Pattern Recognition Letters 36 (2014) 154–160

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Scene transformation for detector adaptation

Liwei Liu [a,*], Junliang Xing [b], Genquan Duan [a], Haizhou Ai [a]

[a] Computer Science and Technology Department, Tsinghua University, Beijing, China
[b] Institute of Automation, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 17 January 2013
Available online 16 October 2013

Communicated by A. Shokoufandeh

Keywords:
Vehicle detection
Detector adaptation
Scene transformation
Automatic parameter estimation

ABSTRACT

This paper focuses on detecting vehicles in different target scenes with the same pre-trained detector which is very challenging due to view variations. To address this problem, we propose a novel approach for detection adaptation based on scene transformation, which contributes in both view transformation and automatic parameter estimation. Instead of modifying the pre-trained detectors, we transform scenes into frontal/rear view handling with pitch and yaw view variations. Without human interactions but only some general prior knowledge, the transformation parameters are automatically initialized, and then online optimized with spatial–temporal voting, which guarantees that the transformation matches the pre-trained detector. Since there is no need of labeling new samples and manual camera calibration, our approach can considerably reduce manual interactions. Experiments on challenging real-world videos demonstrate that our approach achieves significant improvements over the pre-trained detector, and it is even comparable to the performance of the detector trained on fully labeled sequences.

1. Introduction

Vehicle detection in video sequences is of fundamental importance which provides strong observation models for many high level traffic surveillance applications such as traffic analysis, abnormal trajectory detection and collision avoiding. The difficulties behind vehicle detection, however, are also pronounced because of variations in views, resolutions, illuminations and backgrounds. In practical application, a pre-trained vehicle detector often performs worse in general scenes than in the training scene, and furthermore, sometimes it doesn't work at all. The major reason that results in performance decrease is the view variation between training and testing scenes.

In recent years, the fast development of object detection techniques has resulted in many promising methods for detecting particular object classes, e.g., faces (Viola and Jones, 2004; Huang et al., 2007), pedestrians (Wu and Nevatia, 2007; Dalal and Triggs, 2005) and vehicles (Kuo and Nevatia, 2009; Song et al., 2008; Liu et al., 2012). Compared with techniques based on background subtraction (Kamijo et al., 2000), detection based techniques are more robust to lighting variations. But robust solutions of object detection for piratical applications need further research due to view variations in diverse scenes. Aiming at modeling the characteristics of samples from target scenes with few manual interactions, a popular trend is to design a labeler to select positive and negative samples from a target scene to retrain a scene specific detector (Ali et al., 2011; Kalal et al., 2010; Wang and Wang, 2011) or modify a general detector (Jain and Learned-Miller, 2011).

Besides the above approaches, there are a few attempts based on sample transformation to adapt the pre-trained detectors. In Li et al. (2008), the proposed 3D search approach significantly improved detection performance. At each grid point, a rectified subimage is generated to approximate the orthogonal projection of the samples in which the pre-trained detector can be applied. As for most surveillance videos, the camera parameters are unknown which does not meet its requirement. Our work is inspired by the approach of Li et al. (2008). However, we focus on rigid targets such as vehicles in general scenes without camera parameters, and hence there are more difficulties should be handled with: (1) beside the pitch view variation like pedestrians, vehicles also varies a lot in different yaw views; (2) vehicle detection is potentially applied in general surveillance scenes of which the camera parameters might be unknown or imprecise.

In order to address the above-mentioned problems, we propose a novel approach based on scene transformation, rather than training a scene specific detector or modifying a pre-trained detector. The system overview can be summarized as follows. First, preliminary camera calibration is performed by exploiting scene information. With the camera parameters, the scene transformation framework is initialized. And then, the spatial–temporal voting guides the procedure of parameter optimization iteratively. Finally, we can obtain the optimal parameters and their scene transformation framework for the pre-trained detector. Accordingly, our system consists of two key components: scene transformation modeling and camera parameter initialization and optimization.

[*] Corresponding author. Tel.: +86 10 62795495; fax: +86 10 62795871.
E-mail address: llw09@mails.tsinghua.edu.cn (L. Liu).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.10.004

The main contributions of this paper include: (1) Scene transformation is carried out for detector adaptation in general scenes, which is both robust to pitch and yaw view variations (some training and testing samples shown in Fig. 1). (2) Automatic parameter estimation fuses parameter initialization and optimization without manual interactions, which guarantees the best utilization of the pre-trained detector in our transformation framework.

The remainder of the paper is organized as follows. Section 2 will provides the details of scene transformation modeling. And then we will introduce the camera parameter initialization and optimization in Section 3. Experiments are carried out in Section 4 and finally the conclusion is given in the last section.

2. Scene transformation modeling

View transformation in an unknown scene is really a challenging task due to the following reasons: (1) Even in the same scene the view variations of vehicle are different in different positions. (2) In most cases, a vehicle has more than one view variations all of which can be decomposed into pitch and yaw variations. Our objective is to transform the current views of target scenes to the referential views of training scenes which the pre-trained detector can capture.

To address these problems and follow the idea, we propose the scene transformation modeling of which an overview is shown in Fig. 1(d). We actually transform sub-images in searching grids with different parameters. Given searching points, referential views will be selected in training scenes. And then corresponding pitch and yaw view transformation are performed on source images to approximate to the views of training samples. In the next step, the pre-trained detector can work well in the transformed sub-images. Finally, the detection results are projected back to the source images.

