CS771A Project
Report

Object Detection and Classification in Surveillance Video

Submitted in partial fulfillment of
the requirements for the course CS771A

Submitted by
Group 34

13508	Preetansh Goyal
13767	Vandana Gautam
13680	Shubham Gupta
13788	Vikas Jain

Under the guidance of
Prof Harish Karnick

Department of Computer Science and Engineering
Indian Institute of Technology Kanpur
Kanpur, U.P. , India – 208 016

April 15, 2016

Abstract

The design of a Video Surveillance system is directed on automatic identification of events of interest. A video surveillance system generally an integration of three phases of data processing which is foreground extraction, object recognition and decision about the action. In our project we have focused on the former two aspects. The main aim of our project was to classify different object to classes pedestrian, vehicles, etc. To achieve this we have used background subtraction methods like Mixture of Gaussians, Optical flow, etc. along with classifiers like Support Vector Machines, Decision Trees, Random forest, etc. To enhance the classification process we have looked upon various feature representation of images like HOG, BOVW and CNN.

Acknowledgments

Every year CS771A - Machine Learning - Tools, Techniques and Applications Course is offered in IITK. The course involves a course project to be completed in team of four students under the guidance of the course instructor. The course project gives the opportunity to the students to have experience in latest Machine Learning techniques and its fields. It would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them.

The authors would like to express their gratitude towards **Prof Harish Karnick** for his kind co-operation and encouragement which help us in completion of this project.

The authors would like to express their special gratitude and thanks to course TAs for giving us such attention and time.

Preetansh Goyal
Shubham Gupta
Vandana Gautam
Vikas Jain

April 2016
Indian Institute of Technology

Contents

1 Introduction

In today's world, CCTV footages and other surveillance data prove to be a major witness in resolving any crime or mis-happening. With a huge number of CCTV cameras installed all over and a huge amount of data being generated by them, it becomes essential to interpret and make sense of that data. Our project aims at detection and classification of objects in CCTV camera videos into various categories like bicycle, car, person etc through various machine learning techniques. In this project, we have experimented with several classifiers with varied metrics on various feature extraction methods to generate the best results on the available data. Techniques like Background Subtraction and Optical flow have been implemented for detection of moving frames in the test videos.

2 Motivation

Object Detection and Classification in surveillance videos have various applications in real life. The system can be used for security purposes at public places. It can also be used for tracking purposes of objects. Fine grained classification and recognition will help in real-time person identification and number plate recognition of vehicles.

3 Our Model Description and Work

The project mainly consisted of three parts – Feature Extraction, Classifier Learning and Object Localization.

3.1 Feature Extraction

To generate the feature vector for the images, we used the following methods.

Bag of Visual Words[1]
We computed approximately 300 SIFT points per image in the dataset. Each SIFT point is represented by a 128 dimensional descriptor. Using K-means mini-batch clustering algorithm, 700 clusters of sift points are created. Then, the SIFT descriptors of each image are used to generate a vector of dimension 700 which becomes the feature vector of that image.

Histogram of Gradients(HOG)[2]
For obtaining the HOG feature vectors, each image is re-sized to 128 x 128. This generated HOG feature vectors of 15678 dimension for each image.

Convolution Neural Network(CNN) features
The VGG16 [3] model for extracting features from the images. We used learned model of VGG16, trained on ImageNet dataset[6].
The last layer(softmax) from the network is removed and the output of layer before it(FC-4096) is used as feature vector of each image. The dimension of each feature vector is 4096.

3.2 Classification

After obtaining the feature vectors of images in the dataset using the above methods, we tried with various classifiers for testing and training purposes. Some of them are[5]:
Support Vector Machines (SVM) with Linear Kernel, SVM with Gaussian kernel, Decision tree and Random Forest Classifier with approx 200 estimators (decision trees). We tried with varied number of estimators but the result seemed to remain fairly the same on increasing the number of estimators beyond 200.

3.3 Object Localization

We primarily used three techniques to extract objects of interest from the test videos.

Background Subtraction
This technique calculates the foreground mask performing a subtraction between the current frame and a background model, containing the static part of the scene[5]. We used the Mixture of Gaussians (MOG) background subtractor to generate the foreground mask for our purpose.

Optical Flow Measurement
Optical flow is the pattern of apparent motion of image objects between two consecutive frames caused by the movement of object or camera[5]. It is 2D vector field where each vector is a displacement vector showing the movement of points from first frame to second.

Sliding Window Method
A sliding window is rectangular region of fixed width and height that slides across each frame.
For each of these windows, we take the window region and apply an image classifier to determine if the window has an object that interests us.

