2010 IEEE-RAS International Conference on Humanoid Robots
Nashville, TN, USA, December 6-8, 2010

Object segmentation and learning through feature grouping and manipulation

Eva Stergaršek Kuzmič[1] and Aleš Ude[1,2]

[1]*Jožef Stefan Institute, Ljubljana, Slovenia, eva.stergarsek@ijs.si, ales.ude@ijs.si*

[2]*ATR Computational Neuroscience Laboratories, Kyoto, Japan*

***Abstract*— In this paper we present a method for learning new objects situated in uncontrolled and unstructured environments. Visual information only is usually not sufficient for a reliable segmentation and learning of unknown objects without any a priori information. We propose an approach in which the robot introduces additional information by manipulating the entities in the scene, thus generating sufficient information to identify objects and accumulate knowledge about them. Our approach involves the extraction of local feature ensembles that provide hints about the existence of an object, the generation of pushing movements to confirm or reject the initial hypothesis, and the fusion of features that satisfy the assumed motion constraints. To ensure the robustness of the system, probabilistic methods such as RANSAC (RANdom SAmple Consensus) are used in several computational stages. Our experimental results show that the system is successful at segmenting objects in complex scenes. The segmented features can be accumulated across different views to extract more comprehensive knowledge about the objects.**

I. Introduction

Image segmentation methodologies differ in their assumptions and in the level of prior knowledge they utilize. When complete object models are known, the system is in fact performing object detection and recognition. When no *a priori* knowledge about the data present in the image is available, we need some kind of model-free approach to segmentation. However, visual information is often not sufficient for a reliable and accurate segmentation of random heaps of unknown objects [36]. Such a situation is presented in Fig. 1.

Humans are very successful at finding objects in complex scenes, even if they have never seen them before. This ability has proven to be very difficult to replicate by passive vision systems, mainly because it is hard to define what exactly constitutes an object. The meaning of the word "object" is very broad and dependent on semantics and context [6]. While many different principles can be found, e.g. closure, connectedness, bilateral symmetry [16], co-planarity and co-linearity of contours [14], etc., counterexamples can be found for each of them. Hence such basic principles can be applied only to generate hypotheses about the existence of the objects, whereas to confirm such hypotheses, additional paradigms need to be applied. The hierarchical composition of simpler features into more complex entities that become sparser in the image, thus enabling fast and robust classification, has been explored for this purpose in the vision literature [8].

Besides passively observing a scene, a robot can explore the world and learn from the effect its actions have on the external world [21], much like humans explore objects to control the visual input. Active exploration should enable the separation of objects from the background, something that is not trivial when passively observing an object in scenes with a highly cluttered background [13][22]. For example, moving an object can help to extract object boundaries, which is useful for segmentation, even when the motion is short and poorly controlled [10].

To move an object, the robot can either apply pushing movements or transport it in its hand, which assumes that the object can be grasped. While grasped objects can be controlled better than pushed objects, grasping is a more expensive operation. Nevertheless, systems exist that demonstrate such behavior [32]. These systems attempt to identify features points in two (or more) snapshots of an object corresponding to good locations at which the object can be grasped. In [14], the rigid body motion principle is used to fuse features across views. The information gained by the manipulation of grasped objects is utilized in [38] to support model-free segmentation and accumulation of snapshots from different viewpoints.

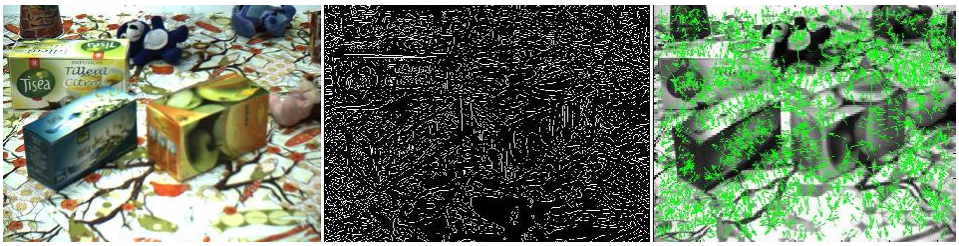

Fig. 1: Segmentation of unknown objects. Original image (640x480) on the left, extracted edges (Canny edge detector, threshold 0.01) in the middle, and extracted SIFT features on the right.

In many cases it is difficult to grasp an unknown object, either because of its size or because of the peculiarities in its shape. Pushing is a suitable alternative in such cases because it is easier to apply but, if successful, also causes the object to move, thus providing additional information for object segmentation. Pushing (or poking) has been used to acquire affordances of previously unknown objects [21].

Here we focus on the extraction of geometric information supported by pushing actions. Pushing actions are synthesized from object hypotheses generated from clusters of local image features. We show that applying pushing actions to the generated object hypotheses provides a sufficient amount of information to separate the objects from the background and acquire its features. The implemented system can cope with multiple objects, overlap and clutter. Finally, our system can accumulate object features from multiple views.