Before introducing the scene transformation, we will first list the denotations. Given the camera position at $\mathbf{P}_c$, at each point of the searching grid, we generate a transformed sub-image Q from the source image I. W is their corresponding 3D coordinate, which is also the bridge between Q and I. Homogeneous coordinates are employed to denote points in Q, I and W by $\mathbf{q} = (u_Q, v_Q, 1)^T$, $\mathbf{p} = (u_I, v_I, 1)^T$ and $\mathbf{P} = (x, y, z, 1)^T$. In our problem, their relations can be denoted as

$$\mathbf{p} = \mathbf{HP} = \mathbf{HMq} = \mathbf{HM_pM_yq}, \tag{1}$$

where $\mathbf{H}$ is the camera matrix, $\mathbf{H} = \mathbf{A}[\mathbf{R}|\mathbf{T}]$. $\mathbf{M}$ is the transformation matrix which contains the pitch view transformation matrix $\mathbf{M}_p$ and yaw view transformation matrix $\mathbf{M}_y$. In our framework, all the transformed sub-images are size-fixed. We enumerate all the pixels of the transformed image, use Eq. (1) to get their corresponding positions of the source image and obtain the color values, finally we can get the transformed image. In the following, we will mainly introduce the pitch view transformation and the yaw view transformation.

Pitch view transformation is carried out to approximate the orthogonal projection of samples. With the camera position $\mathbf{P}_c = (x_c, y_c, z_c, 1)^T$, the search grid point $\mathbf{P}_o = (x_o, y_o, 0, 1)^T$ and its desired projected position $\mathbf{q}_o = (u_o, v_o, 1)^T$ in Q, the pitch view transformation matrix $\mathbf{M}_p$ in the search grid can be written as

$$\mathbf{M}_p = \begin{pmatrix} \cos\theta & 0 & -u_o\cos\theta + x_o/\alpha \\ \sin\theta & 0 & -u_o\sin\theta + y_o/\alpha \\ 0 & -1 & v_o \\ 0 & 0 & 1/\alpha \end{pmatrix}, \tag{2}$$

where α is the size ratio between the real world object and the transformed image patch size of height (after normalization). In our experiments, we set α to be 1700 mm / 50 pixel since vehicles can be normalized to about 50 pixels (higher than 1700 mm has more than 50 pixels and vice versa) which facilitate the detection procedure for appropriate search range. Derived from Eq. (1) and Eq. (2), we can obtain the relation between the 3D coordinate and the transformed image coordinate,

$$\mathrm{P} = \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = \mathrm{M}_p\mathrm{q}' = \begin{pmatrix} \alpha(u' - u_0)\cos\theta + x_0 \\ \alpha(u' - u_0)\sin\theta + y_0 \\ \alpha(v_0 - v') \\ 1 \end{pmatrix},$$

where $\mathrm{q}' = \mathrm{M}_y\mathrm{q} = (\mathrm{u}', \mathrm{v}', 1)$. We get x and y coordinates by rotating the transformed image and z coordinate from the y coordinate of the transformed image. And θ is the angle between Q and the x–z plane in world frame,

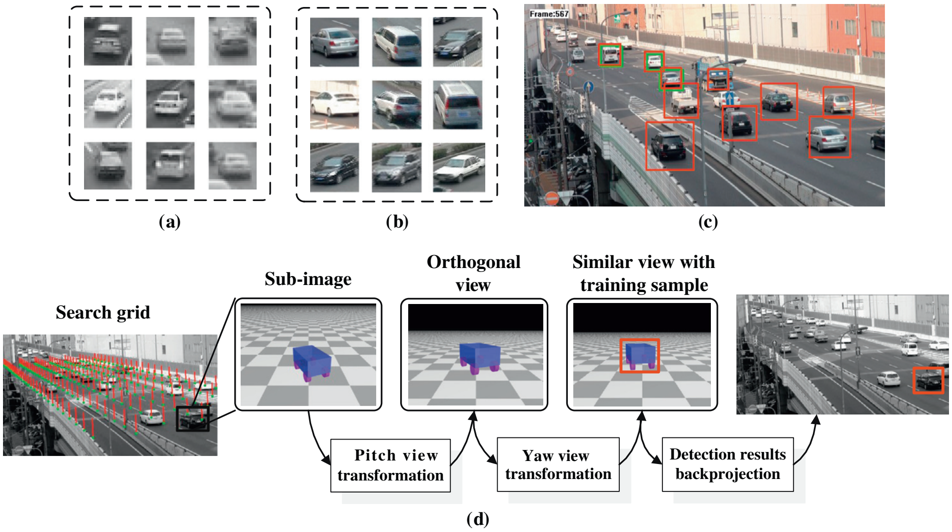
(a)
(b)
Frame:567
(c)
Search grid
Sub-image
Orthogonal view
Similar view with training sample
Pitch view transformation
Yaw view transformation
Detection results backprojection
(d)

Fig. 1. (a) training samples; (b) testing samples; (c) detection in target scenes with large view variations. (green: results of the pre-trained detector; red: results of our approach with the same detector). (d) the scene transformation framework: transform the testing samples' views to the training samples' in search grids, then detect vehicles with the pre-trained detector, finally back-project to source images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)