Overview of the 2007 Semantic Robot Vision Challenge Competition

Alyosha Efros and Paul E. Rybski
The Robotics Institute
Carnegie Mellon University

Abstract

In July 2007 the first contest of the Semantic Robot Vision Challenge (SRVC) was held at the annual conference for the Association for the Advancement of Artificial Intelligence (AAAI) in Vancouver, Canada. Four robot teams and two software-only teams entered the event. The performance of the top two robot competitors and the top software-only competitor showed that the technical goals of the competition were within reach of the research community. This paper summarizes the competition and suggests new directions for the future.

Introduction

The first Semantic Robot Vision Challenge (SRVC) was held on July 25-27 2007 in Vancouver, Canada, as part of AAAI 2007 conference. Four teams participated in the challenge: 3 in the robot league and 2 in the software league (one team entered both). The competition consisted of two rounds: a practice round on July 24th and a final round on July 25th. Only the results of the final round were used for scoring. In what follows, we will briefly describe the particular decisions made in the competition setup, the final results, comments on the teams performance and some lessons learned.

Overview of the Competition

Robotics competitions held at venues such as AAAI (Smart *et al.* 2005) and RoboCup (Veloso *et al.* 2000) have been held for over a decade and have been host to many interesting research efforts. These events are designed to attract researchers and encourage them to push the envelope of the state of art. Competitions such as these provide a standardized testbed on which to evaluate their systems and compare results against others participating in the same task.

The SRVC competition is designed to push the state of the art in image understanding and automatic acquisition of knowledge from large unstructured databases of images, such as those generally found on the web. This competition seeks to fuse robotic and computer vision research. In this competition the robots are given a textual list of physical objects that they are expected to find in an environment.

Very little formal structure is given to the names of objects in this list. For instance, if a pen were to be one of the items on the list, the name might be "pen", "blue pen", "ballpoint blue pen", or any other combination. Objects which have specific titles, such as books, movies, or music, have are prefixed with the title "book", "DVD", or "CD" and then the title and potentially the authors at the end. The format of these names is deliberately left to be somewhat ambiguous.

Once the list has been downloaded to the robots via a USB drive, the robots are connected to the Internet and are given an hour to *autonomously* search for visual examples of each of the items on the list. The expectation is that the robots would use the images acquired from this search to build a visual classification database of the items in question. Note that when the robots are given the text list of items, the human team members who brought the robot to the competition are no longer allowed to touch the hardware unless under the direct supervision of a contest organizer (to avoid the temptation of providing "help" to the robot).

After an hour of searching, the robots are disconnected from the Internet and placed into the competition arena. They have 15 minutes to autonomously search the environment and find as many objects as they can. When the time has expired, the robots are removed from the environment and given another 15 minutes to process the data that they collected. Finally, the robots generate a set of output images, at most one per object in the list, where each image is labeled with the name of the object. To receive points for each image, the actual object must be present in the image AND a bounding box must be drawn around that object. The quality of the match of the bounding box to the silhouette of the object determines the score for that object.

Teams were allowed to participate in a robot league as well as a software league. In the robot league, teams brought all of the hardware and software that they needed to do the competition (robots, laptops, cameras, other sensors, etc...) In the software league, the teams only brought their computers. In the robot league, the teams collected their own image data with their robots and cameras. In the software league, the organizers of the competition ran a robot through the arena and collected a set of images in a pseudo-random fashion. These images were given to the competitors of the software league as the official data set.

For further details, please see the competition web page[1] for the most up-to-date rules.

Competition Setup in 2007

The choice and the placement of objects to be used in the challenge is, of course, very important and should be handled with care. Because in this competition we are interested in both, particular object recognition as well as category recognition, we wanted a good mix of particular vs. generic objects. Our goal was to select about 20 random everyday objects and place them in such way as they are likely to be found in normal everyday life. However, the constraints of the current state of computer vision, the availability of data on the Internet, and the physical constraints of the competition space placed some limitations on our selection. As a result, the following was the protocol we used for finding and placing objects:

