Integrating perception and problem solving to predict complex object behaviours

Damian M. Lyons[a], Sirhan Chaudhry[a],
Marius Agica[b] and John Vincent Monaco[b],
[a]*Fordham University, Robotics and Computer Laboratory, Bronx NY 10458;*
[b]*Pace University, Department of Computer Science, New York NY 10023*

ABSTRACT

One of the objectives of Cognitive Robotics is to construct robot systems that can be directed to achieve real-world goals by high-level directions rather than complex, low-level robot programming. Such a system must have the ability to represent, problem-solve and learn about its environment as well as communicate with other agents. In previous work, we have proposed ADAPT, a Cognitive Architecture that views perception as top-down and goal-oriented and part of the problem solving process. Our approach is linked to a SOAR-based problem-solving and learning framework. In this paper, we present an architecture for the perceptive and world modelling components of ADAPT and report on experimental results using this architecture to predict complex object behaviour.

A novel aspect of our approach is a 'mirror system' that ensures that the modelled background and foreground objects are synchronized with observations and task-based expectations. This is based on our prior work on comparing real and synthetic images. We show results for a moving object that collides and rebounds from its environment, hence showing that this perception-based problem solving approach has the potential to be used to predict complex object motions.

Keywords: cognitive robotics, problem-solving, simulation, computer vision, sensory fusion.

1. INTRODUCTION

Cognitive Robotics aims to build robot systems capable of reasoning about all the kinds of complex physical phenomena that occur in everyday, real-world interactions. We have developed an approach to this problem based on using an open source 3D game engine. Informally, the simulation can act as the 'imagination' of the robot, allowing it to carry out a particular kind of thought experiment: allowing the simulated world to run faster than the real world for the purposes of prediction. A key novelty of our approach is that the output of the game engine is a synthetic 'image' of the predicted world, which can be compared directly to the image from the robot's visual sensor. In this way, problem solving becomes an integrated part of the robot's perceptual processes.

Comparing a graphical, synthetic image and real image of the same scene poses many problems; in previous work [9], we propose an approach called the *match-mediated difference* (MMD) approach that allows for effective comparisons of synthetic and real images of similar scenes, even scenes with objects added or removed. We present a schema-based architecture that builds on MMD to synchronize the 3D simulation with visual observations of the surrounding scene and moving object behavior. We refer to this as a 'mirror system' because independent of task actions it causes the simulation to mimic visual observations within the context of the task. Section 2 briefly reviews prior work. Section 3 introduces the minimal subscene and architecture of the mirror system. Section 4 shows example results for misaligned background scenes and for stationary and moving objects.

2. PRIOR WORK

Cognitive functions such as anticipation and planning operate through a process of internal simulation of actions and environment [14]. Indeed there is a history in the field of Artificial Intelligence of using 'simulated action' as an algorithmic search procedure, e.g., game trees, though such an approach typically had problematic computational complexity. Shanahan [14] proposes a large-scale neurologically plausible architecture that allows for direct action

(similar to a behavior-based approach) and also 'higher-order' or 'internally looped' actions that correspond to the 'rehearsal' or simulation of action without overt motion. The Polybot architecture proposed by Cassimatis et al. [5], and based on his Polyscheme cognitive architecture, implements planning and reasoning as sequences of 'mental' simulations that include perceptive and reactive subcomponents. The simulations include not just the effect of actions, but also the understood 'laws' of physics (e.g., will a falling object continue to fall) and are implemented as a collection of specialist modules that deliberate on propositions of relevance to the robot.

In previous work we have introduced ADAPT[2][3] an architecture for cognitive robotics. ADAPT merges RS [10], a language for specifying and reasoning about sensory-based robot plans with Soar [8], a widely used cognitive architecture. RS, based on Arbib's 'schema theory' [1], represents robot plans as networks of perceptual and motor schemas. We also proposed adding a 3D simulation engine that allows physical scenarios to be simulated as part of planning and learning.

The integration of simulation into the reasoning process has been investigated for assembly and task planning [15]; the integration was achieved by allowing the planning module access to the internal data structures of the simulation. However, that approach is difficult to use in robotics because there is no general way to link the data structures of a simulation with the sensory apparatus of the robot. This is strongly related to the problem of the perceptual anchoring of symbols [6].