A. The outline of the approach

The basic building blocks of our system are the following procedures:

- Generation of object hypotheses: Features are extracted from stereo images and put into correspondence by comparing feature vectors and using epipolar geometry. These matches allow us to calculate 3-D feature points. The set of 3-D points is investigated for occurrences of regular surfaces such as planes. The extracted and ranked hypothetical surfaces are used to generate actions for manipulation.
- Object manipulation: An action, e.g. a push, is applied to the hypothetical surface to provide additional information for building complete and verified object descriptions.
- Evaluation of hypotheses: This process establishes correspondences between 3-D feature points, estimates the parameters of the induced object motion, and verifies whether the hypothetical object features move as expected for features associated with an object. Some features can be excluded and other added based on the results of the verification process.

Note that if any of these processes fails, the system can generate and verify additional hypotheses. Hence we do not need to assume that the outcomes of our image processing algorithms are perfect, which makes our system robust.

Studies have shown that systems based on local, partially invariant features are able to recognize objects from multiple viewpoints and also detect these objects in cluttered scenes [7][17][30][27]. Object recognition can often be understood as a feature matching problem with essentially three phases: detecting, describing and matching features. There are plenty of options for a feature detector. Existing detectors are based on affine normalization around Hessian and Harris points [24], difference of Gaussians (DoG) [17], edges [37], intensity extrema [37], 'maximally stable extremal regions' [20] or 'salient regions' [28]. Evaluation of feature detectors [25] shows that performance of detectors depends on the application and the associated image content, which determines for instance the robustness, accuracy, and density of features. Evaluation of descriptors [23] such as steerable filters [11], differential invariants [34], complex filters [33], moment invariants [39] and cross-correlation for different types of interest points, Scale Invariant Feature Transform descriptor (SIFT) [17] shows that the latter often performs best. Similarly, evaluation on 3-D objects [27] shows that SIFT descriptor performs well compared to other descriptors. For this reason, we selected SIFT descriptors with the difference of Gaussians as a feature detector as basis for the processing of the acquired images. The choice of detectors and descriptors does not influence the implementation of the methods proposed in this paper. Other detectors and descriptors could be used if required by the application.

In the proposed approach, object hypotheses are formed by estimating parameters of hypothetical planes in the scene and evaluated by verifying whether associated features moved as a rigid body would after the application of the pushing action. Both processes, i.e. generation and evaluation of hypotheses can be subject to a significant amount of outliers in the data set. To ensure robustness, we employ the RANSAC algorithm [9] known for its ability to perform robust estimation of model parameters.

In our system we match the local features acquired before and after the pushing action. In principle it is possible to continuously track the pushed object features, which minimizes the differences between the consecutive views, thus reducing the possibility of loosing the features due to large viewpoint changes. Unfortunately, in our application there is a high probability that the robot's arm occludes the object while executing the pushing motion. Since the complete geometry of the object to be pushed is not available, it is impossible to ensure pushing without occlusion, i.e. only side pushing. The generation of hypotheses is based on visible features, therefore the manipulation is focused on these features and occlusions are highly probable, which makes continuous tracking impractical.

The rest of the article is organized as follows. Section II describes the formation of hypotheses about the objects. In Section III we describe the planning of robot motion to generate pushing actions. This is followed by the evaluation of hypotheses based on the resulting changes in the scene. In Section V we present the results of experiments performed to evaluate the implemented system. We also tested the accumulation of object knowledge across viewpoints after consecutive pushing actions. Section VI discusses the potential and limitations of the proposed algorithm and concludes the paper.

II. OBJECT HYPOTHESES

We start by proposing a methodology towards the formation of hypotheses for the generation of manipulative actions that can be applied to induce object motion. For reasons given above, any such methodology needs to make decisions with respect to the type of features and objects that can be expected in the scene. Note that in our system these are just the initial hypotheses that can later be objectively confirmed by manipulation. In the following we focus on household environments. Households contain many objects with planar surfaces, hence one type of objects we are

interested in are objects with some planar surfaces. With respect to the planarity constraint, four points are not in general position if they fall on a plane. Our basic strategy is to look for groups of features that fulfil certain hypotheses, e.g. planarity constraint, and apply robot manipulation actions to move the object associated with the detected ensembles of features and confirm or reject the hypothesis. In this way hierarchical interpretation of visual data can be simplified and objects can be found with less prior assumptions.

We selected SIFT features as a basic visual representation. Besides being powerful interest point detectors [23][27], SIFT descriptors have also proven to be robust against moderate rotations in the scene (up too thirty degrees [19]), which allow the system to match the descriptors before and after the object has moved. A SIFT descriptor is computed from gradient information around the keypoint region. Local appearance is described by histograms with 8 bins for each 4 x 4 subregion of a 16 x 16 region around the keypoint, which results in a 128-dimensional feature vector.

