J. Vis. Commun. Image R. 34 (2016) 187–203

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Vehicle detection in aerial imagery : A small target detection benchmark[☆]

Sebastien Razakarivony [b,a,*], Frederic Jurie [b]

[a] Safran, 1 rue Genevive Aube, 78114 Magny-les-Hameaux, France
[b] University of Caen – CNRS UMR 6602 – ENSICAEN, 14000 Caen, France

ARTICLE INFO

Article history:
Received 28 January 2015
Accepted 7 November 2015
Available online 18 November 2015

Keywords:
Detection
Low resolution images
Vehicles
Database
Aerial imagery
Infrared imagery
Target detection
Computer vision

ABSTRACT

This paper introduces *VEDAI: Vehicle Detection in Aerial Imagery* a new database of aerial images provided as a tool to benchmark automatic target recognition algorithms in unconstrained environments. The vehicles contained in the database, in addition of being small, exhibit different variabilities such as multiple orientations, lighting/shadowing changes, specularities or occlusions. Furthermore, each image is available in several spectral bands and resolutions. A precise experimental protocol is also given, ensuring that the experimental results obtained by different people can be properly reproduce and compared. Finally, the paper also gives the performance of baseline algorithms on this dataset, for different settings of these algorithms, to illustrate the difficulties of the task and provide baseline comparisons.

1. Introduction

Automatic Target Recognition (ATR), which is the task of automatically detecting targets in images, has an history of more than 35 years of research and development in the computer vision community. The basic aim of such ATR systems is to assist or remove the role of man from the process of detecting and recognizing targets and hence to implement efficient and reliable systems of high performance. One typical application is surveillance and reconnaissance, two tasks which need to be more and more automatic, as recent high resolution surveillance sensors produce imagery with high data bandwidth. As explained by Wong [1], a surveillance mission over a 200 mile square area with a one foot resolution (an appropriate size for recognizing many targets), will generate approximately 1.5×10^{12} pixels of data. If the area is split in 10 million pixels images, photo interpreters would have to examine over 100,000 images, which is an impractical workload and results in a delayed or incomplete analysis. In addition, the delay would allow movable targets to relocate so that they cannot be found in subsequent missions. Vehicle detection is hence of crucial matter in defense applications.

Despite the aforementioned very long history of ATR in the computer vision literature, it is still a challenging problem even with the most recent developments of this area. This is demonstrated by the figures given in the experiments section of this article.

The traditional way to address ATR consists in the following pipeline [2]: (i) preprocessing, which consists in improving target contrast and reducing noise and clutter (ii) target detection, *i.e.* the process of localizing the area in an image where a target is likely to be present, often done by computing image regions with high contrasts (iii) segmentation, which consists in accurately extracting the potential targets from the background and (iv) recognition, consisting in extracting visual features from these potential target and finally classifying them.

Modern approaches for automatic object detection uses a rather different paradigm. Indeed, they try to avoid taking intermediate decisions by relating directly the input space with the final decision space and make extensive use of machine learning techniques. Two prototypical examples are the face detector of Viola and Jones [3] based on the use of Haar wavelets and a cascade of boosted classifiers and the Dalal and Triggs's pedestrian detector [4] using Histogram of Oriented Gradients (HOG) combined with Support Vector Machine (SVM) classifiers. The popular bag-of-words model [5] has also been used successfully for object detection [6]. The combination of such efficient machine learning algorithms with discriminative features is the foundation of modern object

[☆] This paper has been recommended for acceptance by M.T. Sun.

[*] Corresponding author at: Safran, 1 rue Genevive Aube, 78114 Magny-les-Hameaux, France.

E-mail address: sebastien.razakarivony@safran.fr (S. Razakarivony).

http://dx.doi.org/10.1016/j.jvcir.2015.11.002
1047-3203/© 2015 Elsevier Inc. All rights reserved.

detection algorithms. More improvement has also been done recently by using more complex object models, such as the Deformable Parts Model [7].

One reason to explain the progress in this field is the release of publicly available datasets allowing the development, the evaluation and the comparison of new algorithms in realistic conditions. PASCAL VOC [8] benchmark provides one of the key datasets for object detection. From 2005 to 2013, yearly evaluation campaigns have been organized. The detection competitions of PASCAL VOC consist in predicting the bounding box and the label of each object from twenty possible target classes in the test image. In 2012, a total of more than 10,000 annotated images were available for training and validation. Several other datasets, presented in the related work section, are available for the evaluation of different detection tasks (*e.g.* person detection, face detection), such as ImageNet [9] or LabelMe [10].

However, none of these datasets is actually adapted to ATR. Indeed, one specificity of ATR is to require the detection of small targets while these dataset includes objects whose size in images is usually bigger than 200 pixels and can be the main topic of the image. These recent datasets are more concerned by the diversity of object appearance, articulated objects, number of categories than by target size, image noise, multi-spectral images, sensor technology.