In [9] we proposed a unique approach to this problem: allowing the simulation to communicate with the robot in a language common to its sensors – a visual image of the world. Integration of visual and 3D graphical imagery has been considered in applications such as predictive teleoperation [4]. Our problem however requires comparing the synthetic and real imagery to look for differences between actual and predicted object behaviours. We have developed an approach called the Match-Mediated Difference (MMD) image that allows effective comparison of real and synthetic views of a scene. The MMD image operation also allows the real and synthetic camera poses to be synchronized. However, if the simulation is to be used to predict complex object motions such as the effects of object collisions, then the simulation needs to be forced to 'mirror' visual observations for objects and actions related to ongoing activities.

3. VISUAL IMAGINATION: MIRRORING REALITY

ADAPT consists of an RS module that represents active, ongoing robot plans and sensing and a deliberation module based on Soar. To provide the 'mirror system' discussed in the previous section, additional structure is added to the RS model of ADAPT and we introduce that structure in this section.

3.1 Minimal Subscene

Itti and Arbib [11] define the *minimal subscene* as the middle ground between language and visual attention. Salient objects, the actions associated with them, and other objects associated with those actions are recursively gathered into the minimal subscene which then provides the context for discourse. We adopt this concept, and in our case, the minimal subscene provides a perceptual, problem solving context (Figure 1).

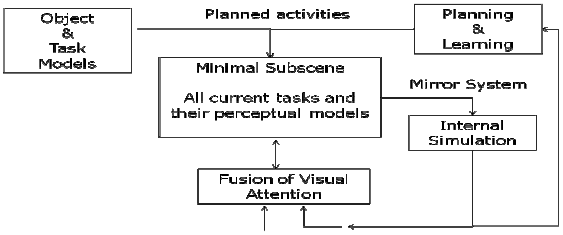


Figure 1: The Minimal Subscene

The minimal subscene is composed of a network of sensory and motor schema, put in place partially by the Soar module (top-down) and partially by ongoing perception (bottom-up). The elements of the subscene have corresponding elements in the simulation module. The focus of visual attention module integrates the visual image generated by the simulation and the image from the video camera. To describe this in more detail, we need to introduce our working example.

3.2 Example Application

Our objective is to allow a cognitive robot system to reason about complex physical actions. For the task where a robot must predict the location of a moving object in order to intercept it, e.g., intercept a moving soccer ball, a

behavior-based approach that includes visual tracking of the ball can yield a robust solution (e.g., [12]). However, if the problem is expanded to include the ball moving towards a wall or another unexpected agent then since the dynamics used in tracking a ball typically does not include information about bouncing off walls or other agents, tracking and prediction becomes more challenging. While a fast tracking system might reacquire the ball target after the bounce, it certainly will not be able to predict the bounce, and any action that the robot takes before the bounce will be predicated on the ball continuing its observed path. This puts the robot in the position of always playing 'catch-up' with the ball instead of accurately predicting where the ball will be and moving there. This same issue arises whenever a robot is operating in a complex dynamic environment, for example, an urban search and rescue robot moving on a semi-stable pile of rubble.

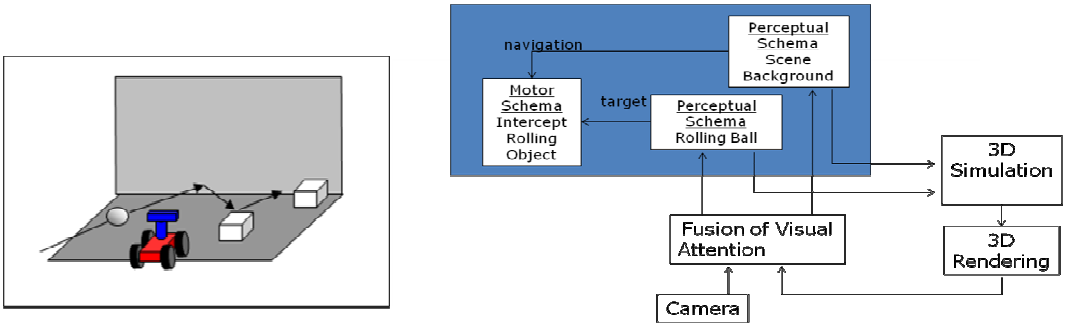


Figure 2: Experimental Scenario; (A) bouncing scenario, (B) The Subscene & Mirror System.

