Neurocomputing 150 (2015) 214–226

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

On the purity of training and testing data for learning: The case of pedestrian detection

Matteo Taiana*, Jacinto Nascimento, Alexandre Bernardino

Institute for System and Robotics – Lisbon, Portugal[1]

ARTICLE INFO

Article history:
Received 31 January 2014
Received in revised form
5 July 2014
Accepted 7 September 2014
Available online 14 October 2014

Keywords:
Sample purity
Pedestrian detection
Machine learning
Partial occlusion
INRIA person data set
Labelling

ABSTRACT

The training and the evaluation of learning algorithms depend critically on the quality of data samples. We denote as *pure* the samples that identify clearly and without any ambiguity the class of objects of interest. For instance, in pedestrian detection algorithms, we consider as pure samples the ones containing persons who are fully visible and are imaged at a good resolution (larger than the detector window in size). The exclusive use of pure samples entails two kinds of problems. In training, it biases the detector to neglect slightly occluded and small sized samples (which we denote as *impure*), thus reducing its detection rate in a real world application. In testing, it leads to the unfair evaluation and comparison of different detectors since slightly impure samples, when detected, can be accounted for as false positives. In this paper we study how a sensible use of impure samples can benefit both the training and the evaluation of pedestrian detection algorithms. We improve the labelling of one of the most widely used pedestrian data sets (INRIA) taking into account the degree of sample impurity. We observe that including partially occluded pedestrians in the training improves performance, not only on partially visible examples, but also on the fully visible ones. Furthermore, we found that including pedestrians imaged at low resolutions is beneficial for detecting pedestrians in the same range of heights, leaving the performance on pure samples unchanged. However, including samples with too high a grade of impurity degrades the performance, thus a careful balance must be found. The proposed labelling will allow further studies on the role of impure samples in training pedestrian detectors and on devising fairer comparison metrics between different algorithms.

1. Introduction

Machine Learning (ML) is the field of science researching how computers can learn from data. ML has been successfully applied to many areas of knowledge, from medical [1,2] to financial [3,4], to scientific applications in general [5–7]. The results obtained by a ML system, however, depend heavily on the quality of the data used in its training [8,9]. Moreover, the evaluation and comparison of such systems depend on the quality of the data used for their testing [10]. In this paper, we focus on the application of ML to the detection of people in images.

Detecting humans in images is a challenging task that attracts the attention of the scientific community and industry alike. The problem assumes different contours depending on whether the sensor used to capture the images is fixed or mobile, whether the detection is performed on a single image or on a sequence of images, and whether the sensor is a single camera or a richer sensor providing depth information. One further distinction can be made between the methods that do and do not restrain the articulation of the persons. This work concentrates on the detection of pedestrians, i.e., people assuming poses that are common while standing or walking, in images acquired by a mobile camera. Detecting pedestrians is important as it enables the estimation of the presence and the position of humans in the vicinity of a vision sensor. The task is complex mostly because of the high variability that characterizes the pedestrians projections on the camera image plane. The appearance of a pedestrian on the image is influenced by the person's pose, his or her clothing, occlusions, and the atmospheric conditions that contribute to the illumination of the scene. Background clutter also plays a role in making the detection difficult.

The publication of data sets is an important step towards a fair comparison of the performances of Pedestrian Detection (PD) systems, but it is not enough. Standard evaluation code is also needed as different evaluation procedures can lead to discrepancies in the reported performances. Data sets are created not only with the intent of comparing the performance of algorithms, but

* Corresponding author.
E-mail address: mtaiana@isr.ist.utl.pt (M. Taiana).
[1] http://www.isr.ist.utl.pt

http://dx.doi.org/10.1016/j.neucom.2014.09.055
0925-2312/

also with the goals of exposing the limitations of contemporary algorithms and stimulating advances in the state of the art. As such, data sets have a limited life span: as the understanding of the problem by the scientific community grows, hurdles are conquered and data sets become obsolete.

The missed detection rate for the INRIA data set [11] at 0.1 False Positives Per Image (FPPI) has dropped from around 50% to around 20% since its publication (see [12]). There is still room for improvement, which explains why that data set is still widely used as a benchmark [13–16]. The same data is also very popular for training: 13 out of 16 algorithms reviewed in [12] are trained on it. The labelling, like in many other PD data sets, consists of rectangular bounding boxes each one of which tightly enclosing a person. For the purposes of this work, we choose to enrich and extend the labelling of the INRIA person data set.

