CSE 803 Final Project report

Xinge Ji and Raghunandan Pasula

November 24, 2014

1 Introduction

Automated image classification is a classic problem in computer vision [1][2][3]. It has major applications in content based image retrieval and automatic image annotation. The aim of this project is to automatically classify a given food image into one of the 14 known classes. The food classes used in this report are 14 salad, pasta, hotdog, frenchfry, burger, apple, banana, broccoli, pizza, egg, tomato, rice, strawberry and cookie. Most methods require a model learning process that learns the representation of each class based on a training data. In this work, the training images are collected from ImageNet database [4]. Sample images in the dataset are shown in Figure 1.

The objective is to come up with a classification framework that maximizes the average of true detection rate and true rejection rate. It is also possible for a single image to have multiple classes associated with it. For example, image of a serving tray containing a burger with french fries is categorized into two classes namely *burger* and *frenchfry*.

2 Descriptors

For the given task, we extract three types of features from each image.

- Color Histograms
  we chose the normalized color space described in [5] due to its accuracy in image retrieval tasks. The following equation show how this normalized color space is computed

$$R = \frac{r}{r+g+b}$$

$$G = \frac{g}{r+g+b}$$

  We quantized R and G into 32-bin histograms to generate a 64 dimensional color descriptor.

- PHOW descriptor histograms
  SIFT is a 128 dimensional descriptors. SIFT has been successfully used in a wide variety of image classification tasks. PHOW features are a variant of SIFT descriptors [6]. We extracted these features in a 4×4, 6×6, 8×8 and 10×10 grid in an attempt to remain scale-invariant. Then they are normalized to unit L_1 norm so that they can be compared and clustered in the same feature space.

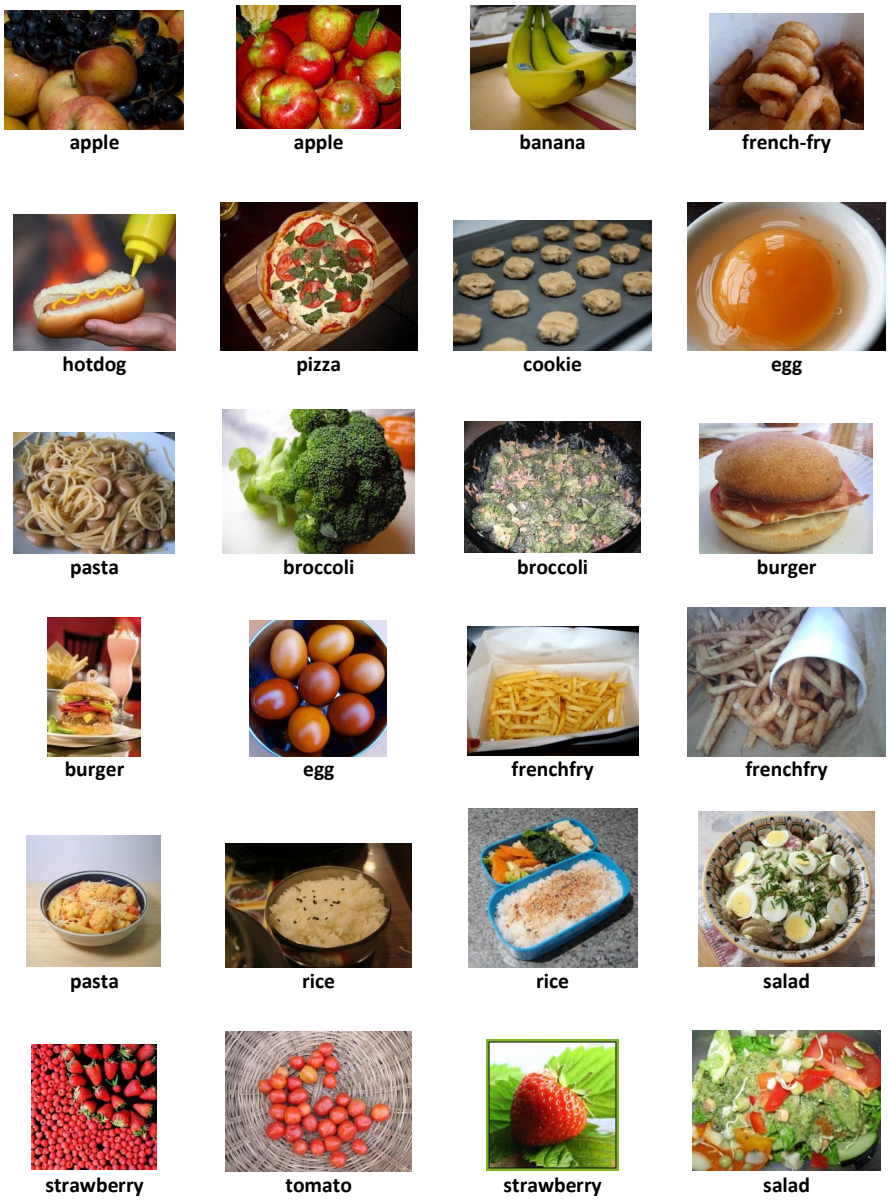

Figure 1: Sample images in ImageNet database

- LBP descriptor histograms
  Uniform LBP [7] is a 59 dimensional histogram where each bin represents a distinct local pattern. We extracted LBP in a 5×5, 10×10 and 15×15 grid.

