COMP5421 Computer Vision

Homework Assignment 3
Spatial Pyramid Matching for Scene Classification

Hartanto Kwee Jeffrey
jhk@connect.ust.hk — SID: 20851871

1 Representing the World with Visual Words

1.1 Extracting Filter Responses

Q1.1.1

1. Gaussian: in a Gaussian filter, each pixel will store the gaussian-weighted average of its surrounding pixels, so it basically picks up the characteristics of neighbouring pixels. Roughly speaking, each pixel "tells whats happening around it".
2. Laplacian of Gaussian: edges and corners
3. derivative of Gaussian in the x direction: vertical edges
4. derivative of Gaussian in the y direction: horizontal edges

We use multiple scales because we want to capture different sizes of surrounding responses to cater for different sizes of features and images. For example, a large sigma may only capture features of neighbours one or two pixels away, while a small sigma may capture wider and large-scale features spanning tens of pixels. For edge detection, larger sigmas are more sensitive to local image gradients and will be more noisy, while smaller sigmas create smoother gradients.

Q1.1.2

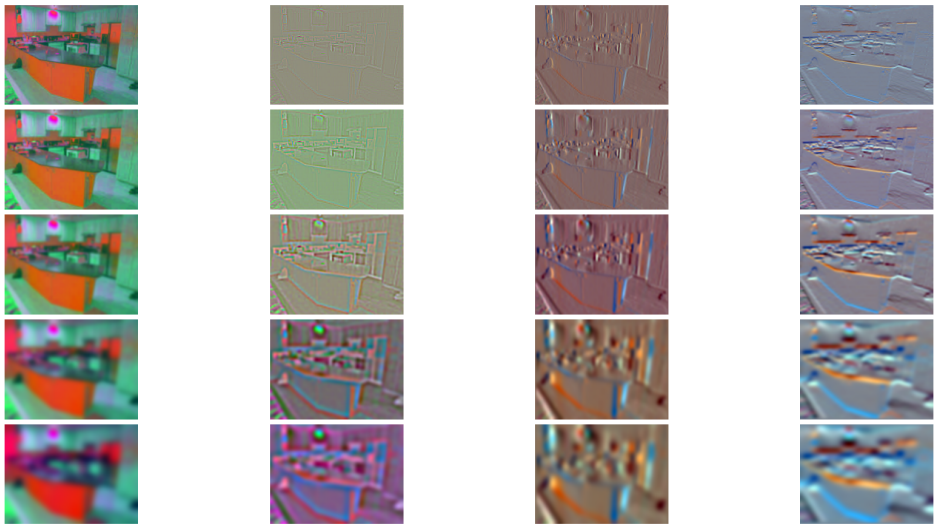

Figure 1: 20 filter responses resulting from the filter bank

Q1.2

This implementation uses `alpha` = 200 and `K` = 150.

Q1.3

Refer to Figure 2 for the wordmaps.

Observing the wordmaps, we see there are clusters of pixels mapped to the same visual word, such as the wooden surfaces of cupboards and white counter top in a kitchen, the ripple pattern of sand in the desert, blue sky with occasional strips of smoky cloud. At the same time, we also see that edges in images are clearly mapped to the same wordmap, showing the effect of the LoG and Sobel Gaussian filters. (If we didn't have these filters and only had Gaussian filters, then these edges may only be mapped different words based on merely its color, and we would see noise instead of the uniform edge colors we see now.) Edges include the counter top and sand ripples. These wordmaps allow images to "speak their features" using visual words, as we see different visual words show up in different frequencies in different images.

2 Building a Recognition System

Q2.5

Confusion matrix:

```
                Predicted Class
      [[13.  0.  0.  0.  2.  2.  1.  2.]
       [ 0. 12.  1.  2.  0.  1.  1.  3.]
       [ 1.  2.  7.  6.  1.  1.  0.  2.]
Actual [ 0.  2.  1. 15.  0.  0.  0.  2.]
 Class [ 3.  0.  1.  0. 11.  3.  0.  2.]
       [ 1.  0.  0.  1.  8. 10.  0.  0.]
       [ 2.  0.  0.  0.  0.  2. 16.  0.]
       [ 0.  1.  0.  5.  0.  0.  0. 14.]]
```

Accuracy: 0.6125

Q2.6

Refer to Figure 3 for the failed cases.

If we refer to the confusion matrix, the most frequently misclassified paris are [kitchen, laundromats] (11 cases), [desert, highway] (7 cases) and [highway, windmill] (7 cases).

Kitchen and laundromates are both indoor places and share common features such as wooden cupboards, white concrete walls and metallic silver (sinks, oven doors, doors of washing machine etc.).

