Cognitive Vision for Cognitive Systems

Barbara Caputo, Marco Fornoni
Idiap Research Institute
http://www.idiap.ch/~bcaputo
http://www.idiap.ch/~mfornoni
bcaputo@idiap.ch
mfornoni@idiap.ch
Object Recognition -- the robot vision way
What is an Object for a Robot?

- A visual landmark for helping localization, mapping and navigation
- An obstacle to be avoided
- Something to grasp/manipulate
How to recognize it?

- Robots perceive visual information in space and time (video sequences)
- In order to manipulate an object, 3D information is needed
- So... all that is learned in computer vision is useless?
The Semantic Robot Vision Challenge
THE SEMANTIC ROBOT VISION CHALLENGE

Overview

The Semantic Robot Vision Challenge seeks to fuse the state of the art research from various vision and mobile robotics communities. In this challenge, robots must perform a scavenger hunt in a previously-unknown indoor environment. The robots will be given an electronic text file that will contain a list (in English) of objects that must be located in the environment. Before entering the competition arena, the robots will be given time to access the Internet and search it for examples of the objects in order to build a classification database. Once the robot has its database, it will be required to explore the competition arena and perform a computer vision-based search for these objects. At the end of the event, the robots are required to generate a set of images of the objects (taken from the robot’s on-board cameras) where the locations of the objects are highlighted with a bounding box and labeled appropriately. Teams will be scored by how many objects have been correctly located.

**robot league:** for people having their own robot

**software league:** for people not having their own robot

two hours for accessing the Web and learn 20 objects
2007

robot league winner: UBC

software league winner: UIUI
(L. Fei Fei, L. Jia, J. C. Niebles, R. Socher, R. Mehta, B. Collins)
2007

robot league winner: UBC
SRVC Phases

- **Training phase**: Web-crawling and classifier/memory construction.
- **Exploration phase**: Collect photos of potential objects.
- **Classification phase**: Match collected photos and web-crawling images.
Curious George – the Hardware Platform

- ActiveMedia PowerBot.
- SICK LMS200 laser range finder.
- Directed Perception Pan-Tilt Unit PTU-D46-17.5.
- PointGrey Research Bumblebee colour stereo camera.
- Canon PowerShot G7 10MPix 6x optical zoom.
Web-crawling and Image Ranking

- Google image search and other sites are scanned for images that match the list of query objects.
- The obtained images are re-ordered by a ranking algorithm using visual and URL cues.
- Features for obtained images are computed.
SIFT-based Image Matching

- SIFT features (Scale Invariant Feature Transform)
- Feature consistent cliques in training set are found.
- Training set, and exploration phase photos are matched.
- Geometric consistency checking through voting
SIFT-based Image Matching
Perception and Attention System

- **Potential objects** ("proto objects") are detected using spectral saliency and depth from stereo.

- **Saccades**: The PTU then saccades to centre the potential object in the still-image camera. Extent determines zoom.

- **Storage**: A photo is taken and stored for analysing later.

- **Object Persistence**: Location of object is tagged to map to maintain which images are views of the same object.
Saliency

- Fast saliency computation (0.1 sec/frame) based on spectral saliency (Hou et al. CVPR07) and MSER (Matas et al. BMVC02)
Depth from Stereo

- Bundled code from PointGrey Research
- Used to detect objects and obstacles above ground plane
Saccadic Gaze Control

- Fast, since no visual servoing is used.
- Learned via exploratory actions and observation of SIFT features (Forssén CRV07).
Saccadic Gaze Control

- Fast, since no visual servoing is used.
- Learned via exploratory actions and observation of SIFT features (Forssén CRV07).
Usage of Map Data

- **SLAM**: Simultaneous Localisation and Mapping is performed actively, by going to unexplored territory.

- **Stereo Obstacle Avoidance**: Objects detected in the disparity map are temporarily added to the occupancy grid to aid planning.

