping known objects, grasping of unknown objects is acknowledged to be a challenging problem. The main difficulty is that visual features extracted from the scene (which in general are afflicted with a large degree of noise) need to be related to grasping actions. Here we show that 3D multi-modal contours can be a powerful trigger for such grasping actions. In particular, we show that a

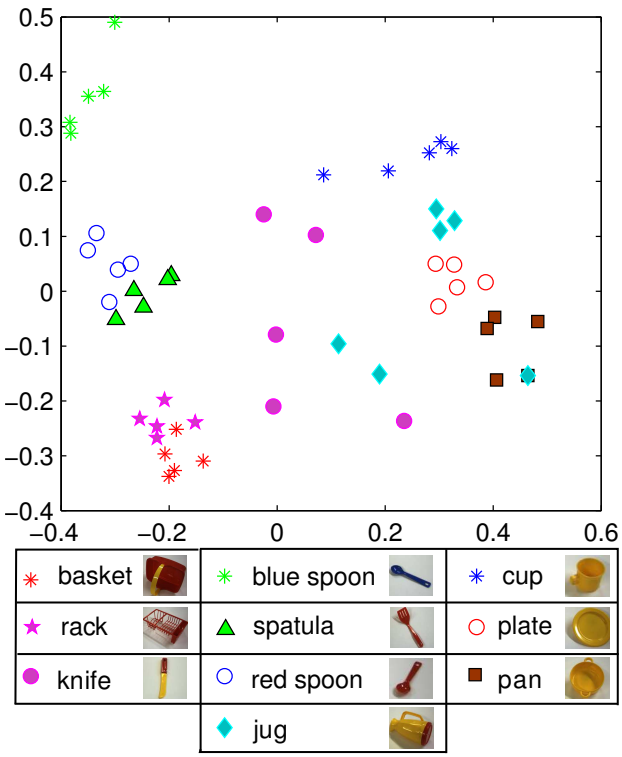

Fig. 5. Similarity maps between different poses.

more global approach based on contours outperforms a local approach based on local primitives in terms of stability and robustness.

Grasps for a two-finger gripper are defined through a 3D location and two directions (x,r_1,r_2) as shown in Fig. 6(a). Making use of the co-planarity relation as defined in Section II-B, we can associate a number of grasps to two co-planar contours or primitives (see Fig. 6(b)). In this way, we can compute a large variety of grasping hypotheses (see Fig. 6(e)) for any given object.

The grasps associated to contours are defined by using the 3D location of the central primitive of one of the contours (x) and the normal of the plane fitted to the contours (r_2). Furthermore, the second direction of the grasp, r_1, is calculated as the cross product of r_2 and the tangent vector of the contour at x (see Fig. 6(f) for an example). Those grasps that are associated to contours are specifically defined between contours that are co-color and co-planar. It has been shown that by such a straightforward association of co-planar contours to grasps, without making use of any prior object knowledge, it is already possible to achieve grasp success rates of about 40% (for details, see [30]).

In the following, we compare the performance of grasps defined by relations between contours and grasps defined by relations between local features. For the local approach, the grasping hypotheses are calculated from the location and orientation of two primitives that are co-color and co-planar as described in [29] (see Fig. 6(b) for a brief explanation).