After extracting the required images from the test video frame we extract the features and used our trained classifier to predict the label of the image in the video frame.

4 Dataset Used

The labelled dataset[7] of surveillance videos of IITK gate are provided. We extracted the labelled bounding boxes for the 6 classes from the video frames. Every 20th frame was used for the images extraction from the frame. The 6 classes used for classification purpose are – Person, Bicycle, Car, Motorcycle, Rickshaw, Autorickshaw.

The above extracted bounding boxes for the 6 classes are used for feature extraction and learning the classifier.

For testing part, we use any video as raw input.

5 Experiment and Results

5.1 Experiment for the best classifier for the features obtained

- A major obstacle before us was the quality of labeled data available. After separation of labeled images from the video input data we found out that many of the images were ill suited for the classification process, thus we manually screened the images from all the images obtained and removed only the wrongly labeled ones.
- We used a test of $\sim$ 1600 training images and $\sim$ 450 test images.
- The training set was well balanced with around 300 images from each class.

5.2 Accuracy results

Accuracy Obtained for variuos combinations of features and classifier algorithms

Classification Accuracy Obtained			
Classifier	BOVW	HOG	CNN
SVM(linear kernel)	0.8535	0.9129	**0.9873**
SVM(Gaussian kernel)	0.879	0.756	0.9809
Random Forest	0.8322	0.9023	0.9745
Decision Tree	0.6475	0.6072	0.8917

The CNN feature results outperformed other features on all the classifiers used and also linear kernel SVM performed best across almost all features. Thus we used linear kernel SVM with CNN features for the classification process.

5.3 Confusion Matrix for SVM linear kernel

5.3.1 with CNN features

Car	Auto rickshaw	Bicycle	Motorcycle	Person	Rickshaw
51	0	0	0	0	0
1	99	0	0	0	0
0	0	99	0	0	1
0	0	0	99	1	0
0	0	2	0	98	0
0	0	1	0	0	19

5.3.2 with HOG features

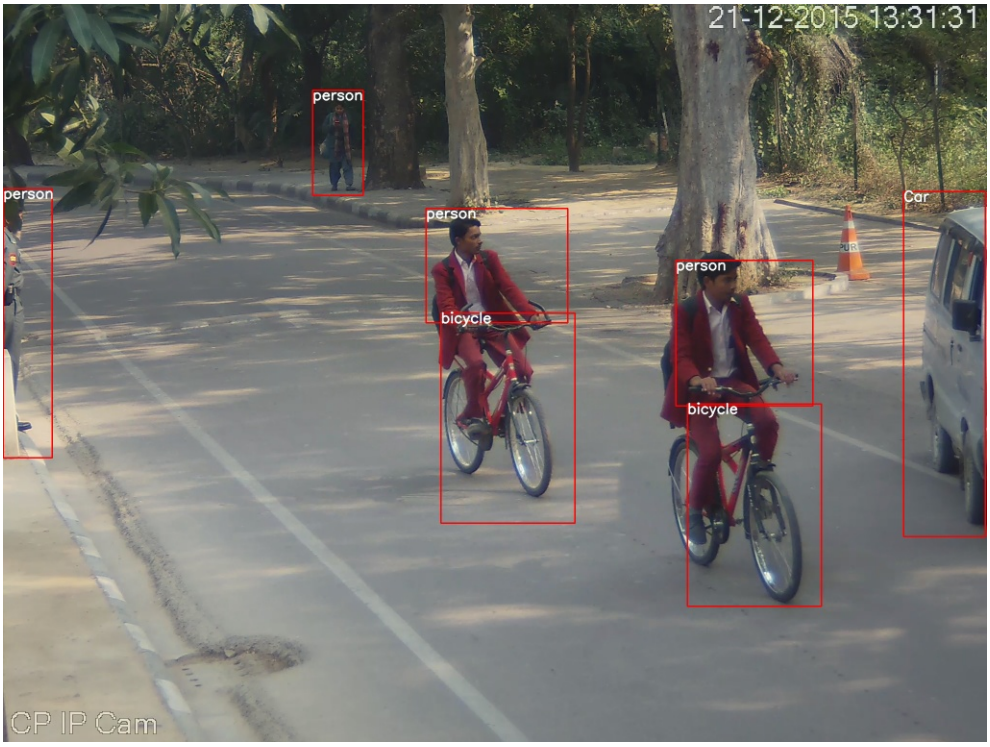
Car	Auto rickshaw	Bicycle	Motorcycle	Person	Rickshaw
49	2	0	0	0	0
1	91	4	2	0	2
1	1	92	4	1	1
1	1	8	86	3	1
0	0	2	3	95	0
2	1	0	0	0	17

