16-720B Computer Vision: Homework 1

Vigneshram Krishnamoorthy

27th September 2017

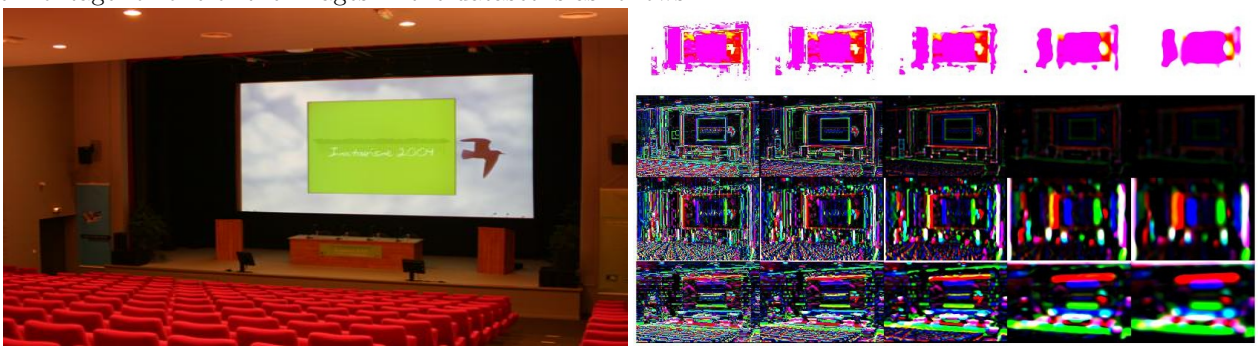
1 Answer to Q1.0

Each filter in the given multi-scale filter bank picks up different features in an image at different scales.

- The first row corresponds to a rotationally symmetric Gaussian lowpass filter at different scales. It removes noisy components in the image, removing local discontinuities and picking up the similarities.
- The second row corresponds to rotationally symmetric Laplacian of Gaussian filter at different scales. It does the opposite of the first row, picking up local discontinuities.
- The third row of filter banks pick up vertical edges (by computing vertical gradients using edge operators such as Prewitt or Sobel) at 4 different scales.
- The fourth row picks up horizontal edges (by computing horizontal gradients using edge operators such as Prewitt or Sobel) at 4 different scales.

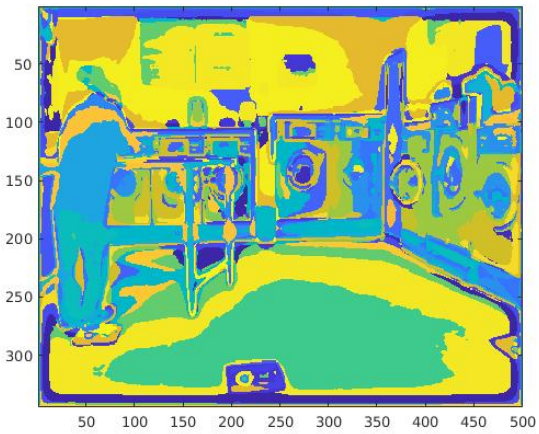
2 Answer to Q1.1

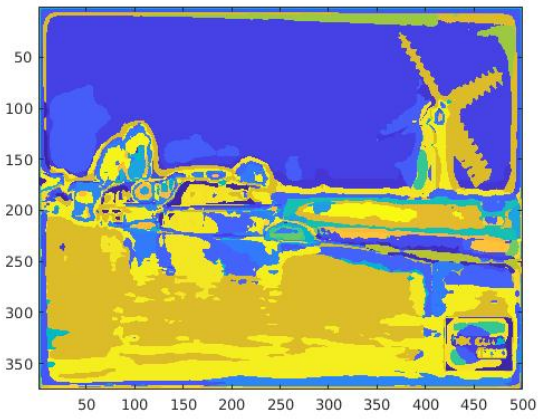
The montage for one of the images in the dataset is as follows

