Neurocomputing 149 (2015) 800–810

Contents lists available at ScienceDirect

Neurocomputing

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

Cognitive pedestrian detector: Adapting detector to specific scene by transferring attributes

Xu Zhang, Fei He, Lu Tian, Shengjin Wang*

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 2 January 2014
Received in revised form
18 June 2014
Accepted 28 July 2014
Communicated by M. Wang
Available online 21 August 2014

Keywords:
Detector adaption
Transfer learning
Sparse coding
Pedestrian detection

ABSTRACT

Training a reliable generic pedestrian detector on different scenes is still a very challenging problem. In this paper, we propose a novel transfer learning framework for improving the performance of a generic detector by adapting the detector to a scene specific detector. The main contributions come from 2 aspects: (1) instead of hand-crafted ad-hoc rules, a scene based auxiliary attribute classifier and a position priori map are automatically trained from target scene to collect confident samples; (2) conditional distribution transfer sparse coding is presented to match the conditional distributions of the source and the target samples. Experiments show our approach significantly improves the performance of the generic detector and outperforms the state-of-the-art adapting approaches in benchmark datasets. Comparing with the state-of-the-art methods, the improvements are 6% on the CUHK square pedestrian dataset and 33% on the ETH pedestrian dataset which is considered quite hard because the background is dynamic.

1. Introduction

Pedestrian Detection has achieved significant improvements in the past decades [1]. However, it is still very challenging to train a reliable generic detector for all kinds of situations, due to various view points, illuminations, cluster backgrounds and so on. Training such a detector requires both a sufficient large dataset to cover all the variations and a very complex model to fit all the samples. Performances of the state-of-the-art detectors trained on one dataset may drop significantly on another [2]. However, considering possible variations for a specific application is much more practical. For example, in a particular surveillance video, the camera is always static, the variations of the view points and the backgrounds are limited. Even for dynamic cases, such as video captured by cameras located on cars, the view points of pedestrians are limited. As a result, it is much easier to train a reasonable detector using samples from the specific scene, but repeating labeling samples in every scene requires much effort. Automatically collecting positive and negative samples from target scene and using those samples to enhance the performance of the generic detector in specific scene attracts more and more attentions. Many existing methods present hand-crafted ad-hoc rules (such as background subtraction and pedestrian path model) to collect confident samples. However, hand-crafted ad-hoc rules still strongly depend on the scene itself. In this paper, we tackle this problem by proposing a knowledge based transfer learning framework which concentrates on finding shared attributes of the pedestrian samples and matching the conditional distributions of the samples in different datasets. Bridging source (generic training samples) and target (specific scene samples) samples with semantic feature is a known phenomenon in human learning. Knowledge based transfer leaning tries to imitate this function of human brain.

2. Related work

Existing works of adapting a generic detector to specific scenes always train an automatic labeler to select positive and negative samples from target scene and then enhance the generic detector with those samples. The automatic labeler should have high accuracy and offer additional information. Rosenberg et al. [3] chose confident samples classified by the generic detector to retrain the generic detector. Since it could not offer auxiliary information, the algorithm was easily to drift. Background subtraction method is widely used to distinguish backgrounds and moving pedestrians [4,5]. However, the accuracy of the background subtraction labeler is low, especially when the scene is crowded and there are other moving foregrounds (such as cars) in the scene. Wang et al. [6] designed multiple context cues including pedestrian path models [7], locations, sizes and appearances of the

* Corresponding author. Tel.: +86 010 62773710 303.
E-mail address: wgsgj@tsinghua.edu.cn (S. Wang).

http://dx.doi.org/10.1016/j.neucom.2014.07.054
0925-2312/

pedestrians to select confident samples. The main disadvantages of their approach were that some critical context cues were invalid when the background was moving and the well-designed ad-hoc rules only worked for traffic scenes. Co-training framework [8,9] trained two detectors with different features. Results of one detector could be training samples of another detector. The two detectors are trained iteratively until convergence. Co-training framework requires two detectors that are highly independent, otherwise it might cause drifting problem. The design of the independent detectors is very difficult. Dalal et al. [10] has already proved that appearance detector and motion detector are highly correlated.