For a stereo pair of images, a set of matching features can be determined on the basis of Euclidean distance between SIFT descriptors by best-bin first algorithm [17]. The best candidate match for a SIFT feature is its nearest neighbor, i.e. the feature with a minimum Euclidean distance between descriptor vectors. Some of these initial matches could be incorrect due to 1) ambiguous features, 2) features that arise from background clutter or 3) features that were not detected in the first image. We refine the search for corresponding points by taking into account the epipolar geometry [12]. Intrinsic camera parameters are acquired by a standard calibration procedure; hence we can capture this geometric constraint in an algebraic representation known as the essential matrix

$$\boldsymbol{u}_i^T E_{ij} \boldsymbol{u}_j = 0 . \qquad (1)$$

$\boldsymbol{u}_i = [x_i, y_i, 1]^T$, $\boldsymbol{u}_j = [x_j, y_j, 1]^T$ are the image locations of the SIFT descriptor and $\boldsymbol{E}$ is the essential matrix.

The next step is finding features that belong to planar parts of the scene. Various algorithms for plane detection have been proposed. While it is possible to detect planes without computing the 3-D information explicitly, for instance with disparity maps [35], our approach anticipates 3-D information to generate manipulation actions and we therefore exploit 3-D information for plane detection. 3-D information is also needed for fusing features across views.

Occurrences of co-planarity are determined with the RANSAC algorithm. Tolerance t_P indicates the maximum allowable absolute distance between a point and the hypothetic plane:

$$ax + by + cz + d = 0 . \qquad (2)$$

Since no prior assumptions about the environment were made, the planes can incorporate features that in fact belong to different objects or background. To ensure successful manipulation, we group the features of every plane using X-means clustering [30], which is an extended K-means algorithm that includes the estimation of the number of clusters. This process is presented in Fig. 2.

To evaluate the generated hypotheses we need to determine specific robot movements that are likely to cause the object associated with the detected features to move. We rank the clusters of planar features by the number of features associated with them. Such hypotheses are more likely to lead to success and we therefore start with a cluster with the highest number of features and generate a suitable pushing motion (see next section). Not every pushing motion will be successful, e.g. attempts to push the planes belonging to objects that cannot be moved will fail. If the generated pushing motion is not successful, i.e. the verification of the objectness failed, we can select the next hypothetic cluster with less features, determine and execute manipulative movements, and repeat doing so until manipulation results in

Input: all detected 3-D points

while there exists more than F_P points and maximum number of steps not reached
  repeat N_P times
    select 3 points at random
    estimate plane parameters
    find inliers - features that fit the estimated plane parameters with a predefined tolerance t_P
  select parameters associated with the plane with the biggest number of inliers F
  if ($F > F_P$)
    save parameters and remove these inliers from the set
    cluster features within the detected plane

Output: the hypothetic planes and point ensembles belonging to each plane

Fig. 2: Generation of object hypotheses. In our experiments we chose the number of samples for RANSAC to be N_P =1000, error tolerance t_P = 0.005m, threshold F_P = 20 and the maximum number of steps 6.

Fig. 3: Original image (640x480) on the far left with all extracted features on the far right. First hypothesis (middle left) and second hypothesis (middle right), after the manipulation of the first failed.

success or no other hypothesis is available.

Fig. 3 shows an example scene, the hypothesis with the highest number of feature points and the additional hypothesis.

III. OBJECT HYPOTHESIS MANIPULATION

As explained in the introduction, there exists no fully general definition about what features characterize an object. The term object normally refers to the result of a grouping process, which tends to be hierarchical regardless of the underlying features that serve as the basis for the visual representation. To find an object, we need to characterize the degree of regularity between visual features at each abstraction layer of the hierarchical grouping process. Trees have been proposed as a suitable representation for expressing such perceptual interpretations [6]. If the interpretation process is purely visual, then objects are defined by visual interpretations that are both coherent and complete. A humanoid robot, however, can apply manipulative actions on the perceived feature groupings. In this way the robot acquires additional information, e.g. based on the rigid body motion principle, which can be used to generate object representations. Object motion is a powerful cue; features associated with the same object normally move in a coherent way. Rigid body motion is the most common assumption, but it is possible to consider more general cases such as articulated or deformable motions. By moving an object the robot can acquire a reliable description of the visible part of the object without needing to fully interpret the scene, which can only be subjective due to the lack of a proper definition of objectness.

There are various possibilities for how to generate robot hand movements that induce the hypothesized object points to move. One possibility is to attempt to grasp and move the object hypothesis as suggested in [14]. While very appealing, especially because in this way we can move an object on a well-defined trajectory, grasping of unknown objects is a difficult and error-prone process. To avoid difficulties with grasping, we explored the possibility of applying pushing actions to induce objects to move. Pushing actions are easier to generate and the probability of success is higher. Moreover, objects that cannot be grasped can often still be pushed. On the other hand, it is more difficult to move an object along the desired trajectory when applying pushing movements. Since precise object movement is not important for our application – additional information about the object can be acquired as long as the object moves sufficiently regardless of the actual movement – we selected pushing as the most suitable hand motion for the verification of the generated object hypotheses.