Since the SIFT and LBP descriptors are computed over multiple grid windows, the size of the descriptors is very large. The descriptors in the training

set are clustered to result in a set of objects labeled as the *vocabulary*. Each descriptor is then projected onto the objects in the vocabulary to result in a compact *histogram* feature. Figure shows the generation of these vocabulary sets for SIFT and LBP features.

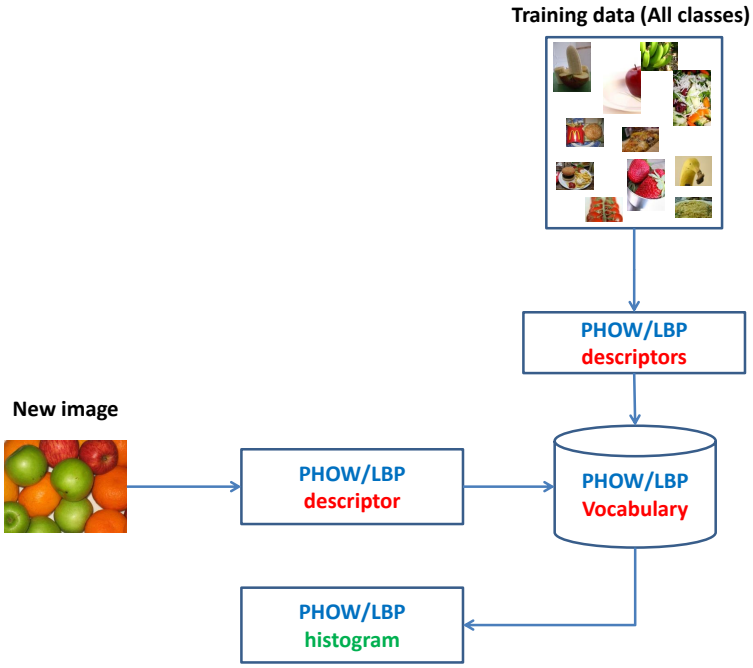

Figure 2: PHOW/LBP histogram generation via vocabularies. This helps in dimension reduction of the original descriptors.

3 Model generation

As mentioned in the previous section. Each image has 3 sets of histogram features namely Color, PHOW and LBP histograms. A SVM classifier is built for each class by training it on the entire training dataset. Images belonging to that specific class are labeled +ve instance while all the other images are labeled as -ve instances. There are two such SVM classifiers built, one with RBF (Radial Basis Function) kernel and Intersection kernel. Figure 3 shows the models built for each class.

From the above figure, we have 6 SVM models are generated (combination of 3 descriptors and 2 kernels). Hence for any given image, we can have 6 decisions for each class. For example, a given image will have 6 decision outputs corresponding to six models for apple class. It might read as [-1,1,1,1,-1,1]. However we need to come up with a single decision for the input image for it to be classified as apple/non-apple. Hence, an ensemble classifier is built by training it on the entire dataset. See Figure 4.