5.3.3 with BOVW (SIFT) features

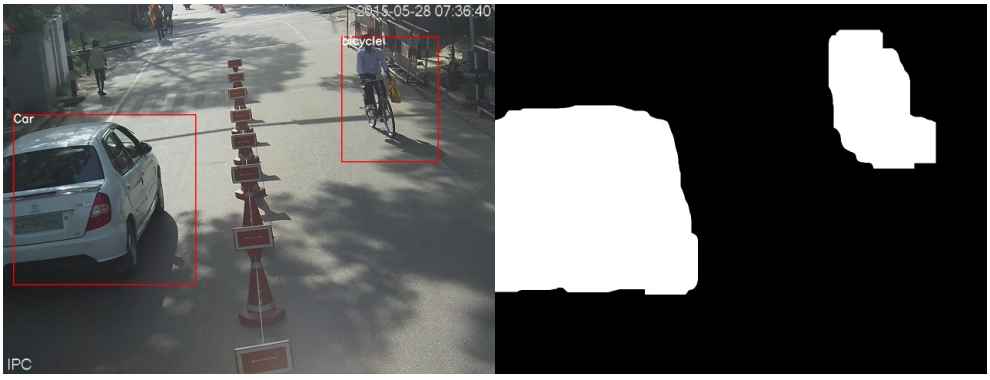
Car	Auto rickshaw	Bicycle	Motorcycle	Person	Rickshaw
49	1	1	0	0	0
1	90	3	2	1	3
1	2	78	10	3	6
1	4	8	82	5	0
1	1	4	4	87	3
0	2	0	1	1	16

5.4 Using classifier on video data

- Using video file with object bounding boxes present as input we first extracted the CNN features from the labeled frames and subsequently classified them using linear kernel SVM. The results obtained were excellent as the classifier performed very well.



- Using raw video file as input we first used our object localization to obtaining the images of interest. Then we extracted the CNN features for the images obtained through localization and subsequently classified them using linear kernel SVM. he results suffered slightly due to relatively weak localization output but still the results were very promising as over the duration of the video objects were correctly identified and labeled.



6 Conclusion

Through this project, we have been able to reach our objective of classification of object after localization with very good results. The object classifier classifies the object with good accuracy as we have mentioned in the reports. We have also classified the object images obtained from localization into classes like Pedestrian, Rickshaw, Auto-Rickshaw, car, etc in real time.
We were able to localize objects based on motion detection with reasonable accuracy. We have also been able to model the background using MOG and perform foreground segmentation in real time.

7 Future Work

7.1 Localization

BGS, Optical flow , sliding window do not work good very well for localization. We can use a more comprehensive learning method for localization.

Also, instead of sliding window method, better methods for finding windows can be used by finding gradient of windows[8].

7.2 CNN

The classifier with CNN features works the best but it requires around 1-2 second per frame for the classification. For real time classification, *Fast R-CNN*[4] can be used.

7.3 Using averaged labels

Currently the classifier considers every image independent of the others for prediction but we can exploit the fact that we are provided a video input and using the continuity of frames we can improve prediction accuracy as we can average out the predicted labels over the video for objects.

7.4 Dataset improvement

The quality of labeled dataset and labels is noisy and with improvements in the dataset results are likely to improve.

References

[1] Object recognition from local scale-invariant features. *Lowe, David G.* Proceedings of the International Conference on Computer Vision. pp. 1150–1157, 1999.

[2] Histogram of oriented gradients for human detection. *Dalal, Navneet and Bill Triggs* Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

[3] Very Deep convolutional Networks for Large-Scale Image Recognition. *K. Simonyan, A. Zisserman.* arXiv.1409.1556

[4] Fast R-CNN *Ross B.Girshick.* Proceedings of the IEEE International Conference on Computer Vision. 2015.

[5] Libraries Used
OpenCV(i/o, bgs, optical flow, sift),
sklearn(k-means, SVC, random forest, decision tree, etc classifiers),
sklearn-image(hog)

[6] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[7] "IIT Kanpur Surveillance Video Dataset" Computer Science Department IIT Kanpur, Class of CS771 Winter 2016 (list)
Tool Used : Video Annotation Tool from Irvine, California (VATIC)
MetaData : Annotation MetaData Table

[8] Lampert, Christoph H., Matthew B. Blaschko, and Thomas Hofmann. "Beyond sliding windows: Object localization by efficient subwindow search." Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008.