- **Object Persistence**
CD “Hey Eugene” by Pink Martini
DVD “Gladiator”
yogurt Kettle chips
banana
2007

software league winner: UIUI
(L. Fei Fei, L. Jia, J. C. Niebles, R. Socher, R. Mehta, B. Collins)
Our Approach

1. **Object List**

2. **Robot images**

3. **Get candidate windows**

4. **Crawl the web for data**

5. **Build object template/model database**

6. **Classify and Annotate**
Our Approach

Object List

Crawl the web for data

Build object template/model database

Classify and Annotate

Get candidate windows

Robot images
Our Approach

Specific Objects:
- i.e. branded items, books, cd's
- Little intra class variation
- Can be covered with a few images
- Template matching

General Objects:
- i.e. apple, pear, suitcase, spoon
- High intra class variation
- Need a more general model
Our Approach

Crawl the web for data

Object List

Robot images

Get candidate windows

Classify and Annotate

Build object template/model database
Crawl the web for data

Specific Objects ~
15 Images

Image search engines
Crawl the web for data

General Objects ~ 40 Images

Image search engines
Our Approach

Object List

Robot images

Crawl the web for data

Get candidate windows

Classify and Annotate

Build object template/model database
General object model

Kadir & Brady interest point detector

Compute SIFT descriptor

[Low'96]

Histogram Representation

Vector quantization
Nonparametric topic model

- Hierarchical Dirichlet Process (HDP)

Teh, et al. 2004; Sudderth et al. CVPR 2006;
Wang et al. CVPR 2006
Specific Objects

- SIFT matching (Lowe)

- Find best matching between all templates and testing images independently
Our Approach

Google
Crawl the web for data
Build object template/model database

Object List

Robot images
Get candidate windows
Classify and Annotate
Finding candidate windows

- Kadir&Brady saliency detector
- low resolution image

windows from full resolution image
Our Approach

Object List

Robot images

Crawl the web for data

Get candidate windows

Classify and Annotate

Build object template/model database

Google

picsearch™
Official results
15 min break!
2008

**robot league winner:** UBC
(D. Meger, S. Helmer, S. McCann, M. Muja, M. Dockrey)

**software league winner:** National University of Singapore
(A. Mishra)
The 2008 SRVC
The Final List of Objects

- CD "Retrospective" by Django Reinhardt
- apple
- saucepan
- book "Paris to the Moon" by Adam Gopnik
- remote control
- Spam
- digital camera
- Ritz crackers
- upright vacuum cleaner
- book "Big Book of Concepts"
- DVD "I, Robot"
- banana
- DVD "300"
- game "Crysis"
- eyeglasses
- Doritos Blazin' Buffalo Ranch
- fax machine
- uulu
- Kiwi Strawberry Snapple
- frying pan

- **Missing object**
  - Eyeglasses
- **Distractor objects**
  - Dole Pineapple
  - Reese's Peanutbutter Cups
  - Bounty paper towels
SRVC Team Placements

- **Robot league**
  - 1\textsuperscript{st} Place
    University of British Columbia
  - 2\textsuperscript{nd} Place
    IMRA-Europe

- **Software league**
  - 1\textsuperscript{st} Place
    National University of Singapore
  - 2\textsuperscript{nd} Place
    Universidade de Aveiro
University of British Columbia

- 1st Place @ 12 points
- 5 scoring objects: 8 pts + 4 bonus points
- 1 no-point object
National University of Singapore

- 1st Place @ 12 points
- 6 scoring objects: 12 points
- 2 non-scoring objects

<table>
<thead>
<tr>
<th>Object</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>saucepan</td>
<td>3 points</td>
</tr>
<tr>
<td>Ritz</td>
<td>3 points</td>
</tr>
<tr>
<td>I, Robot</td>
<td>3 points</td>
</tr>
<tr>
<td>Fax</td>
<td>1 point</td>
</tr>
<tr>
<td>Spam</td>
<td>1 point</td>
</tr>
<tr>
<td>Vacuum</td>
<td>1 point</td>
</tr>
<tr>
<td>banana</td>
<td>0 points</td>
</tr>
<tr>
<td>ulu</td>
<td>0 points</td>
</tr>
</tbody>
</table>
Method Overview

- Training data filtering
- Robot hardware
- 3 steps in locating objects:
  - **Exploration**: covering the environment and building a geometric map
  - **Visual attention**: quickly identifying potentially interesting regions
  - **Viewpoint planning**: focus attention on these regions from numerous angles
- Information from each step is added to a **spatial-semantic** map
Data Sources

- **Google Image Search**
  - Large numbers of images
  - Lowest noise ratio of common search engines
- **Walmart product images**
  - Fewer images per category
  - Very high image quality
Training Data Filtering
Hardware
Visual Coverage

- Peripheral camera covers environment to identify promising objects
Visual Attention