We start with a relatively simple scenario (see Fig. 2(A)): A robot is positioned facing a wall. A ball is rolled across the field of view of the robot ultimately bouncing from the wall. The robot needs to intercept the ball after the bounce. Additional objects are placed by the wall so that ball bounces in a complex manner.

3.3 The Subscene Schema Assemblage & Mirror System

The minimal subscene for this problem involves two perceptual schemas and a motor schema (Figure 2(B)). The *Scene Background* perceptual schema is monitoring the distant or background parts of the environment: the wall, floor, etc. The *Rolling Ball* schema is monitoring the state of the moving ball. The motor schema *Intercept Rolling Object* uses the information from both to predict where the rolling object will go and to move the robot to intercept it.

The interception of the rolling ball is an easy enough problem until the ball collides with the wall and rebounds. Figure 2(B) shows the connections between the subscene and the simulation to implement the 'mirror system' discussed earlier. Each perceptual schema is responsible for both a part of the visual image corresponding to its visual focus of attention, and a part of the simulation, corresponding to the model for its visual focus of attention. The *Scene Background* schema is responsible for the appearance of the area around the robot and the pose of the virtual camera in the scene. For the experiments described in this paper, the background schema was manually constructed by taking camera imagery, unwarping it, and applying it as texture to walls in the simulation. This is a reasonable assumption to make, since the literature contains several approaches to extracting depth and visual information from the environment and using it to construct a 3D model [13]. The principle activity of the mirror system for the scene background is therefore a localization problem: maintaining the simulated camera view of the scene background to be the same as the real camera view of the real world.

For the foreground object – the rolling ball in this case – the situation is similar. We will not address here the issue of creating the simulation object and adding the video texture to it so that it appears similar to the observed object. For the results reported in this paper, we manually added the simulation object and the video texture. We focus instead on the problem of correcting the simulation object behaviour so that it remains synchronized with the observed behavior.

4. SYNCHRONIZING REAL AND SIMULATED WORLDS

4.1 Scene Background

The problem of synchronizing the real and simulated cameras comes down to comparing the image generated by the simulation renderer with the camera image and determining what is the change in position and orientation of the camera between the two. We use the Match-Mediated Difference (MMD) to compare images effectively.

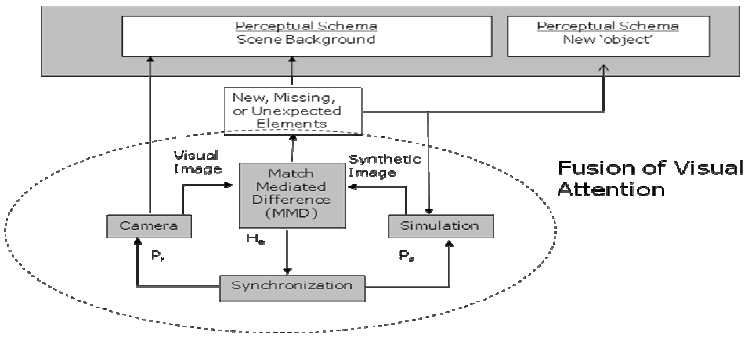


Figure 3: Scene Background Synchronization

Figure 3 shows the mechanism for *Scene Background* synchronization. The real and synthetic images are compared using the MMD and a pose correction H_e generated. The synthetic camera pose is iteratively modified using this correction. If any unexpected areas of difference are generated – that is, any area of difference not being monitored by a perceptual schema, then the *Scene Background* schema trigger a new perceptual schema to model and monitor and area.

4.2 MMD Synchronization

We briefly review the MMD operation here, see [9] for more details. Real and synthetic images of even identical looking scenes produce a large difference image because of the different methods of image generation. The MMD approach looks for common corner features between both images. These matched features are used to first generate a homography mapping one image to the second – this gives the camera orientation correction H_e. Secondly, the matched points are used to generate an MMD mask – if the points really correspond to the same features in both images, then we expect that the difference image should be zero close to these points.

$$I_m(p) = \frac{1}{|P|}\sum_{p' \in P} \frac{q(p')}{S_{p'}} e^{\frac{(p-p')^2}{2v}} \qquad I_d(p) = \frac{|I_s(p) - I'_r(p)|}{I_m(p)} \qquad \textit{eq.1.}$$