Deserts are frequently misclassified as highways, perhaps because both only consists of a blue sky and a grey concrete road. There are not many edges in highways, so usually large areas of grey are identified as highways. In the two desert images in the failed cases, both have little edge-like features and are predominantly plain gray. This may be the reason why deserts are frequently misidentifed as highways. Conversely, highways are not frequently missclassfied perhaps because of the presence of road signs.

Highways are frequently misclassified as windmills. In the three failed cases shown, there are roads in addition to the windmills. Since the windmill only takes up a small area in the image, the effect of the road-like features dominate the prediction, and causes these images to be misclassified as highways.

3 Deep Learning Features

Q3.2

Confusion matrix:

```
                Predicted Class
      [[19.  0.  0.  0.  1.  0.  0.  0.]
       [ 1. 16.  1.  0.  0.  0.  1.  1.]
       [ 0.  0. 19.  1.  0.  0.  0.  0.]
Actual [ 0.  0.  0. 20.  0.  0.  0.  0.]
 Class [ 0.  0.  0.  0. 19.  1.  0.  0.]
       [ 0.  0.  0.  0.  1. 19.  0.  0.]
       [ 0.  0.  1.  0.  0.  0. 19.  0.]
       [ 0.  0.  0.  1.  0.  0.  0. 19.]]
```

Accuracy: 0.9375

The accuracy of deep features is much higher than bag of words.

In our discussion of the bag of words failed cases, we observe that a lot of misclassification stems from sharing features with other classes. If the unique feature of the actual class takes only a small area, our histogram-based matching would easily neglect this feature and misclassify.

Deep features are more robust as they capture both wide-scale and small-scale features, and places different weighting on these features as opposed to histogram matching. Therefore, unique features may lead to greater activations and dominate the network output.