- Potential objects identified with **multi-scale visual saliency** and 3-D structure
- Required to be **fast** since this computation is done on all peripheral pixels
Multi-Scale Saliency
Multi-Scale Saliency

- Saliency computation based on spectral residual saliency [Hou et al. CVPR '07] and Maximally Stable Extremal Regions (MSER) [Matas et al. BMVC '02]
Depth from Stereo

- Scene structure provides an additional cue and helps in segmentation
Spatial Semantic Maps

- Potential objects added to geometric map
- Registration is challenging
Viewpoint Planning

- Move robot to obtain foveal images from a range of angles to aid recognition
- Accomplished with a **viewpoint utility function**
  - Encourages collection of novel views
  - Aggregate utility from different objects
Detection Results
Detection Results
Detection Results
Internet images for CD
"Retrospective" by Django Reinhardt
Training Phase

- Images from the internet have a lot of unrelated items/outliers.
- For specific objects:
  - Determine similarity between them using SIFT features.
  - Cluster them using hierarchical clustering approach; the largest related cluster has the correct templates.
Learning Shapes

- The contour of the extracted foreground is encoded using global shape descriptors

- Shape descriptors

  - Convexity
  - Principal axes
  - Compactness
  - Variance
  - Elliptic variance
Training Phase

- For category objects:
  - Extract global shape descriptors for the contours of extracted foregrounds
Testing Phase

- Saliency detector to extract patches
  - Size proportional to the scale
Testing Phase

- Foreground extraction
  - Color based segmentation the patches
  - Background modeling
  - Foreground/background extraction using Graph Cut
Detection

- **Specific objects:**
  - Match object features with foreground features
- **Category objects:**
  - Match shape of the foreground with object shapes
Results
Learning about images from keyword-based search

Search engines already index images based on their proximity to keywords
  + easy to collect examples automatically
  + lots of data, efficiently indexed
The Challenge

- Lots of images unrelated to the category could be returned
- More variety in terms of viewpoint, illumination, scale, etc
- Could be as few as one "good" image from which anything about the category can be learned

Example results for Google image search for "Face"

Example images from a labeled dataset
Related work

Cluster to find visual themes (e.g., with topic models such as pLSA, HDA) [Sivic et al. 2005, Fergus et al. 2005, Li et al. 2007]

Apply models known to work well with correctly labeled data [Fergus et al. 2004, Schroff et al. 2007]
Our approach

- A *multiple-instance visual category learning* scenario to directly obtain discriminative models for specified categories from sets of examples
  - Assumes that as little as one example in a set or "bag" of images could be positive
  - Obtains a large-margin solution with constraints to accommodate this assumption
  - Iteratively improve multiple-instance classifier by automatically refining the representation

[Vijayanarasimhan & Grauman CVPR 2008]
Multiple-Instance Learning (MIL)

[Diagram showing positive and negative instances in traditional supervised learning versus multiple-instance learning]

[Note: Dietterich et al. 1997]
MIL for Visual Categorization

- Obtain sets or bags of images from independent sources.
- Each set should contain at least one good example of the category being learned.
- Enforce MIL constraint and obtain a classifier such that in each bag, at least one example is classified as positive.
Sparse MIL

Let $\mathcal{X}_p$ denote the set of positive bags and $\mathcal{X}_n$ denote the set of negative bags.

Let $X$ denote a bag, $x$ denote an instance, and let $\mathcal{X}_n = \{x \mid x \in X, X \in \mathcal{X}_n\}$ be all negative instances.
Sparse MIL

To begin, we solve a large-margin decision problem with constraints as suggested in [Bunescu & Mooney, 2007]:

\[
\text{minimize: } \frac{1}{2} ||w||^2 + \frac{C}{|\mathcal{X}_n|} \sum_{x \in \tilde{\mathcal{X}}_n} \xi_x + \frac{C}{|\mathcal{X}_p|} \sum_{X \in \mathcal{X}_p} \xi_X
\]

subject to:

\[
w \phi(x) + b \leq -1 + \xi_x, \quad \forall x \in \tilde{\mathcal{X}}_n
\]

\[
w \frac{\phi(X)}{|X|} + b \geq \frac{2 - |X|}{|X|} - \xi_X, \quad \forall X \in \mathcal{X}_p
\]

where \( \phi(X) = \sum_{x \in X} \phi(x) \)

\[
y(\hat{x}) = +1, \hat{x} \in X, \quad y(x) = -1, x \in X \setminus \{\hat{x}\}
\]

