Classification of Pedestrian Using Convoluted Neural Network

Rohini. A. Chavan, Sachin. R. Gengaje, Shilpa. P. Gaikwad

***Abstract*: We present our work based on classification of pedestrians into a single person and group of people using Convoluted Neural Network (CNN). Major work was done on classification-based feature extraction techniques before CNN is applied to it. CNN can classify objects without extracting the features. Here, we have set up a complete channel for pedestrian detection using sliding window approach and classification using a CNN network. Alex Net and ResNet are the two architectures used in CNN for implementing the classification algorithm. Performance is evaluated on the PET and Caltech dataset which consists of a number of people who are walking with a group or separately in the scene. We got the optimistic results in case of small dataset used for testing. We have also tested our algorithm over large dataset to verify its performance with the help of performance evaluation metrics.**

***Index Terms*: classification, detection, Convoluted Neural Network, sliding window**.

I. INTRODUCTION

Classification of moving objects into semantic meaningful categories is significant for automatic visual surveillance applications. However, this is a very challenging task because of different object related factors such as limited object size, large intra-class variations of objects in a same class, illumination change and real-time performance prerequisite in real-world applications. The purpose of this kind of research is to build up intelligent systems of video surveillance which is capable of detecting, classifying, tracking and even analyzing the behavior of the people from capture video scene in real time and report the doubtful situations to the authorized person. This smart system completely replace the traditional video surveillance systems which is not efficient when the number of cameras exceeds the number of human operators. In recent years, many researchers have given much attention on classifying objects after motion extraction from background. Object segmentation is the important step before classification of moving objects. Most of the previous approaches in this field often uses shape and motion-based information such as maximum area, dimension, location, compactness, bounding box, velocity, etc.

Revised Manuscript Received on July 05, 2019.

Rohini. A. Chavan, Research Scholar, Electronics Dept., Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, India.

Sachin. R. Gengaje, Electronics Dept., Walchand Institute of Technology, Solapur, India.

Shilpa. P. Gaikwad, Electronics Dept., Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, India.

However, silhouette and location of the objects are always changing drastically under different camera view angles. In addition, shadow is also detected as a foreground object which affect on the accurate classification of objects. In a smart video surveillance system, an object classifier must have the following properties [1]:

1. Random and large deviation in weather conditions including snow, fog , rain etc.
2. Effect of change in illumination of light as artificial and natural for indoor and outdoor environment respectively.
3. The movement of camera because of strong wind gives noisy image at the output.
4. Complete or partially occlusions among objects while moving in the scene [2].
5. Change in colors of pedestrian's clothes.
6. Swaying of tree leaves and wavering of water gives false object detection.
7. Posture variations of humans [3].

This paper presents a robust pedestrian detection and classification system using sliding window based Convoluted Neural Network (CNN) approach [4]. We provide first few frames of video as a training frames to the sliding window approach to convert the single image into multiple sub images at the output called as sub sampling. These set of cropped images are given as a input to CNN network to classify the pedestrians into a single person or group of person. Further, the post-processing technique is applied to compute the bounding boxes around classified objects.

II. RELATED WORK

In previous years, many researchers were proposed feature based and motion-based techniques for object classification.

A. Feature based techniques

In feature-based method, supervised and unsupervised learning techniques are used for classification problem [6]. First time, under the category of feature-based method, Rowley [4] proposed neural network-based approach for face detection. The work done by Viola and Jones [5] was main landmark in pedestrian detection and classification. They have used AdaBoost classifier for selecting the features and classification of moving objects based on these features. Lun Zhang et al. [6] proposed appearance-based method to accomplish robust objects classification in varied camera view angles. He has proposed a novel descriptor such as Multi-block Local Binary Pattern (MB-LBP) to confine the large-scale structures in object appearances. An AdaBoost algorithm is introduced based on MB-LBP features to choose a subset of discriminative features and

build the powerful two-class classifier. At last, the Error Correcting Output Code (ECOC) is initialized to realize robust multi-class classification presentation in diverse scenes [6]. The work of Dalal and Triggs [7] based on Histograms of Oriented Gradients (HOG) for human classification and detection resulted in further enhancements to feature based prediction and detection system. However, most of the researchers applied feature-based detection along with SVM regression methods to execute pedestrian classification [9]. The survey paper by Dollar et al. 2012[2] recapitulate various feature based methods in brief and compare their performances on the Caltech pedestrian dataset [4].

