- On Seeing Stuff: The Perception of Materials by Humans and Machines, 
  By Adelson

- Semantic Texton Forests for Image Categorization and Segmentation, 
  By Shotton et al.

Presented by 
Mani Golparvar-Fard
On Seeing Stuff

• Perception of Object vs. Materials

• Examples of Material Importance:
  – Robotics
  – Construction

• Humans infer material properties using all the senses (e.g., look and feel)
Different illumination and viewing directions

Plaster-a
Crumpled Paper
Concrete
Plaster-b (zoomed)

Source: Leung and Malik, ICCV '99, Corfu, Greece
Common Vocabularies for material visual appearances

- **Luster** (the optical quality of the surface), **Resinous** (Like Plastic), **Adamantine** (like Diamond), **Greasy**, **Pearly**, **Silky**, **Vitreous** (Glassy), **Metallic**, **Sub metallic**, **Dull**, **Earthy** or **Chatoypant** (like a cat’s eye)
- When broken, may be uneven, **Conchoidal** (shell-like), **Hackly** (like cast-iron), or **Splintery** (like broken wood).
- **Habits**: Prismatic, massive (no form), acicular (needle-like), reniform (kidney-like spherules), bladed, dendritic, granular, fibrous, encrusting, colloform, porous, concretionary, botryoidal (grape-bunches), foliated (leaves or layers), scaly, felted, hairlike, stalactitic, nodular, columnar, plumose (feathery), microcrystalline, platy (flat thin plates), reticulated, lamellar, mammillary, saccharoidal (like sugar), ameboid, oolitic, or pisolitic.
Relevancy to concrete: 96%

Material-Based Image Retrieval Engine

Check Material

Process/Result

Materials Database
(Concrete, Forms, Steel, etc.)

Schedule Information

Check Time

Under Progress Material

As-planned Material

Other Material
How vision determines materials?

• Image of an object = \( \sum \) (Surface Shape, Surface Reflectance, Distribution of light in the environment and observer’s point of view)

• Perception of Material? A Hard Problem

• Does appearance depend on environment?
Does Appearance depend on environment?

Photographed in the Same room with the same lighting

- Every sphere depends on the environment in which it is viewed
- Sometimes seem hopeless to make sense of the spheres reflectance properties without knowing the environment first
Configuration and Context

• Reflectance properties fully characterized by BRDF (bi-directional reflectance distribution function),
  – in simple form Lambertian Surface
  – Albedo = Percent of light reflected

• How easily Albedo can be calculated?
  – A great number of configural cues about points and their shadows need to be known.
Importance of Context

Shiny sphere (with and without specularities), generated by computer graphics

Visual cues tell more than Optical Qualities – Maybe mechanic property of material?

Blobs of Hand cream vs. Cheese cream
Optical and Mechanical Aspects of World as well as Optical and Mechanical Aspects of Environment

• In addition to these aspects of a material, existence of light in the environment
  – Reflection, Refraction as well as Absorbance
Habits = Shape + Texture?

Figure 8: Left: the habit of a pin oak. Center and right: A melanoma can be visually distinguished from a normal mole. The combination of shape, color, and texture are used to make the diagnosis.
How Images are made?

• Understanding how images are built
• Ecological optics = What forms materials take and what pattern of light illuminate them?
• 3-D Graphics = Researchers use visual tricks
• Traditional Painting = Is portraying material easy?
• 2D Graphics = e.g., Photoshop
• Photography = Light and Camera are in hand of the photographer
Material Appearance = Texture Perception?

- Shows even a simple uniform convolution produces reasonable impression of a roughened metal sphere.
- Infers two things: Intensity Histogram, Frequency Domain
Classification

- Environment tends to contain a broad range of luminances and numerous sharp edges,
  - We expect these properties to manifest themselves in the Specular reflections
Analysis by Synthesis

- **Shape + Lighting + Albedo given a known contour**
  - A grassfire algorithm was used to compute distance from the contour, and then apply a smoothing algorithm.

*Figure 11. (a) A photograph of a penguin shaped bath-oil capsule. (b) synthesized penguin image, based on estimated shape and Ward reflectance parameters. (c) synthesized penguin with synthesized drop shadow. (d) sphere rendered using the same lighting and reflectance parameters.*
Lessons Learned from the paper

• **Mechanical** and **optical** properties of material are the main properties that humans derive from image information.

