People Detection with DSIFT Algorithm

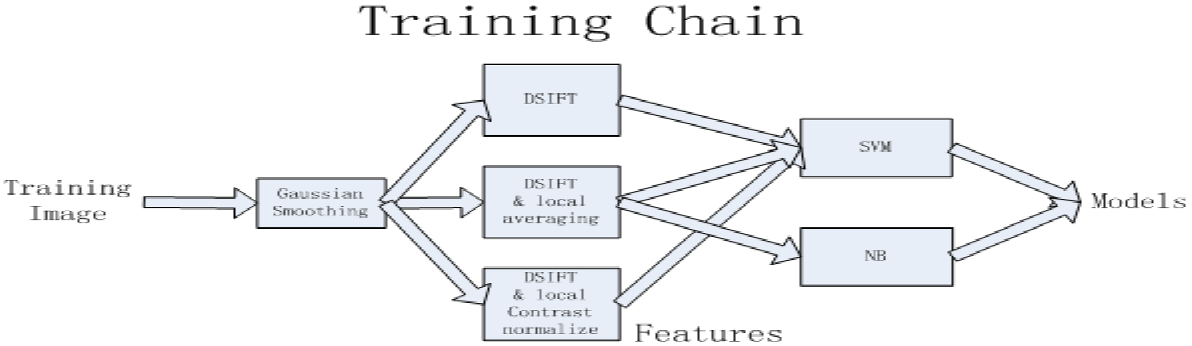
By Bing Han, Dingyi Li and Jia Ji

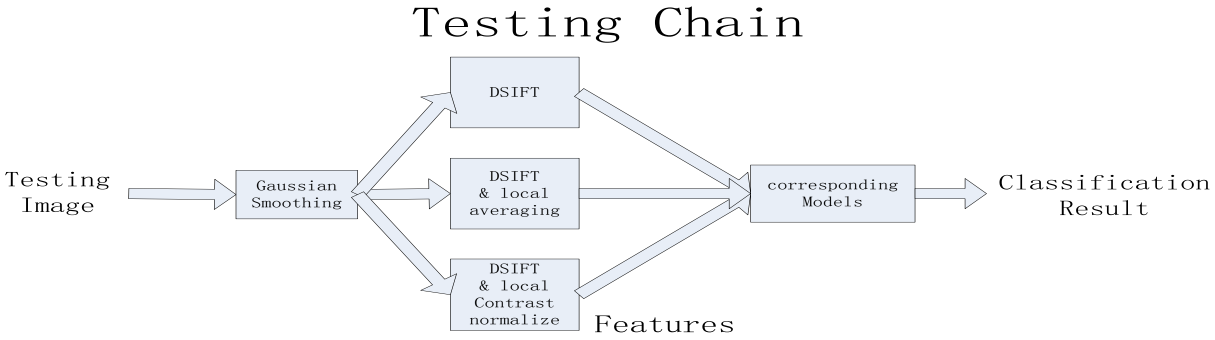
1 Introduction

People detection is an interesting computer vision topic. Locating people in images and videos have many potential applications, such as human computer interaction and auto-focus cameras. There have been much effort on developing people detection algorithms, and among them, the "Dense Scale-Invariant Feature Transform (DSIFT) [1]" algorithms appear to be very effective. To implement our own people detection program, we utilized the DSIFT algorithm to extract feature vectors from images, and used Support Vector Machine (SVM) and Naive Bayes algorithms for classification. We explored different imaging processing and machine learning techniques, and assessed their performance.

2 Algorithms

In this project, we used DSIFT algorithm for feature extraction, local averaging and local contrast normalization for feature modification, and SVM and Naïve Bayes algorithm model training and classification.





DSIFT derives from SIFT algorithm, which is an important keypoint based approach. Given an image, SIFT finds all the keypoints in the image with respect to the gradient feature of each pixel.

Every keypoint contains the information of its location, local scale and orientation. Then, based on each keypoint, SIFT computes a local image descriptor which shows the gradient feature in the local region around the keypoint. Combining all the local descriptors, we get the complete features from the image.

Based on SIFT algorithm, dense SIFT makes some new assumptions: (a) the location of each keypoint is not from the gradient feature of the pixel, but from a predesigned location; (b) the scale of each keypoint is all the same which is also predesigned; (c) the orientation of each keypoint is always zero. With this assumptions, DSIFT can acquire more feature in less time than SIFT does.

Based on the basic DSIFT algorithm, we also explored local averaging and local contrast normalization techniques. In the local averaging technique, we averaged every 4 x 4 block of local pixel's descriptor, and got one averaged descriptor to represent the entire block. Local contrast normalization normalized the descriptors locally in a 4 x 4 block to make the descriptor partially invariant to change in illumination.

