Neurocomputing 184 (2016) 55–65

Contents lists available at ScienceDirect

Neurocomputing

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

Multi-sparse descriptor: A scale invariant feature for pedestrian detection

Yazhou Liu [a,*], Pongsak Lasang [b], Mel Siegel [c], Quansen Sun [a]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[b] Panasonic R&D Center Singapore, 534415 Singapore, Singapore
[c] Robotics Institute, Carnegie Mellon University, Pittsburgh 15213-3891, USA

ARTICLE INFO

Article history:
Received 22 January 2015
Received in revised form
20 May 2015
Accepted 18 July 2015
Available online 25 December 2015

Keywords:
Local descriptor
Sparse coding
Scale invariance
Pedestrian detection
Multi-dictionary learning

ABSTRACT

This paper presents a new descriptor, multi-sparse descriptor (MSD), for pedestrian detection in static images. Specifically, the proposed descriptor is based on multi-dictionary sparse coding which contains unsupervised dictionary learning and sparse coding. During unsupervised learning phase, a family of dictionaries with different representation abilities is learnt from the pedestrian data. Then the data are encoded by these dictionaries and the histogram of the sparse coefficients is calculated as the descriptor. The benefit of this multi-dictionary sparse encoding is three-fold: firstly, the dictionaries are learnt from the pedestrian data, they are more efficient for encoding local structures of the pedestrian; secondly, multiple dictionaries can enrich the representation by providing different levels of abstractions; thirdly, since the dictionaries based representation is mainly focused on the low frequency, better generalization ability along the scale range is obtained. Comparisons with the state-of-the-art methods reveal the superiority of the proposed method.

1. Introduction

Detecting human in images/videos has received more attentions in recent years, and its potential applications include smart surveillance, driving assistance, robotics and content based image/video management. Even though remarkable progress has been made [1–16], finding the human is still considered as one of the hardest task for object detection. The difficulties come from the articulation of human body, the inconsistency of clothes, the variation of the illumination and the unpredictability of the occlusion.

Human detection falls into the general object detection category and typically it contains two main processing steps [6]: feature extraction and classification. The flow of the scan window based pedestrian detection is illustrated in Fig. 1. This work focus on the feature extraction step, as shown in Fig. 1(d). The details of the proposed method is presented in Section 3, as illustrated in Fig. 3. Generally, the target of feature extraction is to map the data from its original space to some feature space so that the data belong to the same category are more compactly distributed; then a classifier is learnt in this feature space to represent the distribution or the boundary of the data that belong to a specific category. Since the difficulties mentioned above lead to a large intra-class variation [17], the feature extraction step is even more important for human detection, and extracting the most informative features is critical for the detectors' performance. There exist extensive literatures on human feature extraction. The works in [2,6,8,10,14,18–21] are the only a few of the promising methods that have been invented recently.

Besides the difficulties mentioned above, scale is another important factor that might affect detection accuracy essentially [22–24]. Sliding window is one of the most successful approaches for object detection. Within this framework, to detect the pedestrian in multiple scales, one can apply a single detector to a multi-scale image pyramid [2,13] as shown in Fig. 1(b), or use a multi-scale detector on a single image [25,26], or do them both [23]. A basic assumption behind these approaches is that a detector is trained at a fixed canonical scale can be generalized to all the resolutions [27]. However, the sensors with finite resolution will lead to serious information loss especially for the low resolution targets. When rescaling small examples into large sizes blurring artefacts appear. This scale invariant assumption might leads to severe performance decrease when people appear at small scales [24].

Some works have addressed this issue from different aspects. Park et al. [28] present a multiresolution representation of pedestrians. Based on the histogram of oriented gradient (HOG) feature [2], they build multiple model templates with different feature granular which is capable of representing the pedestrian

* Corresponding author.
E-mail addresses: yazhouliu@njust.edu.cn (Y. Liu), Pongsak.Lasang@sg.panasonic.com (P. Lasang), mws@cmu.edu (M. Siegel), sunquansen@njust.edu.cn (Q. Sun).

http://dx.doi.org/10.1016/j.neucom.2015.07.143
0925-2312/

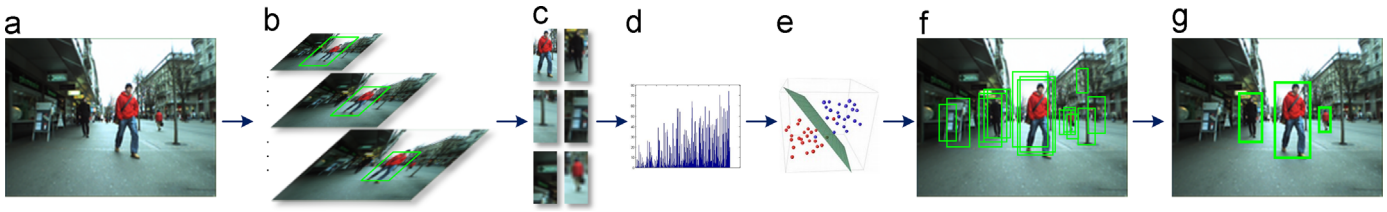

Fig. 1. The pipeline of the scan window based pedestrian detection. (a) Input image. (b) Image pyramid. (c) Scan windows. (d) Feature representation. (e) Classifier. (f) Classification results. (g) Detection results (After grouping and non-maximum suppression).

