2008 IEEE International Conference on
Robotics and Automation
Pasadena, CA, USA, May 19-23, 2008

Object Separation using Active Methods and Multi-View Representations

Kai Welke, Tamim Asfour and Rüdiger Dillmann
University of Karlsruhe (TH), IAIM, Institute of Computer Science and Engineering (CSE)
P.O. Box 6980, 76128 Karlsruhe, Germany
Email: {welke, asfour, dillmann}@ira.uka.de

Abstract— **Daily life objects reveal natural similarities, which cannot be resolved with the perception of a single view. In this paper, we present an approach for object separation using active methods and multi-view object representations. By actively rotating an object, the coherence between controlled path, inner models, and percept is observed and used to reject implausible object hypotheses. Using the resulting object hypotheses, pose and object correspondence are determined. The proposed approach allows for the separation of different object candidates, which have similar views to the current percept. With the benefit of active methods the perceptual task can be solved using even coarse features, which facilitates a compact multi-view object representation. Furthermore, the approach is independent from a specific visual feature descriptor and thus suitable for multi-modal object recognition.**

I. Introduction

In this work, we present an approach, which solves a task that is natural to humans. A humanoid robot which acts in a natural environment has to cope with a large variety of different objects. In order to perceive the surrounding world in a robust manner, the robot has to be able to learn, classify and act on objects encountered in the environment accordingly. The number of different objects imposes a challenging problem for the research in machine vision. Many approaches concerned with object recognition are restricted to a small number of objects and are also restricted to objects that are separable with the feature extraction method deployed. Since visual perception is one of the building blocks of cognition, this restriction hinders the application of cognitive systems in real environments. Inference and reasoning usually require a large set of examples, which covers the variety of percepts required to solve a cognitive task.

One typical problem when dealing with a large number of objects consists in the natural similarities of daily life objects. Many objects can not be discriminated when observed from one unique view point using a single feature descriptor. There are two different approaches to cope with this fact. One possibility consists in the integration of different modalities and sensors to reduce the uncertainty that is imposed by the similarities ([1], [2]). Another possibility consists in the active exploration of objects in order to determine the correct correspondence between acquired object representation and perceived world. In this paper, we present an approach, which uses active vision to reduce the uncertainty deriving from similarities in the world surrounding. Within a coupled action-perception framework, the robot generates as much object views as neccessary to narrow the number of correspondences between object representations and perceived object to one candidate.

Since the active vision paradigm was introduced ([3], [4], [5]), the availability of humanoid robot systems with distinct manipulation capabilities has opened the possibility to study and implement active methods in real environments. Recent research focuses on solving some ill-posed problems in machine vision with active methods. Fitzpatrick et. al use the manipulator of a robot to gain an initial idea of the presence and shape of an object [6]. Omrcen et. al propose a control scheme and an active vision approach, which addresses the problem of figure-ground segmentation [7].

In the work introduced in this paper, an active approach for object separation is presented, which has been designed for the implementation on a humanoid robot platform.

The next section will introduce the underlying principles from cognitive science and neuroscience, which were taken into consideration during the development of the approach. The subsequent section gives a brief overview of the different modules of the proposed system. The algorithm used for interpretation of the current percept will be explained in Section IV. Finally, experimental results are presented and discussed.

II. Guiding Principles

In the following section some of the guiding principles in the development of the system are presented. All principles stand in line with the conviction that perception systems aimed for the application in cognitive systems should agree with the basic findings of cognitive science and neuroscience in the past years. Here we present principles which directly influence the approach, explain to what extent, and give references to work which focuses on similar aspects.

1) Object representation is based on two dimensional views: In our work, we use view-based representations of objects. Psychophysical experiments on humans [8] and on monkeys [9] have lead to a view-based model of how the visual system achieves consistent identification of objects. View-based methods model the object by selected views rather than by constructing a 3D-model to match the object in the scene. Logothetis et al. found cells in the macaque Inferotemporal Cortex (IT) that are tuned to specific views of an object [10]. Psychophysical studies carried on with human subjects indicate that object recognition is performed

978-1-4244-1647-9/08/$25.00 ©2008 IEEE.

