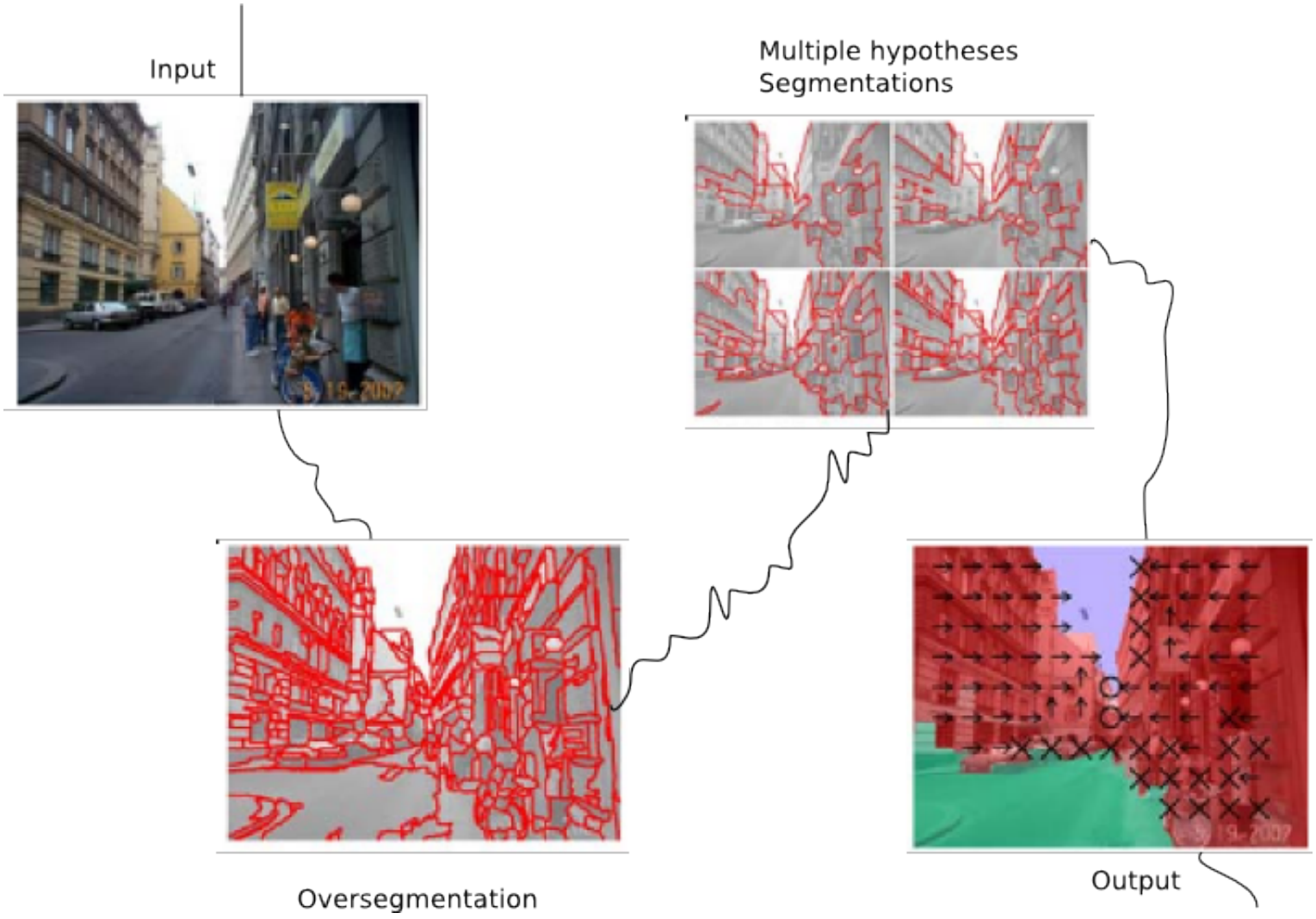
Geometric Context from a Single Image

Derek Hoiem, Alexei A. Efros, Martial Hebert

February 26, 2009

Outline

Big Picture



Hoiem et. al. 2007 (parts)

2005 vs. 2007

Feature Descriptions	Num
Color	**16**
C1. RGB values: mean	3
C2. HSV values: C1 in HSV space	3
C3. Hue: histogram (5 bins) and entropy	6
C4. Saturation: histogram (3 bins) and entropy	4
Texture	**15**
T1. DOOG filters: mean abs response of 12 filters	12
T2. DOOG stats: mean of variables in T1	1
T3. DOOG stats: argmax of variables in T1	1
T4. DOOG stats: (max - median) of variables in T1	1
Location and Shape	**12**
L1. Location: normalized x and y, mean	2
L2. Location: norm. x and y, 10^{th} and 90^{th} pctl	4
L3. Location: norm. y wrt horizon, 10^{th}, 90^{th} pctl	2
L4. Shape: number of superpixels in region	1
L5. Shape: number of sides of convex hull	1
L6. Shape: $num\ pixels/area(convex\ hull)$	1
L7. Shape: whether the region is contiguous $\in \{0, 1\}$	1
3D Geometry	**35**
G1. Long Lines: total number in region	1
G2. Long Lines: % of nearly parallel pairs of lines	1
G3. Line Intsctn: hist. over 12 orientations, entropy	13
G4. Line Intsctn: % right of center	1
G5. Line Intsctn: % above center	1
G6. Line Intsctn: % far from center at 8 orientations	8
G7. Line Intsctn: % very far from center at 8 orient.	8
G8. Texture gradient: x and y "edginess" (T2) center	2

Hoiem et. al. 2005

Location and Shape
L1. Location: normalized x and y, mean
L2. Location: normalized x and y, 10th and 90th pctl
L3. Location: normalized y wrt estimated horizon, 10th, 90th pctl
L4. Location: whether segment is above, below, or straddles estimated horizon
L5. Shape: number of superpixels in segment
L6. Shape: normalized area in image

Color
C1. RGB values: mean
C2. HSV values: C1 in HSV space
C3. Hue: histogram (5 bins)
C4. Saturation: histogram (3 bins)

Texture
T1. LM filters: mean absolute response (15 filters)
T2. LM filters: histogram of maximum responses (15 bins)

Perspective
P1. Long Lines: (number of line pixels)/sqrt(area)
P2. Long Lines: percent of nearly parallel pairs of lines
P3. Line Intersections: histogram over 8 orientations, entropy
P4. Line Intersections: percent right of image center
P5. Line Intersections: percent above image center
P6. Line Intersections: percent far from image center at 8 orientations
P7. Line Intersections: percent very far from image center at 8 orientations
P8. Vanishing Points: (num line pixels with vertical VP membership)/sqrt(area)
P9. Vanishing Points: (num line pixels with horizontal VP membership)/sqrt(area)
P10. Vanishing Points: percent of total line pixels with vertical VP membership
P11. Vanishing Points: x-pos of horizontal VP—segment center (0 if none)
P12. Vanishing Points: y-pos of highest/lowest vertical VP wrt segment center
P13. Vanishing Points: segment bounds wrt horizontal VP
P14. Gradient: x, y center of mass of gradient magnitude wrt segment center

Hoiem et. al. 2007

Location

- 2D position in the image provides a strong cue for recovering the rough 3D scene geometry!