This paper is an extension of our work published in [17], in which we proposed a new labelling for the INRIA *test set* to improve the comparison between different algorithms. In this work we present a new labelling for the INRIA *training set* and show through experimental results the importance of the correct use of pure and impure samples both in the training and in the testing phase. The labelling was conducted following the method proposed in [12]. The proposed annotation is available on the authors' website.[2] We argue that the new test set labelling leads to a better evaluation of PD algorithms, while the new training set labelling enables researchers to analyse the impact of pedestrian height and visibility during training on the detection performance. In this paper we restate, for continuity of exposition, the contributions described in [17] and build on them to present new results. In the section on the evaluation protocol we show that a fair evaluation of detectors with the original labelling of the INRIA test set requires the use of a minimum resolution limit: since only pedestrians taller than 90 pixels are systematically labelled, the evaluation should ignore detections shorter than that limit. Furthermore we show that using the proposed labelling for testing produces a more truthful evaluation of the detectors. The contributions specific to this paper are: the introduction of the notion of sample purity, the elaboration of a new labelling for the training set, the results of Experiment 1, confirming that visibility plays an important role for detectability, the results of Experiment 2, showing that it is worth to include partially occluded pedestrians in the training set, even when testing on fully visible pedestrians, and the results of Experiment 3, showing that it is important to have "short" examples in the training set when the goal is to detect "short" pedestrians.

The remainder of the paper is organized as follows. In Section 2 we introduce the reader to the PD problem. In Section 3 we detail how annotations for data sets are usually compiled, while in Section 4 we describe the principles that guided the proposed labelling. In Section 5 we describe the PD used in this work and in Section 6 we define the evaluation protocol used in the experiments. We relate results in Section 7 and draw conclusions in Section 8.

2. Related work

Advances in Pedestrian Detection (PD) stem mostly from research in the areas of visual feature extraction and Machine Learning, the most common classifiers being based either on AdaBoost [18] or Support Vector Machines [19]. Seminal work in PD was presented in [20,21]. The authors of [22] introduced Integral Images for faster feature computation, AdaBoost for combining many weak classifiers into a strong classifier and a Cascaded Detector for speeding the detection up. That work focused on the recognition of frontal faces and used Haar-like features, which failed to perform as well in the person detection task. The architecture, nonetheless, became very popular for PD algorithms. Dense features, computed on a regular grid over the image, have been very successful. One example of such features, which is ubiquitously used in detection, is the Histogram of Oriented Gradients (HOG). Introduced in [11] and reminiscent of SIFT [23], it represents gradient information in a way that enables robust classification. A recent trend is that of combining multiple features: the Integral Channel Features [24] exploit 10 channels of information based on colour and gradient. The authors of [25] combine Gradient Histograms, Local Binary Patterns (to exploit texture information), Colour Self Similarity (second order statistics of colour) and Histograms of Flow (to exploit movement information). Another line of work concentrates on reducing the detection time, see for instance the Fastest Pedestrian Detector in the West (FPDW) algorithm [26]. One dualism in the literature contrasts monolithic detectors (see [11,13,16]), which compute features at fixed locations on the detection window, to part-based detectors (see [27,28]), which explicitly model the articulation of the human body and use the features where the limbs are estimated to be.

Comparing the performance of PD systems is a fairly complex matter. Many data sets have been published over the years. A first notable example is the MIT pedestrians data set [20], introduced in 1997. It includes frontal and rear views of pedestrian and only positive windows, i.e., fixed-size rectangular images designed to contain a person. The INRIA person data set [11] was introduced by Dalal and Triggs in 2005, it is divided in training set and test set and it provides both positive and negative examples. The ETH pedestrians data set [29] was introduced in 2007. It was recorded with a mobile platform moving along a sidewalk, equipped with a stereo camera. It presents a scenario typical for a mobile robot. The TUD-MotionPairs/TUD-Brussels data set [30] (TUD) and the Caltech pedestrian data set [12] were introduced in 2009 and contain sequences of images taken in automotive scenarios. The size of the data sets has grown over time, from 924 positive examples (MIT data set) to 350 000 labels over 250 000 images (Caltech data set). Each data set can be characterized in a number of ways, one important parameter being the range of sizes of the annotated pedestrians. Most PD algorithms output detections in a selected range of sizes, in order to perform a fair evaluation it is important that such ranges coincide.

The code used to evaluate the performance of a detector on a data set can considerably influence the results. Many parameters, such as the number of classes of labels used for annotating the data and the amount of padding on the candidate images, can influence the reported results. A solution for this problem is to use the same evaluation code on each algorithm. Dollár provides such a code[3] together with a collection of data sets and the detections obtained running several state-of-the-art detectors on such data sets. We adopt that evaluation code and describe its principles in Section 4.

3. Labelling strategies

The purpose of the labelling of a data set for Pedestrian Detection (PD) is twofold. First, the annotation of the training set enables the extraction of the positive and negative examples for training the detector. Second, the annotations of the validation and

[2] Proposed annotation http://users.isr.ist.utl.pt/~mtaiana/data.html

[3] Caltech Pedestrian Detection evaluation code www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/DollarEvaluationCode