A Scene Representation Based on Multi-Modal 2D and 3D Features

Emre Başeski
Syddansk Universitet
Denmark
emre@mmmi.sdu.dk

Nicolas Pugeault
University of Edinburgh
United Kingdom
npugeaul@inf.ed.ac.uk

Sinan Kalkan
Universität Göttingen
Germany
sinan@bccn-goettingen.de

Dirk Kraft
Syddansk Universitet
Denmark
kraft@mmmi.sdu.dk

Florentin Wörgötter
Universität Göttingen
Germany
worgott@bccn-goettingen.de

Norbert Krüger
Syddansk Universitet
Denmark
norbert@mmmi.sdu.dk

Abstract

Visually extracted 2D and 3D information have their own advantages and disadvantages that complement each other. Therefore, it is important to be able to switch between the different dimensions according to the requirements of the problem and use them together to combine the reliability of 2D information with the richness of 3D information. In this article, we use 2D and 3D information in a feature-based vision system and demonstrate their complementary properties on different applications (namely: depth prediction, scene interpretation, grasping from vision and object learning)[1].

1. Introduction

There exist acknowledged differences between visually extracted 2D and 3D information (see, e.g., [2, 4]). In addition to the difference in dimension, two aspects of 2D information can be distinguished [12]: appearance based information (such as pixel color values or contrast transition) and geometric information (such as the position and orientation of a local edge). An overview of such differences is given in Table 1.

Two dimensional geometric information varies significantly with viewpoint changes. Actually, it is only the change of 2D orientation that allows for the reconstruction of a 3D orientation. For many tasks such as object recognition, this imposes the problem to compensate for this variance which can be done for example by invariant descriptors (see, e.g., [10, 11]). However, an invariance to such transformations leads necessarily to a weakening of the structural richness of the representations since properties that the system becomes invariant to can not be represented anymore.

For both types of 2D information, geometric or appearance based, the transformation under viewpoint changes can be computed explicitly or at least approximated once the underlying 3D model is known. Hence, using 3D information reduces the problem of variance under view-point transformation (with the exception of occlusions) and also allows to compute rich geometric information in terms of 3D position and 3D orientation. It also allows for the definition of semantic relations such as the Euclidian distance of visual entities or their co-planarity (see below). Moreover, in the context of robotic systems, the 3D space is closer to the space the action takes place in comparison to the 2D image space. For example in grasping, the transformation between joint co-ordinates and 3D pose is usually trivial [13]; and in navigation, planning is often done in maps representing depth information in an Euclidian way.

However, there are also problems connected to the use of 3D information. First, significantly more complex processing is required: Besides the fact that multiple cameras are required that usually need to be carefully calibrated, correspondences need to be found. For feature based matching, this imposes a number of possible error sources. For example, besides the possibility of a wrong match, it might even be that a feature is extracted in only one of the images. Moreover, when 3D information is extracted by stereo, the quality of information highly varies *with* space since the uncertainties that are associated to reconstructions at different positions in Euclidian space are highly non–isotropic and hence any depth information carries an uncertainty that depends strongly on the viewpoints [15].

We suggest that efficient visual systems should make use of the complementary properties of 2D and 3D information according to the actual context and task. This seems to hold

[1] This work has been supported by EU-Project Drivsco

978-1-4244-1631-8/07/$25.00 ©2007 IEEE

for human vision as well. For example, although 2D information is sufficient for a large number of vision tasks, Edelman and Bülthoff [4] have shown that the existence of 3D information reduces the mean error rate for tasks like recognition. Since 2D information is more reliable but 3D information is richer, one can for example use the complementary aspects of both kinds of information by doing semantic reasoning and hypotheses generation in the 3D space and feed these hypotheses back to lower levels of processing.

In [9], a visual representation, which is based on local symbolic features called multi-modal primitives, has been introduced. These primitives (see Figure 1(a)) represent a local part of the scene in terms of condensed 2D and 3D information covering appearance based aspects of visual information (color and local phase) as well as geometric information in terms of 2D and 3D position and orientation. These primitives allow for switching between 2D and 3D as well as geometric and appearance based information and hence their complementary properties can be used efficiently. Moreover, in [15], a model for the uncertainties of the 3D properties covered by the primitives is derived and is used to facilitate the reasoning processes in 3D space.