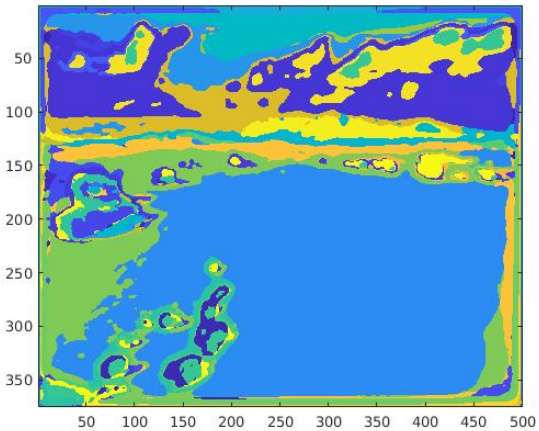


3 Answer to Q1.3

Given below are the wordmaps of 3 different images from 3 different categories in the dataset.







As we can see from these wordmaps,local features in each of the images are mapped to its corresponding dictionary element (visual word). We can see that similar features in the image get mapped consistently to similar visual words. Hence the information (most of it) available over M*N*3 pixel elements is now available to us through the wordmap representation over the image as seen in the above visualizations. Now that we have almost the same information with reduced dimensionality, we can process things much faster.

4 Answer to Q2.5

- The accuracy obtained using the BoW approach on the given test set is 47.5 %
- The confusion matrix obtained is as follows

$$\begin{pmatrix} 7 & 1 & 0 & 0 & 2 & 3 & 0 & 0 \\ 0 & 7 & 1 & 2 & 0 & 0 & 3 & 2 \\ 1 & 3 & 14 & 1 & 0 & 0 & 0 & 1 \\ 2 & 2 & 2 & 9 & 1 & 1 & 1 & 5 \\ 4 & 1 & 0 & 1 & 13 & 4 & 0 & 0 \\ 3 & 0 & 2 & 2 & 2 & 8 & 5 & 1 \\ 2 & 3 & 0 & 0 & 1 & 3 & 10 & 3 \\ 1 & 3 & 1 & 5 & 1 & 1 & 1 & 8 \end{pmatrix}$$

5 Answer to Q2.6

- We can see that the desert class (class 3) gets most accurately classified, perhaps since its visual words don't often correlate with other classes (such as auditorium, baseball field, kitchen etc.). The kitchen class also classifies well under BoW due to image features that don't have correlations with other classes such as baseball field, desert etc.
- The classes auditorium and baseball field get poorly classified. The features of the auditorium class often gets confused with the kitchen class or the laundromat class. Similarly, the features of the baseball field class seems to be correlated with the windmill and waterfall classes (probably due to the presence of objects such as the sky, grass, trees etc. in all these classes).
- The system often confuses a windmill to a highway and vice-versa (5 out of 20 occurrences in each). This can probably be attributed to many samples in the highway scene which are wide-angle, long-range shots covering objects such as grass, trees, sky etc.) along with the primary objects that are unique to highways such as cars, roads, trucks etc.
- We conclude that the bags-of-words performs decently with classes with more unique and distinctive features, and gets confused between classes sharing common features.

6 Answer to Q3.2

- The accuracy obtained using the Deep Learning approach on the given test set is 93.75 %
- The confusion matrix obtained is as follows

$$\begin{pmatrix} 19 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 19 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 19 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 19 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 19 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 18 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 17 \end{pmatrix}$$

1. We can clearly see that the results obtained from the Deep Learning approach is significantly (Almost 2 times better) than the conventional BoW approach discussed earlier.

2. The Deep Networks approach learns much more about any given image in the dataset than the bags-of-words, in terms of the intricate feature details in any given image, due to presence of multiple layers (with non-linearity) extracting more and more information unique to any given class.

3. The Deep Networks approach comes much closer to emulating how a human perceives objects in the real-world than the bags-of-words method and is hence much more robust.

4. The jump in accuracy can also be partially attributed to the presence of much more data, for the pre-trained network than for the BoW system which had only 1440 training examples.