After extracting the features in training images, we used SVM and Naïve Bayes algorithms to learn several models. For all the testing images, we extract their features with the same method. Then putting these features into the corresponding models and making classifications, we acquired the recognition result/detection rate.

3 Experiment

3.1 Data Selection

Two set of datasets were used in our experiment. First, MIT's pedestrian database, which contains 924 pedestrian images, was used as our positive samples [3]. 899 random non-human images from the Internet were selected as our negative samples. Second, a more challenging dataset called INRIA were used [4]. INRIA consists of over 2000 images that can possibly contain multiple people or partial person with various backgrounds. The size of all samples is 64 x 128 pixels. Figure 2 is a snapshot of the images.

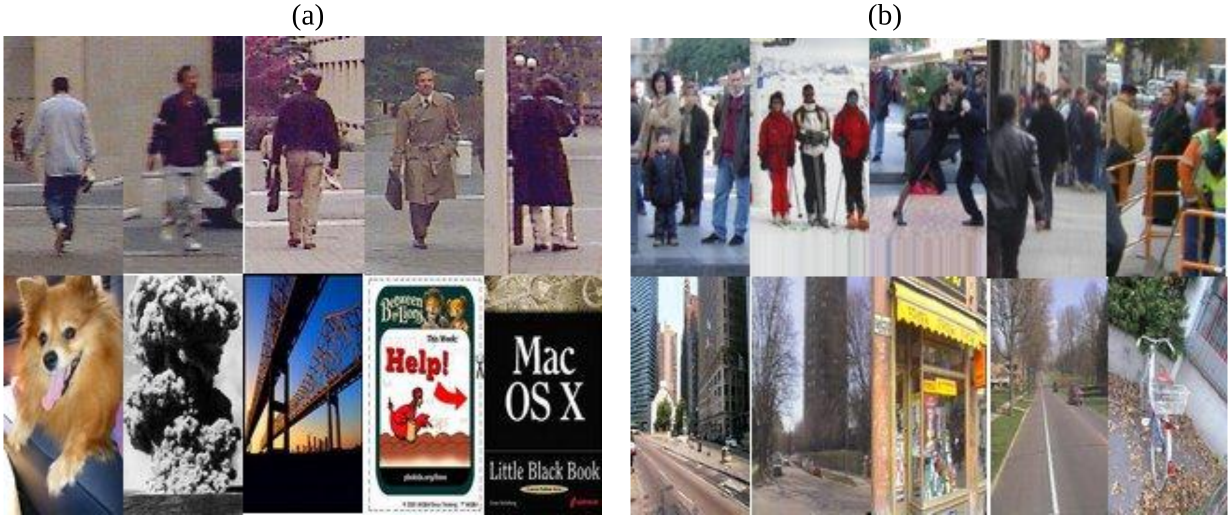


Figure 2 (a) Top: MIT's pedestrian images; Bottom: random images from Internet.
(b) Top: INRIA's positive images; Bottom: INRIA's negative images

3.2 Implementation

We generated a 1 x 772096 feature descriptor vector for each sample image with the Dense SIFT algorithm, which is provided by the VLFeat [5] open source library. To explore the performance, we tuned our algorithms with two image processing techniques separately, which are local contrast normalization and local averaging. Local contrast normalization normalizes the descriptors locally in a 4 x 4 block, which theoretically will robustify our algorithm over different background illumination. The local averaging technique averages every 4 x 4 block of local pixel's descriptor, and gets 1 averaged descriptor to represent the entire block. This way, the descriptors become less noisy and more compressed. As a machine learning project, we also compared different training algorithms, which are SVM and Naïve Bayes algorithms.

We assessed and compared the performance of different image processing and machine learning techniques with 4 criteria, which are classification accuracy, detection rate, false alarm rate, and computation time.

4 Performance Analysis

Figure 3 shows the comparison of accuracy, detection rate, false alarm rate and computation time over the basic DSIFT and the two image processing techniques we discussed above: DSIFT with local averaging and DSIFT with local contrast normalization. Here, all the three sets of data were got using the SVM algorithm for the machine learning part. From the figure, we can find that the local averaging technique gives the best performance in terms of accuracy, detection rate, and computation time, but as a tradeoff, it has the highest false alarm rate. The computation time of basic DSIFT and the local contrast normalization algorithms increase dramatically after size of training sample reaches 180, due to large requirement of computer memory.