On the other hand, and as far as we know, none of the recent approaches for object detection (*e.g.* [4,7,11]) have been evaluated in the context of ATR.

Within this context, the motivation for this paper is twofold. First, the paper introduces VEDAI (Vehicle Detection in Aerial Imagery), a new database designed to address the task of small vehicle detection in aerial images within a realistic industrial framework.[1] This dataset was made to help the development of new algorithms for aerial multi-class vehicle detection in unconstrained environment, giving the possibility to evaluate the influence of image resolution or color band on detection results. Images includes various backgrounds such as woods, cities, roads, parking lot, construction sites or fields. In addition, the vehicles to be detected have different orientations, can be altered by specular spots, occluded or masked. No specific constraints were put on the types of vehicles. This diversity of backgrounds and vehicle appearances will allow to make progress in the field of automatic scene analysis, scene surveillance and target detection. Second, we benchmark some baseline algorithms and show their performance on the proposed dataset, to allow people to have some point of comparison.

The organization of the paper is as follows. After presenting the related works in Section 2, we introduce the dataset (*i.e.* the images, the vehicle classes as well as the background types, the annotations and the organization of the dataset) in Section 3. To make comparisons between algorithms possible, we give in Section 4 the evaluation protocol associated with the dataset. We finally present in the last section (Section 5.2) experiments in which baseline algorithms are evaluated on the dataset, giving baseline results and some analysis of the influence of the parameters on the performance.

2. Related works

Object detection – often considered as being one of the most challenging computer vision task – has a long history in the computer vision literature. This section focuses on three aspects of the problem, namely (i) the datasets publicly available to develop, validate and compare object detectors, (ii) the different ways to measure detection performance, (iii) the current state-of-the-art approaches for object detection.

2.1. Databases for object detection

Modern approaches in computer vision rely on machine learning and require annotated training data. In addition, there is also an increasing need for comparing approaches with each other and establishing what are the most promising avenues. The consequence is that a lot of new datasets have been recently produced and made publicly available. If most of them are related to object/scene recognition (*e.g.* [12–16]) – which is related to our problem but covers different needs – only a few of them specifically address object detection.

More precisely, datasets for detection usually fall into the following categories: (i) pedestrian detection (ii) face detection (iii) detection of everyday objects (iv) vehicle detection. A summary of these datasets is given Table 1.

2.1.1. Person/pedestrian detection

This is one of the very popular detection task, probably because of the large number of applications (surveillance, indexing, traffic safety, *etc.*) that may result. The INRIA person dataset, first introduced in [4], contains several hundreds cropped images of humans with different resolution (64 × 128, 70 × 134, 96 × 160). Images are also provided with the whole background and the base is separated in train and test sets. The images come from various sets of personal photos and a few from the web. The people appear in any orientation and among a wide variety of backgrounds. Many people are bystanders taken from the backgrounds of the input photos, so ideally there is no particular bias in their pose. This dataset was introduced because the previous dataset of reference – the MIT person dataset [17] – was not challenging enough.

The CalTech pedestrian dataset [18] has been introduced once the INRIA person dataset was considered to be too small and too easy and addresses more specifically the case of pedestrian detection. It is a collection of images taken from a vehicle driving through regular traffic in an urban environment. It contains 350,000 labeled pedestrian bounding boxes in 250,000 frames. Occlusions are annotated with a two bounding box system and annotations are linked between frames, forming tracks.

The increase of the number of images between these two datasets (from hundreds to thousands) reflects a current trend in the production of datasets.

2.1.2. Face detection

Face detection is another well known detection task. Contrarily to pedestrian detection and despite the fact that it is often considered as an important task related to interesting applications such as security or safety, only a few datasets exist. Most of the existing face-related databases are indeed oriented toward face recognition (*e.g.* [15]) and not face detection. The CMU-MIT dataset [19], which includes the MIT dataset [20], is one of the dataset of reference, extensively used in the past. It contains only 130 different images for a total of 507 different faces (front view only). Moreover, this dataset is small and the evaluation protocol and the metric are not clearly defined. The results presented by the numerous papers using it cannot be compared in a reliable way, as noticed by Hjelmås and Low [21]. More recently, Kodak has compiled and released a new image database for benchmarking face detection and recognition algorithms [22]. This database has 300 images of different sizes, the size of faces in images varying from 13 × 13 pixels to 300 × 300 pixels. Finally, the most used and well-known dataset in face detection is Face Detection Dataset Benchmark (FDDB, [23]). It is made from images extracted from Faces in the

[1] The database can be downloaded at https://downloads.greyc.fr/vedai/.