For the comparison of the global and local approaches, two

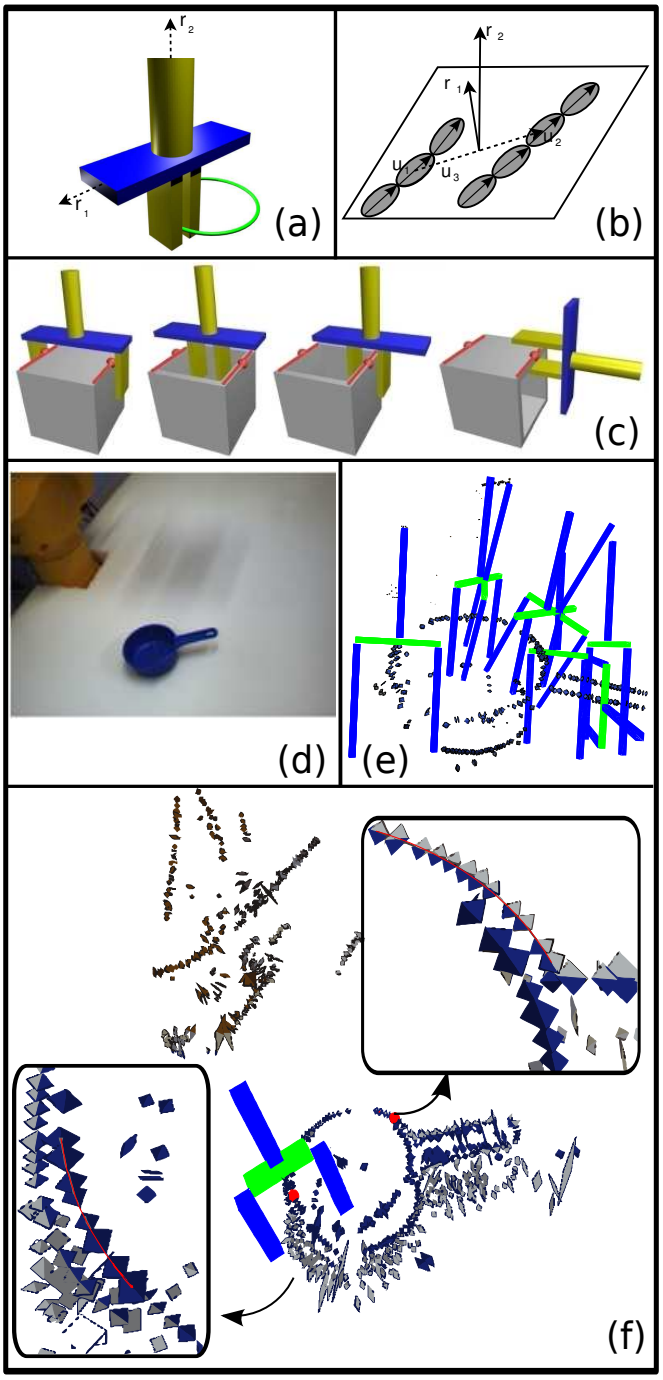

Fig. 6. Grasping with a two-fingered gripper. **(a)** A grasp is defined by a location and two directions (r_1,r_2). **(b)** The direction vectors for local features are defined as $r_2 = (u_3 \times u_1) + (u_2 \times u_3)$ and $r_1 = u_3 \times r_2$. For contours, r_2 is the normal of the common plane, and r_1 is the cross product of r_2 and the tangent vector of the contour at grasping location. **(c)** Grasping hypothesis types based on contours. **(d)** A blue pan to be grasped. **(e)** A set of grasping hypotheses generated to grasp the blue pan. **(f)** An example grasp based on contours.

contours from the brim of the can that is shown in Fig. 7(a) were manually selected to make sure that they belong to the same surface. The grasp hypotheses were calculated using both the local and global approaches while adding random noise within the uncertainty range of the primitives. As shown in Fig. 7(b) and (c), even though the best grasps were chosen for the local approach, the global approach performs significantly better. Note that this fact is comprehensible as the local approach is based on only two primitives, while the global approach makes use of the geometric stability of multiple primitives.

Besides the fact that using co-planar contours instead of co-planar local features increases the robustness and success rate of grasping unknown objects, in [30], it has been shown that the performance can be further increased by learning. To this end, we constructed an artificial neural net that predicts a success likelihood for a grasp depending on the relations of the contours defining the grasp. Three relations between contours in addition to co-planarity – namely, co-colority, normal distance, and angle – were used as input features to the neural network. Each grasp attempt was evaluated haptically by the robot, which resulted in a set of triplets containing the performed grasp, the contours defining the grasp as well as a label for success or failure. From these data, the robot system learned a function predicting the success likelihood of the grasp depending on the values of the four relations. For example, from the large number of potential grasps shown in Fig. 6(e), the robot picked the one with the highest predicted success likelihood. We have shown that by such a selection based on learning we can increase the success rate of the grasping behavior from below 40% to above 60% (for details, see [30]). Note that by learning a success likelihood for a grasp that can be associated to a certain constellation of contours, no prior knowledge pertaining to specific objects is introduced. Rather, the system acquires general knowledge about the chance of grasp success when certain sets of relations occur in the scene.

IV. Conclusions

We demonstrated the potential of using multi-modal 3D contours and their relations in the context of cognitive vision and robotics. We observed that geometrical reasoning in 3D benefits from the use of multi-modal contours and contour relations. We showed how the relational space can be used to encode the structure of a scene and to grasp unknown objects. We also showed that the same space provides a useful basis for learning, by improving the success rate of grasping of unknown objects based on these relations.

Acknowledgment

The work described in this paper was conducted within the EU Cognitive Systems project PACO-PLUS (IST-FP6-IP-027657) funded by the European Commission.

References

[1] E. Kandell, J. Schwartz, and T. Messel, *Principles of Neural Science (4th edition)*. McGraw Hill, 2000.

[2] T. Wiesel and D. Hubel, "Ordered arrangement of orientation columns in monkeys lacking visual experience," *J. Comp. Neurol.*, vol. 158, pp. 307–318, 1974.

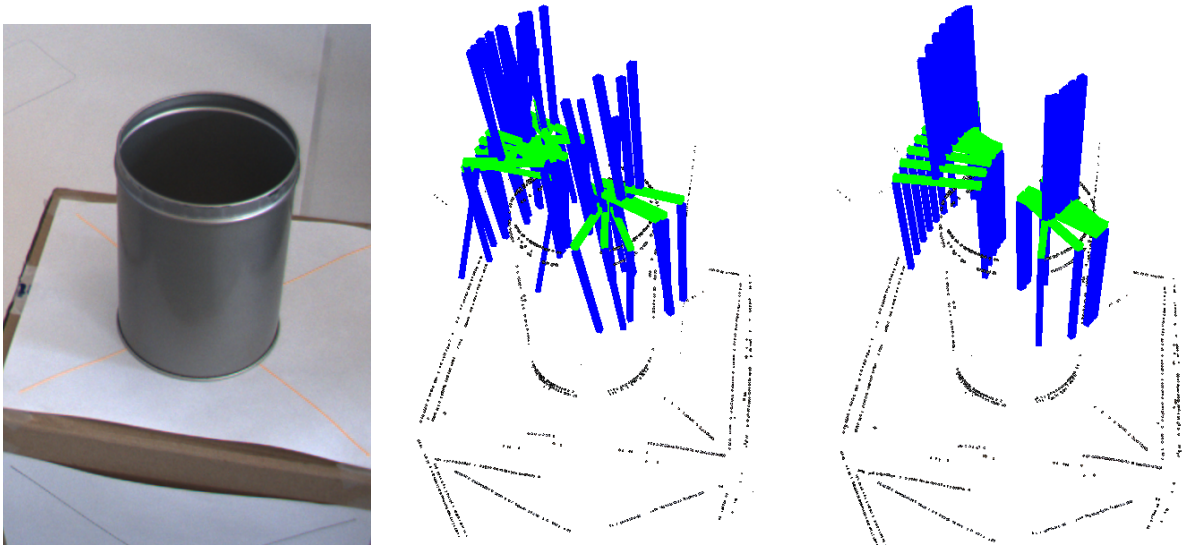

Fig. 7. Grasp locations for different approaches. **(a)** Object to be grasped. **(b)** Grasps that are calculated by using local features for different noise levels. **(c)** Grasps that are calculated by using contours for different noise levels.

[3] N. Krüger, M. Lappe, and F. Wörgötter, "Biologically Motivated Multi-modal Processing of Visual Primitives," *The Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour*, vol. 1, no. 5, pp. 417–428, 2004.

[4] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Proceedings of the European Conference of Computer Vision (ECCV 2006)*, 2006.

[5] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 2, no. 60, pp. 91–110, 2004.

[6] K. Mykolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proceedings of the European Conference in Computer Vision (ECCV 2002)*. Springer–Verlag, 2002.

[7] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, 2005.