The I_m *MMD mask* is composed of Gaussians centered at match points *p*, and weighted by *q(p)* so that good matches produce larger Gaussians and bad matches, smaller Gaussians. The MMD mask is used to weight the difference between the real and warped synthetic images to generate the MMD image I_d. Figure 4(D) shows two examples of MMD masks.

$$H_e = K_s^{-1}\left(R - \frac{\tau n^T}{d}\right)K_r \qquad \textit{eq.2.}$$

The difference between the real camera and synthetic camera pose is given by the vector τ and rotation matrix R in eq.2. In this expression, K_s and K_r are the camera projection matrices for synthetic and real images. We will make the assumption that the position error in small and just use the rule in eq.3. to modify the synthetic camera orientation R_s.

$$R_s(t+1) = R_s(t)\,(g\,K_r^{-1} H_e K_s) \qquad \textit{eq.3.}$$

where $0<g\leq 1$. An algorithm to produce both translation and orientation from eq.2. is given by Guerrero et al. [7].

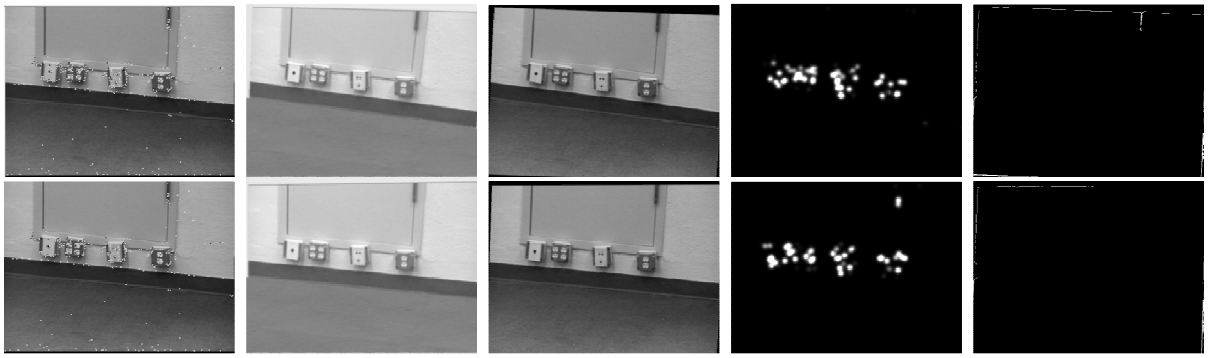
4.3 Results

Figure 4 below shows the real (Fig. 4(A)) and synthetic (Fig. 4(B)) images of the lab wall in our working example. The figure shows the start/end of a sequence of corrections using the approach in section 4.2. It can be seen (Fig. 4(C)) that the camera corrections (eq.3.) align the images and the MMD image shows no difference throughout. Figure 5 shows the results of a sequence of camera corrections Fig. 5(C) through (E) where the image contains a 'valid' difference. The correction proceeds normally and in each case, the MMD shows the 'unexpected' object.

4.4 Scene Foreground

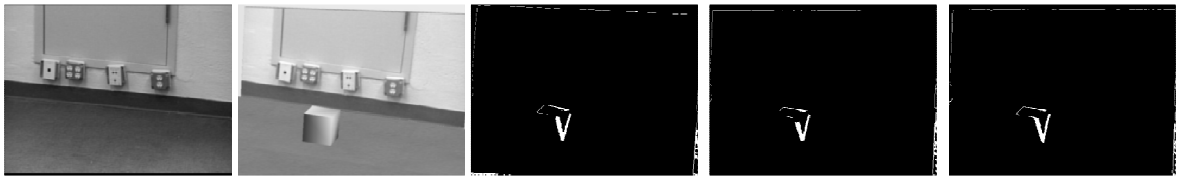
The perceptual schema for a foreground object has the responsibility of both monitoring and modeling: monitoring the visual image for the object and interacting with the simulation to model the object behavior. This is shown in Figure 6, which looks similar to the process in Figure 3. The principle difference is that the output of the MMD is a difference region (e.g., Fig. 8(A)-(C)) and the perceptual schema uses this information to adapt the simulation

parameters of the object so that it more closely follows observed behavior.



(A) (B) (C) (D) (E)

Figure 4: Two steps (t=5, 20) during the 20 step synchronization of real and synthetic images. Column (A) real image, (B) Synthetic image with corner points, (C) warped image, (D) MMD mask, and (E) MMD image.