B. *Motion based*

In recent years, major focus of many researchers has been given to the classification of objects by considering their motion. The interested motion targets can be separated from a stationary background reasonably by background subtraction method and reduces the problem of clutter. Normally, human being follow periodic motion and non human moving objects like bicycle, vehicles follow non periodic motion. Based on these properties, humans can be easily separated from vehicles. Repetitive changes in the shape of object gives recurrent motion behavior which is useful to generate the Recurrent Motion Image (RMI) [10]. The areas of RMI express high motion recurrence which is useful to determine the object's class. For example, the RMI of a walking pedestrian has high recurrence near the hands and legs, whereas the RMI of a moving vehicle has not shown motion recurrence. It is an improved motion-based recognition approach with the use of specific feature vector called Recurrent Motion Image (RMI) to classify moving objects into preset categories, namely single person, group of persons, vehicle and four-legged animals etc. The detected motion targets are classified based on their periodic motion patterns captured with the RMI. Out of all the motion based classification methods, RMI is only one of the approach that produces high recognition rate [11]. RMI is a specific feature vector which estimate repetitive motion behavior of moving objects. Thus, moving objects are classified into single person, group of persons or vehicle based on their corresponding RMI given as ,

$$DSa(x, y, t-1) = Sa(x, y, t-1) + Sa(x, y, t) \quad (1)$$

$$RMIa = \sum_{k=o}^{T} DSa(x, y, t-k) \quad (2)$$

RMI is computed with (1) and (2) to calculate the areas of moving object's shape which undergoes repetitive changes. S_a is a binary silhouette for object 'a' at frame 't', and DS_a is a binary image showing areas of motion for object 'a' between frame t and t-1. RMIa is the RMI for object 'a' determined over T frames.

III. PROPOSED NETWORK

To handle pedestrian classification problem, general flow of processing steps is given in figure 1. The proposed implementing steps are region proposal method, a feature extraction and a region classifier.

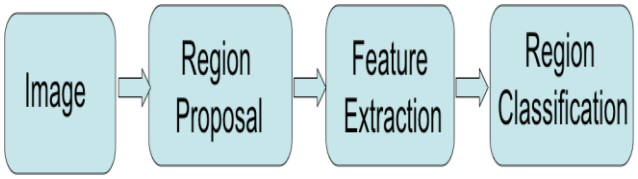


Fig. 1 Block diagram of Proposed Pipeline

A. *Region proposal*

For region proposal, a sliding window approach is implemented using Adaptive Median filter. In this step, one image is converted into multiple sections called as sub images which has different aspect ratios and sizes. This process is known as sub sampling. The training and testing both uses the sliding window approach which runs in the same fashion. These cropped images are further given to CNN network to extract their features.

B. *Feature extractor*

These sub images produced by sliding window, are fed to the feature extractor. The feature extractor can extract the features with the help of Histogram Oriented Gradient (HOG). For this approach the convoluted neural network (CNN) is used as a feature extractor. The cropped image from the median filter sliding window detector is given as a input to the convoluted neural network. During the training phase, the images are preprocessed by reducing their size. The median filter sliding window detector can set the image scale into size 96x48. These sub images are again scaled to essential size of the CNN that is 220 x 220 inputs. The images are then normalized by subtracting the mean and scaling value by a standard deviation. First the group of 32 images are formed as one batch or one group and then training is performed on these batches. CNN having two architectures like AlexNet and ResNet, which functions as a feature extractors. It is observed that the ResNet performs very well on the classification problems compared to the other architectures. In this way, CNN plays very important role in classification of moving objects and successfully applied in computer vision and image processing applications. CNN architectures also minimize the errors occurs in the image net challenge by record levels.

C. Region classification

The last step after feature extraction is classifier. In this paper, we have limited the scope of research by keeping two classes. Class 1 represent identification of single person and class 2 represent as identification of group of person. The softmax classifier is applied here. The formula for the loss is shown in Equation 3.

$$L_i = -\log\left(\frac{e^{f_{yi}}}{\sum_j e^{f_j}}\right) \quad (3)$$

where L_i is the loss associated with sample i, and f_{yi} is the class score of the class y for sample i.