Originally, the multi–modal primitives have been designed to formulate predictions in an early cognitive vision system to disambiguate visual information (see [19]). In this work, we make use of this representation to characterize scenes and objects by 2D and 3D properties of the primitives as well as by a number of relations defined upon the primitives such as parallelism, co-planarity etc. We show that the structural richness of the representations allows for semantic reasoning about object properties and object relations in scenes. The representations are rather generic since they basically cover known attributes of visual information such as orientation, color, local motion as also computed in the first stages of human visual processing [7].[2] Hence, the primitives can be made use of for a variety of tasks.

In this paper, the strength of the approach is demonstrated on a variety of applications such as depth prediction, road interpretation, grasping, and object learning. Here, we focus less on the detailed description of the algorithms but on how the introduced representation facilitates the computation for the different tasks. In that sense, this article has a review character of previous works as well.

The paper is structured as follows: In section 2, the visual representation in [9] is summarized. In section 3, we then briefly describe 4 applications and in section 4, we reflect upon the properties of the representation.

2. Primitives and Relations

In [9], a visual representation has been introduced in terms of local condensed symbolic features called multimodal primitives. We give a brief description of these features in section 2.1. In section 2.2, we introduce perceptual relations on these symbolic features that are applied in the applications described in section 3.

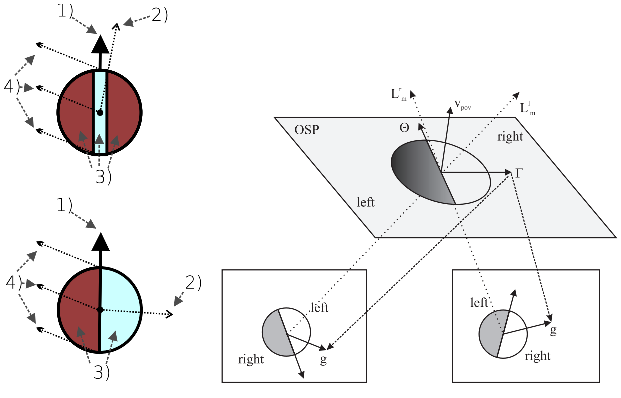
2.1. Multi-modal primitives

In its current state, the primitives discussed can be edge-like or homogeneous and carry 2D or 3D information. For edge-like primitives, the corresponding 3D primitive is extracted using feature based stereo. Since correspondences can not be found for homogeneous image structures, 3D primitives for these image structures can be estimated from the surrounding 3D edge-like primitives (see also section 3.1).

An edge-like 2D primitive (Figure 1(a)) is defined as:

$$\pi = (\boldsymbol{m}, \theta, \omega, (\mathbf{c}_l, \mathbf{c}_m, \mathbf{c}_r), f), \qquad (1)$$

where $\boldsymbol{m}$ is the image position of the primitive; θ is the 2D orientation; ω represents the contrast transition coded in the local phase; $(\mathbf{c}_l, \mathbf{c}_m, \mathbf{c}_r)$ is the representation of the color, corresponding to the left $(\mathbf{c}_l)$, the middle $(\mathbf{c}_m)$ and the right side $(\mathbf{c}_r)$ of the primitive; and, f is the optical flow.



(a) (b)

Figure 1. **(a)** Two types of edge-like 2D primitives [9] 1) represents the orientation of the primitive, 2) the phase, 3) the color and 4) the optic flow. **(b)** Reconstruction of a 3D primitive from two 2D primitives.

As the underlying structure of an homogeneous image patch is different from that of an edge-like patch, a different representation is needed for homogeneous 2D primitives (called *monos*):

$$\pi^m = (\boldsymbol{m}, \mathbf{c}), \qquad (2)$$

where $\boldsymbol{m}$ is the position in the image, and $\mathbf{c}$ is the color of the mono. Note that these different image structures can be distinguished by the intrinsic dimension of the image patch

[2] A more detailed discussion of the biological motivation can be found in [9].

	3D	2D	
pros	Distances and angles are invariant under camera transformations	Distances and angles are variant under camera transformations	**cons**
	Units have physical meaning (distance in millimeters)	Pixel coordinates are not directly usable for physical measurements	
	Relations are richer (coplanarity,proximity)	Restricted to 2D relations	
	Possible to obtain a complete model of an object	To cover all perspectives of an object a high number of images are required	
	Directly relatable to actions	Requires additional computation to become related to actions	
cons	High computational complexity	Low computational complexity	**pros**
	High likelihood of errors and uncertainty	Higher reliability	

Table 1. Different properties of 2D and 3D information. While 3D information has geometric properties (position and orientation), 2D information covers also appearance based properties (color,contrast transition etc.).