\[
\sum_{x \in X} \frac{\phi(x) + b}{|X|} \geq \sum_{x \in X} \frac{y(x)}{|X|} - \xi_x
\]

\[
\geq \frac{y(\hat{x})}{|X|} + \sum_{x \in X \setminus \{\hat{x}\}} \frac{y(x)}{|X|} - \xi_x
\]

\[
\geq +1 + \frac{1 - |X|}{|X|} - \xi_x
\]
Iterative MIL category learning

Compute optimal hyper-plane with sparse MIL

Re-weight positive instances
Training phase

- Face
- Faccia
- Visage

Positive bag 1
Positive bag 2
...
Positive bag N

Bag of words (1000 words)

SIFT on Hessian affine interest points

sMIL-SVM

Negative bag 1
Negative bag 2
...
Negative bag N

Category model

Keyword search results on other categories
## Example bags (Spam category)

<table>
<thead>
<tr>
<th>Engine</th>
<th>Language</th>
<th>Bag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>English</td>
<td><img src="https://via.placeholder.com/150" alt="Images" /></td>
</tr>
<tr>
<td>Google</td>
<td>French</td>
<td><img src="https://via.placeholder.com/150" alt="Images" /></td>
</tr>
<tr>
<td>Google</td>
<td>German</td>
<td><img src="https://via.placeholder.com/150" alt="Images" /></td>
</tr>
<tr>
<td>Yahoo</td>
<td>English</td>
<td><img src="https://via.placeholder.com/150" alt="Images" /></td>
</tr>
<tr>
<td>MSN</td>
<td>English</td>
<td><img src="https://via.placeholder.com/150" alt="Images" /></td>
</tr>
<tr>
<td>MSN</td>
<td>French</td>
<td><img src="https://via.placeholder.com/150" alt="Images" /></td>
</tr>
<tr>
<td>MSN</td>
<td>German</td>
<td><img src="https://via.placeholder.com/150" alt="Images" /></td>
</tr>
</tbody>
</table>
Test phase
Official results (What worked)

Spam

(sMIL Spam Filter)
Official results (What did not)

- saucenpan
- CD “Retrospective” by Djang Reinhart
- Book “Parls to the Moon” by Adam Gopnik
- Remote control
- Digital camera

Note possible confusion between remote control keypad and fax machine keypad.
Unofficial results

The contest allowed 30 minutes to detect the objects, but our program took 37 minutes to finish. Once the program completed, these were the remainder of the results...
Unofficial results

fax machine
15 min break!
2009

**robot league winner:** United States Naval Academy
(J. Searock, J. Piepemeier, K. Kinkade, M. Fick, C. Felps)

**software league winner:** UBC
(M. Baumann, P. Fazli, C. Gamroth, A. Gupta, S. Helmer,
T. Hoffmann, M. Muja, K. Okuma, T. Southey, P. Viswanathan,
W. Wohlkinger, S. Meger, D. Lowe, A. Mackworth, J. Little)
SRVC 2009
Objects

- pumpkin
- orange
- red ping pong paddle
- white soccer ball
- laptop
- dinosaur
- bottle
- toy car
- trying pan
- book "I am a Strange Loop" by Douglas Hofstadter
- book "Fugitive from the Cubicle Police"
- book "Photoshop in a Nutshell"
- CD "And Winter Came" by Enya
- CD "The Essential Collection" by Karl Jenkins and Adiemus
- DVD "Hitchhiker's Guide to the Galaxy" widescreen
- game "Call of Duty 4" box
- toy Domo-kun
- Lay's Classic Potato Chips
- Peperidge Farm Goldfish Baked Snack Crackers
- Peperidge Farm Milano Distinctive Cookies
Distractor Objects

- orange boomerang
- toy carrot
- toy Pluto
- scissors
- toy Mickey Mouse
- book “The Watchmen”
- running shoes
Rankings

- **Robot league**
  - USNA 4 pts
  - KSU 0 pts
  - UBC 0 pts

- **Software league**
  - UBC 32 pts
  - KSU 13 pts
  - UA 8 pts
  - USNA 2 pts
  - Harz 2 pts
Functional Block Diagram

- Test Input
  - Training Portion
    - Conducts web image search
    - Filters for appropriate images
    - Stores training images
  - Exploration Portion
    - Autonomously navigates environment
    - Locates objects
    - Collects images of objects
    - Stores environment images