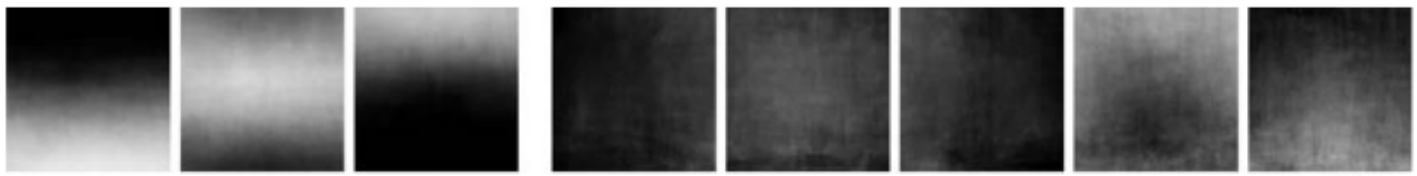


Figure 4. Likelihood (higher intensity is more likely) of each geometric class given location in the image. Location is highly discriminative for the main classes. Porous surfaces (often vegetation) tends to form a canopy around the center of the image, while solid surfaces often occur in the front-center of the image. The left/center/right likelihoods show the tendency to take photos directly facing walls or down passages.

Hoiem et. al. 2007

Color

- Color itself has little to do with 3D in general.
- However, since it can be used to identify objects that correspond to geometric classes, it can be a strong cue...

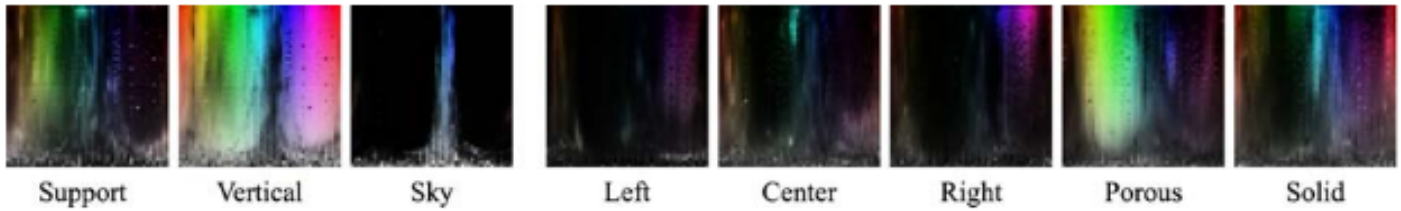


Figure 5. Likelihood of each geometric class given hue and saturation. Each image shows the given hue and saturation, with the label likelihood given by the intensity. Blue sky is easily distinguishable. More confusion exists between the support and vertical classes, though natural tones of brown, green, and blue lend evidence for support. Gray (low saturation) is common in all main classes. Color is not very discriminative among the subclasses, with the exception of porous which tends to be colors of vegetation.

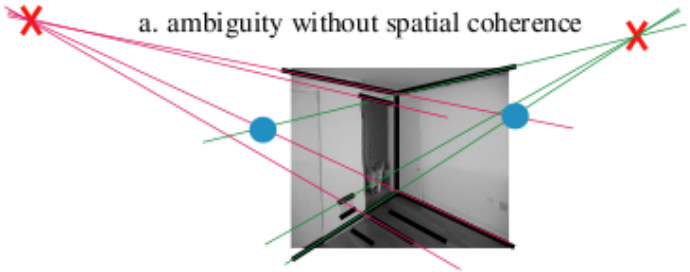
Hoiem et. al. 2007

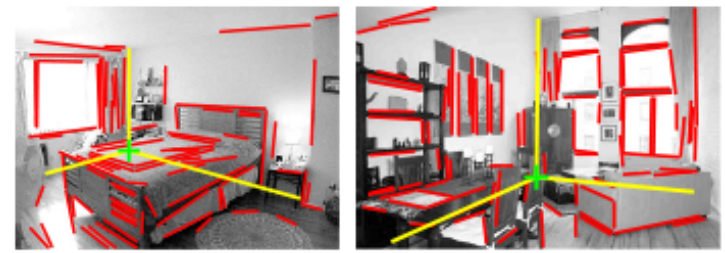
- Two color spaces are used to represent colors, RGB and HSV.

Texture

- In other applications texture can be used as a cue for 3D structure of an object.
- Here it is mainly used for its ability to identify objects that belong to geometric classes (just like color).
- A subset of the filter bank designed by Leung and Malik (2001) is used.
  - 6 edge, 6 bar, 1 Gaussian, 2 Laplacian of Gaussian filters.

Perspective I

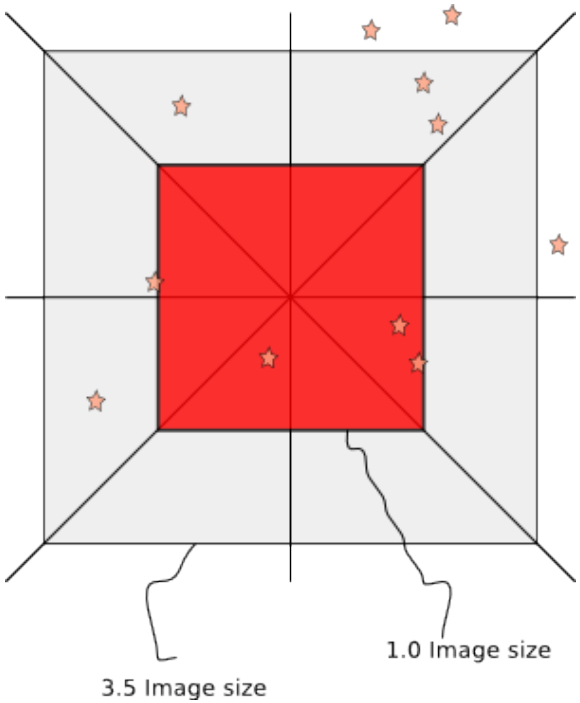




Yu et. al. 2008

- Knowledge of vanishing points can fully specify the relative 3D orientation...
  - ONLY if the image is well structured.

