

# Learning Hierarchical Image Representation with Sparsity, Saliency and Locality

Jimei Yang  
jyang44@ucmerced.edu

University of California, Merced  
California, USA

Ming-Hsuan Yang  
mhyang@ucmerced.edu

---

## Abstract

We present more experimental results and analysis in this supplementary document. Specially, we carry out sensitivity analysis to demonstrate the performance of our HSSL model with respect to parameter settings. We also present some analysis about saliency maps.

## 1 Sensitivity Analysis

We evaluate our HSSL model with different parameter settings on the Caltech101 dataset (where 30 training samples per class are used).

### 1.1 Dictionary Size of the first layer

We first evaluate the performance of HSSL with respect to the size of dictionary in the first layer. We vary the number of atom signals in the dictionary and fix all the other parameters for experiments. Figure 1 shows that the performance decreases slightly when the number of bases is increased. One explanation is that the learned dictionary tend to be more selective when the size is increased. Thus, similar image patches may respond to different atom signals and have different sparse codes (i.e., the diversity within the same class is increased). Nevertheless, the results show that it is reasonably easy to obtain good results with sensible size of dictionary.

### 1.2 PCA dimension in the first layer

We evaluate the effect of PCA dimensions in the first layer on classification accuracy (while fixing all the other parameters). Figure 2 shows the results where one can see that the classification results are relatively stable as the number of PCA dimension is varied within a sensible range. By setting the grid size of saliency pooling to  $4 \times 4$  and the dictionary size to 2048 in the second layer, we find 96-dimensional PCA gives the best result.

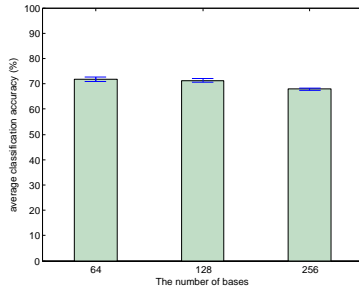


Figure 1: Effect of dictionary size in the first layer.

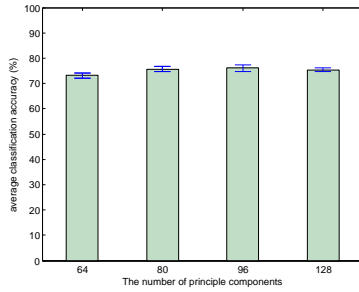


Figure 2: Effect of PCA dimensions in the first layer.

### 1.3 Local Grouping vs. Dictionary Size in the second layer

The smaller the grid size of grouping is, the less complex the output feature of this layer will be. We evaluate the effects of this parameter when varying the size of dictionary of second layer in Figure 3. The results clearly show that our method is relatively robust to different settings of grid size and dictionary size.

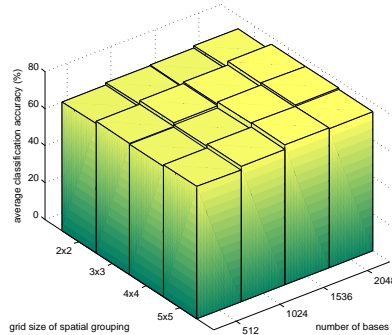


Figure 3: Effect of grid size of local grouping and the dictionary size of second layer on classification accuracy.

## 2 Saliency Maps

We next investigate the effect of saliency map on object recognition tasks. Saliency map algorithms help object “stand out” from background clutter. Thus, when an object appears with background clutter, saliency map can enforce the max pooling operator to select the features from the foreground object region and generate a better representation. However in some cases, the target object is not salient in the image so that saliency map will inadvertently pop out background region and introduce significant amount of noise. Also, when the extent of the target object is very large, some parts of the foreground object may be mistaken as background and suppressed.

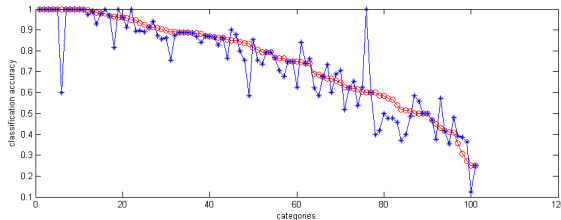


Figure 4: Classification results without (blue line) and with (red line) saliency map.