The group of scaled and normalized frames are given as a input to CNN in training mode. It is trained as 32 images per batch using Nesterov momentum based SGD method [8]. For training the bounding boxes are passed to the scaled images using CNN. In testing phase, the images are passed to the CNN which classifies it into two classes and these class

values are extracted as a output. These values are helpful for testing the accuracy of CNN as well as acting as a input for further post processing action. In this step, the sub images are again regrouped into their parent images and the bounding boxes are created around classified objects based on the classes identified in the image.

IV. FEATURES AND DATASET

The PET 2009, PET 2012 and Caltech pedestrian Database [3] containing more than 2,00,000 frames which is used for training and testing data. These datasets are standard dataset for testing the pedestrian detection and classification algorithm. Samples from the database are shown in Figure 2.

Fig. 2 Caltech Dataset samples for classification

As shown in Figure 2, some of the images contain a single person walking, some contain multiple persons in the scene and some do not contain any moving pedestrian. Moreover, in some of the images the pedestrians are clearly visible and are bigger in size while in some of the images, the pedestrians are occluded by objects and very small in size which are far away from the camera. The dataset consists of 10 sessions collected from video. Out of that, first five sessions are allocated to the training set and the residual are allocated to the testing set. The images are obtained in term of .seq files and those are preprocessed by using matlab code. Also, around the pedestrian, the bounding box is provided as compilation of text files. Each file corresponds to a frame in the dataset. It is empty if pedestrian is not present and if pedestrians are present then it has the coordinates of the bounding boxes. The training dataset is randomly sampled into 2000 sub images (around 120000 images for the CNN) and the testing set is also minimized to 1400 images by random sampling. The Dollar et al. [2] recommended in his paper that every third frame in the training set is considered as a training sample and every 30th frame in the testing set is selected as test sample. However, using sliding window approach, the images are further divided into about 50 sub images. This causes the training time was extremely long and even it was challenging for testing. The performance evaluation metrics measured for estimating the accuracy of the detection system are the Miss Rate (MR) and false rate. If the bounding box drawn by the algorithm and the actual bounding box of ground truth do not overlap with each other for more than 50% of their area, then it is treated as miss or a False Negative. If the bounding box is given by the algorithm for a particular region does not contain a ground truth bounding box, then it is treated as False Positive (FP). This is used to calculate the number of false positives per image (FPPI) and take average of it. The miss rate (MR) is the ratio of the number of False Negatives to the number of accurately identified bounding boxes. The standard and realistic results are obtained by training and testing on a larger dataset.

V. RESULTS AND DISCUSSION

A. *Experimental results*

Pedestrian detection and classification algorithms are evaluated with various database covering different types of critical situations. Their performance is analyzed and evaluated using various metrics. The network is trained by passing the images through sliding window detector which generates set of smaller sub images for training. These are fed into the CNN for training. The Nesterov momentum method is applied for training. Different learning rates are tried for both the architectures and the best performing values are selected by experimentation as 0.1 for the Resnet and 0.01 for the AlexNet and set the momentum value as 0.9. The training set is run for multiple epochs such as 6 for AlexNet, 3 for ResNet with a group size of 32 frames for an iteration of both the networks. Fig. 2 shows the training results of AlexNet and ResNet Networks.