Perspective II



Computing statistics of straight lines


Estimated vanishing points

- 24-bin histogram is used to compute statistics of line intersections.
- Vanishing points are estimated by EM approach of Kosecha and Zhang (2002).
- When there are no parallel lines, texture gradient may be a helpful orientation cue.

Spatial Support

- **Location, color, texture and perspective cues are useful when some kind of structure is known.**
- Incrementally building up the structural information:
  - Pixels to superpixel, superpixels to multiple segmentations.
- Simple cues can be obtained from superpixels, but complex cues such as perspective requires larger regions.
- Multiple segmentations using different number of segments:

$$n_s \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100\} \tag{1}$$

Step-by-step Example: Input

http://www.rotary-austin.org/documents/website/UT%20tower%201.jpg

Step-by-step Example: Oversegmentation

Step-by-step Example: Confidences

Support

Vertical

Sky

Left

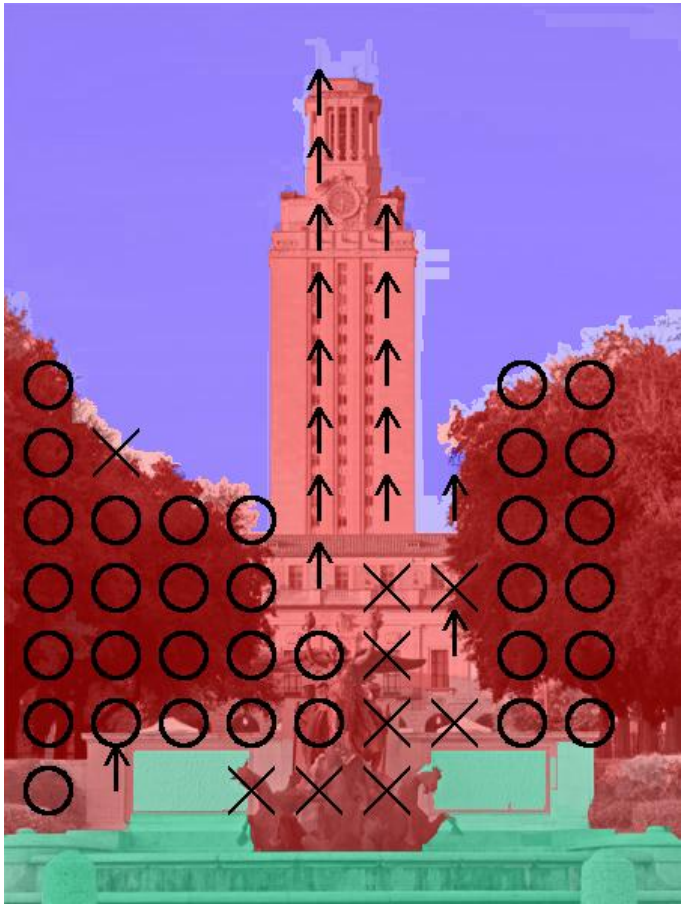
Front

Right

Porous

Solid

Step-by-step Example: Output



Precision vs. Recall

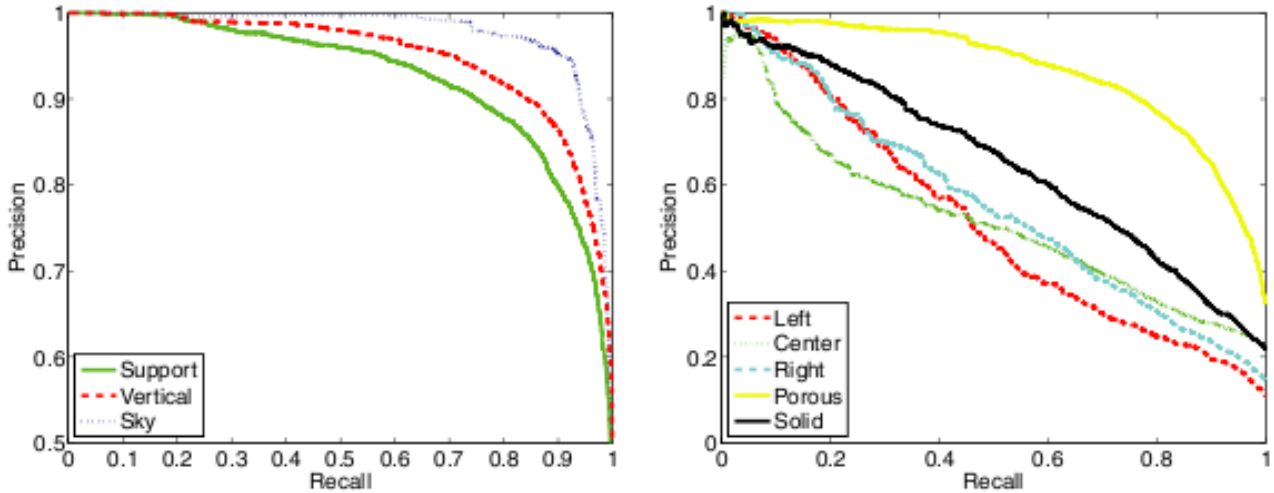


Figure 12. Precision-recall curves. Precision is the percent of declared labels (at some confidence threshold) that are true, and recall is the percent of true labels that are declared. The curve is generated by varying the confidence threshold.

Hoiem et. al. 2007

Pixels or superpixels or segmentations or ...?

Table 3. Confusion matrices (row-normalized) for multiple segmentation method.

Main Class

	Support	Vertical	Sky
Support	**0.84**	0.15	0.00
Vertical	0.09	**0.90**	0.02
Sky	0.00	0.10	**0.90**

Vertical Subclass

	Left	Center	Right	Porous	Solid
Left	**0.37**	0.32	0.08	0.09	0.13
Center	0.05	**0.56**	0.12	0.16	0.12
Right	0.02	0.28	**0.47**	0.13	0.10
Porous	0.01	0.07	0.03	**0.84**	0.06
Solid	0.04	0.20	0.04	0.17	**0.55**

Hoiem et. al. 2007

Table 4. Average accuracy (percent of correctly labeled image pixels) of methods using varying levels of spatial support.