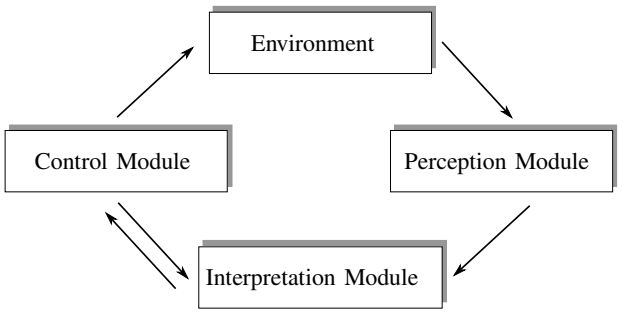


Fig. 1. Overview of the system structure. The system couples control and perception modules via the environment. The interpretation module receives input from perception and control modules and guides the control module.

around views presented while training [11]. In the approach proposed in this paper each object model consists of a set of object views, which are represented by nodes in a spherical graph. This representation of objects is usually referred to as *aspect graph*. In recent research the aspect graph showed to be a feasible representation for multi-view object representations. Aspect graphs usually contain prototypical views of objects, which are used during recognition tasks (see e.g. [12]). Also the extraction of outstanding views of objects as prototypes has been addressed in order to facilitate a more compact representation ([13], [14]). The problem of pose estimation using a combination of local features and aspect graphs is discussed in [15].

2) Representations should allow for multi-modal integration: One possibility to reduce uncertainty during perception consists in the integration of different cues for the task of object separation. There have been many efforts to integrate different visual modalities [16] and modalities from different sensors ([1], [2]) into a unified percept. In this paper an approach is proposed that facilitates the integration of different visual modalities. Feature descriptors can be used with the approach if they are global and rotationally invariant in the viewing plane.

3) Sensory memory is transient and of limited capacity: The model of human sensory memory has been introduced with the Atkinson-Shiffrin memory model [17]. Sensory memory contains rich sensory information and is transient. The proposed system accounts for this model in the sense that rich sensory information is only stored in order to be processed immediately. The approach is designed in a way that, despite the inner models, only the features from the current percept are required for processing.

III. System description

Figure 1 shows the structure of the proposed system. The control module guides the pose of the object using the manipulator of the robot. In simulation, the control module updates the rotation of the object model. The perception module provides the feature extraction methods. In this work global features are extracted from the current percept. The interpretation module validates the coherence between object candidates, percepts and the currently controlled movement. For this purpose path hypotheses are generated on the surface of the viewing sphere and compared with the controlled path. From the path hypotheses the poses between controlled path and object candidates are estimated. Using the pose estimates, the best separating view among object candidates is determined and the controlled movement is adjusted accordingly. Object separation is performed using quality ratings for path hypotheses. Since the movement approaches a view which is ideally valid for one distinct object, only path hypotheses belonging to that object will be plausible and rated accordingly.

All three parts of the system run at different speeds. The interpretation module runs at about 3-5 Hz. Each time an iteration has been completed, the feature of the current percept is requested from the perception module. The timing of the control module is independent. In the experiments a new movement is initiated every second.

As feature descriptor color cooccurrence histograms (CCHs) are used throughout the experiments. CCHs offer some properties, which allow the application in real world recognition tasks as they combine texture information in terms of the distribution of pixel pair colors as well as color information. The resulting description of the objects' appearance is invariant to rotation in the viewing plane and robust to scaling. For a more detailed description the reader is referred to [18]. In our work CCHs based on red and green color channels as well as on the gradient image of red and green color channels are used. The resulting feature vector of 320 dimensions describes the appearance of one object view. With this compact descriptor, the CCHs only exhibit coarse information about the object appearance. Nevertheless, as will be shown in the experimental results, using the combination with active methods, the CCHs are descriptive enough to perform object separation.

A. Building the model

Prior to object separation using the proposed approach, object models in the desired form are generated. Object models consist of aspect graphs with features corresponding to the views associated to each node. The views are generated in simulation by rotating a 3D model of the object and sampling views at equidistant positions. To encode the neighbour relationship between nodes in the aspect graph, Delauney triangulation is performed. All extracted views are processed with the feature extraction method. As the inner model for objects, the extracted features for each view, the node positions, and connections are stored.

B. Perception Module

The perception module extracts descriptors from the appearance of the currently perceived object. Since background subtraction is not necessary in simulation, CCHs are directly extracted from the current view of the object.