Our goal is to apply pushing movements in such a way that the chance that the hypothesized plane moves is maximized. In order to achieve this, we calculate the trajectory according to the plane parameters and the underlying surface. The guiding principle when determining the robot motion for pushing is that this motion should be performed within a surface parallel to the ground surface. In this way the robot can avoid pushing into or away from the ground surface.

The system is calibrated so that the x-axis of the robot base system is oriented opposite to the gravity direction. The actual height of the ground plane can be calculated from the detected SIFT features, if needed. The angle θ between this plane and the selected hypothetical plane is the basis for determining the end-effector trajectories. We define pushing trajectories by specifying a point $\boldsymbol{r}_o$ and vector $\boldsymbol{n}$, which respectively determine the initial point and the direction of push. We consider two cases in relation to angle θ:

- $\theta \geq 45°$: The initial point of pushing $\boldsymbol{r}_o$ is specified as the average value of all hypothetical plane points. To obtain the direction of push, the hypothetical plane containing object features is rotated around the axis of intersection with the ground plane. The normal of the rotated plane is taken as the direction of push. By construction it is guaranteed to be parallel to the horizontal ground plane.
- $\theta < 45°$: In this case we project the detected feature points onto the plane parallel to the ground plane. The height of this plane is determined as in (3). We fit an ellipse through these points and calculate the point at the end of the shorter principal axis closer to the origin of the camera coordinate system. This is the point r_o. Pushing is performed along the principal axis starting at the point r_o in the direction of a vector pointing towards ellipse center r_c.

$$x' = \max\ x_{avg} - \Delta,\ \min\ x_g + \Delta, \frac{x_{avg} + x_g}{2} \qquad (3)$$

In the above equation x_{avg} is the average height of all inliers and x_g is the height of the ground plane. In our experiments Δ was set to 0.02m. The specification of the end-effector trajectory is summarized in Fig. 4.

Input: selected hypothetic plane with parameters (a, b, c, d) and inliers $\{\boldsymbol{x}^H\}$

determine horizontal ground plane in the scene
calculate dihedral angle θ between the hypothetic plane and the horizontal plane
if ($\theta \geq 45°$)
  $\boldsymbol{r}_o$ = mean($\{\boldsymbol{x}^H\}$) (initial point of pushing)
  rotate the selected plane to vertical position
  direction vector $\boldsymbol{n} = [a', b', c']^T$
else
  vertically project points to the horizontal plane at the height $\boldsymbol{x}'$ (see text)
  fit ellipse to points
  determine $\boldsymbol{r}_o$ (see text)
  direction vector $\boldsymbol{n} = \boldsymbol{r}_c - \boldsymbol{r}_o$
manipulation parameters: start at $\boldsymbol{r}_o + \Delta r\boldsymbol{n}$, push for $2\Delta r$ in direction $-\boldsymbol{n}$

Output: starting point and direction of the pushing movement

Fig. 4: Computation of the pushing motion.

IV. Hypothesis Evaluation and Extension of the Object Model

Once the computed pushing movement has been applied to the selected ensemble of features, we can make use of the additional information provided by the induced object motion to confirm or reject the hypothesis. As mentioned before, we assume the rigid body motion and investigate the acquired stereo images before and after manipulation to confirm whether the hypothetical object features has moved as a rigid body. Moreover, other features in the surrounding of the hypothetical object features can be added to the object model if they move in the same way as the rest of the features.

Matching is performed between the set of stereo images before and after the manipulation. The first set of 3-D locations with corresponding descriptors in left and right stereo images $\{\boldsymbol{x}^1, \boldsymbol{d}_L^1, \boldsymbol{d}_R^1\}$ is the basis for forming hypotheses about objects and for planning of the pushing motion for verification. The second set of stereo images and the computed 3-D locations and descriptors $\{\boldsymbol{x}^2, \boldsymbol{d}_L^2, \boldsymbol{d}_R^2\}$ are used to verify whether the features moved as a rigid body.

The initial sets of matches between both feature sets are obtained by comparing descriptors in left and right images.