• Recent work suggests that concepts used in **texture analysis** may be usefully applied to the problem of **material appearance**.
Relevancy to forms: 94%
Concrete Rejections: 20%

Material-Based Image Retrieval Engine

Check Material

Process/Result

Schedule Information

Check Time

Materials Database
(Concrete, Forms, Steel, etc.)

Other Material

Under Progress Material

As-planned Material

Check Material

Schedule Information

Check Time
Eamon

- Reading Adelson led me to consider how the opposing views of direct vs. mediated perception could apply to material properties. It seems strange to think that an observer would build a representation that explicitly contains information about a material's intrinsic mechanics and optics, but it's definitely the case that we have access to this information when we need it. Would focused visual attention be required to "bind" information about a material's shininess and smoothness, or is the character of "stuff" a feature on its own?
Ultimate goal for this paper:

- Simultaneous segmentation and recognition of objects in images or videos in real-time

[shotton-eccv-08]

[shotton-cvpr-06]
Real-Time Semantic Segmentation Demo (Winner of CVRP 2008 Demo Prize)
Overview

• **Motivations:**
  1) Visual words approach is slow
     – Compute feature descriptors
     – Cluster
     – Nearest-neighbor assignment
  2) Conditional Random Fields is even slower
     – Inference always a bottle-neck

• **Approach:** Acts directly on pixel values
  • An efficient and powerful low-level feature approach

• **Result:** works well and efficiently
Overview

• Contributions
  – Semantic Texton Forests
    • Hierarchical clustering into semantic textons and a local classification
  – The Bag of Semantic Textons Model
    • Application in categorization and segmentation
  – Image-Level Prior (ILP)
    • Improving semantic segmentation performance
Quick Overview on Decision Trees

Structure of a decision tree

- Internal nodes correspond to attributes (features)
- Leafs correspond to classification outcome
- edges denote assignment

- Advantages?
- Drawbacks?
Random Forests

- Decision tree show problems related to **over-fitting** and lack of **generalization**.
  - The main motivation behind application of Random Forest

- Random Forests mitigate such problems by:
  - Injecting **randomness** into the training of the trees, and
  - Combining the output of multiple randomized trees into a single classifier.

- Pros:
  - Produce lower test errors than conventional decision trees
  - Performance comparable to SVMs in multi-class problems
  - Maintain high computational efficiency.
Example of a Random Forest

An example $x$ will be classified as $\alpha$ according to this random forest.
Recap on Randomized Decision Forests

• Approach
  – Each node $n$ in the decision tree contains an empirical class distribution $P(c|n)$
  – Learn decision trees such that similar features should end up at same leaf nodes

![Decision Trees Diagram](image)

  – The leaves $L = \{l_i\}$ of a tree contain most discriminative information
    • Classify by averaging

\[
P(c|L) = \frac{1}{T} \sum_{t=1}^{T} P(c|l_t)
\]
Recap on Randomized Decision Forests

– Input: Features describing pixel
– Output: Predicted class distribution

• Another histogram of texton-like per pixel!
STF Features

- Simple Function of image pixels
- Center a \( d \)-by-\( d \) patch around a pixel (5x5)

Potential Features

1. Its value in a color channel (CIELab)
2. The sum of two points in the patch
3. The difference of two points in the patch
4. The absolute difference of two points in the patch

- Feature invariance accounted for by rotating, scaling, flipping, affine-ating training data
Training based on Extreme Random Decision Tree

- Take random subset of training data
- Generate random features $f$ from above
- Generate random threshold $t$
- Split data into left $I_l$ and right $I_r$ subsets according to
- Repeat for each side

\[
I_1 = \{ i \in I_n \mid f(v_i) < t \} \\
I_r = I_n \setminus I_1.
\]

This feature maximizes information gain

\[
\Delta E = -\frac{|I_1|}{|I_n|} E(I_1) - \frac{|I_r|}{|I_n|} E(I_r),
\]

- Advantage: Fast to Learn and Fast to evaluate
• Each **patch** represents one leaf node. It is the average of all the patches from the training data that fell into that leaf.
• Learns colors, orientations, edges, blobs

• [distance = 21 pixels]
Simple model results

- Semantic Texton Forests [Random chance is under 5%] – Poor Segmentation
- Training takes about 15min on 500 feature tests and 10 threshold test per split

<table>
<thead>
<tr>
<th></th>
<th>Global</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>supervised</td>
<td>49.7%</td>
<td>34.5%</td>
</tr>
<tr>
<td>weakly supervised</td>
<td>14.8%</td>
<td>24.1%</td>
</tr>
</tbody>
</table>