C. Control Module

The control module generates views at different viewing angles by rotating the object and makes them available to the perception module. In all our experiments we start

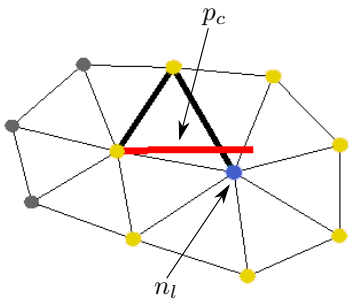


Fig. 2. The initial algorithm connects nodes to the path hypothesis which are neighbours of the last node n_l and similar to the current view generated by the controlled path p_c.

moving from an arbitrary starting position and in an arbitrary direction. Once the most separating view of the object candidates is determined within the interpretation module, control is guided towards this view.

D. Interpretation Module

The interpretation module validates the coherence between object candidates, current percepts and the controlled movement. Path hypotheses are calculated on the surface of the viewing sphere and rated according to their similarity to the controlled path and the current percept. The transformation from controlled path to path hypotheses can be determined and used as a pose estimate. Once the transformation of all object candidates has been calculated, the control movement is guided towards the most separating view. Through the rating of the path hypotheses, the best correspondence for the current percept is calculated within the object candidates.

IV. Path Hypotheses Generation

In order to separate between multiple object candidates the ability of a humanoid robot to actively rotate objects in front of its vision system is exploited. While rotating, the current percept of the object changes and reveals new object views at different aspects. Given the controlled rotation and the sequence of object views, path hypotheses are calculated on the surface of the viewing sphere for all object candidates. In the following, the approach for path hypotheses generation is described.

A. Initial Algorithm

Each path hypothesis p_i is expressed by a sequence of nodes in the following way:

$$p_i = (((\alpha_0, \beta_0), c_0), \ldots, ((\alpha_{N-1}, \beta_{N-1}), c_{N-1})), \quad (1)$$

where α describes the rotation of the node around the vertical axis and β around the horizontal axis. The value c_m describes the number of percepts, which have been valid for the path element m and is referred to as hit counter. From the sequence of rotations executed by the control module, the controlled path p_c is determined in a similar way. The angles α and β describe the currently controlled rotation, the hit counter c_m for each path element of the controlled path p_c is set to one.

In the course of rotating, path hypotheses on the viewing sphere of the object candidates considering the feature of the current view are determined. Algorithm 1 describes the initial idea on how to generate path hypotheses:

Input: current views, object candidate models
Output: path hypotheses $P = \{p_i\}$

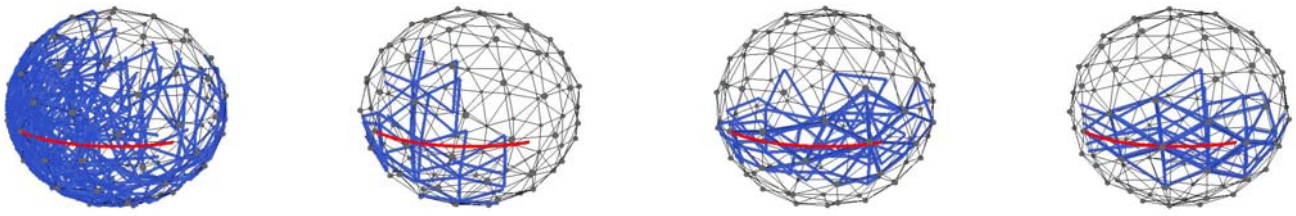
```
1  {p_i} = searchExhaustive(current_view, models);
2  while !converged do
3      {p_i} = selectBest({p_i},N_hypo);
4      foreach Path p_i do
5          n_l = lastNodeOfPath(p_i);
6          {n_j} = neighbours(n_l);
7          {n_j} = {n_j} ∪ n_l;
8          n_b = mostSimilar_w0(current_view, {n_j});
9          if n_b == n_l then
10             increaseHitCounter(n_l);
11         else
12             appendNodeToPath(p_i,n_b);
13         end
14     end
15     converged = determineConvergence({p_i});
16 end
```

Algorithm 1: Initial algorithm for path hypotheses generation

The feature of the current view and all object candidate models are made available as input to the algorithm. In the beginning, all features of the object models are traversed and compared with the extracted feature of the current view. Only the N_{hypo} best matches are stored in the set of path hypotheses P. In each iteration, the last node n_l of each path p_i is considered. With the edges stored during the model building step, the neighbours of n_l are identified. The stored features of all neighbours n_j including the last node n_l are compared to the input feature. The most similar node n_b is determined using the similarity measure w_0 for the feature extraction method used. If the most similar node corresponds to the last node of the path n_l, the hit counter of n_l is increased. Otherwise, the most similar node n_b is appended to the path p_i.