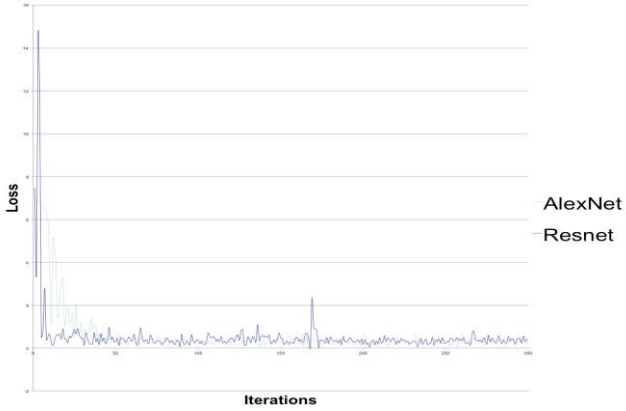
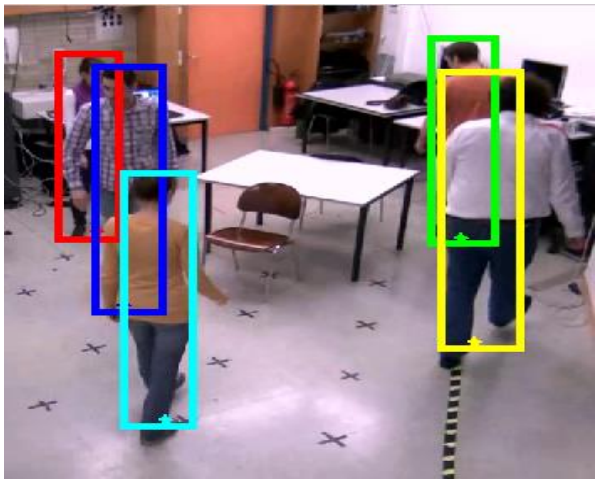
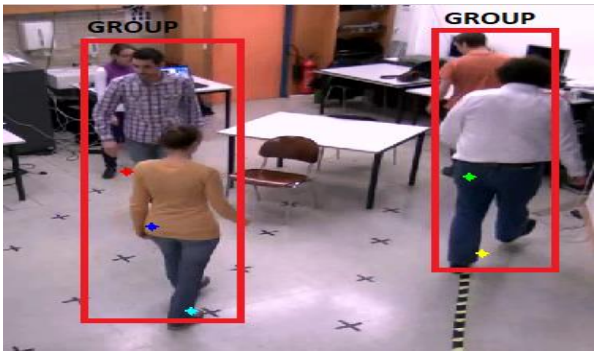


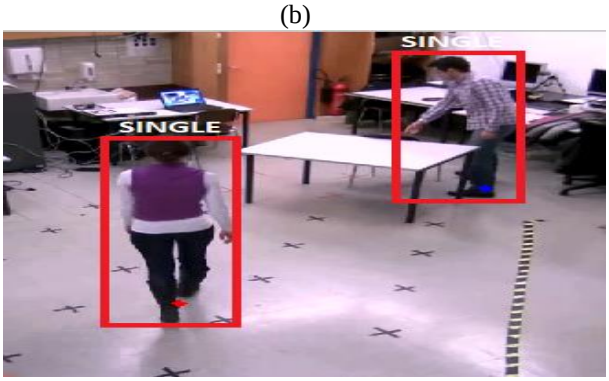
Fig. 2: Training results of the CNNs



(a)



(b)



(c)

Fig. 3 (a): Tracking individual person from video, (b) Classification into group of person, (c) Cassification into single person.

Classification of pedestrians into group of person and single person is shown in Fig. 3 (a) and (b) respectively.While training is carried out for multiple period and we observed that the loss is considerably reduced. Figure 4 shows the development of the loss function after the training for both the CNN architectures. The loss for the ResNet network comes down much faster for first 100 iterations and then did not change much. The ResNet network loss rapidly comes down than the AlexNet network. The training is executed on 2000 training sub images which results of total about 100000 images.

B. Discussion

From the quantitative analysis, it is observed that the miss rate and false rate average values are very less for all tested video sequences. Maximum moving objects are classified accurately in all frames of video and all classified objects are track successfully with bounding boxes. The accuracies of testing for the trained models is found to be extremely high. This appears to be an enormously optimistic measure of performance.

Table I: Quantitative Analysis over various Video set

Video	Video Frame	FPPI (%)	Miss Rate(%)
Multi pedestrians CAVIAR data set		6.23	8.39
Five pedestrians in the corridor (CAVIAR)		9.12	10.48
Sparse crowd (CANDATA)		5.16	9.87
Generated dataset		7.53	4.40
Crowd(PET 2012)		8.43	4.21

Sliding window approach generates training images that are having full visible of multiple people in the scene which are walking single or in group. Very few images are generated by sliding window which consists of pedestrians that are overlap with background. Thus, the network is over fitted for such kind of cases and it works well on the test set also. To avoid such kind of issues, large number of test images are required with larger sample set. It was also observed that using a sliding window approach for feature detection requires large data size so it requires down sampling. Moreover , in some cases, where the people are far away from camera observed in small portion of the whole frame then the sliding window execution is not smart enough to taken out people correctly. This probably gives the over fitting and incorrect test results Thus for future work, moving to an R-CNN based approach where the CNN acts as both the region proposal as well as the feature detection methods and avoid the use of sliding window detector.