1. Obtain a set of objects from local sources (supermarkets, office supplies stores, and colleagues houses). Objects that were too big (e.g. higher than a table) or too small (smaller than a candy bar) were not considered. Objects that did not have texture of any sort, nor a distinct shape (e.g. a piece of white paper) were considered too difficult for current computer vision and also discarded.
2. For each object, we checked that images of that object (for particular objects) or of that object category (for generic objects) were available on the Internet given some reasonable subset of the keyword queries into a few standard image search engines (Google Image Search, MSN Search, Flickr, Amazon, LabelMe, etc). Interestingly, this eliminated about 65% of the originally obtained objects.
3. From the remaining objects, we picked the final set, making sure that half were specific objects (e.g. Twix candy bar), and the remainder generic objects (e.g. scientific calculator).
4. The objects were placed in the area in a way that was at least partially arranged by context. We tried to place food and eating utensils together, for example, or books and DVDs. Likewise, we tried to place objects where they would usually be found (vacuum cleaner on the floor, a DVD on the table or chair). However, due to the limitations of the area, this was not always possible).

Following this protocol, we obtained the final set of 19 objects listed here:

1. scientific calculator
2. Ritter Sport Marzipan
3. book “Harry Potter and the Deathly Hallows”
4. DVD “Shrek”
5. DVD “Gladiator”
6. CD “Hey Eugene” by Pink Martini
7. fork
8. electric iron
9. banana
10. green apple
11. red bell pepper
12. Lindt Madagascar
13. rolling suitcase
14. red plastic cup
15. Twix candy bar
16. Tide detergent
17. Pepsi bottle
18. yogurt Kettle Chips
19. upright vacuum cleaner

[1] http://www.semantic-robot-vision-challenge.org

Figure 1: The competition arena populated with objects as well as the robot entry from the University of Maryland). Objects could be found on the floor, on the chairs, and on the tables. Objects were not found on the tables surrounding the arena.

Figure 1 shows a picture of the 2007 competition arena.

Competition Results

When each team has finished processing the collected image data, the images are given to the organizers for scoring. Each returned image was labeled with the name of one of the listed objects. Within the image, the object in question had to be visible and a bounding box was required to be drawn around the object. The organizers examined each image and drew an “idealized” bounding box around that image. The quality of the team’s score was computed as the ratio of the intersection over the union of their bounding box and the organizer’s bounding box.

For each team, we show the number of points scored (see our scoring rules), the number of images returned (one image per object, 19 image max), and the number of times the returned bounding boxes has a non-zero overlap with the correct object (even if the overlap was too small to score points). Figure 5 shows some images of the top results from the competition.

Figure 2: The University of British Columbia robot

Figure 3: The University of Maryland robot

Robot League

1. University of British Columbia (Meger *et al.* 2008) (Figure 2): 13 points, 15 images returned, 7 objects found with non-zero overlap.
2. University of Maryland (Figure 3): 6 points, 2 images returned, 2 objects found with non-zero overlap
3. Kansas State University (Figure 4): 0 points, 3 images returned, 0 objects found with non-zero overlap

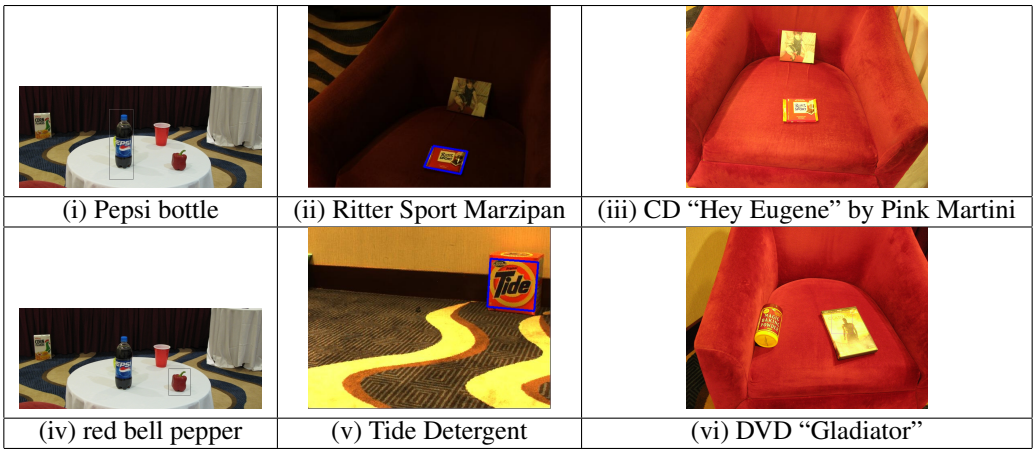
Software-only League

1. Princeton-UIUC: 5 points, 10 images returned, 7 objects found with non-zero overlap
2. Kansas State University: 0 points, 2 images returned, 2 objects found with non-zero overlap

Discussion of the Results

While the number of teams entering the 2007 competition was not high, they contained some of the foremost experts in the field of computer vision as well as robotics. Overall, the fact that even the best-performing team was able to find only about 1/3 of the objects suggests that the level of difficulty set for this challenge was appropriately high.