.

[5]. See [9] for more information about these modalities and their extraction. Figure 2 shows the extracted primitives for an example scene.

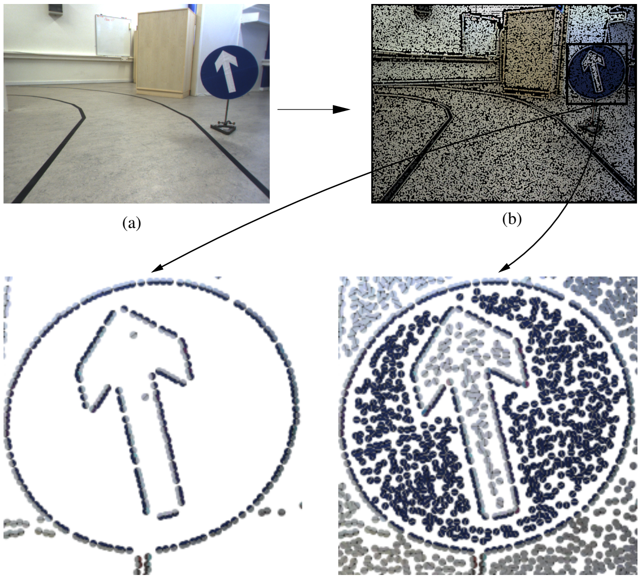

(a)
(b)
(c)
(d)

Figure 2. Extracted primitives **(b)** for the example image in **(a)**. Magnified edge primitives and edge primitives together with monos are shown in **(c)** and **(d)** respectively.

A primitive π is a 2D feature which can be used to find correspondences in a stereo framework to create 3D primitives (as introduced in [16]) which have the following formulation:

$$\mathbf{\Pi} = (\boldsymbol{M}, \Theta, \Omega, (\mathbf{c}_l, \mathbf{c}_m, \mathbf{c}_r)), \tag{3}$$

where $\boldsymbol{M}$ is the 3D position; Θ is the 3D orientation. Appearance based information is coded in the phase Ω (i.e., contrast transition) and $(\mathbf{c}_l, \mathbf{c}_m, \mathbf{c}_r)$ is the representation of the color, corresponding to the left $(\mathbf{c}_l)$, the middle $(\mathbf{c}_m)$ and the right side $(\mathbf{c}_r)$ of the 3D primitive. Both, phase and color, are extracted as a combination of the associated values in the corresponding 2D primitives in the left and right image. The reconstruction of a 3D primitive from two corresponding 2D primitives is examplified in Figure 2(b).

In section 3.1, we estimate the 3D representation $\mathbf{\Pi}^{\mathrm{m}}$ of monos which stereo fails to compute:

$$\mathbf{\Pi}^{\mathrm{m}} = (\boldsymbol{M}, \mathbf{n}, \mathbf{c}), \tag{4}$$

where $\boldsymbol{M}$ and $\mathbf{c}$ are as in equation 2, and $\mathbf{n}$ is the orientation (*i.e.*, normal) of the plane that locally represents the mono.

2.2. Perceptual relations between primitives

The sparse and symbolic nature of the discussed primitives allows for perceptual relations defined on them that express relevant spatial relations in 2D and 3D space. These relations can be applied in rather different contexts such as depth prediction, object learning and grasping (see section 3).

Collinearity: Two spatial primitives $\mathbf{\Pi}_i$ and $\mathbf{\Pi}_j$ are collinear (i.e., part of the same group) if they are part of the same contour. Due to uncertainty in the 3D reconstruction process, in this work, the collinearity of two spatial primitives $\mathbf{\Pi}_i$ and $\mathbf{\Pi}_j$ is computed using their 2D projections π_i and π_j. We define the collinearity of two 2D primitives π_i and π_j as:

$$col(\pi_i, \pi_j) = 1 - \left| sin\left(\frac{|\alpha_i| + |\alpha_j|}{2}\right) \right|, \tag{5}$$

where α_i and α_j are as shown in Figure 3(a).