Method	Main	Sub
Pixels	82.1	44.3
Superpixels	86.2	53.5
Single segmentation	86.2	56.6
Multiple segmentations	**88.1**	**61.5**
Ground truth segmentation	95.1	71.5

Hoiem et. al. 2007

- Even though superpixels allow more useful cues, the perspective cues require better spatial support.
- Better spatial support results in better results (ground truth).

Which cue is more informative? I

Table 5. Average accuracy under different sets of cues.

Cues	Main	Sub
All	**88.1**	**61.5**
Location	82.9	42.1
Color	72.2	43.1
Texture	79.9	54.6
Perspective	68.3	51.8
No Location	84.4	59.5
No Color	87.0	60.4
No Texture	86.7	58.2
No Perspective	88.1	56.6

Hoiem et. al. 2007

- Simple cues (location and color) are effective for main classes, but poor for distinguishing vertical classes.
- Each cue itself is quite effecting. On the other hand removing only one does *not* affect the performance significantly.

Which cue is more informative? II

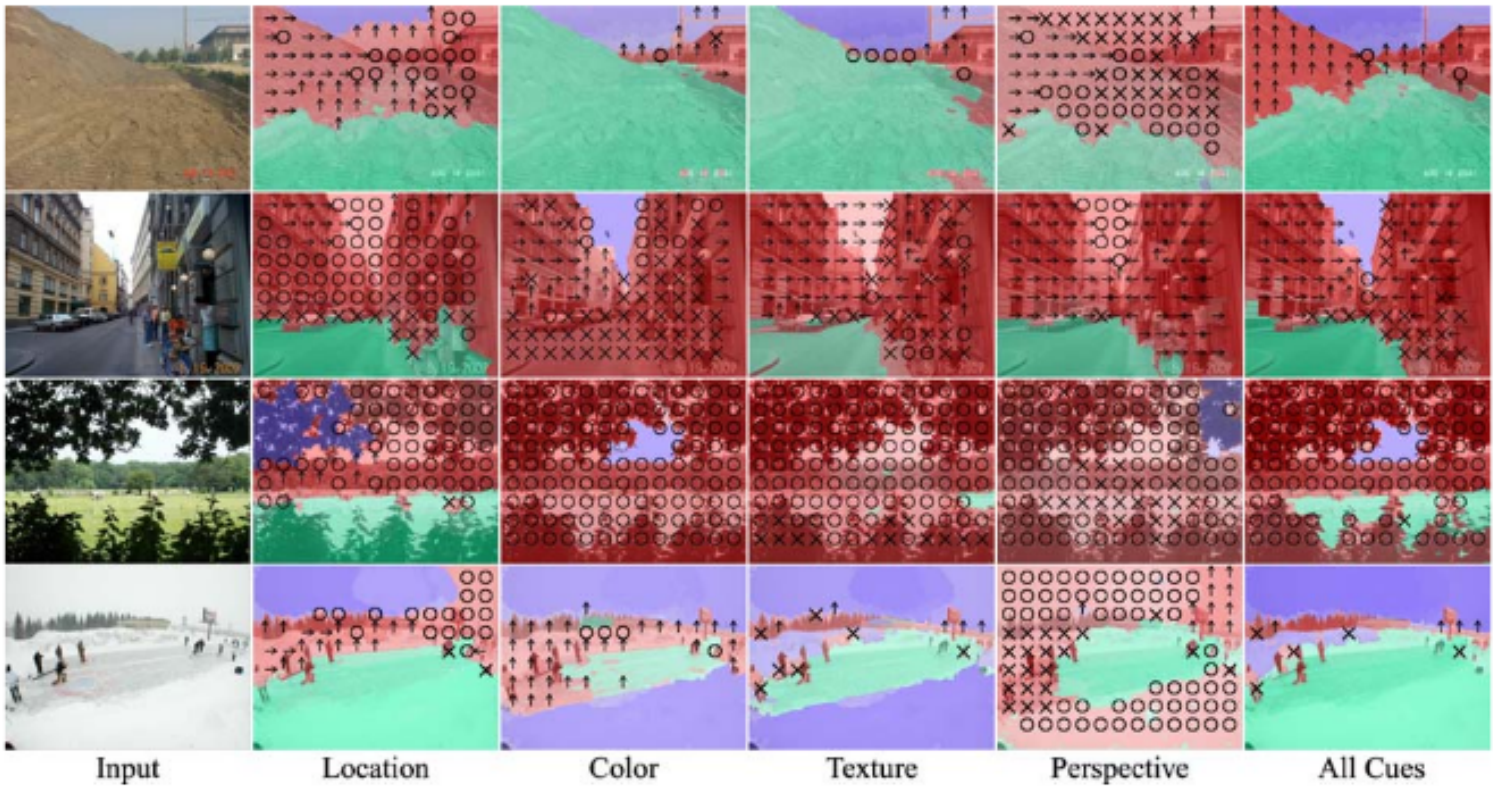


Figure 19. Results when performing classification based on each cue separately and using all cues. In each case, the same multiple segmentations are used (which are based on location, color, and texture), and those segments are analyzed with the given type of cues.

Hoiem et. al. 2007

Will I get better results with more data?

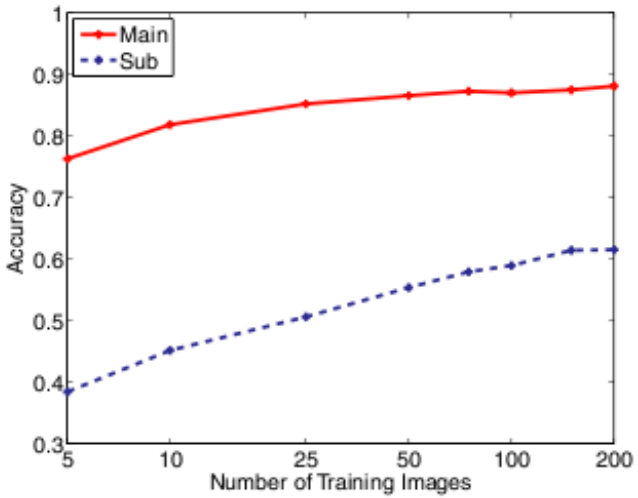


Figure 13. Accuracy while varying the number of training images. According to these trends, much larger training sets would probably result in significantly higher accuracy, especially for the subclasses.