Figure 2 illustrates how the algorithm searches for new nodes. In this example, the object has already been rotated according to the controlled path p_c. During rotation, the path hypothesis is extended from an initial node to two connected nodes. In each iteration, the algorithm searches in the set of nodes connected to the path's last node n_l for the most similar view. If the most similar view is not the last node, the path is extended accordingly.

Figure 3(a) shows the results of the initial algorithm using $N_{hypo} = 50$ after rotating an object 70 degrees around the vertical axis. The generated hypotheses are distributed over the viewing sphere, the control path could not be approximated. This behaviour results from the similarity of CCH features of neighbouring nodes. While the controlled path moves away from the starting node, the current percept is similar to a large amount of nodes, which results in



(a) Initial Algorithm (b) Particles Algorithm (c) Loop Elimination (d) Similarity Accumulation

Fig. 3. All 50 path hypotheses for the four different variations of the algorithm. The object was rotated 70 degrees around the vertical axis. The controlled path p_c is displayed in red. All path hypotheses have been transformed according to the estimated pose.

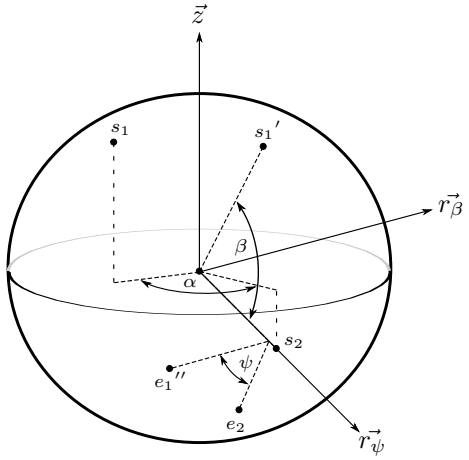


Fig. 4. Transformation of the path (s_1, e_1) into the path (s_2, e_2). First a rotation around the z-axis with angle α is performed until the projection into the xy-plane of the startingpoints s_1, s_2 are collinear with the sphere center. Then a rotation with the angle β is performed in order to tranform s_1 to the position of s_2. Finally the rotation ψ is performed to ensure that the transformed endpoint $e_1{}''$ lies on the same arc as s_2 and e_2. Note that only relevant parameters are depicted in the figure.

the generation of spurious paths. Once a path is generated that leads to an incorrect direction, the algorithm cannot backup to the correct path, since only neighboured nodes are considered.

In the following we will explain how the initial algorithm was adapted in order to deal with coarse features like CCHs. All adpations were accomplished taking into account the guiding principles presented in section II.

B. Hypothesis generation using Particles

In the initial algorithm the knowledge of the controlled path p_c was not considered. Path hypotheses are only generated on base of the similarity of features from the current percept and nodes of the object models. Since the controlled path is known, path hypotheses can be rated according to their similarity not only to the current view, but also to the course of the controlled path. To establish a rating for the quality of the path course, the similarity between controlled path and each path hypothesis is determined.

In order to obtain comparable paths, the rotations required to transform a hypothesis path to the controlled path are calculated. Figure 4 illustrates how a path hypothesis with starting point s_1 and end point e_1 is transformed into the controlled path with starting point s_2 and end point e_2. Three rotations with different rotation axes are performed. The first two rotations with the angles α and β ensure that the starting points of both paths coincide. The third rotation with the angle ψ assures that the transformed starting point s''_1 and end point $e_1{}''$, the starting point s_2, and the end point e_2 lie on the same arc of the viewing sphere.

Overall the complete transformation t_i can be described with the following parameters:

$$t_i = (\alpha, \beta, \psi, \vec{r_\beta}, \vec{r_\psi}), \tag{2}$$

where $\vec{r_\beta}, \vec{r_\psi}$ are the axes for the rotation with β and ψ respectively. The initial rotation α is performed around the z-axis.

