Recognizing groceries in situ using in vitro training data

Michele Merler†, Carolina Galleguillos‡, Serge Belongie‡
Univ. of Trento†, U. California San Diego‡

SLAM 2007
Outline

• Introduction
• Related work
• Grozi -120
• Experiments
• Discussion
• Future work
Introduction

- Object recognition algorithms demand large amounts of training data acquired under different environmental conditions.

- Many real world applications need training data for which the appearance is drawn from different distribution that the test data.
Introduction

- Assistive vision technology for the blind, e.g: Grozi\(^1\) project @ UCSD.

- Object recognition for mobile robots, e.g: Semantic Robot Vision Challenge\(^2\) @ CMU.

\(^1\) [http://grozi.calit2.net](http://grozi.calit2.net)

\(^2\) [http://www.cs.cmu.edu/~prybski/SRVC](http://www.cs.cmu.edu/~prybski/SRVC)
Related Work

Object recognition algorithms:

- Color Histograms (LAB) features with integral image computation [Swain & Ballard '91].
- SIFT descriptor [Lowe '99].
- Boosted Haar-like features [Viola & Jones '01].
Related work

Object recognition databases:

- PASCAL VOC
- Caltech 101(256)
- SOIL-47
- ALOI
- ETH-80
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).
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).
**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 (binary mask)
In situ data

- 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.
**In situ data**

- 29 videos containing all products.
- Product location in coordinates saved every 5 frames.
- A total of 11194 *in situ* images (average 93 per product).
Experiments

• **Color histogram matching (CHM):** Histogram template + integral image + L1 distance.

• **SIFT with bag of features approach (SIFT):** One bag per object + L2 distance.

• **Adaboost (ADA):** Data + synthetic data + Haar-like features + cascades (14 stages).
Experiments

Recognition:

*In vitro* training data and *in situ* testing images.
Experiments

Localization: 

*In vitro* training data and *in situ* testing videos frames (images).

<table>
<thead>
<tr>
<th>Method</th>
<th>%Rec</th>
<th>%Pre</th>
<th>%TP</th>
<th>%FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHM</td>
<td>Mean</td>
<td>15</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Std Dev</td>
<td>28</td>
<td>16</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Best (20)</td>
<td>71</td>
<td>82</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Worst (32)</td>
<td>0.7</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>SIFT</td>
<td>Mean</td>
<td>72</td>
<td>18</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Std Dev</td>
<td>20</td>
<td>17</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Best (34)</td>
<td>14</td>
<td>83</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>Worst (9)</td>
<td>26</td>
<td>0.9</td>
<td>0</td>
</tr>
<tr>
<td>BHaar</td>
<td>Mean</td>
<td>15</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Std Dev</td>
<td>13</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Best (92)</td>
<td>35</td>
<td>74</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Worst (5)</td>
<td>0.5</td>
<td>0.2</td>
<td>0</td>
</tr>
</tbody>
</table>

Rec = Overall Recall, Pre = Overall Precision
Experiments
Discussion

We presented:

• A new multimedia database for studying object recognition in presence of *in vitro/in situ* dichotomy.

• Baseline performance figures for three widely used algorithms.

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

- We intend to include more objects and grow the number of samples per product.
- We plan to elaborate new algorithms that fuse these different approaches in order to improve results.
- We plan to make use of context information based on physical object proximity to improve localization of objects in natural scenes.