CS771 : Final Project Report

Background Foreground Separation and Object Classification

Instructor: Professor Harish Karnick

Group Information :

- Group Number : 6
- Group Members :
    - Avani Samadariya
    - Rahul Kumar Wadbude
    - Ritika Mulagalapalli
    - Shubham Agrawal

Problem Statement :

Classify the objects in a video in four categories : Pedestrians, Two wheelers, Three Wheelers and Four Wheelers. Predict the labels on an input video. We are provided with labelled videos. Thus, we have at-least 100 images in each category.

Problem break down :

1. Choosing Feature Representation
2. Choosing Classifier
3. Final Prediction

Feature Extraction:

Image are stored in the form of matrix. Consider a 50 X 50 sized image. It has 2500 pixels, with each pixel having three values, resulting in 7500 data points. Thus, if used all this information is used directly, size of feature vector will be very large. It has been proved in vision, that there exists some latent attributes for each image, which contains most information about image. Finding proper feature representation was important as whole algorithm depends on data.

We tried two different feature representation :

1. **Deepnet Features** : Used Caffe framework with GPU hardware acceleration for feature extraction. A pretrained **BVLC GoogLeNet** model[1] is used. This model uses deep convolutional neural network architecture to predict labels on images. It was responsible for setting the new state of the art for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC 2014).

2. **SURF Features** : Apart from Caffe features, we also considered SURF features[2] and compared them with deepnet features on Caltech dataset (since only a few data videos were available at early stage of the project) with the same classification algorithm. On comparing the results, it was observed that the deepnet (Caffe) features clearly outperformed the SURF features. We continued with the deepnet features for the rest of the project.

The following are the results obtained for both the features using one-vs-rest svm [3] for comparison:

1. SURF features with tuned parameters : Kernel = Sigmoid, C = 10000; gave an accuracy of **59.94%.**
2. Deepnet features with tuned parameters : Kernel = Polynomial, C = 1; gave an accuracy of **94.4297%**.

Clearly, Deepnet features outperformed SURF features and hence used for rest of the project.

Classification Model:

Once feature representation was decided to be the deepnet features, classification model has to be chosen for prediction of labels. One vs Rest SVM, Decision Trees and Random Forests are known to perform good on multi label classification. There are two main methods for tackling the multi-label classification problem:

---

[1] Szegedy, Christian et al. "Going deeper with convolutions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2015: 1-9.

[2] "Speeded up robust features - Wikipedia, the free encyclopedia." 2015. 15 Apr. 2016 <https://en.wikipedia.org/wiki/Speeded_up_robust_features>

[3] Ramage, Daniel et al. "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* 6 Aug. 2009: 248-256.

problem transformation methods and algorithm adaptation methods. Problem transformation methods transform the multi-label problem into a set of binary classification problems, which can then be handled using single-class classifiers. Example : One vs Rest SVM Algorithm. Adaptation methods adapt the algorithms to directly perform multi-label classification. In other words, rather than trying to convert the problem to a simpler problem, they try to address the problem in its full form[4]. Example : decision trees and random forests.

As Deepnet features gave results with high accuracy, we also attempted *fine classification* into the sets, person, bicycle, motorcycle, rickshaw, autorickshaw, car apart from the coarse classification into pedestrian and 2/3/4 wheelers, which was our original problem statement. Higher accuracies were obtained for coarse classification, so we tested our classification models for coarse classification

Dataset L :

Data is collected from cctv cameras installed at the entrance gates of I.I.T. Kanpur. Crude Data was in the form of videos. From each video, individual objects were located and labelled frame by frame. Throughout this project, we used following dataset with train-test ratio equal to 1:3

S. No.	Objects	Number of Images
1.	Pedestrian	1323
2.	Two Wheeler (Bicycle, Motorcycle)	906
3.	Three Wheeler (Auto Rickshaw, Rickshaw)	81
4.	Four Wheeler (Car)	134

[4] "Multi-label classification - Wikipedia, the free encyclopedia." 2011. 15 Apr. 2016 <https://en.wikipedia.org/wiki/Multi-label_classification>

Decision Tree Classifier:

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility[5]. Multi label classification is inherent to decision tree and hence we decided to consider this for comparision.
Decision Tree was trained on Gini and Entropy impurity functions[6] were used. Following results were obtained :

S No.	Impurity Function	Leaf node	Accuracy
1	Gini	1	84.9918
2	Entropy	1	85.155
3	Gini	3	84.9918
4	Entropy	3	85.155
5	Gini	5	86.7863
6	Entropy	5	86.2969