Hoiem et. al. 2007

- More data is especially good for accurate subclass prediction.

How many segments should I have?

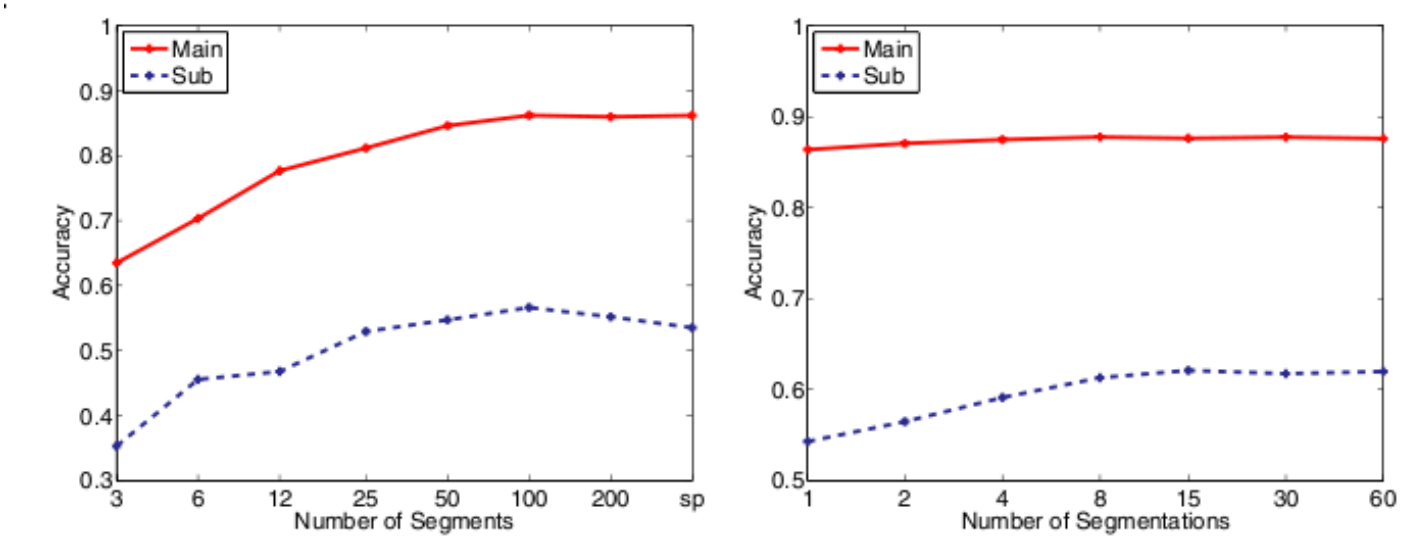


Figure 14. Analysis of segmentation parameters. On the left, classification is performed for single segmentations into varying numbers of segments ("sp" is the segmentation into superpixels). Peak accuracy is at 100 segments, with larger numbers of segments degrading the subclass accuracy slightly. On the right, classification is performed using the multiple segmentation method for varying numbers of segmentations. Although eight segmentations outperforms single segmentations by about 2% and 5% for main and subclasses, increasing the number of segmentations further produces no significant change in accuracy.

Hoiem et. al. 2007

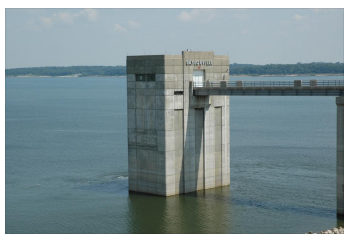
- Who wants to explain the decline in subclass accuracy after 100 segments?

Reminder: Training set looks like ...







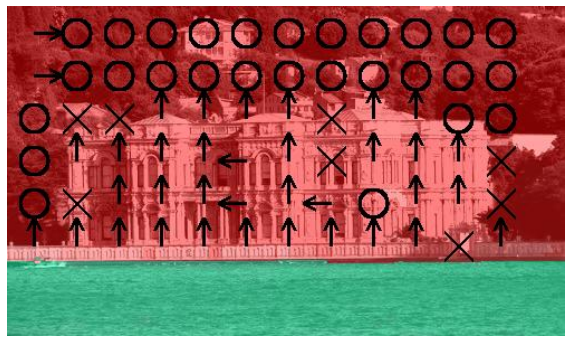





Strengths I

- Multiple cues are used which complement each other.



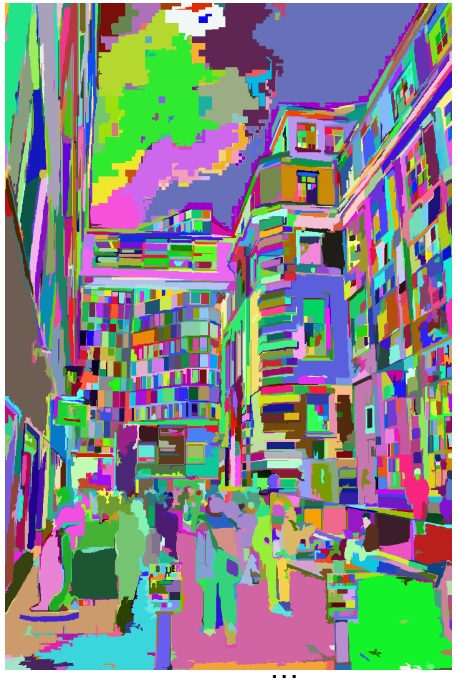
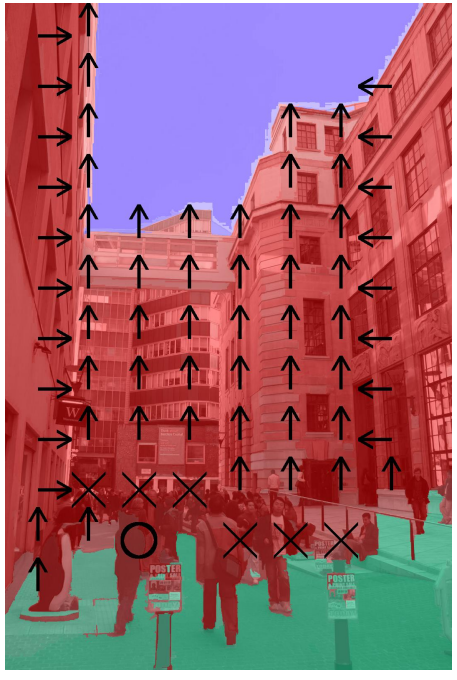


www.lampguesthouse.com/images/istanbul5.jpg

- Sea is properly classified as support even though sky has the same color with a higher probability.