Once the transformation has been determined, the similarity between hypothesis and controlled path is calculated. In order to compare paths according to their elements, it is ensured during the composition of the controlled path that the overall number of hits of hypothesis path and controlled path is identical. This is achieved by adding one element to the controlled path once a new iteration of the hypothesis generation is started and setting the hit counter to one. Let $h(p, e)$ be the function that returns the path element, which is valid at the time of the e-th hit. The similarity of the path's course is calculated in the following way:

$$w_1(p_c, p_i) = \frac{\sum_{e=0}^{H-1} d_{t_i}(h(p_c, e), h(p_i, e))}{H}, \tag{3}$$

where H is the sum of hits over the complete path and d_{t_i} the distance of the transformed element on the path p_i and the corresponding element on the controlled path p_c.

In order to integrate the path course rating, an approach similar to particle filtering is deployed. In a hypothesis generation step, different hypotheses referred to as particles are generated. In a verification step, the best particles are determined and kept for the next iteration of the algorithm. The initial algorithm is altered in the following way. Instead of appending the most similar node to a hypothesis path (Alg. 1 line 12), particles are generated for each neighbour of the last node n_l and the hit counter of the node n_l in the currently considered particle is increased. After all path hypotheses have been generated, the similarity of the path courses with the controlled path $w_1(p_c, p_i)$ is calculated. Furthermore, the similarity of the last new node n_l of each

path hypothesis with the current view is determined using the similarity measure w_0. In order to combine both measures independent from their actual values, all path hypotheses are inserted into one priority queue for each measure. The overall rating is calculated by the mean position of the hypotheses in both queues. Since the number of generated hypotheses is increased with this approach, only the N_{hypo} hypotheses with the best overall rating are stored for the next iteration.

Figure 3(b) illustrates all path hypotheses after applying the adapted algorithm to a rotation of 70 degrees around the vertical axis. The course of the generated paths still differs from the controlled path. Again the similarities between neighbouring nodes lead to the generation of spurious path hypotheses. The similarity measure w_0 causes the paths to loop between similar nodes which results in a poor rating using the measure w_1.

C. Loop Elimination

In order to take this behaviour into consideration, a loop elimination step is integrated into the algorithm. With the restriction to loop free control movements, all segments of a path that have the same start and end node are removed. In a second step multiple occurrences of the same path are eliminated. This is necessary to keep the number of required hypotheses small. Figure 3(c) illustrates the resulting hypotheses after applying loop elimination and the removal of identical paths. The generated hypotheses form a good approximation of the controlled path.

D. Similarity Accumulation

Until this point, only the similarity of the last node of each path hypothesis and the current view were considered for the rating of path hypotheses. In order to improve the approximation of the controlled path, the best similarity encountered while rotating the object is stored for each node. These best similarities can be accummulated in another measure

$$w_2(p_i) = \frac{\sum_{j=0}^{N-1} b(n_j)}{N}, \tag{4}$$

where N is the number of nodes of path p_i and $b(n_j)$ denotes the best similarity of an input view to node n_j as encountered while rotating. The measure w_2 is integrated with the measures w_0 and w_1 in the same way as described above to generate an overall path hypothesis rating. The resulting paths after integrating similarity accumulation are illustrated in Fig. 3(d). The visible difference between the outcome with and without similarity accumulation is marginal. Nevertheless in the results section we will show how the convergence of the algorithm can be improved by introducing the measure w_2.

E. Object Separation

In order to calculate the best separating view between objects, the relative pose between the object candidates is required. This pose is derived from the transformations of the path hypotheses to the controlled path. The calculation of the best separating view is initiated, once the running variance of the mean transformation of all path hypotheses is below a threshold. For all object candidates, the running variance is calculated using the mean rotation angles α, β and ψ over all path hypotheses. A window size of 5 is used for the running variance and the sum of the variances of all three angles is used to establish a threshold for convergence of the transformation. The relative pose is approximated by the mean of all rotation angles for each object candidate.

With the relative pose, the aspect graphs of all object candidates are transformed into one common base coordinate system in order to calculate the similarity graph. The nodes of each object candidate are associated to the closest nodes of the similarity graph with respect to the common coordinate system. In order to derive a measure for the similarity of corresponding nodes, the variance of the features from the set of associated nodes is calculated and stored in the similarity graph. Figure 5(a) illustrates the similarity graph for two object candidates with different textures on their backsides. The most separating view is determined using the variance stored within nodes in the similarity graph. In order to account for inaccuracies in the pose estimation, the mean of the variances of each node including its neighbours is considered and the node with the highest mean variance is chosen as most separating view.