- MSRC-21 dataset

- **Supervised = 1 label per pixel**
  - Increase one bin in the histogram at a time

- **Weakly-supervised = members of the classes in image as training labels per pixel**
  - Increase multiple bins in the histogram at a time
Bag of Semantic Textons

• Extension of bag of words with low-level semantic information

• How can we get a prior estimate for what is in region $r$?
  1) Average leaf histograms in region $r$ together $P(c | r)$
     • Good for segmentation priors

  2) Create hierarchy histogram of node counts $H_r(n)$ visited in the tree for each classified pixel in region $r$
     • Want testing and training decision paths to match
Histogram-based Classification

• Main idea:
  – Have 2 vectors as features
    • (training-tree’s histograms, testing-tree’s histograms)
    – Want to measure similarity to do classification

• Proposed approach: Kernalized SVM
  – Kernel = Pyramid Match Kernel (PMK)
  – Computes a histogram distance, using hierarchy information
  – Train 1-vs-all classifiers
**Review on pyramid match**

*Level 0*

\[ X \]

\[ Y \]

\[ N_0 = 2 \]
\[ w_0 = 1 \]

\[ H_0(X) \]

\[ H_0(Y) \]

\[ I_0 = 2 \]
Review on pyramid match

Level 1

$H_1(X)$

$N_1 = 4 - 2 = 2$

$w_1 = \frac{1}{2}$

$H_1(Y)$

$I_1 = 4$
Review on pyramid match

Level 2

\[ N_2 = 5 - 4 = 1 \]

\[ w_2 = \frac{1}{4} \]

\[ H_2(X) \]

\[ H_2(Y) \]

\[ I_2 = 5 \]
Scene Categorization

- The whole image is one region
  - Using histogram matching approach
  - End result is an **Image-level Prior**

- Comparison with other similarity metric (RBF- radial basis function)
  - **Unfair?** RBF uses only leaf-level counts, PMK uses entire histogram

<table>
<thead>
<tr>
<th></th>
<th>Global kernel $K$</th>
<th>Per-category kernel $K_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBF</td>
<td>49.9</td>
<td>52.5</td>
</tr>
<tr>
<td>PMK</td>
<td>76.3</td>
<td>78.3</td>
</tr>
</tbody>
</table>

Table 2. Image categorization results. (Mean AP).

- Results
  - $K_c$ = An idea to account for unbalanced classes

- Number of trees does not significantly Affect returns after N=5

Figure 5. Categorization accuracy vs number of STF trees.
Improving Semantic Segmentation

- Use idea of shape-filters to improve classification
- Main idea: After initial STF classification, learn how a pixel’s class interacts with neighboring regions’ classes

- Approach: Learn a second random decision forest (segmentation forest)
  - Use different weak features:
    - Histogram count at some level $H_{r+i}(?)$
    - Region prior probability of some class $P(\cdot | r+i)$

- Difference with shape filters:
  - Shape-filters learn: cow is adjacent to green-like texture
  - Segmentation forest learn: cow is adjacent to grass

- Trick: multiply with image-level prior for best results
  - Convert SVM decision to probability
Comparison segmentation results on MSRC-21

<table>
<thead>
<tr>
<th></th>
<th>Without ILP</th>
<th>With ILP</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) only leaves</td>
<td>61.3%</td>
<td>64.1%</td>
</tr>
<tr>
<td>(b) all nodes</td>
<td>63.5%</td>
<td>65.5%</td>
</tr>
<tr>
<td>(c) only region priors</td>
<td>62.1%</td>
<td>66.1%</td>
</tr>
<tr>
<td>(d) full model</td>
<td>63.4%</td>
<td>66.9%</td>
</tr>
<tr>
<td>(e) no transformations</td>
<td>60.4%</td>
<td>64.4%</td>
</tr>
<tr>
<td>(f) unsupervised STF</td>
<td>59.5%</td>
<td>64.2%</td>
</tr>
<tr>
<td>(g) weakly supervised STF</td>
<td>61.6%</td>
<td>64.6%</td>
</tr>
</tbody>
</table>

- In all cases the ILP improves results.
- The region priors alone perform remarkably well.
  - Comparing to the segmentation result using only the STF leaf distributions (34.5%) this shows the power of the localized BoSTs that exploit semantic context.
- Random transformations of the training images improve performance by adding invariance.
- Performance increases with more supervision, but even unsupervised STFs allow good segmentations.
MSRC-21 Results