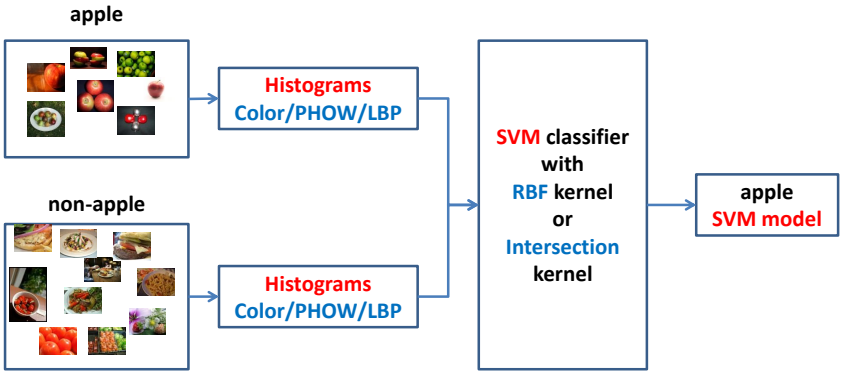

Figure 3: SVM model generation for each class

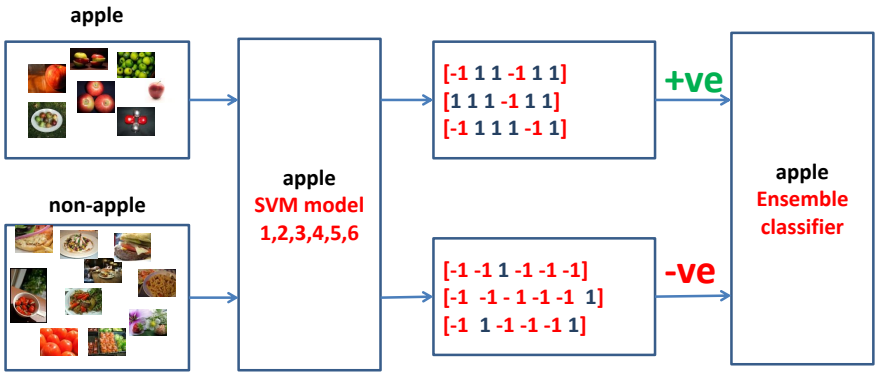

Figure 4: Six individual SVM models are combined to form an ensemble classifier that results in a single decision.

4 Classification

Test image is classified into one or more classes based on the framework shown in Figure 5.

For each test image, three features are extracted and fed into two SVM classifiers each descriptor. Overall, six decisions are obtained from six SVM models. The decisions from the all these classifiers is fed into an ensemble classifier to result in a single decision output.

5 Notes on Classifiers and Parameters

We experimented with several types of classifiers and found RBF-SVMs and Intersection-SVMs outperformed all other classifiers. Therefore, we use them as

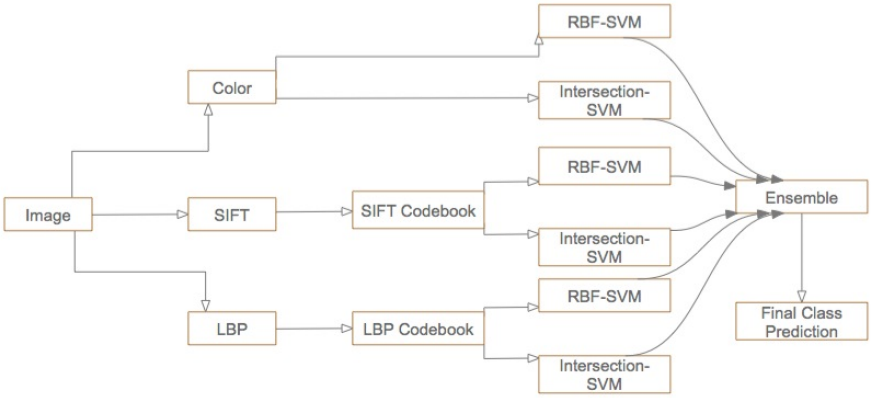

Figure 5: Steps involved in classification of a test image

the base classifiers for this task. Although each SVM classifier may be sufficient, decisions from multiple SVMs could be fused together in various ways to come up with a single decision.

Effectiveness of SVM classifiers depends on the selection parameters. There are two different parameters need to be considered, one is the soft margin parameter C and the other is the kernel width G. To obtain the optimum set of parameters, we did a 5-fold cross-validation grid search over the 2-dimentional space of C and G.

There are six SVM classifiers that are trained on the three different features. We trained a 300 weak-learners over the prediction outputs of the classifiers and yield the final prediction.

6 Results and Discussion

The performance is evaluated as the average of true positive and true negative detection rates. Table 1 shows these values for all the classes in the test dataset.

However, the average accuracy does not present a clear picture. The true positive detection is almost always much lower than the true negative rate. There are three main reasons for this behaviour.

- Intra-class variation is too large in the training dataset
- The choice of feature vectors play a major role in the performance
- Most importantly, we have not performed ***segmentation*** in this work. There are many instances where the background occupies major portion of the image compared to our interest object.

7 Conclusions

We have implemented an automated image classification system that uses color histograms, PHOW and LBP descriptors in combination with SVM and ensem-

Table 1: Performance of the proposed method

Class	Classification accuracy
apple	0.5526
banana	0.6688
broccoli	0.9358
burger	0.6715
cookie	0.4938
egg	0.7002
frenchfry	0.4979
hotdog	0.7797
pasta	0.6681
pizza	0.6023
rice	0.7572
salad	0.6940
strawberry	0.7750
tomato	0.6881

ble classifiers.

References

[1] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, (6):610–621, 1973.

[2] Ruud M Bolle, Jonathan H Connell, Norman Haas, Rakesh Mohan, and Gabriel Taubin. Veggievision: A produce recognition system. In *3rd IEEE Workshop on Applications of Computer Vision*, pages 244–251, 1996.

[3] Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126. ACM, 2003.

[4] http://image-net.org/.

[5] K. E A Van de Sande, T. Gevers, and C. G M Snoek. Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1582–1596, Sept. 2010.

[6] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Image classification using random forests and ferns. 2007.

[7] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.