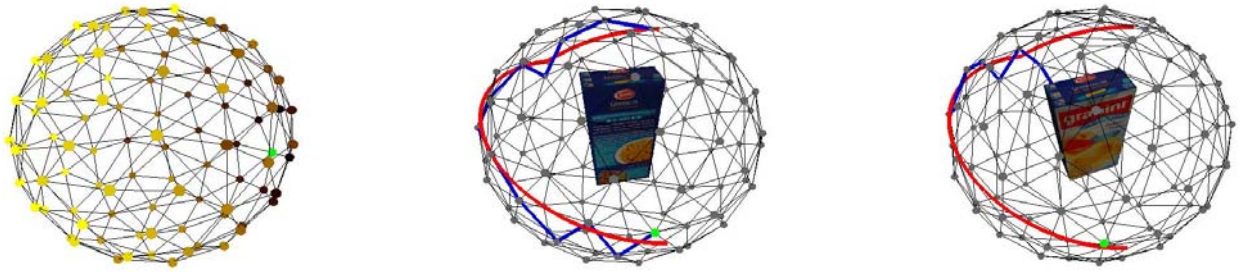
Since the transformations between controlled path and object candidates are known, the position of the best separating view is available in the control coordinate system. The shortest path between the current view and the best separating view is calculated and the movement is adapted accordingly.

In order to separate between object candidates, the rating of the best path hypotheses for each object candidate is considered. Figure 5(b) illustrates the estimated path for a valid object candidate. The controlled path is approximated well, which results in a good rating. In Fig. 5(c) the backside of the object was exchanged compared to the inner model. Once object views are generated that are not corresponding to the inner model, the algorithm will not be capable of finding a good approximation to the controlled path. This results in a poor path rating, which is exploited for object separation.

V. Experimental Results

All results in this paper were achieved in simulation. Block-shaped object models were used to generate views at desired viewing angles. The application of block shapes does not imply a simplification for the approach. Since the algorithm has to cope with planar surfaces, many neighbouring views of one object look similar and thus do not reveal information which can be exploited to generate path hypotheses.

All inner models were represented with aspect graphs consisting of 100 views. The number of maximum hypotheses per iteration was set to $N_{hypo} = 50$. Both parameters were chosen empirically prior to the experiments. The approach was evaluated using randomly selected starting points in spherical polar coordinates in all experiments. All inner models were transformed using random rotation axes and



(a) Similarity graph as calculated from two object candidates and related estimated poses.

(b) The best path hypothesis for the correct object candidate approximates the controlled path p_c.

(c) The best path hypothesis for the incorrect object candidate does not converge to the controlled path p_c.

Fig. 5. Object Separation

angles. The initial direction of the movement was also selected randomly with constant increments in the azimuth and zenith in spherical polar coordinates. By using spherical polar coordinates, it was ensured that not only straight paths were generated.

A. Path hypotheses accurracy

The parameters for path hypotheses convergence were chosen in a way that the generated pose estimates are accurate enough to identify and reach the most separating view. During model building, the viewing sphere was discretised to 100 views. The mean angle between neighbouring views amounts to 22.6 degrees. This gives a benchmark to select the thresholds for convergence accordingly.

Figure 6(a) illustrates how the different variations of the initial algorithm perform in comparison. The results were achieved using views of the same model for the representation as well as for the view generation in order to allow convergence of the path hypotheses. To capture the correct trend for the different approaches, 100 runs with random starting points and transformations were performed for each variation. The graph shows the development of the sum of mean running variances over the three rotations measured during 100 runs. The initial algorithm does not converge to a robust object pose estimation. With the introduction of particles the pose estimation performs better but settles down at a very high variance. The restriction to loop free paths results in much more accurate approximations of the controlled path on the surface of the viewing sphere. Combined with the elimination of similar paths this approach converges very fast. The accumulation of views also improves the convergence of the hypotheses.