Strengths II

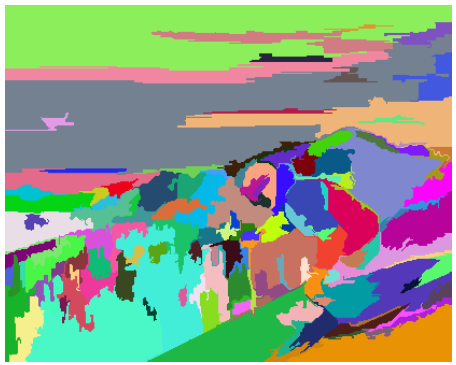
- Very accurate results can be achieved even in subclass categories if the test image is a good representative of the training set.

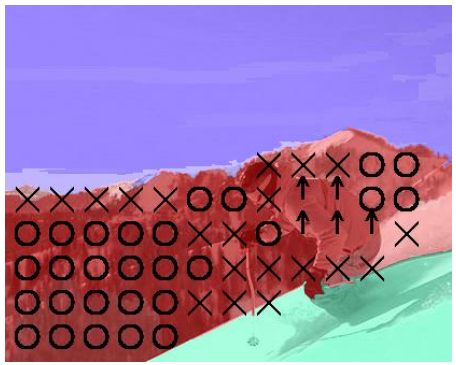
- Lots of lines to capture accurate perspective cues, and object ratios are appropriate (people are much smaller than the buildings).

Strengths III

- Some semantics can be learned from the training data.



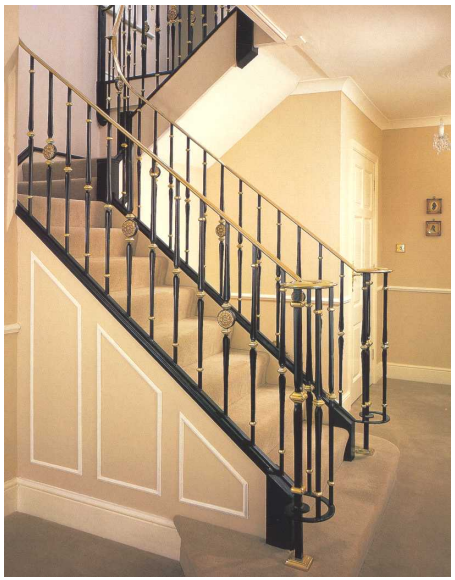
...

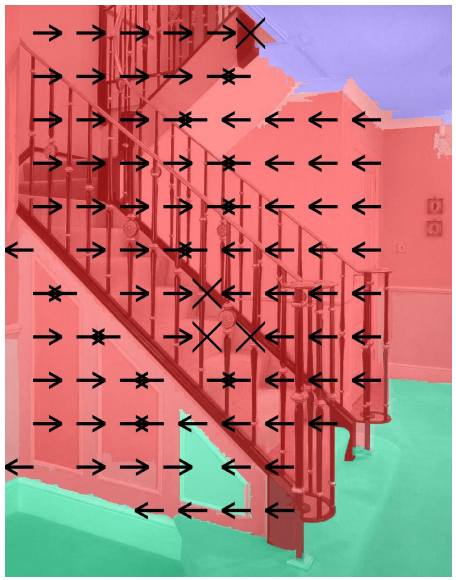


- Skier is classified as a vertical object. The snow, even though white and steep is correctly classified as support.

Strengths IV

- Special purpose training is straightforward.

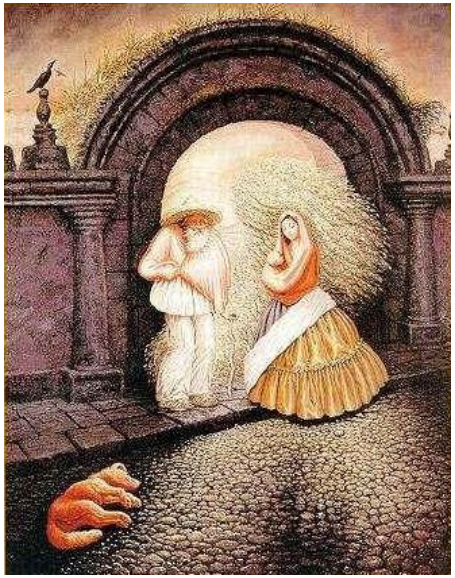
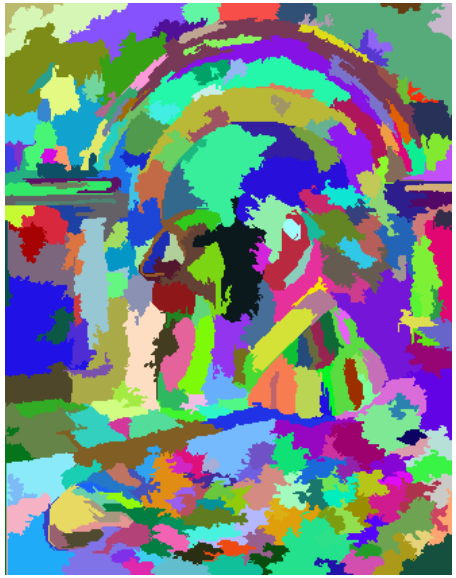




- Segment classifiers are trained with indoor images. Label classifiers are trained with outdoor images. The results are still pretty good.