- Classification Portion
  - Maps objects
  - Identifies objects
  - Processes and compares images
Object Recognition - SIFT

- An algorithm to detect and describe local features in images
  - Finds key points
- Solves problem of object scale and rotation
- Key points of two images can be compared
Figure 1. Initial Sift

Figure 2. Transformation

Figure 3. Object Identification
Elimination of Bad Matches
Filtering: Semantic/Visual Coherence

Image A.) Peach

Image B.) Sun

Image C.) Peach

Image D.) Peach

Fig. 1: Semantic VS Visual Coherence. From: T. Mei, Y. Wang, X. Hua, S. Gong, and S. Li, "Coherent Image Annotation by Learning Semantic Distance," IEEE Computer Vision and Pattern Recognition, CVPR 2006.
Filtering Algorithm

Number of Matches Between Image 1 & 2
Number of Matches Between Image 1 & 3
Number of Matches Between Image 1 & n

--Average Number of Matches for Each Image
--Create Array of Images Based on Average Number of Matches
Filtering Algorithm Example:

---

- Total Number of Matches: 17 Matches
- Average Number of Matches: 4.25
Filtering Results: “I Am A Strange Loop”

Downloaded Images of Book

Filtered Images of Book
Filtering Results: Downloaded Data

- **Attempted 10 images / object:**
  - Number Of Images Downloaded- 141
  - Number Of Good Images Downloaded- 108
  - Number Of Bad Images- 33

- **Total Time ~ 50 minutes**
Filtering Results: Filtered Data

- Number Of Outliers Removed - 30
- Number Of Outliers Not Removed – 3
- Returned 2 Images / Object
  - Filtered Out at Least 1 Good Image / Object
  - Filtered 2 Good Images for Each Object but 3
Functional Block Diagram
Navigation and Control

- Finite state machine using nine possible states.
- Reactive navigation architecture.
- Randomization = a little bit of luck
Exploration to Map Furniture
Table and Object Segmentation

- High quality 3D data from the tilting laser for:
  - Segmentation of furniture using RANSAC and clustering based on surface normals
  - Segmentation of objects on tables by clustering groups of 3D points supported by tables
  - Segmented objects and tables are candidates for picture taking and for classifiers
Table and Object Segmentation
Table and Object Segmentation
Table and Object Segmentation
Semantic Object Graph

- Fusing table and object segmentations from many viewpoints and spatially registered detector hypothesis forms a hierarchical, semantic view of the environment.
- Registering multiple viewpoints of objects allows for improved detection rates by pooling confidences and active vision.
- Behaviour disabled for contest due to time limits.
Object Detection Systems

- 3 (4) object detectors:
  - SIFT matching
  - Contour matching
  - Deformable Parts Model (DPM)
  - Spherical harmonics for matching 3D data (not enabled for SRVC)
  - DPM trained on pre-known objects only, SIFT and contour based on web images
DPM Model for frying pan
DPM Model for frying pan
SIFT Detector

- Similar approach to previous years
  - Nearest neighbour appearance matching between training images and test image
  - Score based upon how much points agree with an optimal affine transformation
  - Bounding box uses the affine transformation to warp the training image "footprint" onto test image
Contours Detection with Stereo

- Shoe w/ Scale Prior
- Shoe w/ Scale Prior
- False Positive w/o Scale Prior
- False Positive w/ Scale Prior
Data Collection

- Multi-threaded downloading from:
- Specific objects:
  - Google
  - Walmart
- Object categories:
  - Imagenet (automated approx. bounding boxes)
  - Labelme (human-annotated)
- Were very close to training DPM model online, but training time was a bit too short
Imagenet Vacuums
Image Download

Image Sources:
- Google, Walmart, ImageNet, LabelMe, (Amazon?)

3D Sources:
- 3D Warehouse

Filtering procedure:
- Homogeneous background with single connected component of valid imagery
- Clustering to find correct meaning
- Hand labeling of pre-known categories
- Validation steps on all classifiers to produce accurate confidence
Software League Results

- 13 partials
- 31 total score
- 3 category objects
Wrapping Up

• Mere plug and play of computer vision algorithms does not seem to work well enough for robot vision

• Challenge I: computational complexity

• Challenge II: generalization (categorization)

• Challenge III: how to go from 2D to 3D?????

...maybe we need to do something smarter....