Problem Statement

• Given images of interior of a building, how much can a robot recognize the building later
• Qualitative Image Localization

I am in Corridor 4 but I do not know the exact location
Global v. Local approach

- **Global - Histogram of Oriented Gradients**
  - Introduced by Dalal & Triggs, CVPR 2005
  - Extended by Bosch et. al., CIVR 2007 - pyramid of HoG - used in the experiments with no pyramids
  - Kosecka et. al., CVPR 2003 uses simpler version of HoG for image based localization

- **Local - SIFT features**
  - Kosecka et. al., CVPR Workshop 2004
Basic HoG algorithm

- Divide the image into cells
  - In our case, every pixel is a cell
- Compute edges of the image
  - canny edge detector
- Compute the orientation of each edge pixel
- Compute the histogram
  - Each bin in the histogram represents the number of edge pixels having orientations in a certain range
Parameters to HoG

- Number of Bins of the Histogram
- Angle - 180° or 360°,
  - 180° - contrast sign of the gradient is ignored
    - used in the experiments
  - 360° - uses all orientations as in SIFT
• Histogram of gradient orientations
  - Orientation
  - Position

- Weighted by magnitude
Different HoGs

• Difference between level 0 of pyramid HoG in Bosch et. al. versus Kosecka et. al. implementation of HoG
  ▪ The vote of each edge pixel is linearly distributed across the two neighboring orientation bins according to the difference between the measured and actual bin orientation - soft voting
  ▪ Eg.: Bins - 10°, 20°, 30°; measured value - 17°,
  ▪ vote for: Bin 10° - .15, Bin 30° - .15, Bin 20°- .75
Chi-Square distance

\[
\chi^2(h_i, h_j) = \sum_k \frac{(h_i(k) - h_j(k))^2}{h_i(k) + h_j(k)}
\]

\(h_i\) and \(h_j\) are histograms of two frames

\(k\) is the number of histogram bins

Kosecka et. al., CVPR 2003
Benefits of HoG

• Computed globally
• Occlusions caused by walking people, misplaced objects have minor effects
• Can generalize well
• Has worked really well for finding pedestrians on the street
Dataset

- Total number of images: 92
- Randomly selected 80% to form the training set
- Rest 20% is the test set
- Number of classes: 12
- Ran HoG and SIFT ten times
HoG Experiments

- Effect of a threshold - how much is the nearest image in the training set far from the next nearest
  - ratio of matching features in both the training images
- Effect of Quantization - One representative or prototype view of every class
- Effect of number of bins
• Effect of varying the threshold
• Number of Bins = 10

For threshold = 0.2,
Undecided but would have been
• correctly classified - 10!!
• wrongly classified - 8

Many images in the training set have nearly
same histogram of oriented gradients
Accuracy - Vary Bins

Effect of varying the number of bins
Threshold = 0

- Less number of bins - Too much quantization of orientations
- Large number of bins - Very less quantization of orientations
Accuracy - Prototype Views

- Threshold = 0, Bins = 10, One prototype image per class
- Prototype image computed by taking mean of images of same class

Accuracy

Prototype Views?

No

Yes
Best Combination

- Threshold = 0
- Bins = 30
- No prototype views
HoG Results

Test

Result

Correct

Correct
Obvious answers

Test

Result

Wrong

Wrong
Some images are just hard to classify…

Test

Result
Guess?

Test

Result
Guess?

Test

Result
Confused?

Test

Result

All are wrongly classified, though they look so similar…
SIFT

• Scale & affine invariant feature detection
  ▪ Combines edge detection with Laplacian-based automatic scale selection
  ▪ Mikolajczyk et. al. CVPR’06, BMVC ‘03

• SIFT descriptor
SIFT Vector Formation

• Threshold image gradients are sampled over 16x16 array of locations in scale space
• Create array of orientation histograms
• 8 orientations x 4 x 4 histogram array = 128 bit vector
Algorithm

• For every test image
  ▪ For every training image
    • Find the nearest matching feature
    • Find the second nearest matching feature
    • If nearest neighbor 0.6 times closer than the second nearest neighbor
      ▪ Number_of_matching_features ++
  ▪ Find the training image with most number of matching features
Let $d_i$ be the minimum distance and $d_j$ be the second minimum distance, then feature $i$ matches feature $j$ if $d_i < 0.6 \times d_j$. 