Input: $(\{\boldsymbol{x}^1, \boldsymbol{d}_L^1, \boldsymbol{d}_R^1\}, \{\boldsymbol{x}^{Hi}, \boldsymbol{d}_L^{Hi}, \boldsymbol{d}_R^{Hi}\}$; i = 1, …, number of hypotheses)

while (hypotheses available) and (object features not verified)
  select hypothesis $\{\boldsymbol{x}^{Hi}, \boldsymbol{d}_L^{Hi}, \boldsymbol{d}_R^{Hi}\}$
  manipulate selected plane (Fig. 4)
  compute $\{\boldsymbol{x}^2, \boldsymbol{d}_L^2, \boldsymbol{d}_R^2\}$ based on images acquired after manipulation
  determine the set of features with matching descriptors $\{\boldsymbol{x}_M^1, \boldsymbol{x}_M^2\}$, where $\{\boldsymbol{x}_M^1\} \subseteq \{\boldsymbol{x}^{Hi}\}$
  repeat N times (RANSAC)
    select *3* matching 3-d locations at random from the set $\{\boldsymbol{x}_M^1, \boldsymbol{x}_M^2\}$
    estimate parameters $\boldsymbol{R}$ and $\boldsymbol{t}$
    find the number of matches in sets $\{\boldsymbol{x}^1\}$, $\{\boldsymbol{x}^2\}$ that fit the rigid motion model with tolerance t_e
  evaluate transformation with the highest number of matches F:
  if $(|\varphi| < \varphi_{\min}$ and $\|\boldsymbol{t}\| < t_{min})$
    select next hypothesis
  else if $(F < F_{min})$
    exit and form a new hypothesis
  else
    determine object features $(\{\boldsymbol{x}_{OF}^1\}, \{\boldsymbol{x}_{OF}^2\}$ within the set of available features $\{\boldsymbol{x}^1\}$, $\{\boldsymbol{x}^2\}$ that fit the model with a tolerance t_e

  Output: verified object features

Fig. 5: Hypothesis evaluation. In our experiments the error tolerance t_e was set to 0.005m, the number of samples for RANSAC N_R to 100, threshold F_{min} to 10 and the tolerances $\varphi_{\min}$ and $t_{\min}$ to 5° and 0.05m, respectively. Tolerance t_e indicates the maximal allowable absolute displacement between features in 3-D space.

The union of locations of matches in both sets is verified using RANSAC and least squares fitting of the rigid body transformation based on SVD [1]:

$$\boldsymbol{x}^2 = \boldsymbol{R}\boldsymbol{x}^1 + \boldsymbol{t}\,. \qquad (4)$$

Here $\boldsymbol{R}$ and $\boldsymbol{t}$ are the rotation matrix and the translation vector, respectively.

Only motion can provide additional information for object verification, therefore it is important to determine if the features have moved. The rigid body motion consists of a rotational and translational part. If the estimated feature translation and rotation imply there was no motion, we have to select a new hypothesis among previously determined hypotheses (Fig. 6) because in this case no verification is possible. To this end we calculate the norm of parameters $\boldsymbol{R}$ and $\boldsymbol{t}$ as estimated by RANSAC. The amount of rotation is determined by calculating the unique axis (up to the sign) and angle of rotation φ associated with the rotation matrix. It is possible to show that the above value defines a distance on the space of all rotation matrices [28]. To conclude whether or not the features moved, we define the minimal values for the angle of rotation $\varphi_{\min}$ and the norm of translation t_{min}. Fig. 5 presents the process of motion matching, estimating the parameters of rigid-body transformation and finally evaluating the hypothesis.

Our goal is to determine at least the minimum number of object features F_R . However, if the number of object features is smaller after verification, we cannot simply choose another hypothesis if the parameters of the rigid-body movement implied that the features have moved and thus the scene has changed. In this case a new set of hypotheses based on newly acquired stereo images must be acquired.

Fig. 6: Three of the detected hypotheses.

A. Accumulation of object knowledge

To form a more comprehensive object model, the robot needs to see the object from different viewpoints. Our algorithm for acquiring images from multiple viewpoints is essentially a modification of the manipulation algorithm presented in Section III. The point $\boldsymbol{r}_o$ and the direction of pushing $\boldsymbol{n}$ are determined as described in Fig. 4. The difference is that the direction of pushing $\boldsymbol{n}$ is rotated around the axis in the direction of gravity by a pre-specified angle α. (In our experiments the angle α was set to 20°). Alternatively we could modify the point of pushing.

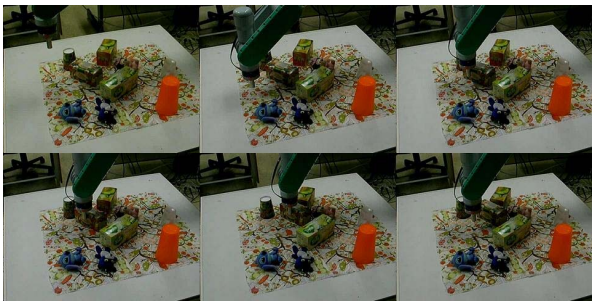

Fig. 7: The execution of pushing movement.

V. Experimental evaluation

We experimentally evaluated the formation of hypotheses about the objects, the generation of pushing actions to confirm or reject the generated hypothesis and the verification of hypotheses based on the resulting changes in the scene.