(A) (B) (C) (D) (E)

Figure 5: Synchronization trace with unexpected object; real (A), synthetic (B), MMD image at t=0,10,20

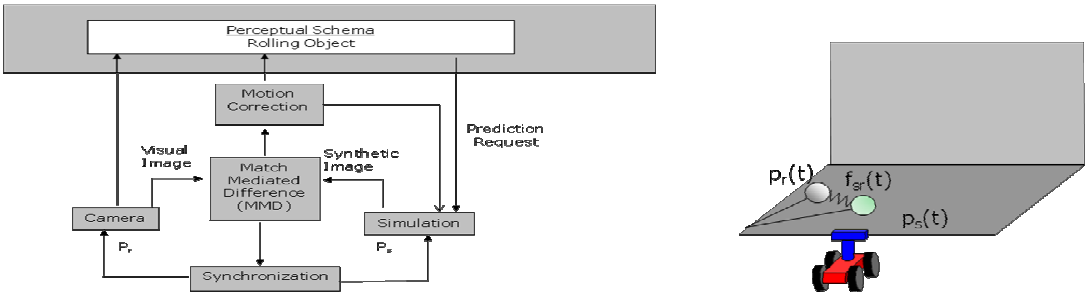


Figure 6: Foreground Object Synchronization (A); Spring correction model (B)

Figure 6(B) shows the correction model used. With knowledge of the camera pose and using a ground plane assumption (if stereo is not available), we calculate the corrective force to apply to the simulated object using a spring rule:

$$f_{sr}(t) = k\,(\,p_r(\,t\,) - p_s\,(t)\,) \qquad \textit{eq.4.}$$

4.5 Results

Figure 7(A) through (C) shows a sequence of (cropped) MMD images from a rolling ball. Each MMD image shows the real object and the synthetic object – a white bounding box has been added around both for clarity. In the top (uncorrected row of images) the two drift apart. In the bottom (corrected row), the simulation is forced to speed up the object and the two remain very close to one another eventually producing a single difference in Fig. 7(C) bottom.

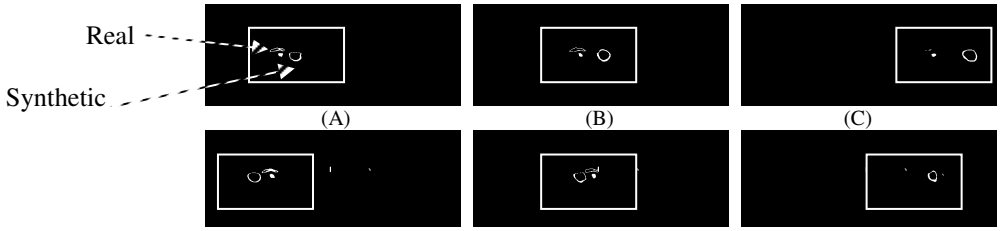


Figure 7: MMD Image sequence for uncorrected rolling object (top); corrected rolling object (bottom).

Figure 8 (A) through (D) shows the result of foreground object motion uncorrected (top) and corrected (bottom) in the case of an object bouncing off the front wall in our working example. Without correction the bounce produces an even worse discrepancy than in rolling (Fig. 8(D) top). However, with correction the two objects almost completely coincide.

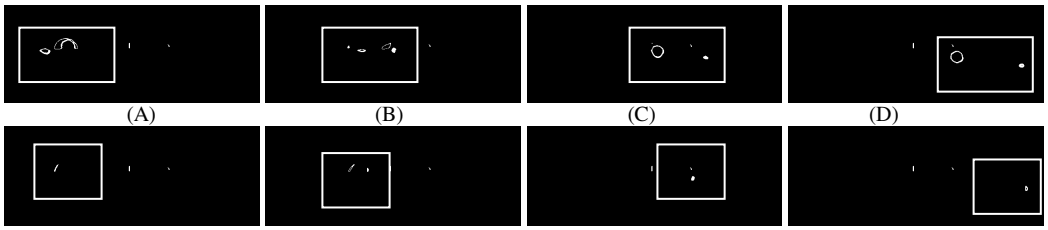


Figure 8: MMD Image sequence for uncorrected bouncing object (top); corrected bouncing object (bottom).

