1) Software Used:
Python version 2.7

Libraries used:
opencv 2.4.6
scikit-learn 0.14.1
scipy 0.11.0
numpy 1.8.0

2) Feature Extraction Process:

SURF (Speeded UP Robust Features 128-dimension) descriptors are extracted from each image in the training set. The extracted SURF descriptors are clustered using mini batch K Means algorithm with K = 32. After clustering, Vector of Linearly Aggregated Descriptors(VLAD) are computed based on the assignment of descriptors to cluster centers for each image. Spatial pyramids of level one are used for forming VLAD at finer levels of the image and then concatenated. Each VLAD has dimension KD (where K = 32, D = 128). These are then used as features for each image. Same process is used for both training and test images.

3) Similarity/Distance metrics:

Euclidean distance is used as the similarity metric for mini batch K-means clustering.

4) Classifier used:

Different classification algorithms(SVM with rbf-kernel, PolySVC, RandomForest) were used and LinearSVC was found to give the best performance on cross-validation. The final labels are generated using LinearSVC.

5) References:

1. Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2169–2178).
2. H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In *Proc. CVPR*, 2010.
3. Herbert Bay , Tinne Tuytelaars , Luc Van Gool, SURF: speeded up robust features, Proceedings of the 9th European conference on Computer Vision, May 07-13, 2006, Graz, Austria.
4. http://pascallin.ecs.soton.ac.uk/challenges/VOC/

Algorithm Description:

1) Feature Extraction :

Local descriptors from the grayscale version of each image in the training and test set are extracted. We used SURF-128 for local descriptors since computation of SURF is highly scalable. This problem requires the extracted features to be scale, rotation, viewpoint invariant and so SURF or SIFT would be a good choice. We tried experimenting with SIFT-128 too, but SIFT finds more key points in the image than SURF. The performance difference wasn't much and hence resorted to using SURF. SURF descriptors from training images are clustered using Mini Batch K Means(an scalable iterative version of K Means). Each image was divided into four sub-images of equal size. VLAD of dimension KD based on SURF descriptors in each of the sub-images and then aggregated to form VLAD for the entire image. VLAD for entire image is concatenated with VLAD for sub-images by scaling using appropriate weights to form the feature vector for each image. Using spatial information is found to considerably increase classification accuracy for Bag-Of-Visual-Words model. Later, we decided to make use higher-order characteristics in the image. We considered the usage of VLAD or Fisher Vector(based on soft-clustering using GMMs). Fisher Vectors require greater memory space since they include second-order characteristics(variance). Moreover, there were too many key points in training set to cluster using GMMs(computationally expensive). Thus, we resorted to the use of VLAD features. SPM(Spatial pyramid matching) could also be combined with VLAD. The feature vector has dimension 5KD. Mini Batch K Means with K Means ++ for initialization are used for clustering. Mini Batch size was chosen based on heuristics relating to number of features and iterations. Number of clusters was chosen based on performance in cross validation.

2) Training Algorithm:

Since previous results show good performance of SVM on known object recognition and image classification challenges, we decided to use SVM. Initially, we used SVM with both linear and non-linear kernels (rbf, poly, histogram_intersection, chi2) on Bag Of Visual Words model. SVM with rbf kernel gave best performance on cross validation with training data. After employing SPM with BoVW on multiple levels, linear SVM showed better precision and recall. The effect of applying SPM with BoVW approximates the use of Histogram Intersection kernel. Thus linear classifiers could be employed afterwards. This is advantageous since linear classifiers require less training time.

Classifier used: LinearSVC (from LibSVM) which uses one vs all classification policy

Input Format: A set of tuples of the form

<Image_Label, feature 1, ….., feature N, Class_Label> (N = 5KD)

Tunable Parameters: C (controls the margin of the classifier and no. of training errors)

Output: The trained classifier

3) Validation and Parameter Tuning:

Validation is done by 10-fold stratified cross validation on the training set. Mean Average Precision and Standard Deviation of the scores are reported. Parameter Tuning was done through Exhaustive Grid Search on the Hyperparameter space. For parameter tuning, the training data is split into training and test and then precision and recall are used as metrics for choosing the best parameter combinations.

C = 4.0(LinearSVC), K = 32, mini-batches = 20000(Mini batch K means)

4) Prediction Algorithm:

The trained LinearSVC classifier with parameter C = 4.0 is used for predicting the output labels on the test data.

Improvements:

Fisher Vectors(FV) can be used in place of VLAD with spatial information in the encoding. A greater value of K could be used. (We chose K to be 32 since higher values required more than available memory). PCA with whitening could be done to reduce the dimensionality of the data. PCA with FV has shown good results in past challenges. Features like HOG, color moments could also be used along with SURF. The algorithm works well in practice because of the ability of SVMs to learn from high – dimensional data.