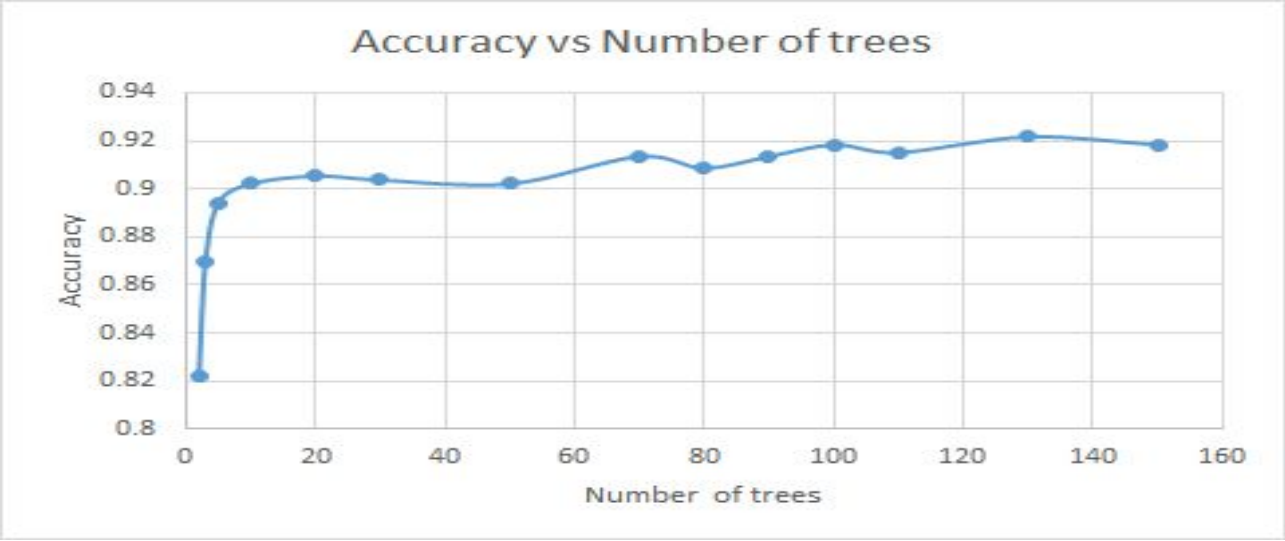
Random forest:

Random forests is a notion of the general technique of random decision forest that are an ensemble learning method for classification,regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification). Random decision forests correct for decision trees' habit of overfitting to their training set[7].

---

[5] "Decision tree - Wikipedia, the free encyclopedia." 2011. 15 Apr. 2016 <https://en.wikipedia.org/wiki/Decision_tree>
[6] Breiman, Leo. "Technical note: Some properties of splitting criteria." *Machine Learning* 24.1 (1996): 41-47.
[7] "Random forest - Wikipedia, the free encyclopedia." 2011. 15 Apr. 2016 <https://en.wikipedia.org/wiki/Random_forest>

We used random forests with maximum depth of tree 5 to ensure weak learners and gini impurity function. Following graph was obtained on varying the number of trees in ensemble :



One can see that around 130 trees accuracy reaches maximum value of 0.92 percent and then saturates. As expected, accuracy saturates after a particular number of trees.

One vs Rest Support Vector Machine :

Support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.[8]
The *one-vs.-rest* strategy involves training a single classifier per class, with the samples of that class as positive samples and all other samples as negatives. This strategy requires the base classifiers to produce a real-valued confidence score for its decision, rather than just a class label ; discrete class labels alone can lead to

[8] "Support vector machine - Wikipedia, the free encyclopedia." 2011. 15 Apr. 2016 <https://en.wikipedia.org/wiki/Support_vector_machine>

ambiguities, where multiple classes are predicted for a single sample.[9] The final label is given by comparing the confidence score given by classifiers.

We tuned svm algorithm for different C's and kernel functions to obtain the best result. Best accuracy 97.063 was obtained at C = 10 and polynomial kernel

Kernal	C = 0.1	C = 1	C = 10	C = 100	C = 1000	C = 10000
Linear	95.758	95.758	95.758	95.758	95.758	95.758
Polynomial	94.943	96.737	**97.063**	97.063	97.063	97.063
RBF	92.659	96.085	96.737	96.737	96.737	96.737

Amongst the three classifiers, SVM performed best and hence used as final classifiers.

Prediction on video input :

One main challenge of this project was to get images out of video for classification. At this point, we have a classifier to whom an image was given as an input and it predicts its label. We were supposed to somehow locate the object in the video and feed it to classifier. For this, Background subtraction was used. Background subtraction is a technique in the fields of image processing and computer vision wherein an image's foreground is extracted for further processing (object recognition etc.). The rationale in the approach is that of detecting the moving objects from the difference between the current frame and a reference frame[10].

Two algorithms from opencv library were tried : MOG and MOG2. MOG2 had better performance among these two. For further refinement we applied various filters like blurring, hole-filling and noise removal.

---

[9] "Multiclass classification - Wikipedia, the free encyclopedia." 2011. 15 Apr. 2016 <https://en.wikipedia.org/wiki/Multiclass_classification>

[10] "Background subtraction - Wikipedia, the free encyclopedia." 2013. 15 Apr. 2016 <https://en.wikipedia.org/wiki/Background_subtraction>

Major Challenges faced:

1. In labelled data, especially auto rickshaw, labels were given incorrectly, which we have to manually correct.
2. Labelled data of two wheeler contain only vehicle . Background subtraction algorithm works on the basis of moving objects. It was unable to separate two wheeler and person sitting on it. Hence, classifier was confused between bicycle and person when a person sitting on bicycle is given to it.
3. Shadow of Trees and objects made background subtraction difficult.
4. Naive implementation of BGS(MOG2 without filters) clubs objects that are close in space.
5. Occluded objects were harder to classify.

Future Extensions:

1. As mentioned in point 2 above, data was labelled by cropping person sitting on it.

    a. Data can be relabelled
    b. It is more intuitive to recognize bicycle and rider separately. This can be solved by sliding window detection technique, where a window of fixed size is slided throughout the frame and at each position of window, classifier predicts whether it contains an object or not.
2. Object localization in video for finding contours
3. We noticed that when object was big in size, classifier predicted label correctly. We can use this in following way :
    a. Trace path of object in video
    b. Predict all labels along the path
    c. For predicting label of object along the path, Increase weightage of that part of path where object size was bigger