Co–planarity: Two 3D edge primitives $\mathbf{\Pi}_i$ and $\mathbf{\Pi}_j$ are defined to be co–planar if their orientation vectors lie on the same plane, *i.e.*:

$$cop(\mathbf{\Pi}_i, \mathbf{\Pi}_j) = 1 - |\mathbf{proj}_{t_j \times \mathbf{v}_{ij}}(t_i \times \mathbf{v}_{ij})|, \tag{6}$$

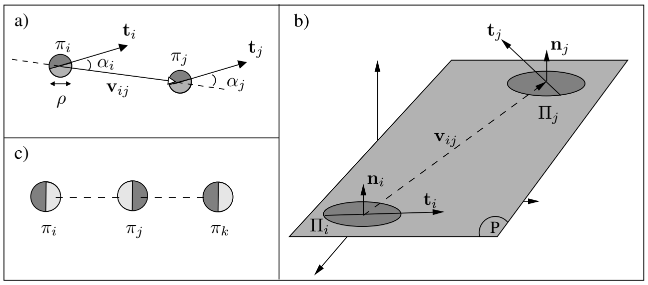


Figure 3. Illustration of the perceptual relations between primitives. **(a)** Collinearity of two 2D primitives. **(b)** Co–colority of three 2D primitives π_i, π_j and π_k. In this example, π_i and π_j are cocolor, so are π_i and π_k; however, π_j and π_k are not cocolor. **(c)** Co–planarity of two 3D primitives $\mathbf{\Pi}_i$ and $\mathbf{\Pi}_j$.

where $\mathbf{v}_{ij}$ is the vector $(\boldsymbol{M}_i - \boldsymbol{M}_j)$; t_i and t_j denote the vectors defined by the 3D orientations Θ_i and Θ_j, respectively; and, $\mathbf{proj_u}(\mathbf{a})$ is the projection of vector $\mathbf{a}$ over vector $\mathbf{u}$. The co–planarity relation is illustrated in Figure 3(b).

Co–colority: Two 3D primitives $\mathbf{\Pi}_i$ and $\mathbf{\Pi}_j$ are defined to be co–color if their parts that face each other have the same color. In the same way as collinearity, co–colority of two spatial primitives $\mathbf{\Pi}_i$ and $\mathbf{\Pi}_j$ is computed using their 2D projections π_i and π_j. We define the co–colority of two 2D primitives π_i and π_j as:

$$coc(\pi_i, \pi_j) = 1 - \mathbf{d}_c(\mathbf{c}_i, \mathbf{c}_j), \tag{7}$$

where $\mathbf{c}_i$ and $\mathbf{c}_j$ are the RGB representation of the colors of the parts of the primitives π_i and π_j that face each other; and, $\mathbf{d}_c(\mathbf{c}_i, \mathbf{c}_j)$ is Euclidean distance between RGB values of the colors $\mathbf{c}_i$ and $\mathbf{c}_j$. Co-colority between an edge primitive π and a mono primitive π^m, and between two monos can be defined similarly (not provided here). In Figure 3(c), a pair of co–color and not co–color primitives are shown.

Rigid-body motion: The rigid body motion $\mathcal{M}_{t\to t+\Delta t}$ associating any entity in space in the coordinate system of the stereo set–up at time t to the same entity in the new coordinate at time $t + \Delta t$ is explicitly defined for 3D–primitives (see Figure 4):

$$\hat{\mathbf{\Pi}}_i^{t+\Delta t} = \mathcal{M}_{t\to t+\Delta t}(\mathbf{\Pi}_i^t). \tag{8}$$

3. Applications

In this section, the framework introduced in section 2 is applied to a variety of tasks such as depth prediction at homogeneous image structures (section 3.1), scene interpretation (section 3.2), grasping (section 3.3) and object learning (section 3.4).

3.1. Depth prediction

Edge primitives represent edge–like structures. It is known that it becomes increasingly difficult to find correspondences between local patches the more they lack structure. On the other hand, it is known that lack of structure also indicates lack of a depth discontinuity [6, 8]. Moreover, we have shown that based on the co-planarity relation, depth at homogeneous image areas can be predicted (see Figures 5 and 6). Such a scheme can be used to ‘fill in’ the representation at homogeneous areas using co–planar relationships between edge–like primitives. In Figure 5, the homogeneous primitives inferred using such a scheme are shown as a disparity map. Results on the same scene are shown for a global dense stereo method (taken from [18]) that uses dynamic programming to optimize matching costs. Figure 5 shows that such depth prediction can be used as a depth cue providing additional information in particular when image structures are too weak to find correspondences. When confronted with an image as in Figure 6, many dense depth estimation algorithms either basically fail or assume implicitly some linearity assumption that leads to rather bad reconstruction. However, our method can ’interpret’ the curved edges of the cylinder in order to reconstruct the round surface.