27- TextonBoost, Shotton et al. 2007
32 – Verbeek and Triggs – Classification with markow field aspect models, cvpr 2007
VOC 2007 Segmentation

<table>
<thead>
<tr>
<th></th>
<th>background</th>
<th>aeroplane</th>
<th>bicycle</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>motorbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv/monitor</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brookes</td>
<td>78</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>5</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>11</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>29</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>11</td>
<td>1</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>33</td>
<td>46</td>
<td>5</td>
<td>14</td>
<td>11</td>
<td>14</td>
<td>34</td>
<td>8</td>
<td>6</td>
<td>3</td>
<td>10</td>
<td>39</td>
<td>40</td>
<td>28</td>
<td>32</td>
<td>8</td>
<td>19</td>
<td>8</td>
<td>24</td>
<td>9</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>Ours + ILP</td>
<td>20</td>
<td>66</td>
<td>6</td>
<td>15</td>
<td>6</td>
<td>15</td>
<td>32</td>
<td>19</td>
<td>7</td>
<td>7</td>
<td>13</td>
<td>44</td>
<td>31</td>
<td>44</td>
<td>27</td>
<td>39</td>
<td>35</td>
<td>12</td>
<td>39</td>
<td>23</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>TKK</td>
<td>23</td>
<td>19</td>
<td>21</td>
<td>5</td>
<td>16</td>
<td>3</td>
<td>1</td>
<td>78</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>23</td>
<td>69</td>
<td>44</td>
<td>42</td>
<td>0</td>
<td>65</td>
<td>30</td>
<td>35</td>
<td>89</td>
<td>71</td>
<td>30</td>
</tr>
<tr>
<td>Ours + DLP</td>
<td>22</td>
<td>77</td>
<td>45</td>
<td>45</td>
<td>19</td>
<td>14</td>
<td>45</td>
<td>48</td>
<td>29</td>
<td>26</td>
<td>20</td>
<td>59</td>
<td>45</td>
<td>54</td>
<td>63</td>
<td>37</td>
<td>40</td>
<td>42</td>
<td>10</td>
<td>68</td>
<td>72</td>
<td>42</td>
</tr>
</tbody>
</table>
More Results
More Results

[Images of various scenes with annotations and segmented images]
And More Results
And More Results
Discussion

• **Pros:**
  – Simple concept
  – Good result
  – Works fast (testing and training)

• **Cons:**
  – Difficult to understand
  – Low-resolution classification
    • Segmentation forest operates at patches
  – Test-time inference is dependent on amount of training
    • Must iterate through all trees in the forest at test time
  – Many “Implementation Details”.

• **Question:**
  • How dependent is the performance on decision tree parameters?
Comments

• Gang
  – I have been to the demo show of the semantic texton forests at CVPR 2008. It was very cool. It could recognize and segment objects in real time and with reasonable accuracy. Random forests is a powerful and efficient tool, even for such a low level feature representation.

• Jianchao
  – For classification, they are using nonlinear kernels, which make it difficult to generalize to training on large amount of data in reality.

• Ian
  – Upon inspection of the segmentation performance results for the background class in Pascal VOC 2007, the "image level prior" decreases performance significantly.

  Ideally, this prior should be used to suppress classes that the image wide statistics don't support. One would expect the background to appear in almost all images, and since modeling a background model is difficult, perhaps this prior could be excluded from the background predictor.
Comments

• Sanketh

1. If each of the ER Trees is being learned on a different subset of the data (with different distributions of class labels), even with the normalization, won't some trees be better at identifying some classes over others? Why average then? Why not weight the output P(C|L) with the confidence in predicting that class label.

2. It has been a while since I visited decision trees but I remember a lot of fuss over pruning them to ensure they do not overfit. In the trees here there is obviously lot of variance. Since the splits made at each stage necessarily increase the "purity" of the children nodes I wonder if there is a danger of overfitting the data, i.e. the decision rules/thresholds chosen may not translate well to new novel examples.

3. It is unclear to me how such simple features can handle the wide variety of variations in viewpoint and appearance from natural categories. If we have more black dogs than black cats in our training won't it infer that black patches => high likelihood of dogs vs. cats?

4. If the decisions at nodes n across trees are different (as are their parent decisions), why bother accumulate statistics at node n across all trees? Don't they represent different things? It doesn't make sense to me.