5. CONCLUSIONS

We have developed a schema-based approach to cognitive robotics in which problem-solving is integrated with perception. It builds on our previously developed ADAPT architecture [2] and MMD [9]. A key novelty of our approach is a 'mirror system' that forces synchronization of the background scene and foreground objects and their behaviour between 3D modelling and visual observations. We refer to this as a 'mirror system' because independent of task actions it causes the simulation to reflect visual observations within the context of the task.

This paper has focused on the synchronization of behaviours rather than appearance. However, we need to include automatic appearance modelling of the background scene and foreground objects. We are developing an approach to this using the computer graphics technique of variable levels of detail modelling.

Although the paper shows how the mirror system allows for predictive behaviour, no results are shown for predictive behaviour and that is a second area of ongoing work. A key difficulty there is understanding when sufficient corrections have been made to allow valid prediction.

Finally, the 'visual imagination' functionality described here demands full details for each scenario – though in fact many of the details may not be relevant to the problem. There would be an advantage to allowing the visual simulations to have more of a cartoon quality, concise only in the relevant task details. One avenue we are exploring to implement this is to extend the MMD to mask not only corner features but other, unimportant appearance features also.

References

[1] Arbib, M.A.,"The Handbook of Brain Theory and Neural Networks" (Ed. M.A. Arbib) MIT Press (2003).

[2] Benjamin, D.P., Lyons, D.,Lonsdale, D., ADAPT: A Cognitive Architecture for Robotics**.** *2004 Int. Conf. on Cognitive Modeling,* Pittsburgh PA July 2004.

[3] Benjamin, D.P., Lonsdale, D., and Lyons, D.M., "Embodying a Cognitive Model in a Mobile Robot," Proceedings of the SPIE Conference on Intelligent Robots and Computer Vision, Boston, October (2006).

[4] Burkert, T., and Passig, G., "Scene Model Acquisition for a Photo-Realistic Predictive Display," Proceedings of the Third International Conference on Humanoid Robots, October (2004).

[5] Cassimatis, N., Trafton, J., Bugajska, M., Schulz, A., "Integrating cognition, perception and action through mental simulation in robots" *Robotics and Autonomous* Systems N49, pp13-23 (2004).

[6] Coradeschi, S., Saffiotti, A., "Perceptual Anchoring of Symbols for Action" *Int. Joint. Conf. on AI*, Seattle WA (2001).

[7] Guerrero, J., Martinez-Cantin, R., Sagues, C., "Visual map-less navigation based on homographies" *J. Robotic Systems* V22 N10, pp.569-581 (2005).

[8] Laird, J., Newell, A., Rosenbloom, P., "Soar: An Architecture for General Intelligence", *Artificial Intelligence* **33**, 1987.

[9] Lyons, D.M., and Benjamin, D.P., "Locating and Tracking Objects by Efficient Comparison of Real and Predicted Synthetic Video Imagery," *SPIE Conf. on Intelligent Robots and Computer Vision*, San Jose CA, Jan. (2009).

[10] Lyons, D., and Arkin, R.C., "Towards Performance Guarantees for Emergent Behavior", (Submitted) IEEE Int. Conf. on Robotics and Automation, New Orleans LA, April 2004.

[11] L. Itti, M. A. Arbib, Attention and the Minimal Subscene, **In:** *Action to Language via the Mirror Neuron System,* (M. A. Arbib **Ed.**), pp. 289-346, Cambridge, U.K.:Cambridge University Press, 2006.

[12] Mantz, F., Pieter Jonker, "Behavior Based Perception for Soccer Robots," in: Goro Obinata, Ashish Dutta, Nagoya University (Eds),*Vision Systems* Advanced Robotic Systems, Vienna, Austria, April (2007).

[13] de la Puente, P., Rodriguez-Losada, D., Valero A., and Matia, F., 3D Feature Based Mapping Towards Mobile Robots' Enhanced Performance in Rescue Missions, *IEEE/RSJ Int. Conf. on Int. Robots & Systems*, 2009 St. Louis, USA.

[14] Shanahan, M.P., A Cognitive Architecture that Combines Internal Simulation with a Global Workspace, *Consciousness and Cognition*, vol. 15 (2006), pages 433-449.

[15] Xiao, J., Zhang, L., "A Geometric Simulator SimRep for Testing the Replanning Approach toward Assembly Motions in the Presence of Uncertainties," *IEEE Int. Symp. Assembly and Task Planning*, (1995).