Shallow-learning methods for pedestrian classification and detection in images

Master's degree in Computer Engineering for Robotics and Smart Industry
Machine Learning and Artificial Intelligence course, a.a. 2021/2022

Lorenzo Busellato[1]

[1]VR472249, lorenzo.busellato_02@studenti.univr.it

Abstract

Pedestrian detection is a key task in the computer vision field, given its obvious applications in surveillance systems, autonomous driving and robotics. This work presents a pipeline for shallow-learning based pedestrian detection. The Histogram of Oriented Gradients (HOG) method for feature extraction is described and applied to the Daimler Mono Pedestrian Classification Benchmark Data Set. Three classification techniques, Support Vector Machines (SVMs), K-Nearest Neighbors (KNN) and Naive Bayes, are described and benchmarked with different sets of hyperparameters. The benchmarks are then compared against each other to determine the best performing method in terms of accuracy in classification. Finally the best performing method is used to implement a full pedestrian detection pipeline.

CONTENTS

I Introduction 2

II Objectives 2

III Dataset 2

IV Methodology 2
- IV-A Histogram of Oriented Gradients . . . 2
- IV-B Naive Bayes . . . 3
- IV-C K-Nearest Neighbors . . . 3
- IV-D Support Vector Machine . . . 3
- IV-E Hard-negative mining . . . 3
- IV-F Pedestrian detection . . . 3
- IV-G Non-maximum suppression . . . 3

V Experiments and results 3
- V-A Adopted metrics . . . 3
- V-B Performed experiments . . . 3
- V-C Results . . . 3

VI Conclusions 3

I. Introduction

THE task of pedestrian detection, i.e. the task of localizing pedestrians within an image, is one of the most tackled instances of the broader class of object detection problems. This is due to its natural applications in autonomous driving, surveillance and robotics. For instance, in the US alone about 5000 of the 35000 annual car accident fatalities involve pedestrians [1], therefore the need for automated methods for pedestrian detection is apparent.

The task of pedestrian detection presents a series of complexities regarding the high variability in scale and size of the person within the image, occlusions with the rest of the scene, susceptibility to lighting conditions and the real time detection requirement.

The main models that are used for pedestrian detection are hand-crafted models and deep learning models. Hand-crafted models are based on classification of hand-crafted features. Deep learning models on the other hand typically use convolutional neural networks for the extraction of features, therefore producing high-level semantic representations of objects. In this work the first class of models will be analyzed and implemented. Specifically, the Histogram of Oriented Gradients (HOG) method will be used to extract features from images, because it is known[2] that it is the best performing non-deep learning based feature extractor for human detection. The classification of images will then be implemented using non-deep learning methods, namely Support Vector Machines (SVMs), K-Nearest Neighbors (KNN) and Naive Bayes. Finally the best performing classifier will be used to implement a pedestrian detector.

The main objectives of this work are described in section II. The data set and methodology employed are described in sections III and IV. The experimental evaluation is presented in section V. The conclusions are summarized in section VI.

II. Objectives

The main objective of this work is the implementation of a pedestrian detection pipeline, which is made up of three main components: a feature extractor, a classifier and a detector.

Regarding the classifiers, the objective is to obtain an in-depth comparison between the performance of some well-known shallow learning techniques with respect to the problem of pedestrian detection, as well as the relation between hyperparameters and classifier performance.

Finally, the best performing classifier will be used to implement a detector, therefore completing the pedestrian detection pipeline. The pipeline will then be applied to a video sequence in order to showcase its performance.

III. Dataset

The data set used in this work is the Daimler Mono Pedestrian Classification Benchmark Data Set. The data set consists of four separate subsets: the base set and three additional sets. The base set is divided into five subsets, three for training (named "1", "2" and "3") and two for testing (named "T1" and "T2"). All five sets contain 4800 18×36 grayscale images containing pedestrians and 5000 18×36 grayscale images not containing pedestrians. The three additional sets each contain 1200 640×280 grayscale images for the purpose of extracting additional negative samples for the hard-negative mining procedure.

Fig. 1. Example figures from the training data set. On the left a non-pedestrian sample, on the right a pedestrian sample.

In this work, data sets 1, 2 and 3 will be grouped together in a singular data set for the training of the models, as will data sets T1 and T2 for the testing.

IV. Methodology

A. Histogram of Oriented Gradients

Given an input image I, for each pixel $(x, y) \in I$ the gradients G_x and G_y are computed. This is typically done via a convolution operation:

$$G_x = \omega_x * I \qquad G_y = \omega_y * I \tag{1}$$

where ω_x and ω_y are the kernels associated to the convolution operation. The most common kernels used in the context of pedestrian detection are:

$$\omega_x = \begin{bmatrix} -1 & 0 & 1 \end{bmatrix} \qquad \omega_y = \omega_x^T$$

They are the most commonly used because experimentally it is observed[2] that they produce the least amount of detected false positives per moving window (FFPW) compared to other kernels.

After the convolution operation, two gradient matrices are obtained, G_x and G_y. For each pixel $(x, y) \in I$ then the magnitude μ and angle θ are computed:

$$\mu(i,j) = \sqrt{G_x(i,j)^2 + G_y(i,j)^2} \tag{2}$$

$$\theta(i,j) = \left| \tan^{-1}\left(\frac{G_y(i,j)}{G_x(i,j)}\right) \right|$$

In practice the angle is computed with $atan2$, therefore it is in the range $(\pi, \pi]$.

The resulting gradient magnitude and gradient angle matrices are then divided into 8×8 patches, for each of which a 9-point histogram of oriented gradients is computed. A histogram is a series of bins labeled with angle values. The number of bins is determined by the angle step between the bins, for instance with an angle step of 20 degrees the histogram will have $180°/20° = 9$ bins.

For each pixel in the patch, the corresponding gradient angle determines two bins on the histogram that the pixel votes for. The two bins selected are those which represent the two closest

angles to the given gradient angle value. The value added to the selected bins is proportional to the distance of the gradient angle to the bin angle, the gradient magnitude and the step between the bins:

$$v_{closest} = \frac{(binAngle - \theta)\mu}{angleStep}$$

$$v_{2nd-closest} = \frac{(\theta - binAngle)\mu}{angleStep}$$

Repeating the process for each 8 × 8 patch, a 16×8 matrix of histograms is constructed. To reduce the variability in gradient intensities due to lighting conditions and foreground-background contrast, a block normalization procedure is introduced. The 16×8 histogram matrix is scanned with a 2×2 window producing 15×7 blocks, for each of which a normalized histogram vector is computed by concatenating the four histograms contained in the block and normalizing the resulting vector. Finally, the feature descriptor for the image is constructed by concatenating the normalized histogram vectors. From each 8×8 patch a 1×9 vector is collected, which is then concatenated with another three 1×9 vectors composing a 1×36 vector, which is finally concatenated with another fourteen 1×36 vectors, resulting in a 1×3780 feature vector.

B. Naive Bayes

C. K-Nearest Neighbors

D. Support Vector Machine

E. Hard-negative mining

F. Pedestrian detection

G. Non-maximum suppression

V. Experiments and results

A. Adopted metrics

B. Performed experiments

C. Results

VI. Conclusions

References

[1] Umesh Shankar. *Pedestrian roadway fatalities*. Tech. rep. 2003.

[2] Navneet Dalal and Bill Triggs. “Histograms of oriented gradients for human detection”. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*. Vol. 1. Ieee. 2005, pp. 886–893.