Transfer learning [11] provided new sights into this area, this kind of methods tried to match the distributions of source and target datasets. Domain adaption achieved that goal by re-weighting source samples according to the similarities with the target samples, and then enhancing the influences of the source samples which were similar to the target samples and reduced the influences of the unlike ones. Cao et al. [12] presented ITLAdaboost to adjust the weights of the source samples. However, their method required a small set of labeled target samples. Wang et al. [13] proposed a unsupervised transfer learning framework to tackle this problem. The proposed Confidence-Encoded SVM provided a more reliable way to re-weight the source and the target samples and proved to have a faster convergence speed than method in [6]. However, their approach still depended on context information [6] to collect samples. The main disadvantage of domain adaption methods is that it requires the source samples set to cover as many ambiences as possible thus to be able to always find an effective subset of samples having similar appearance or attribute for any specific target scene. Collecting such a large source dataset needs great effort and only a few of the source samples are taken into consideration for a specific target scene. Our experiments in Section 4.3 also show that samples with low confidence scores are the most informative samples and reducing their influences might lower the performance of the final detector.

Sparse coding has been a hot research topic in computer vision recently. It is a powerful tool for finding effective representations of images. Sparse coding can represent images with only a few active elements thus can extract high level semantics in data. In the past few years we observed some literatures integrating sparse coding to transfer learning framework. Quanz et al. [14] explored sparse coding to extract effective features for knowledge transferring. Long et al. [15] proposed adding Maximum Mean Discrepancy (MMD) as a optimization term to the objective function of sparse coding to reduce the divergence between two distributions. MMD requires universal kernel, however [15] chose linear kernel to simplify the problem. Linear kernel only provides a very week measurement, which just matches the means of two distributions. In pedestrian detection problem, positive samples and negative samples may have totally different conditional distributions, just matching the joint distributions does not guarantee small discrepancy between the conditional distributions. There have been very few works done on detector adaptation by sparse coding based transfer learning. Liang et al. [16] proposed sparse coding to re-weight the source and the target samples. However, they still filtered target samples with ad-hoc rules.

Visual attributes are human-understandable properties shared among object categories and are treated as "mid-level" features bridging the gap between "low-level" image features and "high-level" object categories. Classemes [17] is one of the most popular attribute features trained on external image data on the Web. It contains 2659 dimensions and each one is an output of a weakly classifier representing a semantic concept. Recent researches have shown appealing results in image representation [18], image retrieval [19] and zero-shot transfer learning [20,21] by applying attribute features. Although the pedestrian samples in two different datasets may vary a lot in appearance, they can share a lot of properties in attribute space. Successful approach of attribute-based transfer learning can be found in human pose estimation [22].

Transferring a generic detector turns even more challenging when the target scene is dynamic. Most existing works require static background assumption when designing ad-hoc rules. Moving background makes it impossible to collect confident positive and negative samples. Chang and Cho [23] proposed online boosting to improve the performance of a generic detector. However, they only used an ideal "strong detector" to get true labels of target samples. CovBoost proposed by Pang et al. [24] shared similar idea with our work. It transferred the detector between different view points by adapting weights of weak classifiers. Since the SVM based methods always perform better than the boosting methods for pedestrian detection (only a few works [25] reported the exception), the overall performance of their final pedestrian detector is much lower even than our generic detector.

The main contributions of this paper come from 2 aspects: (1) instead of hand-crafted ad-hoc rules, a scene based auxiliary attribute classifier and a position priori map are automatically trained to collect confident samples in specific scene. (2) since positive and negative samples can have quite different conditional distributions, we present conditional distribution transfer sparse coding which extends transfer sparse coding to multiple classes cases and matches the conditional distributions of samples in corresponding categories. By using conditional distribution transfer sparse coding, a more reliable attribute classifier can be trained. Experiments show our approach can significantly improve the performance of the generic detector even better than the detector trained with manual labeled target samples. Our approach also outperforms the state-of-the-art automatic adaptation methods and has much wider application scenarios. We call the detector which can learn from circumstance as cognitive detector.

3. Cognitive pedestrian detector

3.1. Overview

Our algorithm starts with a generic HOG + SVM pedestrian detector F_g learned from a general pedestrian dataset. In transfer learning problem this general dataset is called source dataset $\mathbf{D}_s$. Then the generic detector F_g is applied to some unlabeled frames captured from the target scene. All the samples from the target scene with positive response or very low negative response (close to the classification plane) are selected to form the target dataset $\mathbf{D}_t$. Since the generic detector is far from perfect, the target dataset is very noisy. Other than designing the hand-crafted ad-hoc rules, an auxiliary attribute classifier and a position priori map are learned to label the target samples.

The position priori map is learned to label the samples by selecting out the similar patterns appear in the background. And an attribute classifier is used to label the rest of the ambiguous samples. In order to get a reliable attribute classifier, we choose classemes [17] as the original attribute feature and present conditional distribution transfer sparse coding (CDTSC) to find a shared attribute subspace in which the source and the target samples can have similar distributions. Finding shared features can greatly improve the performance of the classifier [26]. After all the samples are coded, the attribute classifier can be trained. The newly labeled target samples and all the source samples are used to train the final detector. The diagram of our approach is shown in