In all our experiments a predetermined number of hypotheses (set to 6) were generated. Pushing movements (see Fig. 7) were actually carried out in 96% of experiments; for the rest of experiments the determined actions could not be conducted due to the limited workspace.

To test the evaluation of hypotheses, we considered two cases. Firstly, the features included in the generated object hypothesis belong to one object. In such cases, the robot should be able to recognize these features as object features and possibly find more object features, which were not included in the initial hypothesis. Secondly, the hypothesis contains features stemming from more than one object. The robot should be able to a) remove features not belonging to the object that was moved and b) possibly add more object features. In order to test the performance of our approach in both cases, i.e. adding and removing features, we encouraged the formation of hypotheses involving more than one object. Objects with planar surfaces were purposely similar in size and placed together during experiments to increase the probability of joining their planar surfaces in the initial object hypotheses. As a result the hypotheses involving more objects represent 42% of all experiments.

Successful performance in case of "one-object" hypotheses is presented in the first row of Fig. 8. If the hypothesis involves more than one object (as in the second and third row of Fig. 8), it is only possible to distinguish between features of different objects if the objects move in different ways. In the experiment in the second row the features of the object, which moved as a result of pushing, were added to the group of object features, while the features belonging to another object were discarded. In the third row the matched features belong to objects moving in unison.

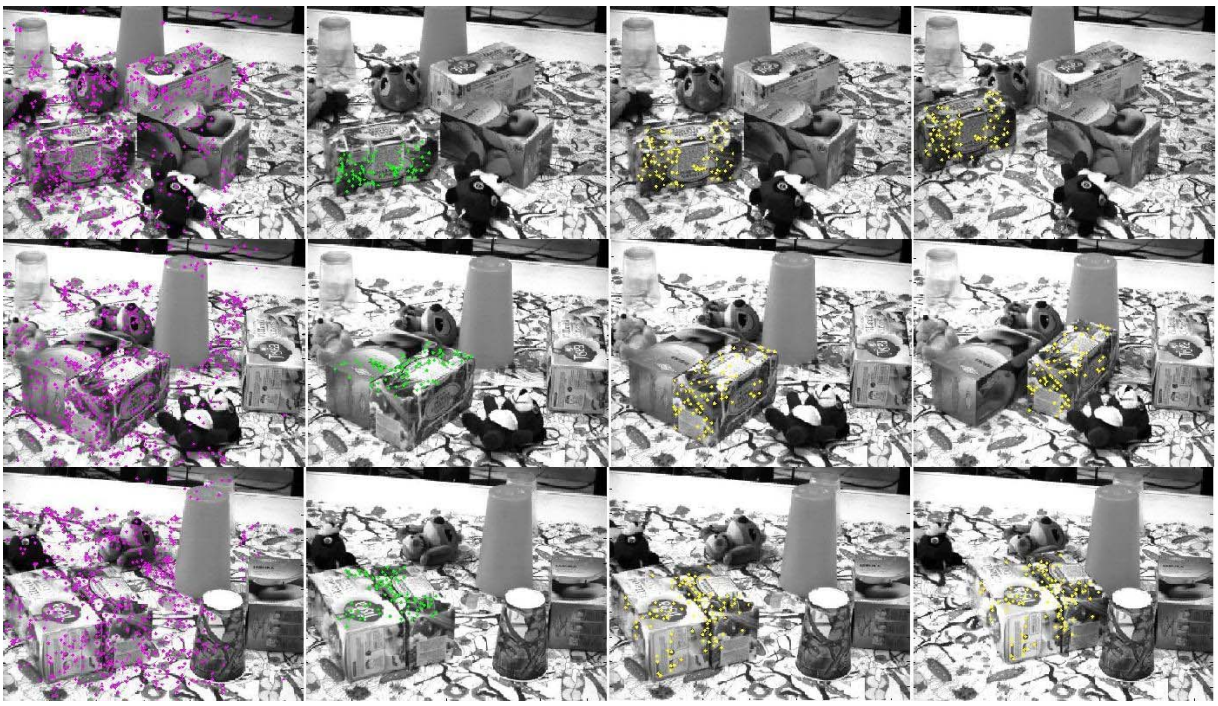

Fig. 8: Experiments. First column presents all 3-d features. Hypotheses are depicted in the second column. Third and fourth column present the determined object features in images before and after manipulation. Adding features to "one-object" hypothesis is presented in the first row. Other rows show determining object features in case of hypotheses involving more than one object.

Success	more than F_R correctly determined 3-D features	58%
Form new hypotheses	less than F_R determined 3-D features	36%
Select next hypothesis	no motion	6%

Tab. 1: Results for 50 experiments. F_R was set to 10. Average number of verified object features was 38.

The built-in logic of our approach is that all features that move in accordance with the rigid body motion principle belong to the same object. Thus the two objects cannot be separated in such cases.