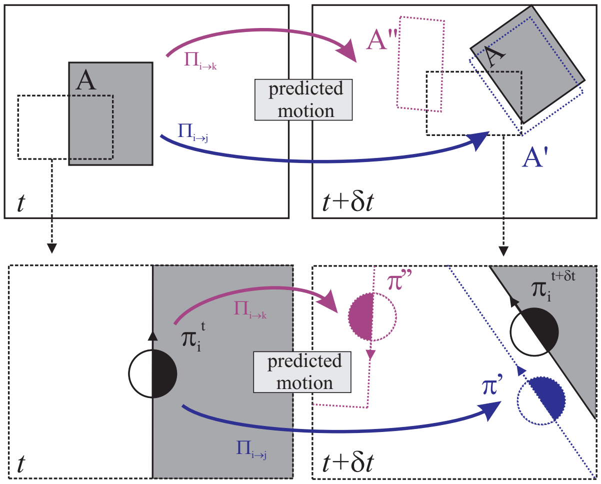
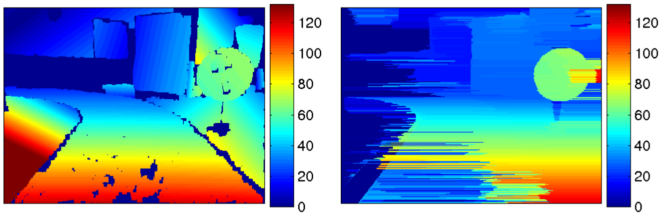
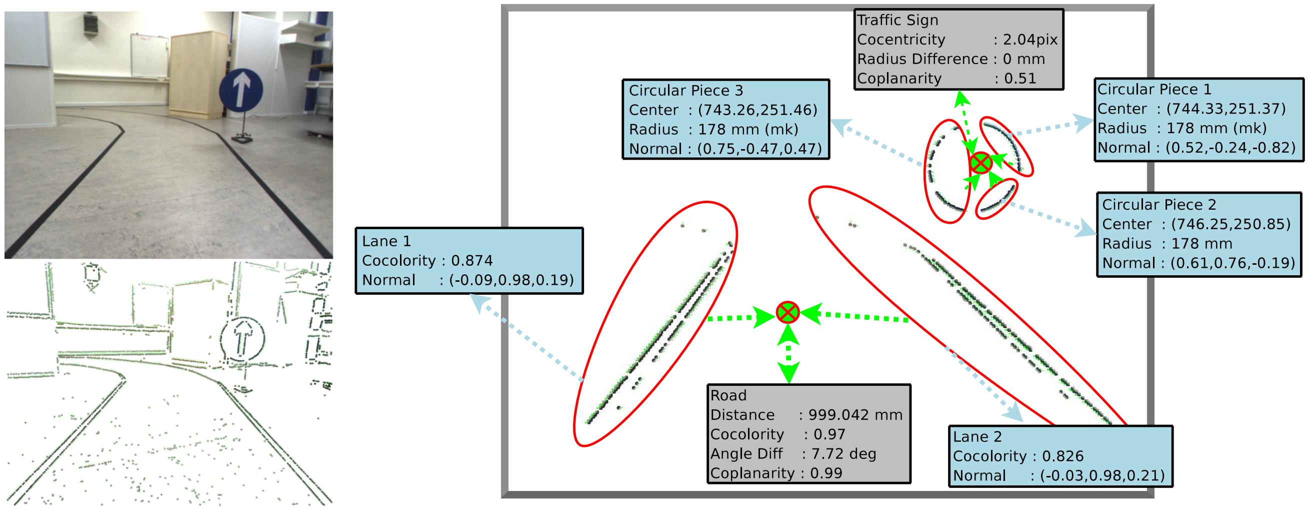


Figure 4. Example of the rigid-body motion of a primitive (see text).