This can be illustrated with the following diagram:

- **Test Images**
- **Each training image**

The distances $d_1, d_2, \ldots, d_n$ are shown between the features in the test image and the training images.
Two Types of Threshold

- One is to check whether there is a matching feature in the given training image or not
  - Fixed - 0.6
- One is to check whether the nearest image is far away from the next nearest image or not
  - Experimented for various values
Results - Numbers

• SIFT
  - Correctly Classified - 99
  - Wrongly Classified - 81
  - Accuracy - 55%

Better than HoG!
SIFT - One bad image ruined the accuracy!
Reason
New Results for SIFT

• Removed the image
  ▪ Avg. no. of images correctly classified: 134
  ▪ Avg. no. of images wrongly classified: 46
  ▪ Accuracy 74.4%
  ▪ Earlier accuracy 55%
  ▪ 19.44% higher accuracy!!
Result

- Varying the threshold
Threshold is not good

- Undecided but were correct with 0 threshold
- Undecided but were wrong with 0 threshold
Modified feature matching in SIFT

• For every test feature, find nearest and second nearest feature from ALL the training images’ features
• A feature is matching if
  nearest_distance < 0.6*second_nearest_distance
• Find the training image that has most features matching with the test image
• Call this one SIFT\textsubscript{2} and the earlier one SIFT\textsubscript{1}
Modified feature matching

Test Image

Training Images

d_1

d_{n-1}

d_n

...
Result of SIFT$_2$

• Threshold = 0
  ▪ Correct - 163
  ▪ Wrong - 17
  ▪ Accuracy - 90.5%
  ▪ Accuracy of SIFT$_1 = 74.4\%$ -- 16.1% higher!!

• Also, the one bad image problem gets removed!
Vary Threshold in SIFT$_2$

Number of training images

![Bar graph showing the number of correct, wrong, and undecided images at different thresholds.]

- Correct
- Wrong
- Undecided but were correct for 0 thresh
- Undecided but were wrong for 0 thresh

Threshold:
- 0
- 0.8
- 1.2
- 1.6

Y-axis: Number of images
X-axis: Threshold
Another dataset

- Till now we had images of the SAME building in our training set

- What if Robot is shown a DIFFERENT building?
  - Can it recognize if an image is a corridor or an office?

- Test dataset has images from different floor and different buildings
  - ACES 5th floor and Taylor hall’s corridor
  - Removed the Taylor Hall’s corridor images from the training set
Dataset - II
No clear winner but $\text{SIFT}_2 = -1$
No clear winner, but $\text{SIFT}_2 = -2$
Results

Test

HoG

SIFT$_1$

SIFT$_2$

\[ \text{HoG} = 1; \quad \text{SIFT}_1 = 1, \quad \text{SIFT}_2 = -2 + 1 = -1 \]
Results

Test  

HoG  

SIFT_1  

SIFT_2

HoG = 2; SIFT_1=1, SIFT_2=-1+2=1
Results

Test

HoG

SIFT\(_1\)

SIFT\(_2\)

HoG = 3; SIFT\(_1\) = 1, SIFT\(_2\) = 2
Results

\( \text{HoG} = 4, \ SIFT_1 = 1, \ SIFT_2 = 2 \)

\( \text{HoG better than SIFT!} \)
• HoG captures the global distinctiveness of a category
• Lets see histograms of some of the images
Result of HoG

Of same class as 1

Test

Result of SIFT

Note

• 3 is similar to 1
• 3 is not similar to 4
• 1 is not very similar to 2
SIFT Explanation

- 20 matching points between test and result images

Test

Result of SIFT$_1$
• Only 6 matching points between test image and the result produced by HoG (correct)

Test

Result by HoG
Conclusion

• SIFT performs better than HoG in previously seen building
  ▪ Local descriptor - gets the distinguishing local features
• HoG performs better than SIFT in previously unseen building!
  ▪ Global descriptor - gets the essence
  ▪ Better than SIFT in formal setting of the environment -- Buildings are never at 30°!!
  ▪ Rotation invariance of SIFT results in worse accuracy
Conclusion

• Matching features across all the training images ($SIFT_2$) is better than matching features image by image ($SIFT_1$)

• $SIFT_2$ performs better than $SIFT_1$ in both previously seen and unseen buildings

• Quantization by taking mean in HoG gives poorer performance

• If we are performing 1-NN approach in classification using $SIFT_1$, then one bad image can deteriorate the results
Discussion Points

- Will threshold for selecting nearest images over next nearest image work when we quantize the image?
  - Since only one image per class
- Modify the threshold criteria by calculating ratio of number of matching features of nearest neighbor and for next nearest neighbor of different class
- Rotation invariance of SIFT is sometimes hurting the performance. Can we make it partially invariant for this task?
- What can be other matching algorithms than SIFT and HoG?
References and Resources

- Kosecka et. al., Qualitative Image Based Localization in Indoor Environments, CVPR 2003
- Dalal and Triggs, Histograms of Oriented Gradients for Human Detection, CVPR 2005
- Kosecka et. al., Location Recognition and Global Localization Based on Scale-Invariant Keypoints, CVPR Workshop 2004
- Pyramid of Histogram of Oriented Gradients
  - http://www.robots.ox.ac.uk/~vgg/research/caltech/phog.html
- Local features detector and descriptor
  - http://www.robots.ox.ac.uk/~vgg/research/affine/detectors.html