Strengths V

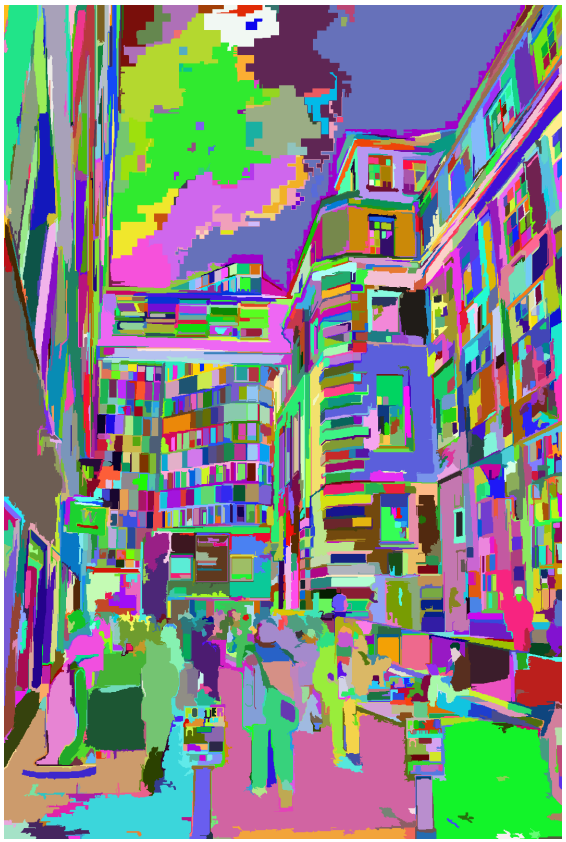
- Amazingly works for some very different test images.

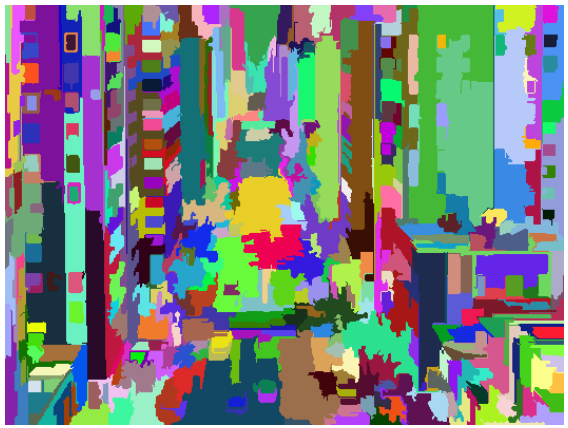




- Support/vertical plane separation is pretty good even though image is challenging.

Weaknesses I

- Can you imagine preparing the training data set?





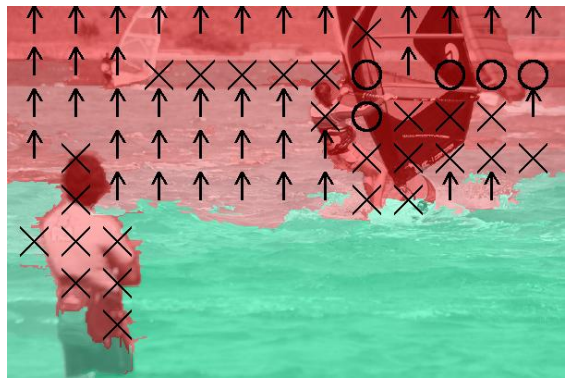
...

- Manual ground-truth labeling is extremely costly. They manually labeled 150.000 segments!

Weaknesses II

- There is a significant location bias.



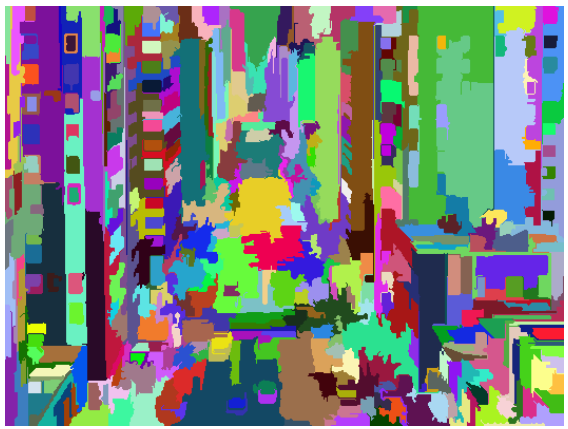
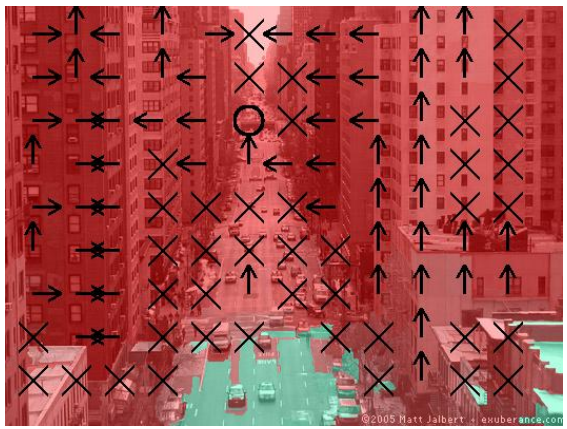


Personal photo.

- Even though color and texture consistencies are fine, due to unexpected class positions, the results are very bad.

Weaknesses III

- There is a significant viewpoint bias.

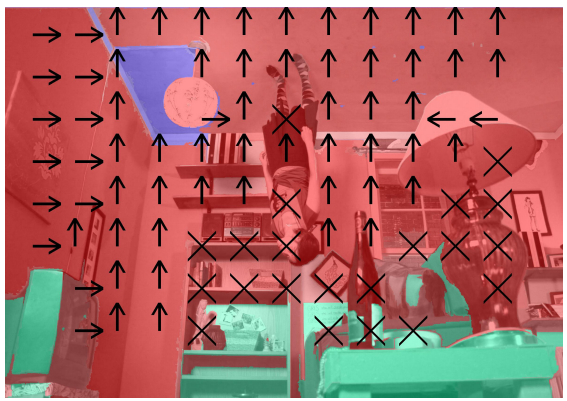

...


- The photo is taken from a position slightly higher than the ground plane resulting in very bad segmentation, even though there are lots of lines to support the perspective cues.

Weaknesses IV

- There is a significant viewpoint bias (second example).



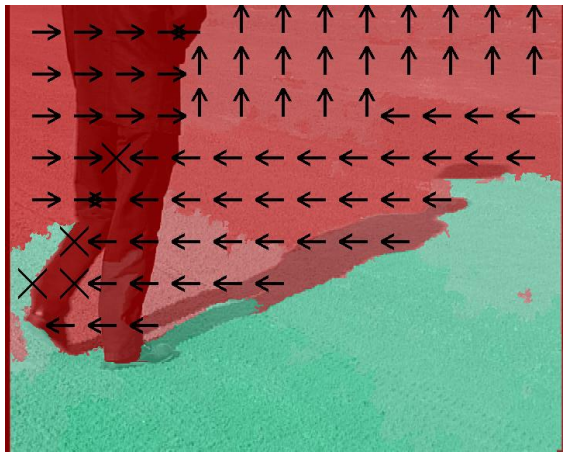


- The photo is taken from a very low point resulting in a bad segmentation (indoor training set is used).
- Fine, the girl is upside down, but should it matter?