(a) (b)

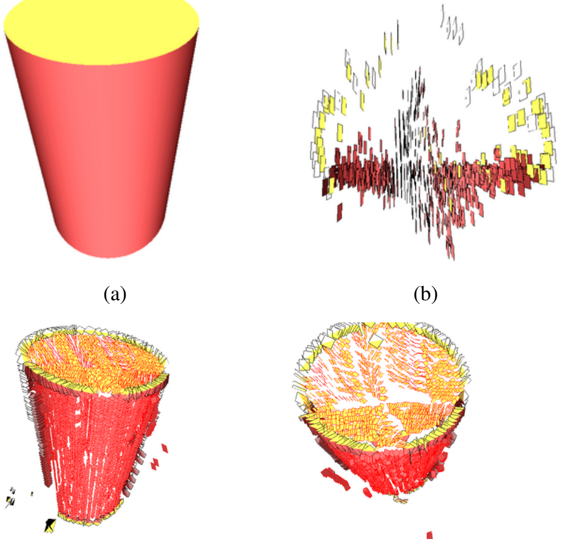
Figure 5. Depth prediction at homogeneous image areas using perceptual relations between primitives. **(a)** The results, shown as a disparity map only at the predictions, are from the scene in Figure 2. **(b)** A global dense stereo method (taken from [18]) that uses dynamic programming to optimize matching costs.



(a)

(b)

Figure 7. Interpretation of a road and a circular traffic sign. **(a)** Input image from a stereo pair and the corresponding 2D primitives **(b)** Interpretation of the scene.



(a)

(b)

(c)

(d)

Figure 6. Depth prediction for a round object. **(a)** Left stereo image. **(b)** The top view of the results of 3D reconstruction from a dense method (taken from [17]). The dense method estimates a planar surface. The dynamic programming method from [18] produces similar results. **(c)-(d)** Two views of the results of our depth prediction method. Note that **(b)-(d)** are snapshots from our 3D visualization software.

3.2. Scene interpretation

Based on the co-linearity relation defined in section 2.2 we can define higher level entities, in the following called groups, as sets of co–linear primitives (for details see [16]). Although the groups of multi-modal primitives have higher semantic meaning than individual primitives, they are not enough to define an object or give an idea about the structure of a scene. Therefore, combinations of groups are more suitable for interpreting a scene. As an example (see Figure 7), one lane of a road can be defined by a group of primitives but this group is not qualified as a road, unless it is not combined with the group that represents the opposite lane. In that sense, the opposite lane is the one that lies on the same plane with a certain distance and similar color. With a similar reasoning, a circular traffic sign is interpreted by the combination of circular pieces that shares the same center and the plane with a similar enough color.

In this way we can make use of the appearance based as well as geometric information in the primitives. Interestingly, this allows for a close to textural description of objects and scenes, e.g., the particular traffic sign in Figure 7 can be described by its geometric properties (curved and co-planar groups with a certain proximity) as well as its appearance based aspects (being blue). In this way, the introduced representations can be seen as as an intermediate step towards high level representations in which by expressing the semantic relations introduced in section 2.2, abstract statements about the scene structure can be made.

3.3. Grasping

In [1], it has been shown how geometry, appearance and spatial relations between multi-modal features can guide early reactive grasping which is an initial *”reflex-like”* grasping strategy. A simple parallel jaw gripper was used and five elementary grasping actions, called EGAs, were associated to co-planar primitives. Two samples are shown in Figure 8(a). The EGAs were tested in a simulation environment [1] as well as in a real environment. It has been shown that with a rather weak assumption of co-planarity and hence without any a-priori object knowledge, successful grasps could be generated which can then be haptically verified and used further in a cognitive system (see section 3.4). Basically, plane hypotheses based on co-planar features (as discussed in section 2) become associated to grasp hypothesis (see Figure 8(b)). By making use of the additional relations co-colority and co-linearity, the number of potential grasp hypotheses could be further reduced.

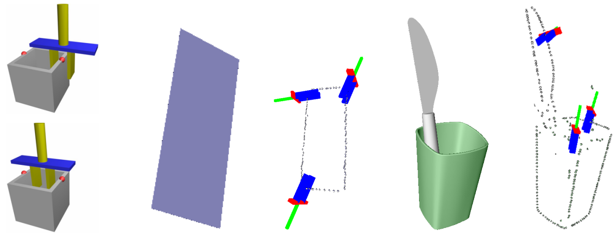


Figure 8. Sample elementary grasping actions and grasping hypothesis from [1] **(a)** Two sample EGAs **(b)** Two sample grasp hypotheses.