To measure the accuracy of the pose approximation process, the mean pose error was calculated for 100 runs. Each run was limited to 400 iterations. If the path hypotheses did not converge after 400 iterations, the run was marked as failure and the pose estimation was not considered. In eight out of the 100 runs failures occurred because the randomly chosen controlled path was close to the singularities of the spherical polar coordinate system. In this case, only few very similar views were presented to the system, which are not sufficient to allow detemining a stable estimate. The resulting mean pose error using the remaining 92 runs amounts to 14.89 degrees, which lies within in the desired accuracy of 22.6 degrees.

B. Object Separation

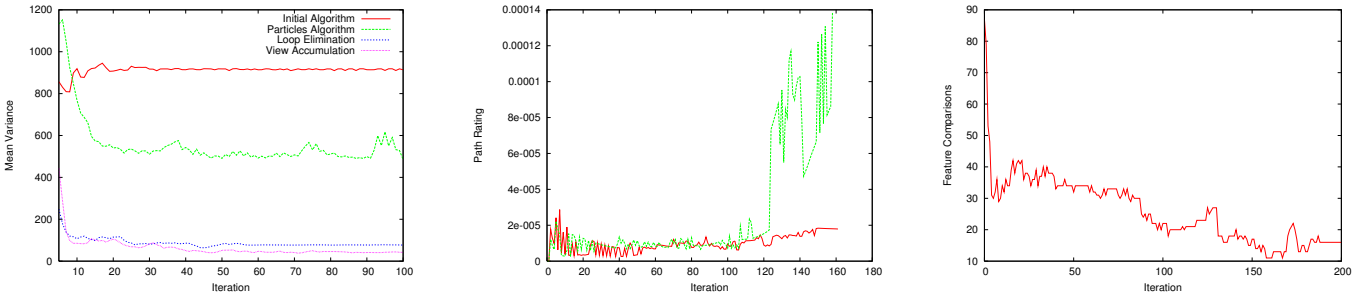
To evaluate object separation, two pairs of object models were deployed. In each pair, the backside of the objects differed. One of the objects from each pair was used as input and both objects of one pair were stored as inner models. The task consisted in the identification of the correct correspondence for the input object.

In Fig. 6(b) the path rating measure w_1 for the best path hypothesis is shown for both object candidates over the iterations until the movement reached the most separating view. Once a view is revealed which is not coherent with one of the object candidates, the distance between controlled path and best path hypothesis increases immediately. This circumstance is deployed to discriminate between the object candidates. We assign a perceived object to an object candidate if its overall path rating is 5 times better than the path rating of the remaining object candidates.

Using this threshold, 100 separation tasks were performed. For each task the convergence of the pose estimation and the calculated correspondence were determined. In the 100 test cases, the object was assigned to the correct object models with only the path course rating w_1 in 88% of the cases. In the remaining 12% of the cases, the path course rating alone was not sufficient to separate between the object candidates. With the help of the similarity rating for the current percept w_0 and the view accumulation rating w_2, 93% of presented objects could be assigned to the correct object candidate. The remaining 7% of tries failed because of controlled movements close to singularities resulting from the random choice of starting view and movement. Singularities can be avoided by restricting movements to be far from singularities, which reveal the necessary amount of object views.

C. Performance

All experiments were accomplished using an Intel Centrino Duo 2.0 GHz notebook. The interpretation module achieved a cycle time of about 3-5 Hz. As in most vision



(a) Convergence of the mean variance of path hypotheses for all four different variations of the algorithm. The best results could be achieved using the particles approach with similarity accumulation.

(b) The path rating measure w_1 is shown for two object candidates with different backsides. Once the backside is revealed, the rating of the invalid hypothesis increases.

(c) Number of feature comparisons required in relation to the iterations of the algorithm. With convergence of path hypotheses, the number of required comparisons decreases.

Fig. 6. Experimental results using the proposed approach.

applications, feature extraction and comparison is the most time consuming task. Compared to a brute-force algorithm, where in each iteration the 100 nodes of the complete inner model are compared with the current view, the hypotheses generation reduces the necessary feature comparison. Figure 6(c) illustrates the number of comparisons between different features required in relation to the number of iterations of the algorithm. Initially the complete neighbourhood of all candidate views has to be examined. Once path hypotheses develop, only the neighbouring nodes of the last path element have to be examined. Since the pose converges towards the same controlled path, many hypothesis paths have the same last nodes and examine the same neighbours. This helps reducing the required feature comparisons to about 14 when the hypotheses converge.