VI. CONCLUSION AND FUTURE SCOPE

We have implemented pedestrian detection using Adaptive median filter and classification using CNN network. The sliding window approach is used for region proposal and both CNN networks like AlexNet and 18 layer ResNet are used for feature detection with the softmax function acting as a classifier. The problem is formulated as single person or group of person classification. This system gives optimum results with great accuracy over Caltech and PET pedestrian data set . The model works moderately well for cases where people are undoubtedly and clearly present in the frame but does not work properly during occlusion condition and when the pedestrians are very small in size (away from the camera). The next step is post processing step which covers all the classified objects with bounding boxes. Experimental results shows that this proposed system gives optimum classification results over small dataset used. In future, the next step is to perform training on a much larger training set and also testing the performance on a larger test set. Moreover, in order to overcome the difficulty occurred because of too many samples required for implementing the sliding window approach, R-CNN approach is selected which eliminate the use of sliding window as feature detector.

REFERENCES

1. Tome D., Monti, F., Baroffio, L., Bondi, L., Tagliasacchi, M., Tubaro, S., “Deep convoluted neural netwoks for pedestrian detection.”, Preprint. Elsevier Journal of Signal Processing: Image communication, 2015.
2. Dollar,. P, Wojek, C., Schiele, B., Perona, P.,“Pedestrian Detection: An Evaluation of the State of the Art.”, IEEE Transactions on Pattern Analysis and Machine Intelligence, ISSN:0162-8828, August 2011.
3. N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection”, IEEE Conference on Computer Vision and Pattern Recognition, 2005.
4. Rowley, A. H., Baluja, S., Kanade.,T, ”Neural Network-Based face Detection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, January 1998.

5. Viola, P., Jones, M., "Robust Real-time Object Detection",2nd International Workshop on Statistical and computational theories of vision, modeling ,learning, computing and sampling, 2001.
6. Lun Zhang, Stan Z. Li, Xiaotong Yuan, " Real-time Object Classification in Video Surveillance Based on Appearance Learning" Center for Biometrics and Security Research & National Laboratory of Pattern Recognition Institute of Automation, Chinese, Academy of Science.
7. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y. " Pedestrian Detection with unsupervised Multistage Feature Learning", International Conference on Computer Vision and Pattern Recognition, 2013.
8. R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000.
9. Fukui, H., Yamashita, T., Yamauchi, Y., Fujiyoshi., H,Murase, H., "Pedestrian Detection based on Deep Convolution Neural Network with ensemble Inference Network", IEEE Intelligent Vehicles Symposium, 2015.
10. Molin,D., "Pedestrian Detection using convolution neural networks", Phd Thesis, Department of Electrical Engineering, Linkopings University, Sweden, 2015
11. Krizhevsky, A, Sutskever, I , Hinton, G.E. " Image Net classification with deep convolutional neural networks", Advances in neural information processing systems, 2012
12. Szegedy, C., Liu, W, Jia, Y., Sermanet, P., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.,
13. "Going Deeper with Convolutions",ILSVRC 2014 learning for image recognition", ILSVRC 2015
14. https://github.com/facebook/fb.resnet.torch
15. http://torch.ch/blog/2016/02/04/resnets.html
16. Li, J., Liang, X., Shen, S., Xu, T., Yan, S., "Scaleaware Fast R-CNN for Pedestrian Detection", arXiv preprint, arXiv:1510.08160
17. http://cs231n.github.io/linear-classify/
18. https://github.com/soumith/convnet-benchmarks

AUTHORS PROFILE

Mrs. Rohini A. Chavan is pursuing Ph. D in electronics from Bharati Vidyapeeth Deemed University College of Engineering Pune. She has completed M.E (E & TC) from Savitribai Phule Pune University. She has more than 10 years of teaching experience. She is working as Assistant Professor in the Department of Electronics and Telecommunication.

Dr. Sachin R. Gengaje has total 27 years of academic experience and adept industrial and research experience. His area of interests includes Image & Video Processing, Soft Computing and Outcome Based Education

Dr. Shilpa P. Gaikwad has obtained Ph.D. in E & TC from SPPU University. She has received M.E. Electronics degree from the Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune and she has more than 15 years of teaching experience. She is working as Associate Professor in Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune.