Real-Time Headgear Detection in Videos Using Deep Learning Based Feature Extraction with A Supervised Classifier

Rajdeep Pal^{*#}, Ranjana Seshadri^{*#}, Swarnashree Mysore Sathyendra^{*#}, Dr. S. Natarajan[†]

^{*}Department of ISE, PES Institute of Technology, Bangalore

[†]Department of CSE, PES University, Bangalore

Abstract—This paper presents an approach for detection of headgear in real-time video footages. The approach used is Feature Extraction followed by Classification. To pick specific features required for the aforesaid problem, a pre-trained deep learning model is used. For the classification, a supervised classifier (SVM with different kernels and k-NN) is used. A comparison of both the possibilities is also mentioned in the paper. The classifier is implemented hierarchically with two levels. The first level is a binary classifier which determines if a person is wearing headgear. The second level determines the kind of headgear (Hat, Construction Hat or Helmet) or his hairstyle (Bald or with hair) in accordance with the results of the first level. Due to the lack of a considerably sized dataset for headgear detection, the dataset used in this paper is collected manually from online sources and surveillance footages available from a video-analytics company. The dataset compiled has 2423 images of 5 different classes. As part of the pre-processing, a person localization and a customized body segmentation model is also implemented. This paper suggests a solution that can be used in real-time with prediction time of 6ms per image and pre-processing time of 165ms per frame (as measured on i7 7700HQ and GTX 1060).

Index Terms—headgear, object detection, AlexNet, SVM, k-NN, deep learning, feature extraction, transfer learning

I. INTRODUCTION

There is a broad range of applications of headgear detection. From the perspective of safety concerns, wearing helmets becomes an important rule to be followed for road safety for riders. In case of construction sites, trespassing could be stopped by ensuring that all the workers with the work permit for the site are wearing their construction hats which not only serves as a safety gear but also as an identification mechanism. As in case of caps/hats, they could be used as one of the distinguishing features of people like in case of a search for missing persons, victims, suspects etc. The main bottleneck in working on real-time video footages for any type of problem is the performance. An extremely accurate solution/prediction for the problem at hand might become irrelevant if the processing time for it is high and vice-versa.

In this paper, we provide a proof of concept for using a deep learning feature extractor along with a classifier to obtain results on video footages within an acceptable time limit with an accuracy and confidence of the predictions good enough for the use case discussed - Headgear detection. The headgear

module can be used as a standalone module or in integration with other functionalities to solve different problem use cases.

All of the training and testing for this research has been done on systems with the following specifications 2 machines with Intel core i7, 7th Generation processor, 1060 NVIDIA GEFORCE GTX GPU with 6GB memory and a RAM of 16GB, one machine with Intel core i7, 7th Generation, 1050 NVIDIA GEFORCE GTX GPU with 4GB memory and RAM of 8GB. While computer vision techniques are viable for tackling a problem such as helmet detection, this problem statement can be remodelled as an object detection problem which is well addressed in the deep learning field with many famous object detection networks like Single Shot Detector, YOLO and AlexNet. In this paper, we discuss pre-processing a given scene or frame of a video footage which involves human figures detection followed by segmentation of those human figures body into three parts namely the head, upper body and lower body. Followed by this, the head segment of each of such human figures can be passed on to a feature extractor which is basically AlexNet infused with transfer learning. Using this feature map, classification could be carried out further.

The rest of the paper deals with the extensive literature survey done, followed by an implementation of our proof of concept concludes with the results and discussions of the same and the future scope of the solution.

II. LITERATURE SURVEY

A lot of work has been done on traffic surveillance systems. Helmet detection is one of the major problems which has been dealt with. In this context Romuere Silva et al. [1] proposes a circular Hough transformer and HOG descriptor-based feature extractor to identify a motor cyclist and SVM to predict if the rider is wearing a helmet.

The dataset used for the aforesaid problem is limited to only 255 images and with respect to a single background scene. A dataset of such a small size may not only increase the possibility of over-fitting but might have a lower accuracy on new unseen data. Since the testing is done on static images the approach used might not work well for the proposed use case since it requires real-time video footages.

Thepnimit Marayatr et al. [2] proposes a method using background subtraction and Hough transforms. They report

TABLE I
DATASET DESCRIPTION

Class	Number of Instances
Bald	643
With hair	663
Construction hat	499
Cap/hat	136
Helmet	482

an accuracy of 77% with a limited number of 20 validation images.

Similarly, for the detection of hats, Man-Woo Park et al.[3] presents a HOG descriptor based classifier at construction sites. This is again a binary classifier and location specific. Rui Min et al. [4] uses dynamic time warping to get a dissimilarity matrix for extracted features and Hierarchical Agglomerative clustering for detection of caps. The accuracy achieved is around 87.5 % and is a binary classifier.