[8] O. I. Camps and T. Kanungo, "Hierarchical organization of appearance-based parts and relations for object recognition," in *In Proc. IEEE Conf. Computer Vision and Pattern Recognition*. IEEE, 1998, pp. 685–691.

[9] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR03*, 2003, pp. 409–415.

[10] D. G. Lowe, *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, 1985.

[11] S. Ullman, *High-level Vision*, Bradford, Ed. MIT Press, 1996.

[12] Y. Lerner, T. Hendler, D. Ben-Bashat, M. Harel, and R. Malach, "A hierarchical axis of object processing stages in the human visual cortex." *Cereb Cortex*, vol. 11, no. 4, pp. 287–297, April 2001.

[13] K. Tanaka, "Mechanisms of visual object recognition: monkey and human studies." *Current Opinion in Neurobiology*, vol. 7, no. 4, pp. 523–529, Aug 1997.

[14] A. Sha'asua and S. Ullman, "Structural saliency: The detection of globally salient structures using a locally connected network," *Computer Vision., Second International Conference on*, pp. 321–327, Dec 1988.

[15] J. Shotton, A. Blake, and R. Cipolla, "Multiscale categorical object recognition using contour fragments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1270–1281, 2008.

[16] X. Wang, J. Keller, and P. Gader, "Using spatial relationships as features in object recognition," *Fuzzy Information Processing Society, 1997. NAFIPS '97., 1997 Annual Meeting of the North American*, pp. 160–165, Sep 1997.

[17] O. Henricsson, "Inferring homogeneous regions from rich image attributes," in *Automatic Extraction of Man-Made Objects from Aerial and Space Images*. Birkhuser Verlag, 1995, pp. 13–22.

[18] S. Dickinson, A. Pentland, and A. Rosenfeld, "From volumes to views: an approach to 3-d object recognition," *Automated CAD-Based Vision, 1991., Workshop on Directions in*, pp. 85–96, Jun 1991.

[19] E. Saykol, U. Güdükbay, and Özgür Ulusoy, "A histogram-based approach for object-based query-by-shape-and-color in image and video databases." *Image Vision Comput.*, vol. 23, no. 13, pp. 1170–1180, 2005.

[20] B. Huet and E. R. Hancock, "Relational histograms for shape indexing," in *ICCV*, 1998, pp. 563–569.

[21] M. Swain and D. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[22] N. Pugeault, *Early Cognitive Vision: Feedback Mechanisms for the Disambiguation of Early Visual Representation*. Verlag Dr. Muller, 2008.

[23] 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," *Special Issue on "Cognitive Humanoid Robots" of the International Journal of Humanoid Robotics*, vol. 5, pp. 247–265, 2008.

[24] N. Pugeault, S. Kalkan, E. Başeski, F. Wörgötter, and N. Krüger, "Reconstruction uncertainty and 3d relations," in *Proceedings of Int. Conf. on Computer Vision Theory and Applications (VISAPP'08)*, 2008.

[25] N. Pugeault, F. Wörgötter, and N. Krüger, "Multi-modal Scene Reconstruction Using Perceptual Grouping Constraints," in *Proc. IEEE Workshop on Perceptual Organization in Computer Vision (in conjunction with CVPR'06)*, 2006.

[26] Y. Kanazawa and K. Kanatani, "Reliability of Fitting a Plane to Range Data," *IEICE transactions on information and systems*, vol. 78, no. 12, pp. 1630–1635, 1995.

[27] R. Hunt, *Measuring Colour. 3rd edition*. Fountain Press, Kingston-upon-Thames, 1998.

[28] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and Applications (Springer Series in Statistics)*, 2nd ed. Springer, Berlin, September 2005.

[29] D. Aarno, J. Sommerfeld, D. Kragic, N. Pugeault, S. Kalkan, F. Wörgötter, D. Kraft, and N. Krüger, "Early reactive grasping with second order 3d feature relations," *The IEEE International Conference on Advanced Robotics, Jeju Island, Korea*, 2007.

[30] L. Bodenhagen, "Adaptive Grasping based on Second Order Visual Feature Relations," Master's thesis, University of Southern Denmark, 2009.