Weaknesses V

- There is a significant bias on continuous segments, and
- Reflections and/or shadows can be misleading.





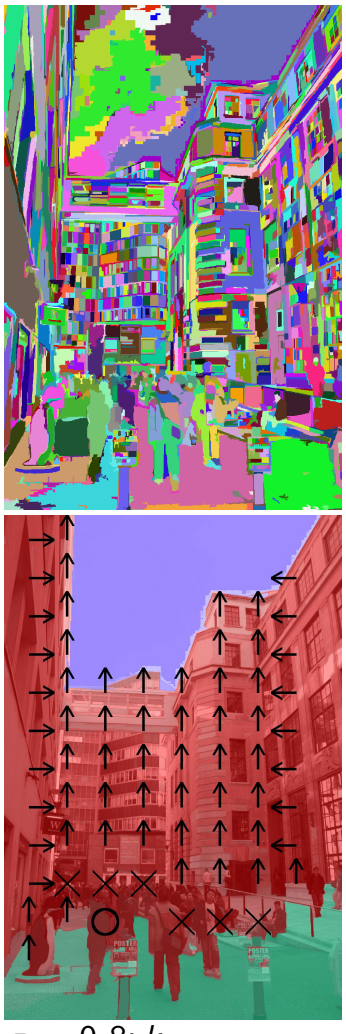
- However, they can actually be used as additional cues...

Weaknesses VI

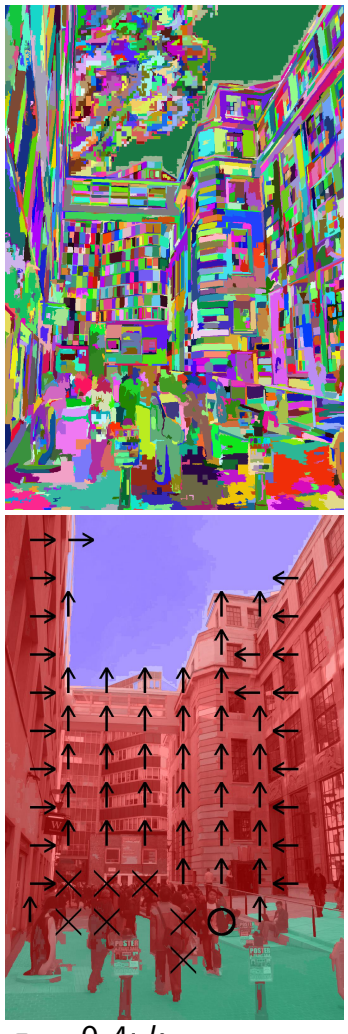
- What is the problem in this example?

- Color? Perspective cues? Texture?

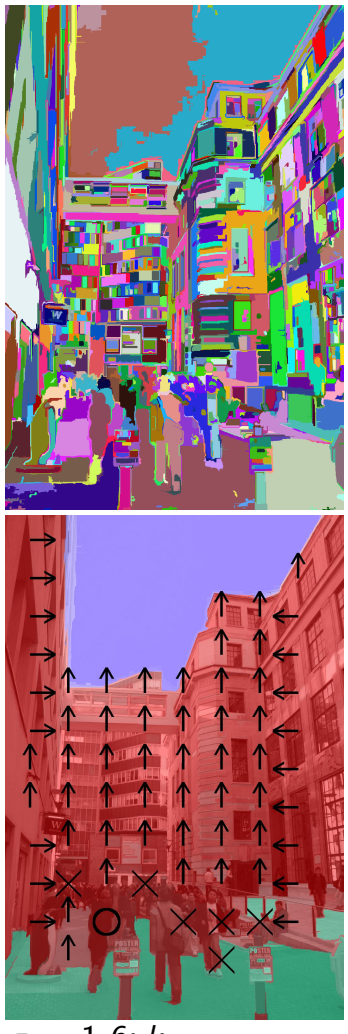
Effects of Oversegmentation Parameters



$\sigma = 0.8$; $k =$ 100; $min = 100$

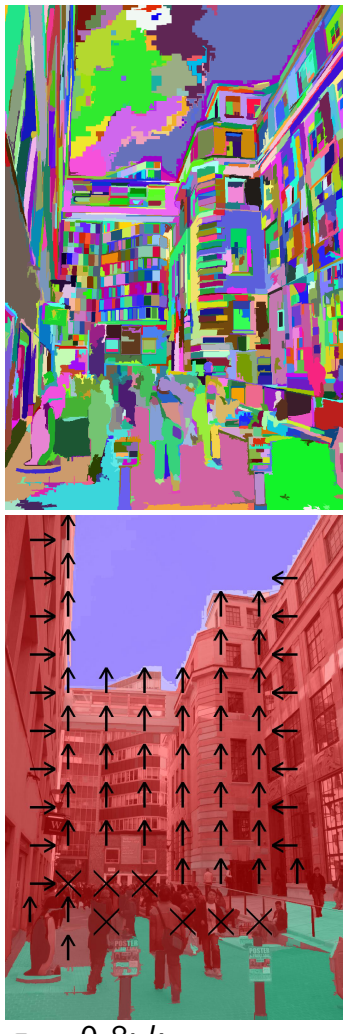


$\sigma = 0.4$; $k =$ 100; $min = 100$



$\sigma = 1.6$; $k =$ 100; $min = 100$

$\sigma = 0.8$; $k =$ 200; $min = 100$



$\sigma = 0.8$; $k =$ 100; $min = 200$

Final Word

- Works pretty well if a similar image is in the training set.
- Removing a single cue does NOT affect the performance a lot.
- Especially location and viewpoint biases are significant.
- Oversegmentation parameters do NOT affect the performance dramatically.

References

- Hoiem, D. Efros, A. A. and Hebert, M. 2005. Geometric context from a single image. In *Proc. ICCV.*
- Hoiem, D. Efros, A. A. and Hebert, M. 2007. Recovering surface layout from an image. *IJCV*, 75(1):151-172.
- Yu, S. X. Zhang, H. and Malik, J. 2008. Inferring spatial layout from a single image via depth-ordered grouping. *Workshop on POCV*.