Carlos A. Luna et al.[5] presents a multi class classification using depth sensors. They use histograms of pixels for feature extraction. Here the images are captured in a certain angle and might not work well in all scenarios where camera is not calibrated accordingly. There is no use of any deep learning model in the problem.

III. DATASET COLLECTION AND PREPROCESSING

Around 2423 images of people who are bald, with hair, wearing helmets, hats and construction helmets (hard hats) are scraped from different online sources. The dataset obtained is divided into training and validation sets with 451 (around 20% of the dataset) images used as validation and 1972 images used for training. Table. I gives a brief description of the dataset and the classes it comprises with number of instances of each.

To classify and annotate these images, 2 levels of pre-processing is done. They are:

A. Person Localization

Each image is padded to form a 227x227 image. This is then sent to a pre-trained SSD [6] network to extract every person in the frame. Each of these extracted images are then passed into the Body Segmentation Module individually.

B. Body Segmentation

The body segmentation module is trained using the YOLOv2 [7] network using the PASCAL-PARTS dataset. The masks are combined to suit our problem and is trained on three classes - HEAD, UPPER BODY and LOWER BODY.

For Annotation, the images with head are padded to create square images. They are then manually annotated and validated to form the following five classes: BALD, CONSTRUCTION HAT, WITH HAIR, HAT and HELMET for compiling the dataset.

In the example Fig. 2 shown, original image a. is converted to b. by passing through the person localization module and then padding it to get 227x227 image (picture is maximized

Fig. 1. Example instances of each class used in the dataset

for clarity). b. is sent to the Body Segmentation Module to detect the head of the human figure. This is now included in the dataset for headgear detection using head.

IV. IMPLEMENTATION

As part of pre-processing, the Person localization and Body segmentation modules are used. A multi class classifier is created by first extracting specific features of the image and then classifying the images using a supervised classifier.

A. Feature Extraction

The images in the dataset are resized to 227x227 images. These images are passed through Alex Net [8] for feature extraction. A 4096x1 feature vector is extracted at the FC-6 layer and also a 4096x1 feature vector at the FC-7 layer.

B. Classification

A k-NN classifier and a Support Vector Machine (SVM) with different kernels - linear , polynomial and RBF(Radial Bias) are used for the classification. A hierarchical approach with two levels is also used for the classification. The first level determines if a person is wearing a head-gear or not and the second level predicts the type of headgear.

V. RESULTS AND DISCUSSIONS

A total of 438 unseen images are used for the validation of binary classifier. The performance of different classifiers (k- NN and SVM (Linear, Polynomial and RBF kernel)) are compared.

For the binary classification SVM with RBF (Radial Bias Function) performed very well. Table. II gives the results of using different SVM Kernels for binary classification to classes with headgear and no headgear.

Fig. 2. Pre-processing pipeline for data annotation

Fig. 3. Overview of helmet detection pipeline

TABLE II
RESULTS OF USING DIFFERENT SVM KERNELS FOR BINARY CLASSIFICATION

SVM Kernel Function	Number of correctly classified images	Accuracy %
Radial Basis	403	92.00
Polynomial	393	89.72
Linear	392	89.49
Total Number of Instance	438	

TABLE IV
RESULTS OF USING DIFFERENT SVM KERNELS FOR MULTI CLASS CLASSIFICATION

SVM Kernel Function	Accuracy %
Radial Basis	91.84
Polynomial	87.50
Linear	91.35
Total Number of images	451

TABLE III
CONFUSION MATRIX OF BINARY CLASSIFIER USING SVM WITH RBF KERNEL

Class	Predicted number of instances	Actual number of instances
Headgear	255	267
No Headgear	147	171

TABLE V
CLASS PRECISIONS FOR MULTI CLASS CLASSIFICATION USING SVM WITH RBF KERNEL

Class	Correctly classified as Class	Wrongly classified as Class	Precision
Bald	108	11	0.908
Construction hat	69	16	0.812
With hair	119	51	0.700
Cap/hat	8	2	0.800
Helmet	55	12	0.821

The highest accuracy achieved was 92.00%. A recall of 0.955 and precision of 0.914 was obtained.

For the multi class classification 451 images are used and a highest accuracy of 91.84% is achieved using radial basis function. Tables. IV and V give the results for multi class classification.

The time taken for each process per video frame (image) is

shown in Table. VII recorded and processed on an i7 7700HQ and GTX 1060.