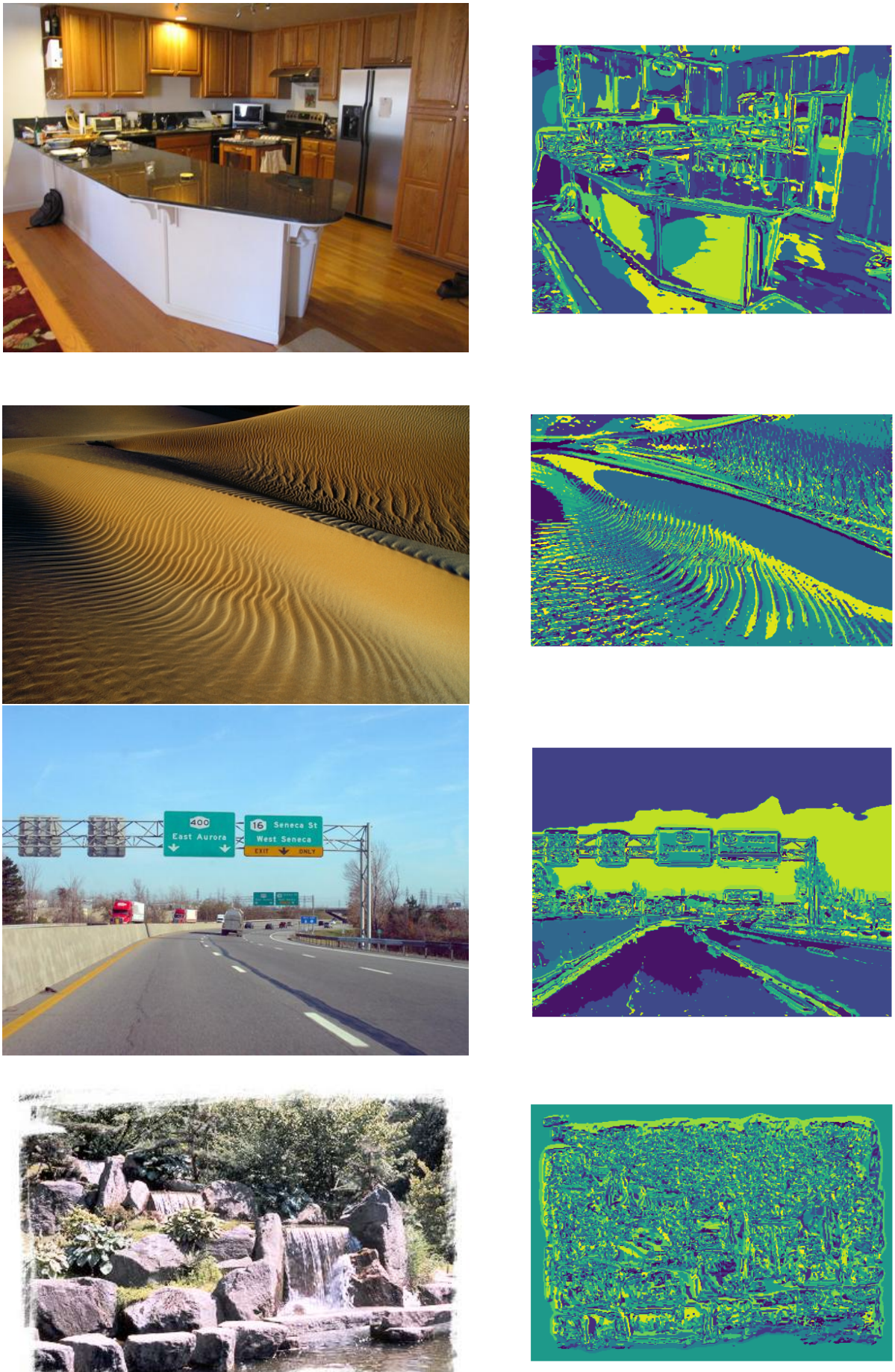

Figure 2: Visualized wordmaps of four images.

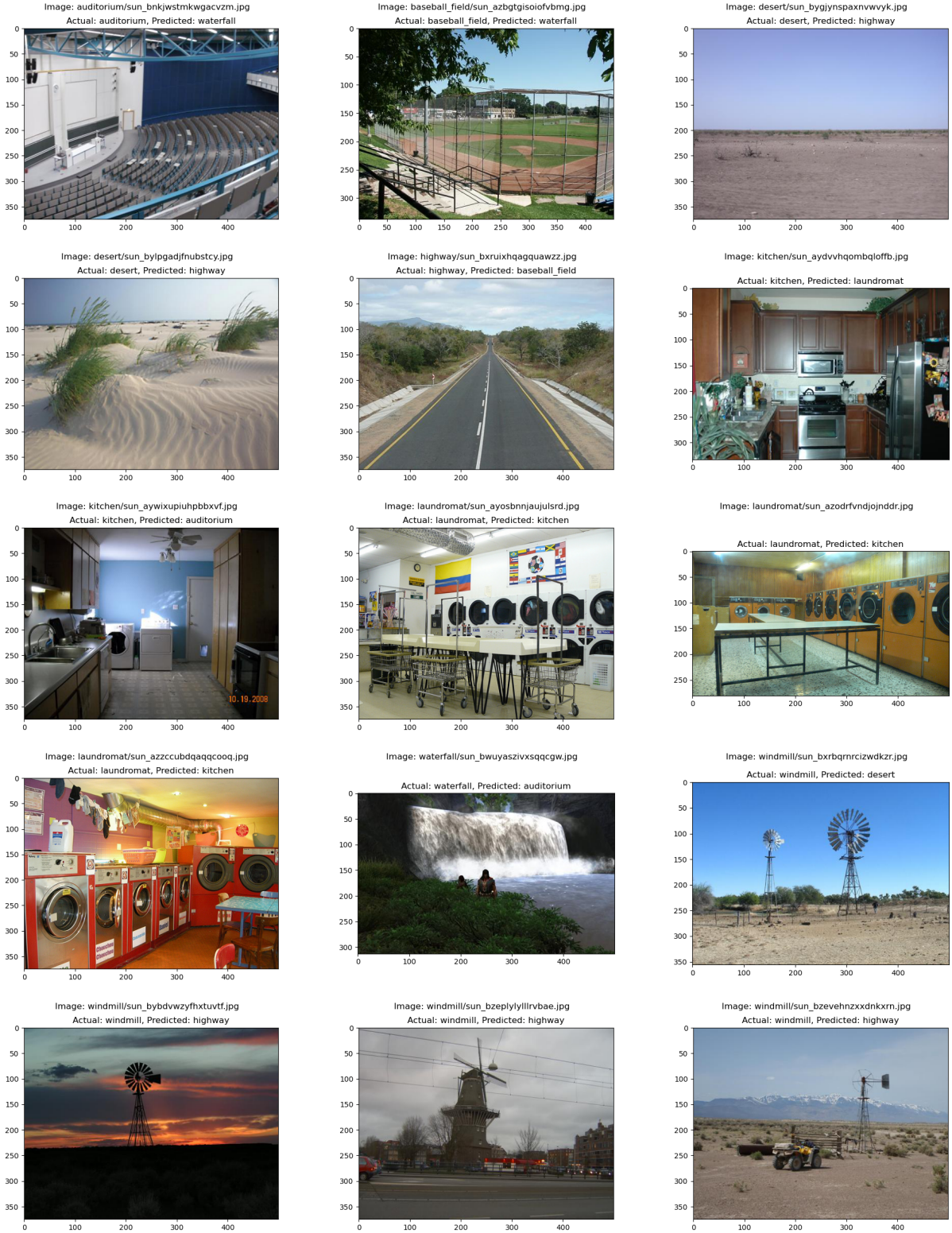

Figure 3: Failed cases with their predicted and actual class.