Recognizing Groceries *in situ*
Using *in vitro* Training Data

Michele Merler

Advisor: Prof. Serge Belongie
Outline

• Motivations
• Related Work
• Grozi -120
• Experiments
• Conclusions
• Future work
Motivations

Applications

• Assistive vision technology for the blind
e.g: Grozi project @ UCSD

• Object recognition for mobile robots
e.g: Semantic Robot Vision Challenge @ CMU.
Motivations

Problem Statement

APPLICATIONS

- Assistive vision systems for the blind
- Mobile robots navigation-interaction

STATE OF THE ART OBJECT DETECTION & RECOGNITION ALGORITHMS

Get from the real world

TESTING DATA

Use

TRAINING DATA

Need
Motivations

Problem Statement

APPLICATIONS

- Assistive vision systems for the blind
- Mobile robots navigation-interaction

STATE OF THE ART
OBJECT DETECTION & RECOGNITION ALGORITHMS

Use

• acquired under different environmental conditions
• appearance drawn from different distribution than the test data

Get from the real world

TESTING DATA

Need
Motivations

Problem Statement

APPLICATIONS

- Assistive vision systems for the blind
- Mobile robots navigation-interaction

Get from the real world

USE

STATE OF THE ART OBJECT DETECTION & RECOGNITION ALGORITHMS

Need

WHERE TO GET THEM?

TESTING DATA

APPLICATIONS

- Assistive vision systems for the blind
- Mobile robots navigation-interaction

Get from the real world

STATE OF THE ART
OBJECT DETECTION & RECOGNITION ALGORITHMS

Need

WHERE TO GET THEM?

TESTING DATA
Motivations

Problem Statement

STATE OF THE ART
OBJECT DETECTION & RECOGNITION ALGORITHMS

Where to get them?

Need

Use

World Wide Web
Related work

Object recognition databases

• PASCAL VOC
• Caltech 101(256)
• SOIL-47
• ALOI
• ETH-80
• LabelMe
Related work

Object recognition databases

Improvements needed*:

- Multiple object class instances within a single image
- Partial occlusion and truncation
- Size, viewpoint and orientation variations
- High degree of intra-class variability
- Exclude pre-segmented objects

Related work

Object recognition databases

Improvements needed*:

- Multiple object class instances within a single image
- Partial occlusion and truncation
- Size, viewpoint and orientation variations
- High degree of intra-class variability
- Exclude pre-segmented objects

Related work

Object recognition databases

Improvements needed*:

- Multiple object class instances within a single image
- High degree of intra-class variability
- Exclude pre-segmented objects

Training and testing data often come from the same distribution

Related work

Object detection/recognition algorithms

- Color Histogram Matching (Lab) [Swain & Ballard '91] with Integral Histogram [Porikli '06]
- SIFT Descriptor [Lowe '99]
- Boosted Haar-like Features [Viola & Jones '01].
Grozi-120

- Multimedia database of 120 grocery products
- Objects vary in color, size, opacity, shape and rigidity. They are found in different lighting conditions and in presence of clutter and occlusion
- *In vitro* and *in situ* image representations (for training and testing data respectively)
Grozi-120

*In vitro* data

- Isolated images captured under ideal imaging conditions (e.g. stock photography studio or lab)

- They can be found in the web (e.g. Froogle, Amazon, etc)
Grozi-120

*In vitro* data

- 676 training images (average 6 images per object)
- Obtained from the web (Froogle, Shopwiki, Amazon Groceries, Yahoo images) using a list of 4000 UPC codes
- Clear foreground–background distinction via binary mask (if desired)
Images from objects captured in natural environments (real world)

They were shot inside a grocery store, using a MiniDV camcorder and includes every *in vitro* object