TABLE VI
CLASS PRECISIONS FOR HIERARCHICAL MULTI CLASS CLASSIFICATION USING SVM WITH RBF KERNEL

Class	Total Number in Dataset	Total Number passed to Level II	Correctly classified as Class	Wrongly classified as Class	Precision
Bald	127	126	111	13	0.895
Construction hat	88	82	71	18	0.798
With hair	130	129	119	33	0.783
Cap/hat	24	18	8	0	1.000
Helmet	82	74	63	16	0.797

TABLE VII
TIME TAKEN FOR ALL MODULES PER VIDEO FRAME/IMAGE

Pipeline Stage	Module	Time Taken in ms
Pre-Processing	Person Localization	90
	Body Segmentation	55
Processing	Feature Extraction	20
	Classification	6

TABLE VIII
CLASS PRECISIONS FOR MULTI CLASS CLASSIFICATION USING K-NN WITH K AS 5

Class	Correctly classified as Class	Wrongly classified as Class	Precision
Bald	101	12	0.894
Construction hat	70	14	0.834
With hair	118	72	0.621
Cap/hat	10	5	0.667
Helmet	43	7	0.860

A. Comparison between SVM and k-NN Classifiers

Table. VIII gives the results of using a k-NN classifier. While the SVM with both linear and radial basis show accuracies higher than that of with a k-NN classifier with $k=5$, the polynomial kernel SVM performs lower than the k-NN classifier.

The decision of which supervised classifier to use, can be dependent on the required use case along with the computational resources available. SVM in general can be considered to be more computationally intensive as compared to k-NN classifiers. However, k-NN is more easily implementable.

VI. CONCLUSIONS AND FUTURE WORK

As shown, the time taken for each of the module is within an acceptable range. Since real-time video processing requires fast modules, the proposed solution can be used for detection of headgear in real-time video footages.

The pre-processing and the headgear detection modules are independent of each other and modular. Hence, they can be hosted on high-capability servers, which can also help in increasing the speed of processing of each of the modules. The approach used in this proof of concept is to identify all kinds of headgears and if there is no headgear. The module can be fine-tuned further based on certain specific use cases by changing the training dataset. For example, if the system has to be installed in a construction site for identification of workers

not wearing the construction hat, the module can be trained with Level I of the hierarchy as Yes, wearing construction hat and No, not wearing the hat.

The future work under progress is to increase the dataset size further. The solution can be extended further for person re-identification and tracking of the person. This might help in situations like traffic control where it becomes important to track a person not wearing helmet. An other important prospective work is to increase the number of levels in the hierarchy which can include divisions like type of helmet (half helmet, full helmet etc), hat or a cap etc.

ACKNOWLEDGEMENT

We would like to extend our heartfelt appreciations to our mentors at AllGoVision Technologies Pvt Ltd, Bangalore, India for having helped and guided us with valuable advice throughout the course of this paper.

The continued support and resources provided by our own institution PES Institute Of Technology and the faculty.

Last but not the least, we are extremely grateful to our professor and guide Dr. S. Natarajan for constantly counselling and believing in us and helping us in bringing out our ideas through this research.

REFERENCES

- [1] Romuere. Rodrigues Veloso e Silva K. Teixeira Aires R. De Melo Souza Veras "Helmet detection on motorcyclists using image descriptors and classifiers" SIBGRAPI pp. 141-148 2014.
- [2] T. Marayatr and P. Kumhom, "Motorcyclist's Helmet Wearing Detection Using Image Processing", Advanced Materials Research, Vols.931-932, pp. 588-592, 2014
- [3] Man-Woo Park, A.M.ASCE ; Nehad Elsafty2 ; and Zhenhua Zhu,A.M.ASCE,. Hardhat-Wearing Detection for Enhancing On-Site Safety of Construction Workers, Journal of Construction Engineering and Management,2015I. S. Jacobs and C. P. Bean, Fine particles, thin films and exchange anisotropy, in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271350.
- [4] Min, Rui, and Jean-Luc Dugelay. "Cap detection for moving people in entrance surveillance." In Proceedings of the 19th ACM international conference on Multimedia, pp. 1253-1256. ACM, 2011.
- [5] Luna, Carlos A., Javier Macias-Guarasa, Cristina Losada-Gutierrez,Marta Marron-Romera, Manuel Mazo, Sara Luengo-Sanchez, and Roberto Macho-Pedroso. "Headgear Accessories Classification Using an Overhead Depth Sensor." Sensors 17, no. 8 (2017): 1845.
- [6] Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. "Ssd: Single shot multibox detector." In European conference on computer vision, pp. 21-37. Springer, Cham, 2016.
- [7] Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger" arXiv preprint (2017).
- [8] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In NIPS, pp.10971105,2012.