The statistical evaluation is presented in Tab. 1. We labeled the experiment a success if we succeeded to verify more than F_R object features. More than half of our experiments (58%) ended in success. Included as success are also the experiments, where the generated pushing action induced more than one object to move in unison (6% of all experiments).

36% of experiments ended with less than F_R object features. Low number of features was due to difficulties in feature matching over big viewpoint changes. Due to the only partial object information, the manipulation can result in uncontrolled movement of the objects and this can cause big viewpoint changes. According to the authors of [27], no feature detector-descriptor combination performs well with viewpoint changes of more than 15-30° and only a small fraction of all features can be matched for viewpoint changes beyond 30°. In such cases it is often not possible to match the features before and after manipulation. Since the movement has occurred and thus the scene has changed, it is necessary to acquire and process new images to form a new set of object hypotheses.

For 6% of the experiments it was established that the group of features did not move. The features did not move 1) because the pushing actions were not conducted (limited workspace) and 2) because the action did not generate changes in the scene. In such cases, the next hypothesis from the selection of lower-ranked hypotheses is selected.

A. Accumulation of feature points

Accumulation of object features depends on changing the direction of view with respect to the object. On the one hand, the object rotation must be big enough to change the view sufficiently to make previously unseen features visible, but on the other hand, it must be small enough to allow feature matching across different views. Fig. 9 presents the successful accumulation of object knowledge across viewpoints after two consecutive pushing actions.

Fig. 9: Building object model. Hypothesis (left) with object features after first push (middle) and after second push (right).

VI. Discussion and Future Work

In this work we focused on detecting and learning about objects that contain some planar surfaces. We discover object hypotheses by detecting planes in the images processed by the SIFT feature detector. These hypotheses serve as the basis for the generation of pushing actions, which are used to confirm or reject the existence of the object after a successful push. The application of a number of consecutive pushing actions allows us to accumulate object knowledge across viewpoints. Planar surfaces are not the only possible smooth surfaces that can be utilized for this purpose. In the literature, other modeling schemes such as superquadrics [15], generalized cylinders [26], geons [3], etc. have been proposed to reconstruct 3-D surface models. We chose to perform this study with planar surfaces because planes can be characterized by a small number of points, which is important for algorithms such as RANSAC. In addition, we are primarily interested in household environments, which contain many objects with planar surfaces. The detection of other types of smooth surfaces and the generation of the associated pushing actions for the verification of object hypotheses is an important topic of our future research. This will require us to solve larger initial reconstruction problems because other surfaces are normally represented by more than 3 points.

We note that other researchers attempted to support object segmentation by manipulation. The work of Fitzpatrick [10] showed that poking can provide additional cues for segmentation. The segmentation process was based on the first contact between the robot and the object, detected when image motion caused by the robot arm spread across a wider distance than the arm could possibly have moved in the time available. Unlike our approach, this work did not include systematic algorithms for identifying initial object hypotheses and determining optimal pokes to confirm or reject the current hypothesis. Another distinguishing feature of our work is that we provide techniques for accumulating 3-D object data based on knowledge provided by consecutive pushing actions.

Acknowledgment

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

References

[1] K. S. Arun, T. S. Huang, and S. D. Blostein, Least-Squares Fitting of Two 3-D Point Sets, Pattern Analysis and Machine Intelligence, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 9, no. 5., pp. 698-700, 1987.

[2] T. Asfour, K. Welke, P. Azad, A. Ude, and R. Dillmann, The Karlsruhe Humanoid Head, in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, Daejeon, Korea, pp. 447-453, 2008.

[3] I. Biederman, Recognition-by-components: A theory of human image understanding, *Psychological Review*, vol. 94, no. 2, pp. 115-147, 1987.

[4] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern classification (2nd edition)*, Wiley, New York, 2000.

[5] C. Dune, A. Remazeilles, E. Marchand, and C. Leroux, Vision-based grasping of unknown objects to improve disabled people autonomy, in *Proc of Robotics: Science and Systems*, *Manipulation Workshop: Intelligence in Human Environments*, Zurich, Switzerland, 2008.

[6] J. Feldman, What is a visual object? *Trends in Cognitive Sciences*, vol. 7, no. 6, pp. 252–256, 2003.

[7] V. Ferrari, T. Tuytelaars, and L. Van Gool, Simultaneous object recognition and segmentation from single or multiple model views, *Int. Journal of Computer Vision*, vol. 67, no. 2., pp. 159-188, 2006.

[8] S. Fidler and A. Leonardis, Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts, in: *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, Minneapolis, MN, pp. 1-8, 2007.

[9] M. A. Fischler, R. C. Bolles, and R. C., Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, *Communications of the ACM*, vol. 24, no. 6., pp. 381-395, 1981.

[10] P. Fitzpatrick, First contact: An active vision approach to segmentation, in: *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Las Vegas, Nevada, pp. 2161-2166, 2003.

[11] W. T. Freeman, E. H. Adelson, The design and use of steerable filters, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, no. 9., pp. 891-906, 1991.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2004.