VI. Conclusion

In this work, an active vision approach for object separation and pose approximation is presented. The experimental results show that the proposed approach allows to reliably separate between multiple object candidates. Also it has been shown that the transformation from the path hypotheses to the controlled path gives a good approximation of the object pose. Both results, pose and object correspondence, are achieved within the same framework only on basis of the coherence between controlled movement and percept. With the proposed approach good performance is achieved using a coarse feature descriptor, which allows to keep the inner models of objects compact. Each object is described by only 32000 feature values. This reveals the benefit that can be obtained by using active methods in solving visual tasks.

Acknowledgment

The work described in this paper was conducted within the EU Cognitive Systems project PACO-PLUS (FP6-2004-IST-4-027657) funded by the European Commission.

References

[1] T. Cooke, S. Kannengiesser, C. Wallraven, and H. H. Bülthoff, "Object feature validation using visual and haptic similarity ratings," *ACM Transactions on Applied Perception*, vol. 5, pp. 1–23, 2006.

[2] C. Beltrán-González and G. Sandini, "Visual attention priming based on crossmodal expectations," in *IEEE Int. Conf. on Intelligent Robots and Systems (IROS 2005)*, 2005.

[3] J. Aloimonos, I. Weiss, and A. Bandopadhay, "Active vision," *International Journal on Computer Vision*, pp. 333–356, 1987.

[4] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 996–1006, 1988.

[5] D. H. Ballard, "Animate vision," *Artif. Intell.*, vol. 48, no. 1, pp. 57–86, 1991.

[6] P. Fitzpatrick and G. Metta, "Grounding vision through experimental manipulation," *Royal Society of London Philosophical Transactions Series A*, vol. 361, pp. 2165–2185, Oct. 2003.

[7] D. Omrcen, A. Ude, K. Welke, T. Asfour, and R. Dillmann, "Sensorimotor processes for learning object representations," in *Proceedings of the IEEE/RAS Int. Conf. on Humanoid Robots (Humanoids 07)*, 2007.

[8] M. Tarr, P. Williams, W. Hayward, and I. Gauthier, "Three-dimensional object recognition is viewpoint dependent," *Nature Neuroscience*, pp. 275–277, 1998.

[9] N. Logothetis, J. Pauls, H. Bülthoff, and T. Poggio, "View-dependent object recognition by monkeys," *Current Biology*, vol. 4, pp. 401–414, 1994.

[10] H. Bülthoff and S. Edelmann, "Psychophysical support for a two-dimensional view interpolation theory of object recognition," in *Proceedings of the National Academy of Sciences*, vol. 89, 1992, pp. 60–64.

[11] N. Logothetis, J. Pauls, and T. Poggio, "Shape representation in the inferior temporal cortex of monkeys," Current Biology 5, pp. 552–563, 1995.

[12] C. M. Cyr and B. B. Kimia, "A similarity-based aspect-graph approach to 3d object recognition," *Int. J. Comput. Vision*, vol. 57, no. 1, pp. 5–22, 2004.

[13] H. Yamauchi, W. Saleem, S. Yoshizawa, Z. Karni, A. Belyaev, and H.-P. Seidel, "Towards stable and salient multi-view representation of 3d shapes," in *Proceedings of the IEEE International Conference on Shape Modeling and Applications 2006 (SMI'06)*. Washington, DC, USA: IEEE Computer Society, 2006, p. 40.

[14] K. Welke, E. Oztop, G. Cheng, and R. Dillmann, "Exploiting similarities for robot perception," in *Proceedings of the IEEE Int. Conf. on Intelligent Robots and Systems (IROS 2007)*, 2007, pp. 3237–3242.

[15] G. Peters, "Efficient pose estimation using view-based object representations," *Machine Vision and Applications*, vol. 16, no. 1, pp. 59–63, 2004.

[16] N. Krüger and F. Wörgötter, "Multi-modal primitives as functional models of hyper-columns and their use for contextual integration," in *BVAI*, 2005, pp. 157–166.

[17] R. Atkinson and R. Shiffrin, "Human memory: A proposed system and its control processes," *The psychology of learning and motivation*, vol. 8, 1968.

[18] S. Ekvall and D. Kragic, "Receptive field cooccurrence histograms for object detection," in *Proceedings of the IEEE Int. Conf. on Intelligent Robots and Systems (IROS 2005)*, 2005.