29 videos containing all products

Product location in coordinates saved every 5 frames

A total of 11194 *in situ* images (average 93 per product)
Grozi-120

\textit{In situ} data
Grozi-120
*In situ* data

<table>
<thead>
<tr>
<th>Item</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honey Nut Cheerios</td>
<td>016000665903</td>
<td>04 Honey Nut Cheerios Cereal Cup</td>
</tr>
<tr>
<td>Colgate plus toothbrush</td>
<td>035000553003</td>
<td>103 Colgate plus toothbrush</td>
</tr>
<tr>
<td>GM HNY NUT CHEERIOS CEREAL CUP</td>
<td>016000141551</td>
<td>95 GM HNY NUT CHEERIOS CEREAL CUP</td>
</tr>
<tr>
<td>Morton Salt, Iodized</td>
<td>024600010030</td>
<td>27 Morton Salt, Iodized</td>
</tr>
<tr>
<td>Chap Stick Lip Balm</td>
<td>036600813313</td>
<td>33 Chap Stick Lip Balm</td>
</tr>
</tbody>
</table>
Grozi-120

*In situ* data

- Truncation
- Multiple object class instances within a single image
- Objects from different classes within a single image
- Partial occlusion
Grozi-120

*In vitro* data
Training

*In situ* data
Testing
Grozi-120

**In vitro** data
Training

**In situ** data
Testing

Big difference in quality!
Grozi-120

**In vitro** data
Training

**In situ** data
Testing

- Illumination
- Deformations
- Clutter
- Occlusion
- Truncation

Big difference in quality!
Experiments

- **Color histogram matching (CHM):**
  Histogram template (16 bins in ab from Lab) + integral histogram + L1 distance

- **SIFT:** bag of features approach
  One bag per object + L2 distance

- **Boosted Haar-like features (BHaar):**
  Data + synthetic data + Haar-like features + cascades (OpenCv)
Experiments

Recognition:

- 10 *in vitro* samples per object for training (200 for BHaar)
- 10 *in situ* images per object for testing
Experiments

Localization:

- 10 *in vitro* samples per object for training (200 for BHaar)
- 14 *in situ* frames per object for true positives
- 100 frames with no product on the database as true negatives

<table>
<thead>
<tr>
<th></th>
<th>CHM</th>
<th>SIFT</th>
<th>BHaar</th>
</tr>
</thead>
<tbody>
<tr>
<td>%Rec</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>15</td>
<td>72</td>
<td>15</td>
</tr>
<tr>
<td>Std Dev</td>
<td>28</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td>Best (20)</td>
<td>71</td>
<td>14</td>
<td>35</td>
</tr>
<tr>
<td>Worst (32)</td>
<td>0.7</td>
<td>26</td>
<td>0.7</td>
</tr>
<tr>
<td>%Pre</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>17</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>Std Dev</td>
<td>16</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>Best (20)</td>
<td>82</td>
<td>83</td>
<td>82</td>
</tr>
<tr>
<td>Worst (32)</td>
<td>0.2</td>
<td>0.9</td>
<td>0</td>
</tr>
<tr>
<td>%TP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>18</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>Std Dev</td>
<td>35</td>
<td>26</td>
<td>19</td>
</tr>
<tr>
<td>Best (20)</td>
<td>100</td>
<td>93</td>
<td>50</td>
</tr>
<tr>
<td>Worst (32)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>%FP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>65</td>
<td>62</td>
<td>65</td>
</tr>
<tr>
<td>Std Dev</td>
<td>32</td>
<td>28</td>
<td>24</td>
</tr>
<tr>
<td>Best (20)</td>
<td>4</td>
<td>25</td>
<td>38</td>
</tr>
<tr>
<td>Worst (32)</td>
<td>100</td>
<td>64</td>
<td>92</td>
</tr>
</tbody>
</table>

Rec = Overall Recall, Pre = Overall Precision
Localization:

- 10 \textit{in vitro} samples per object for training (200 for BHaar)
- 14 \textit{in situ} frames per object for true positives
- 100 frames with no product on the database as true negatives