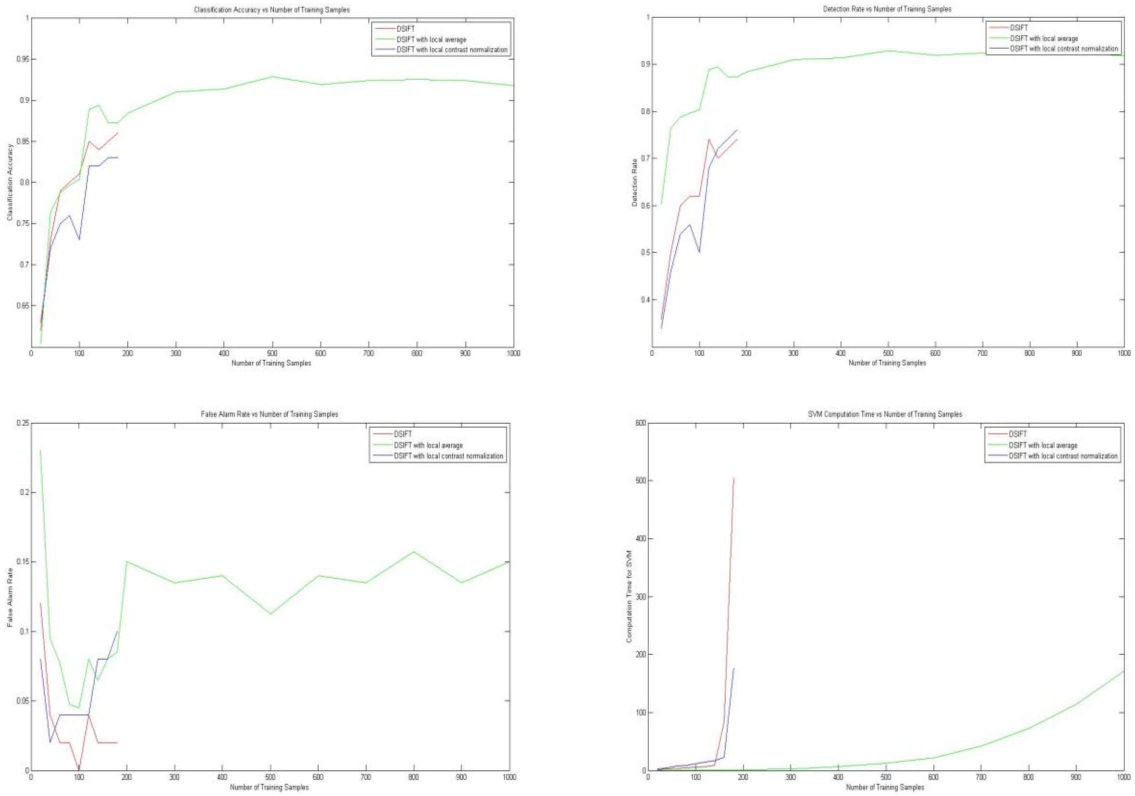


Figure 3. Accuracy, Detection Rate, False Alarm Rate, Computation Time Comparison over Three Image Processing Techniques

Figure 4 shows the comparison of accuracy, detection rate, false alarm rate and computation time over SVM and Na ïve Bayes algorithms. For the image processing part, we used the DSIFT algorithm with local averaging technique for both sets of experiments. Comparing the performance of human detection with SVM and Na ïve Bayes algorithms, we discovered that SVM out-performs Na ïve in terms of accuracy, detection rate and false alarm rate. But the computation time for SVM increases quadratically as training sample size increases, while the time for Na ïve Bayes only increase linearly.