Even more reliable grasping hypothesises can be associated to object parts (see, e.g., [2]). To grasp cylindric or conic objects, grasping options can be associated to a circle (see Figure 10). Here, instead of using second-order relations between multi-modal primitives, 3D locations of circles have been used to generate grasping hypothesises.

To extract a 3D circle, it is important to switch between the 2D and the 3D aspects. The first step is locating the 3D circle by using the fact that a circle in 3D can be approximated by an ellipse in 2D. Although fitting an ellipse to 2D data is easier than fitting a circle in 3D, an ellipse does not give sufficient information about the center, radius and the plane normal of the 3D circle. At that point, it is possible to switch the dimension and obtain the missing information by processing the 3D features that correspond to the 2D features which form the ellipse. Fitting a plane to the 3D features determines the normal of the circle. Finally, the intersection of this plane and the line that passes from the camera center and the multiplication of the pseudo-inverse of the projection matrix and 2D ellipse center gives the center of the circle. An example of the procedure is given in Figure 9 (a-c).

Once a circle is found in 3D, four different grasp hypothesis can be generated (see Figure 10). The first one uses the center and the normal of the circle to place the gripper inside the circle and uses the radius to grasp the object from inside. For the second hypothesis, a point on the circle is calculated and this point is used to grasp the object from its brim. For the third hypothesis, the center and the normal of the circle is used for placing the gripper orthogonal to the circle normal, the radius is used to open the gripper and the object is grabbed from the side. The last hypothesis is similar to the first one but instead of inner side, the circle is grasped from outer side. A sample grasp of the second type is presented in Figure 9 (e-f).

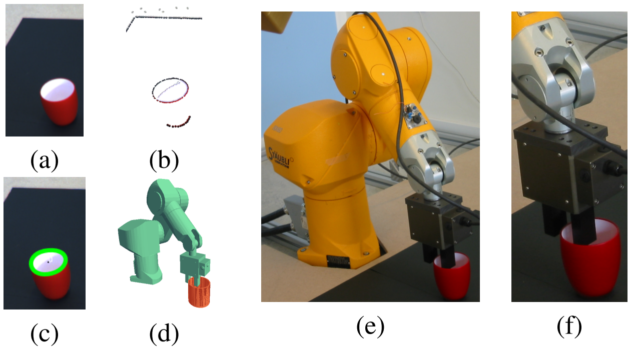


Figure 9. Grasping of a cylindrical cup **(a)** Input left image **(b)** Corresponding 2D primitives **(c)** Detected circle **(d)** Model of the robot **(e-f)** The cup is grasped by the robot with respect to the extracted information.

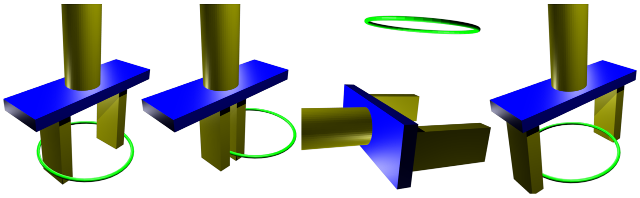


Figure 10. Four different grasp hypothesises for circles

3.4. Learning objectness and object shape

The detection of features belonging to one individual object is not a trivial task when a stereo system only observes a scene since there is no decision criterion that a set of features actually can be separated from the rest of the scene. However, having achieved a successful grasps (as explained in section 3.3), the robot has physical control over a potential object, and it can try to move it (see Figure 11). Since the change of primitives under a rigid–body motion can be described analytically (see section 2.2), predictions about the change of primitives can be derived. Only primitives that change according to these predictions are supposed to be part of the object.[3] In Figure 12, a number of representations are shown that have been extracted by this method (for details, see [14]). First steps in using these object representations for pose estimation and grasping are made in [3].

[3] Note that the primitives belonging to the grasper change according to the robot motion but they can be eliminated using the model of the grasper.

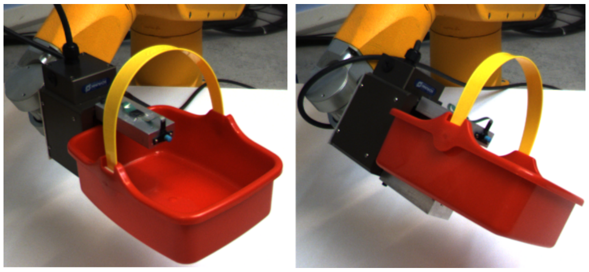


Figure 11. The robot is doing a rotation to extract the 3D model of a basket.