Figure 4 shows the classification accuracy on object category with and without saliency map (using 30 training samples). We sort the classification accuracy with pooling on the saliency map from the best to the worst, and plot the results with the red line. The blue line denotes the corresponding classification accuracy without saliency map. As evident in Figure 4, the blue line basically lies below the red one except two obvious outliers. The average classification accuracy without using saliency map is 73.1%, which is 3% lower than the result with saliency map. The first outlier appears in the mayfly category, where we obtain perfect (100%) accuracy with saliency map and only 60% without saliency map. The second one appears in the octopus category, where we obtain 60% accuracy with saliency map and perfect (100%) accuracy without saliency map. Figure 5 explains these two cases in details. Figure 6 and Figure 7 show the results with saliency maps from the Caltech101 and Oxford Flowers datasets.

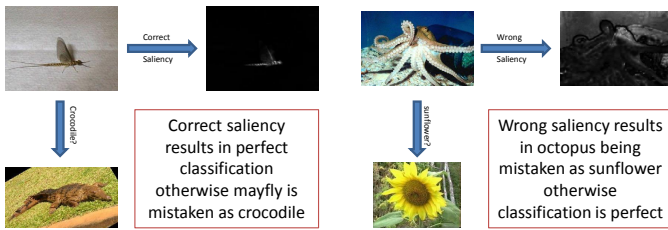


Figure 5: Examples where saliency map works well (left) or not (right).



Figure 6: Saliency maps from the Caltech101 dataset. The left column shows the original images and the right column shows the corresponding saliency maps (where strong filter responses are indicated by high intensity values). The saliency map selects the pixels of the dominant object in an image.

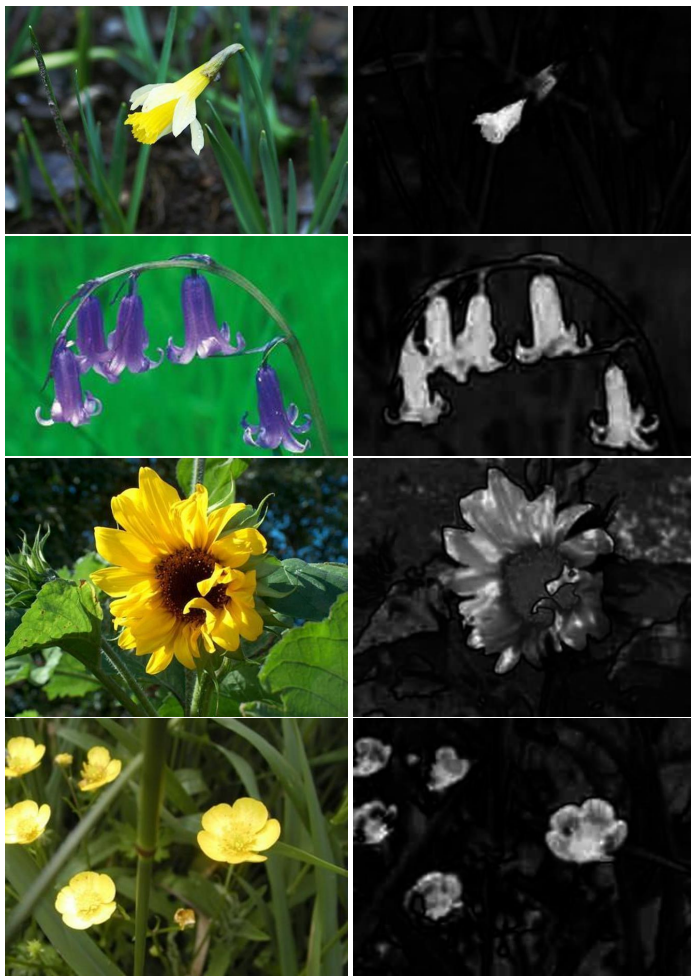


Figure 7: Saliency maps from Oxford flowers. The left column shows original images; the right column shows corresponding saliency maps (where strong filter responses are indicated by high intensity values). The saliency map selects the pixels of the dominant object in an image.