<table>
<thead>
<tr>
<th></th>
<th>CHM</th>
<th>SIFT</th>
<th>BHaar</th>
</tr>
</thead>
<tbody>
<tr>
<td>%Rec</td>
<td>%Rec</td>
<td>%Rec</td>
<td>%Rec</td>
</tr>
<tr>
<td>%Pre</td>
<td>%Pre</td>
<td>%Pre</td>
<td>%Pre</td>
</tr>
<tr>
<td>%TP</td>
<td>%TP</td>
<td>%TP</td>
<td>%TP</td>
</tr>
<tr>
<td>%FP</td>
<td>%FP</td>
<td>%FP</td>
<td>%FP</td>
</tr>
<tr>
<td>Mean</td>
<td>15</td>
<td>72</td>
<td>15</td>
</tr>
<tr>
<td>Std Dev</td>
<td>17</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>Best (20)</td>
<td>18</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Worst (32)</td>
<td>65</td>
<td>62</td>
<td>65</td>
</tr>
</tbody>
</table>

Rec = Overall Recall, Pre = Overall Precision
Experiments

Localization:

- 10 *in vitro* samples per object for training (200 for BHaar)
- 14 *in situ* frames per object for true positives
- 100 frames with no product on the database as true negatives

<table>
<thead>
<tr>
<th></th>
<th>CHM</th>
<th>%Rec</th>
<th>%Pre</th>
<th>%TP</th>
<th>%FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>15</td>
<td>17</td>
<td>18</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>Best (20)</td>
<td>71</td>
<td>82</td>
<td>100</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Worst (32)</td>
<td>0.7</td>
<td>0.2</td>
<td>0</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SIFT</th>
<th>%Rec</th>
<th>%Pre</th>
<th>%TP</th>
<th>%FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>72</td>
<td>18</td>
<td>22</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>Best (34)</td>
<td>14</td>
<td>83</td>
<td>93</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Worst (9)</td>
<td>26</td>
<td>0.9</td>
<td>0</td>
<td>64</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>BHaar</th>
<th>%Rec</th>
<th>%Pre</th>
<th>%TP</th>
<th>%FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>15</td>
<td>17</td>
<td>18</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>Best (92)</td>
<td>35</td>
<td>74</td>
<td>50</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>Worst (5)</td>
<td>0.5</td>
<td>0.2</td>
<td>0</td>
<td>92</td>
<td></td>
</tr>
</tbody>
</table>

Rec = Overall Recall, Pre = Overall Precision
Experiments
Conclusions

We presented:

- A new multimedia database for studying object recognition in presence of *in vitro*/*in situ* dichotomy
- Baseline performances for three widely used algorithms

The results suggest the need of more precise and elaborate recognition algorithms.
Future Work

- Develop new algorithms that fuse these different approaches in order to improve results
Future Work

• Develop new algorithms that fuse these different approaches in order to improve results

• Make use of context information based on physical object proximity to improve localization of objects in natural scenes
Future Work

• Develop new algorithms that fuse these different approaches in order to improve results

• Make use of context information based on physical object proximity to improve localization of objects in natural scenes

• Use Shared features
Future Work

• Develop new algorithms that fuse these different approaches in order to improve results

• Make use of context information based on physical object proximity to improve localization of objects in natural scenes

• Use Shared features

• Use temporal correlation (tracking)
Future Work

- Develop new algorithms that fuse these different approaches in order to improve results

- Make use of context information based on physical object proximity to improve localization of objects in natural scenes

- Use Shared features

- Use temporal correlation (tracking)

- Include more objects and grow the number of samples per product

- Sprite detector

- Dr.Pepper detector

- Tide detector
Acknowledgments

Part of this work has been funded by the UCSD division of the California Institute for Telecommunications and Information Technology (Calit2) through the 2006 Summer Undergraduate Scholarship Program.
Michele Merler, Carolina Galleguillos and Serge Belongie, Recognizing groceries *in situ* using *in vitro* training data, SLAM 2007, Minneapolis

Questions?

http://grozi.calit2.net