Indoor Scene Depth Reconstruction From Monocular Images

Rick Fulton
Evan Rosen

Mentors:
Steve Gould
Geremy Heitz

Professor:
Daphne Koller

July 24, 2008
DAGS
Motivation

Scene Understanding
- Depth Reconstruction as a key task
- Explore the role of structural features in reconstruction
Summer Goal

- Coarse 3D depth reconstruction of indoor scenes from single images
- Framework for exploring structural features
Why Monocular?

- Why Monocular?
  - Important component for scene understanding
    - has been shown to help object recognition
  - Important component of human vision
  - Large repositories of single images
    - Google, flickr
Prior Work

- Similar algorithms for outdoor scenes
  - Hoiem, 2007

- Older approaches using Constraint Satisfaction
  - Waltz, 1975

- Difficulties Encountered
  - not robust to clutter/real world scenes
  - difficulty with foreground objects

- Features are not as intentionally designed
Our Approach

- Estimate structural features, human intuition
- Use 3D features
- Long range interactions
- Using modern machine learning techniques

Raw Images → Estimators → MRF → Depth Map
Single Point Information

- What can we do with a single pixel?

- Location
- Color
- Depth
Two Point Information

- What can we do with two pixels?
  - Difference in grayscale value
  - Difference in location
  - Presence of edges between points
  - Difference in depth

- Can perceive shape independently of scale
Three Point Information

What can we do with three points?

- Presence of Edges
- Orientation (colinear/right angle)
- Colinearity
- Angles
Dataset

- Indoor images, and corresponding depth maps
- Originally collected for object recognition
- discretize each training image into \( n \) depth bins \( d'_1 - d'_n \)
- Subsampling each image, yielding a variable granularity \( m \times n \) square grid
Step One: Creating Estimates

- Depth Estimators $d_i$
- Depth Difference Estimators $\Delta_{ij}$
- Colinearity Estimators $c_{ijk}$
Depth Estimators

- Raw Features
  - compute a set of template responses for each grid point
    - randomly selected patches of dimensions between 30 and 50 pixels
    - computed the cross correlation with the patch $p_{ij}$ surrounding each grid point $X_{ij}$
  - include values of the $p_{ij}$ itself
Boosting

- One-vs-all boosted classifier for each discretized depth $d'_n$
Discontinuity Estimators

- One-vs-all boosted classifier for each discretized depth difference $\Delta'_n$

- Raw Features
  - template response for the pixel at the midpoint between $X_i$ and $X_j$
  - $\text{mean}(P_i) - \text{mean}(P_j)$
  - $s(P_i)$
  - $s(P_j)$
  - number of edges between $P_i$ and $P_j$
  - presence of edge(s) between $P_i$ and $P_i$
Boosting

- One-vs-all boosted classifier for each discretized depth difference, $\Delta'_n$

- Each pixel pair $(X_i, X_j)$ has a score for every discretized depth, which can be thought of as the log probability that $(\Delta_{ij} = \Delta'_n)$
Colinearity Estimators

- Triplets which we believe to be colinear in 3D accompanied by score according to distance between endpoints
- Principle:
  - Any three points which lie in a visible plane that are colinear in 2D are colinear in 3D.
Step Two: Combining the Estimators

- Given a set of constraints, \( f \), how to learn coherent model
- Learn using a CRF

\[
P(X|I) = \frac{1}{z} \exp \left( \sum_i -w_{x_i}^T f(x_i, I) + \sum_{ij} -w_{x_i,x_j}^T f(x_i, x_j, I) + \sum_{ijk} -w_{x_i,x_j,x_k}^T f(x_i, x_j, x_k, I) \right).
\]
MRF Model - Singleton

- each node $j$ in each image $I$ has an associated set of singleton features values, $f(X_j, I)$
- features are a combination of baseline bias and indicators, and previously described boosted estimator
  - we learn a weight $w_c$ for each singleton feature

<table>
<thead>
<tr>
<th>$f(x,I)$: feature values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i$: node $i$</td>
</tr>
<tr>
<td>$I$: image $I$</td>
</tr>
<tr>
<td>$w_{x's}$: weights given $x$’s</td>
</tr>
</tbody>
</table>
MRF Model – Pairwise

- Each adjacent pair of nodes $X_i$, $X_j$ in each image $I$ have an associated set of pairwise features $f(X_i, X_j, I)$
  - We learn the weights of all possible depth differences between the two nodes
  - Each difference in depth should have the same feature
  - Helps capture intuition that pairs of nodes at the same relative depth have similar structural significance

<table>
<thead>
<tr>
<th>$d(x_i)$</th>
<th>$d(x_j)$=1</th>
<th>$d(x_j)$=2</th>
<th>$d(x_j)$=3</th>
<th>$d(x_j)$=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
</tr>
</tbody>
</table>
MRF Model - Colinear

- Nodes we think are colinear $X_i, X_j, X_k$ (using edge detection) have a feature $f(X_i, X_j, X_k, I_c)$ for each possible depth combination for the three nodes in image $i$
  - each feature value is a measure of the L2 difference between the two line segments formed by the three points
MRF Model

- Need to:
  - learn weights
  - predict depths

- learn the weights using pseudolikelihood
  - Learning maximum likelihood objective is intractable
  - Pseudolikelihood is fast, good approximation
  - Has worked well in other vision tasks

- use max product loopy for inference in predicting depths
  - Converges most of the time
  - Limit on maximum number of messages passed
## Results

- Results using three fold cross validation:

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th></th>
<th>Test</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMS</td>
<td>Accuracy</td>
<td>RMS</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Baseline (bias +</td>
<td>.492</td>
<td>.430</td>
<td>.500</td>
<td>.414</td>
</tr>
<tr>
<td>indicator)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline + boosted</td>
<td>.375</td>
<td>.564</td>
<td>.400</td>
<td>.490</td>
</tr>
<tr>
<td>singleton features</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline + boosted</td>
<td>.320</td>
<td>.615</td>
<td>.351</td>
<td>.536</td>
</tr>
<tr>
<td>singleton + pairwise</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline + boosted</td>
<td>.313</td>
<td>.620</td>
<td>.345</td>
<td>.540</td>
</tr>
<tr>
<td>singleton + pairwise + colinear features</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis

- Finer detail needed (multi-resolution approach?)
- RMS score more informative
- Dataset is highly correlated
- Details still difficult (translucent object)
Future Work

- more long range cliques
- more robust singleton and pairwise features
- 3 clique triangle features
- try to predict the normal vector as well
- combine results with stereo cues
- integrate with other components of scene understanding
- different dataset
- different quantization of the depths
- different grid granularity