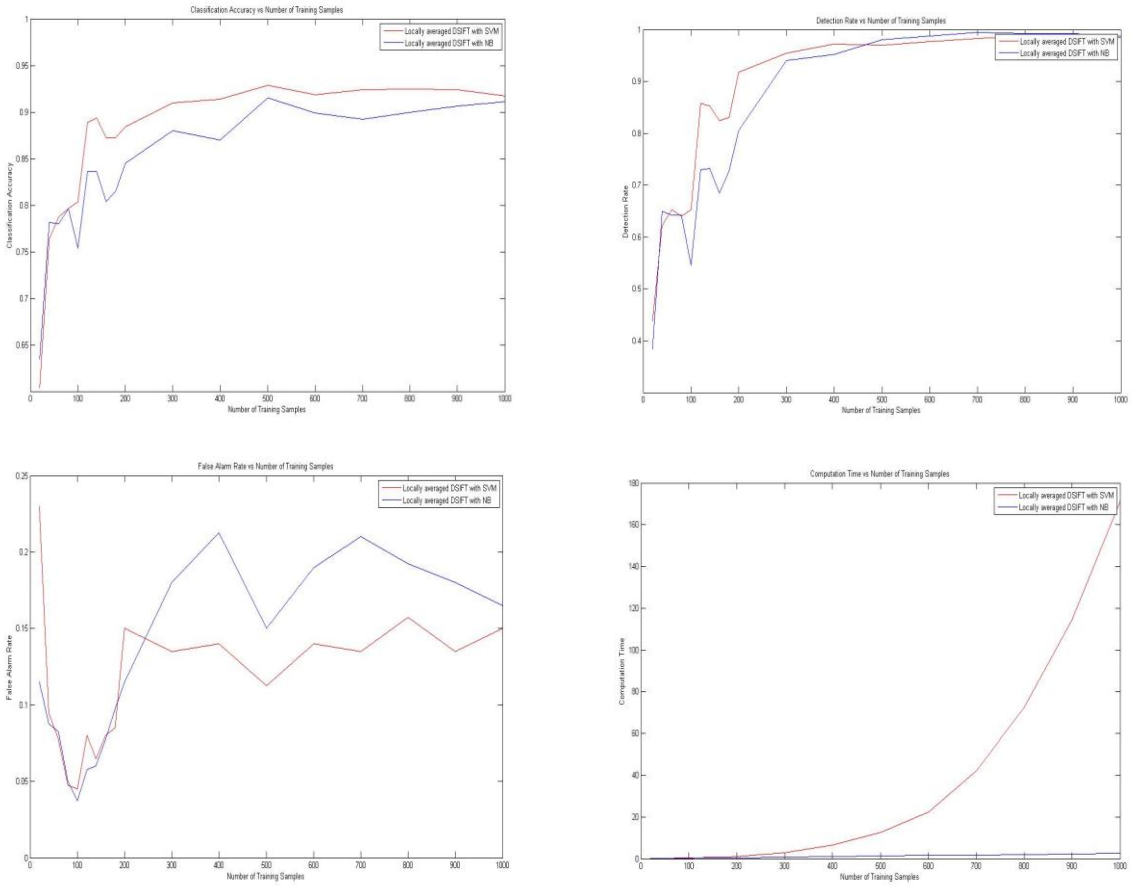


Figure 4. Accuracy, Detection Rate, False Alarm Rate, Computation Time Comparison over SVM and Na ïve Bayes Algorithms

Figure 5 shows the accuracy comparison over the MIT pedestrian dataset and INRIA dataset. We can observe that the accuracy of classification with MIT dataset remains very high even with very small training size and it can reach 99.12%. For the more challenging dataset, we can get higher accuracy with increasing training size, which finally reaches 92.87%.

From the data measured, we can find that using the DSIFT algorithm with local averaging technique and SVM algorithm, the accuracy of human classification accuracy can reach 99.12%; the detection rate can reach 99.75%; the false alarm rate can maintain below 2.75% for the MIT Pedestrian dataset. On the other hand, for the INRIA dataset, the accuracy of classification can reach 92.87%; the detection rate can reach 98.75%; the false alarm rate can maintain below 15.75%.

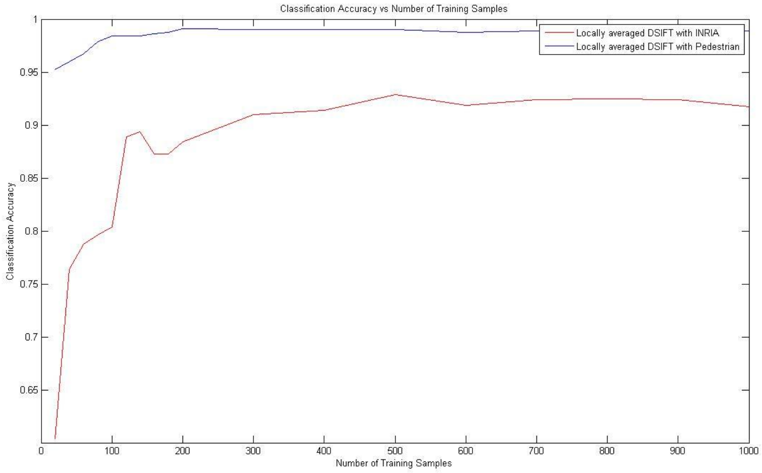


Figure 5 Accuracy Comparison over INRIA and MIT Pedestrian Dataset

5 Conclusion

In this project, we implemented the people detection program with the Dense SIFT algorithm. Different image processing and machine learning techniques were explored and assessed, which are local contrast normalization vs. local averaging, and SVM vs. Naive Bayes. We conclude that the DSIFT with local averaging and SVM algorithm yields the best classification accuracy, which is 99.12% for MIT dataset and 92.87% for INRIA dataset. However, if computation time is of concern, the local averaging with Naïve Bayes algorithm can also give reasonably good accuracy with much less computation time.

6 Acknowledgment

The authors wish to thank Professor Andrew Ng, Will Zou and the teaching assistants of cs229 for their feedback and guidance.

Reference

[1] http://www.vlfeat.org/overview/dsift.html
[2] D. G. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60(2):91–110, 2004.
[3] http://cbcl.mit.edu/software-datasets/PedestrianData.html
[4] http://pascal.inrialpes.fr/data/human/
[5] http://www.vlfeat.org/