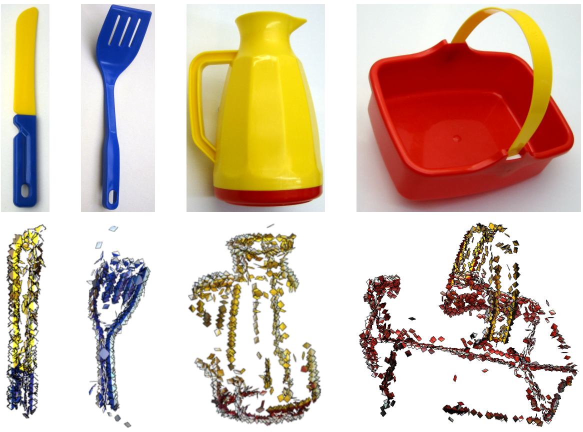


Figure 12. Sample objects and their related accumulated representation [14].

4. Discussion

The advantages of using a 2D or a 3D scene representation is highly dependent on the application and the context. Both have their own advantages and disadvantages as presented in Table 1. By keeping these properties in mind, we described a representation that preserves relevant aspects of 2D and 3D information to allow for switching between the dimensions according to the actual requirements. We exemplified the potential of this approach in four applications of rather different nature, covering depth estimation at homogeneous areas, semantic scene description, grasping and extraction of object representations.

References

[1] 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. *IEEE International Conference on Robotics and Automation (ICRA), Workshop: From features to actions - Unifying perspectives in computational and robot vision*, 2007.

[2] I. Biederman. Recognition by components: A theory of human image understanding. *Psychological Review*, 94(2), 1987.

[3] R. Detry and J. Piater. Hierarchical integration of local 3d features for probabilistic pose recovery. *Robot Manipulation: Sensing and Adapting to the Real World, 2007 (Workshop at Robotics, Science and Systems)*, 2007.

[4] S. Edelman and H. H. Bulthoff. Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Research*, 32:2385–2400, 1992.

[5] M. Felsberg and N. Krüger. A probablistic definition of intrinsic dimensionality for images. *Pattern Recognition, 24th DAGM Symposium*, 2003.

[6] W. Grimson. Surface consistency constraints in vision. *CVGIP*, 24(1):28–51, 1983.

[7] D. Hubel and T. Wiesel. Brain mechanisms of vision. *Scientific American*, 241:130–144, 1979.

[8] S. Kalkan, F. Wörgötter, and N. Krüger. Statistical analysis of local 3d structure in 2d images. *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1114–1121, 2006.

[9] N. Krüger, M. Lappe, and F. Wörgötter. Biologically motivated multi-modal processing of visual primitives. *Interdisciplinary Journal of Artificial Intelligence & the Simulation of Behaviour, AISB Journal*, 1(5):417–427, 2004.

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2(60):91–110, 2004.

[11] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[12] J. L. Mundy, A. Liu, N. Pillow, A. Zisserman, S. Abdallah, S. Utcke, S. Nayar, and C. Rothwell. An experimental comparison of appearance and geometric model based recognition. In *Object Representation in Computer Vision*, pages 247–269, 1996.

[13] R. Murray, Z. Li, and S. Sastry. *A mathematical introduction to Robotic Manipulation*. CRC Press, 1994.

[14] N. Pugeault, E. Başeski, D. Kraft, F. Wörgötter, and N. Krüger. Extraction of multi–modal object representations in a robot vision system. 2007.

[15] N. Pugeault, E. Başeski, N. Krüger, S.Kalkan, and F. Woergoetter. Reconstruction accuracy and relations. In *Signal Processing, Pattern Recognition, and Applications (SPPRA)*, submitted.

[16] 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.

[17] S. P. Sabatini, G. Gastaldi, F. Solari, J. Diaz, E. Ros, K. Pauwels, K. M. M. V. Hulle, N. Pugeault, and N. Krüger. Compact and accurate early vision processing in the harmonic space. *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2007.

[18] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Technical Report MSR-TR-2001-81, Microsoft Research, Microsoft Corporation, November 2001.

[19] F. Wörgötter, N. Krüger, N. Pugeault, D. Calow, M. Lappe, K. Pauwels, M. V. Hulle, S. Tan, and A. Johnston. Early cognitive vision: Using gestalt-laws for task-dependent, active image-processing. *Natural Computing*, 3(3):293–321, 2004.