[13] G. Kootstra, J. Ypma, and B. de Boer, Active exploration and keypoint clustering for object recognition, in: *Proc. IEEE Int. Conf. Robotics and Automation*, Pasadena, CA, pp. 1005-1010, 2008.

[14] D. Kraft, N. Pugeault, E. Baseski, M. Popovic, D. Kragić, S. Kalkan, F. Wörgötter, and N. Krüger, Birth of the object: detection of objectness and extraction of object shape through object–action complexes, *Int. Journal of Humanoid Robotics*, vol. 5, no. 2., pp. 247-265, 2008.

[15] A. Leonardis, A. Jaklič, and F. Solina, Superquadrics for segmentation and modeling range data. *IEEE Trans. on Pattern Recognition and Machine Intelligence,* vol. 19, no. 11, pp. 1289-1295, 1997.

[16] W. H. Li, A. M. Zhang, and L. Kleeman, Bilateral Symmetry Detection for Real-time Robotics Applications, *The Int. Journal of Robotics Research*, vol. 27, no. 7., pp. 785-814, 2008.

[17] D. G. Lowe, Object recognition from local scale-invariant features**,** in: *Proc. Int. Conf. Computer Vision,* Corfu, Greece, pp. 1150-1157, 1999.

[18] D. G. Lowe, Local feature view clustering for 3d object recognition, in: *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Kauai, Hawaii, pp. 682-688, 2001.

[19] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[20] J. Matas, O. Chum, M. Urban, and T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: *Proc. of British Machine Vision Conference, vol. 1*, pp. 384-393, 2002.

[21] G. Metta and P. Fitzpatrick, Grounding vision through experimental manipulation, *Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Sciences*, vol. 361, no. 1811, pp. 2165-2185, 2003.

[22] G. Metta and P. Fitzpatrick, Early integration of vision and manipulation, *Adaptive Behavior*, vol. 11, no. 2., pp. 109-128, 2003.

[23] K. Mikolajczyk and C. Schmid, A performance evaluation of local descriptors, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615-1630, 2005.

[24] K. Mikolajczyk and C. Schmid, Scale & Affine Invariant Interest Point Detectors, *Int. Journal of Computer Vision*, vol. 60, no. 1, pp. 63-86, 2004.

[25] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, A comparison of affine region detectors, *Int. Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43-72, 2005.

[26] R. Mohan and R. Nevatia, Using Perceptual Organization to Extract 3D Structures, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 1, pp. 121-139, 1989.

[27] P. Moreels and P. Perona, Evaluation of features detectors and descriptors based on 3d objects, in *Proc. IEEE Int. Conf. Computer Vision*, Beijing, China, pp. 800-807, 2005.

[28] R. M. Murray, Z. Li., and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulation*, CRC Press, 1994.

[29] T. Kadir, A. Zisserman, and M. Brady, An affine invariant salient region detector, in: *Proc. European Conf. Computer Vision*, Prague, Czech Republic, pp. 228-241, 2004.

[30] D. Pelleg and A. Moore, X-means: Extending K-means with Efficient Estimation of the Number of Clusters, in: *Proc. Seventeenth Int. Conf. Machine Learning*, San Francisco, CA, pp. 727-734, 2000.

[31] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, 3d object modeling and recognition using affine-invariant patches and multi-view spatial constraints, in: *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. II-272-7, 2003.

[32] A. Saxena, J. Driemeyer, and A. Y. Ng, Robotic grasping of novel objects using vision, *The Int. Journal of Robotics Research*, vol. 27, no. 2, pp. 157-173, 2008.

[33] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets. in: *Proc. European Conf. Computer Vision*, Copenhagen, Denmark, pp. 414–431, 2002.

[34] C. Schmid and R. Mohr, Local Greyvalue Invariants for Image Retrieval, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, pp. 530-535, 1997.

[35] E. Trucco, F. Isgro, and F. Bracchi, Plane detection in disparity space, in *Proc. Int. Conf. Visual Information Engineering*, Guildford, UK, pp. 73–76, 2003.

[36] C. J. Tsikos and R. K. Bajcsy, Segmentation via manipulation, *IEEE Trans. on Robotics and Automation*, vol. 7, no. 3, pp. 306-319, 1991.

[37] Tuytelaars, L.Van Gool, Matching Widely Separated Views Based on Affine Invariant Regions, *Int. Journal of Computer Vision*, vol. 59, no. 1, pp. 61-85, 2004.

[38] A. Ude, D. Omrčen, and G. Cheng, Making object learning and recognition an active process, *Int. Journal of Humanoid Robotics*, vol. 5, no. 2, pp. 267-286, 2008.

[39] L. Van Gool, T. Moons, and D. Ungureanu, Affine / photometric invariants for planar intensity patterns, in: *Proc. European Conf. Computer Vision*, Cambridge, UK, pp. 642.651, 1996.