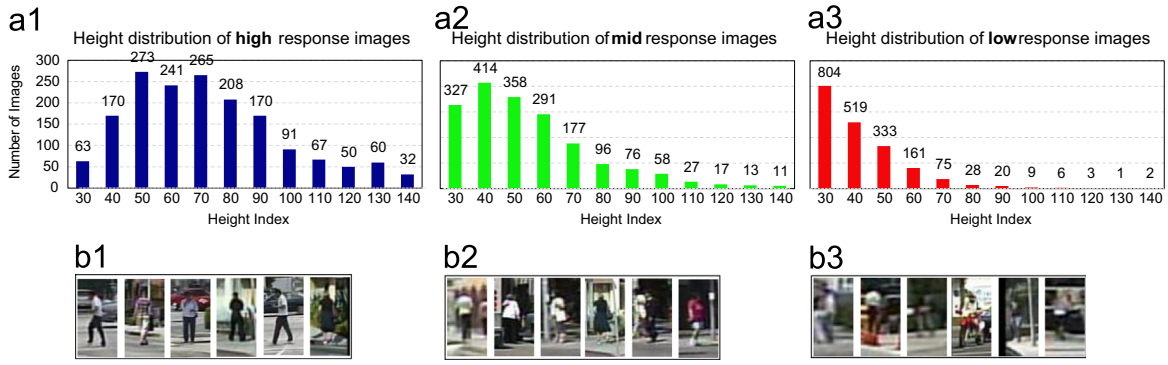

Fig. 2. The relation between HOG detection responses and sample scales. (b1) are the samples from Caltech dataset [24] with *high response* of the HOG detector and (a1) is the scale distribution of these *high response* samples. (a2)–(b2) and (a3)–(b3) are the results of mid and low response samples.

with different details. Recently, Yan et al. [27] take pedestrian detection in different resolutions as different but related problems, and propose a multi-task model to jointly consider their commonness and differences. Based on the HOG feature and deformable part model (DMP) [4], they map the features of different resolutions into the common subspace, so that the features in this space share the same detector.

All of these methods utilize some forms of gradient histogram. As one of the most successful descriptor, HOG is original developed and extensively evaluated on a relatively high resolution image dataset INRIA [2]. However, the feasibility of extending this descriptor to low resolution data is actually questionable. HOG is a statistical representation of the gradients, which corresponds to the high frequency information of the data. For the low resolution data, because of the limited sampling, the high frequency components are either degenerated by aliasing or blurred by resizing. Therefore, representing the low resolution objects by HOG might not yield satisfiable results.

An example is shown in Fig. 2. A HOG detector trained on INRIA dataset [2] is tested on the Caltech pedestrian dataset [24]. All the pedestrian samples in the Caltech dataset are ranked according their detection scores/responses. Higher detection score indicates higher probability of being detected as a human. Then we randomly select 2000 positive samples with high scores, middle scores and low scores respectively and plot the corresponding scale distribution for each category. (b1)–(b3) are the sample images from high to low response categories. Intuitively, high response images have sharp edges while the low response images are more likely to be the blurred ones. (a1)–(a3) illustrate the scale distribution of each category. These figures further verify our previous observation. Statistically, the high score samples tend to have larger scale and vice versa. These results indicate that the scale variation has big impact on the performance of HOG, especially for the small scale objects.

The motivation of this paper is to find a new descriptor that is less affected by the scale variation. Comparing with HOG, this descriptor ideally should have following properties. Firstly, it should have comparable or even better representation ability as HOG. Secondly, it should have better generalization ability along the scale range. Thirdly, it can capture the complementary information of HOG and yield better performance when combining with HOG.

Based on above motivation, we present a multi-sparse descriptor (MSD) for pedestrian detection. For this new descriptor, the local structures of the pedestrian are encoded by multiple dictionaries that are learnt in an unsupervised way from the pedestrian data. During the unsupervised learning phase, each dictionary is constraint to represent the object with specific abstraction level which can be explicitly controlled by a sparsity parameters. By varying$s$, a series of dictionaries with different representation abilities are learnt. The benefit of this multi-sparse descriptor is that it can adapt to different abstraction level to represent the varying characteristics of the complex objects. In addition, unlike the HOG, the information captured by the MSD is mainly focused on the low frequency domain. Based on these properties, better generalization ability along the scale range is obtained.

The rest of the paper is organized as follows: Section 2 gives a brief summarization of the state-of-the-art human detection methods; Section 3 introduces the basic idea and formulation of the proposed method; Section 4 gives the extensive experimental evaluation; and the conclusion is presented in Section 5.

2. Related works

In this section, we review the pedestrian detectors with a focus on their feature extraction approaches. Comprehensive surveys and evaluations can be found in [24,29–31]. Especially, a recent work by Dollar et al. [24] gives extensive evaluations and new insights of the state-of-the-arts. 16 representative detectors have been evaluated using a unified evaluation framework on six public pedestrian data sets. Their study also shows the further research directions which include small scales and occlusion.