Figure 4: The Kansas State University robot



(i) Pepsi bottle | (ii) Ritter Sport Marzipan | (iii) CD "Hey Eugene" by Pink Martini

(iv) red bell pepper | (v) Tide Detergent | (vi) DVD "Gladiator"

Figure 5: Several of the top-scoring images from the 2007 teams. Images (i) and (iv) are courtesy of the first-place University of British Columbia robot team. Images (ii) and (v) are courtesy of the second-place University of Maryland robot team. Images (iii) and (vi) are courtesy of the first place UIUC-Princeton software-only team.

As can be seen from the resulting images, all teams had a much easier time with specific objects rather than object categories. In fact, among all the competitors, only two generic objects were found (both by UBC team): red bell pepper and red plastic cup. Interestingly, in the post-competition workshop, the Princeton-UIUC team reported that after running their code on the data overnight, they were also able to find scientific calculator a rather difficult object category.

Most teams reported results as bounding boxes. However, due to the sparsely of features on some of the objects (and the fact that segmentation was not used), several correct detections received 0 points because of small overlap score. An exception was the Maryland team which returned segmented objects rather than simple bounding boxes.

Regarding the robot search strategies, most robots did not use any active vision approaches in effect using the robots for data collection and then switching into software league-mode to analyze the data. The Maryland team went furthest into experimenting with various on-board perception approaches, such as map-building, and surveying. The University of British Columbia team made use of both a monocular camera as well as a stereo camera to assist with the segmentation of the objects in the environment.

Lessons Learned

There are a number of interesting lessons that we learned from this first year of running the challenge. They include:

1. While finding specific objects seems like a problem that is going to be solved within a few years, category recognition appears to be extremely hard.
2. Even for specific objects, using Internet search engines for locating images of objects has many limitations. For example, the same book might exist under several printings, each using a different cover art, and the search engine such as Amazon will likely only retrieve the latest printing. This problem is particularly acute for standard household brads, such as cereals, which might subtly change the design of their boxes every few months or so. Figuring in differences between countries (even USA vs. Canada) means that the same exact brand item might have a high variance in its appearance.
3. Despite the teams best efforts, there still seems to be limited integration between the robot part (navigation, exploration, object scouting, data acquisition) and the actual object recognition. Hopefully, this will improve in the next year. The organizers are considering how to adjust the rules to encourage this.
4. Reasoning about scene context was also quite limited. Most robots were able to find the floor, but did not make any further inferences regarding the surfaces in the scene. Hopefully, this will also improve in the coming years.
5. To make the challenge an entertaining event to watch, in the future the teams will be urged to provide some visual feedback of the robots thinking. This would also argue for moving from the gather-compute mode of operation into an integrated search-to-find mode, allowing for more interaction with the physical environment.

Acknowledgments

The organizers are very grateful to the National Science Foundation for their encouragement and generous financial support for this competition.

References

Meger, D.; Forssen, P.; Lai, K.; Helmer, S.; McCann, S.; Southey, T.; Baumann, M.; Little, J. J.; and Lowe, D. G. 2008. Curious george: An attentive semantic robot. *Robotics and Autonomous Systems special issue - From Sensors to Human Spatial Concepts*.

Smart, W. D.; Tejada, S.; Maxwell, B.; Stroupe, A.; Casper, J.; Jacoff, A.; Yanco, H.; and Bugajska, M. 2005. The AAAI 2004 mobile robot competition and exhibition. *AI Magazine* 26(2):25–35.

Veloso, M.; Kitano, H.; Pagello, E.; Kraetzschmar, G.; Stone, P.; Balch, T.; Asada, M.; Coradeschi, S.; Karlsson, L.; and Fujita, M. 2000. Overview of robocup-99. In Veloso, M.; Pagello, E.; and Kitano, H., eds., *RoboCup-99: Robot Soccer World Cup III*. Berlin: Springer Verlag. 1–34.