Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

# Segmentation by weighted aggregation and perceptual hash for pedestrian detection ☆

Yifeng Liu [a], Lian Zou [a,*], Jie Li [a], Jia Yan [a], Wenxuan Shi [b], Dexiang Deng [a]

[a] *School of Electronic Information, Wuhan University, Wuhan 430072, China*
[b] *School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China*

ARTICLE INFO

*Article history:*
Received 26 May 2015
Accepted 15 January 2016
Available online 23 January 2016

*Keywords:*
Pedestrian detection
Intra-class variation
Segmentation by weighted aggregation
Probability measure
Perceptual hash
BING
Estimation proposals
HASP

ABSTRACT

Main challenges of pedestrian detection are caused by the intra-class variation of pedestrians in clothing, scales, deformations, occlusions, and backgrounds. The prevalent detection frameworks employ a series of dense sliding windows, which are time-consuming. In this work, we equip the detection framework with another new strategy, and extract the new features, to eliminate the above requirements. Segmentation by weighted aggregation (SWA) provides a probability measure to segment objects from complex backgrounds. Perceptual hash (pHash) has shown its power in similar image retrieval because it is modification-tolerant and scale-invariant. The proposed approach uses binarized normed gradients (BING) to efficiently generate a small set of estimation proposals, and formulates SWA and pHash into a joint descriptor, called HASP, to improve the detection performance significantly. Experimental results both on INRIA dataset and ETH dataset have demonstrated the effectiveness and efficiency of the proposed approach.

## 1. Introduction

Over the last few years, pedestrian detection has attracted considerable attention in computer vision community due to its wide industrial applications and important scientific values. Pedestrian detection represents an important task in a wide range of applications. The most important of them are: intelligent video surveillance [1], automotive safety [2], biometrics [3], traffic [5], mobile robots [4], and video retrieval [6].

Deformation, occlusion and scale variation are common problems in objection detection; however, detecting pedestrian is more challenging because of more variations. Variations arise not only from changes in illumination and viewpoint, but also due to non-rigid deformations and intraclass variability in shape and other visual properties. For example, people stand in complex backgrounds, wear different clothes and take a variety of poses, as shown in Fig. 1. Segmentation by weighted aggregation (SWA) [18] provides a probability measure to assess whether or not two neighboring regions should be included in the same segment, and this graph based method can effectively reduce the influence by the clothing and backgrounds. Perceptual hash (pHash) [19] is a popular method for similar image retrieval. It can tolerate the minion modifications, and extract the principal component from image with any scales. Therefore, we consider applying them to extract the features. In other words, we perform some potential cues aggregation at the feature level.

Besides, there is a remaining technical issue in the testing: the prevalent detection frameworks employ a series of dense sliding windows, which have a lot of overlaps among the neighbors. Obviously, the way of sliding windows is time-consuming, because it repeatedly applies the feature computation to the raw pixels of thousands of common regions per image. Our original motivation for using regions was born out of a pragmatic research methodology: move from image classification to object detection as simply as possible. Binarized normed gradients (BING) proposed in [20] can efficiently estimate objectness at very high speed. So we merely compute the proposal regions.

In this paper, we show that our joint approach can run the detection only in several regions, and then extract the effective features by our joint descriptor. An overview of our proposed approach is shown in Fig. 2. The steps of our approach are summarized as follows.

1. BING is utilized to propose a small set of candidate bounding boxes instead of sliding windows. BING can be used for efficient objectness estimation at 300 fps, which requires only a few atomic operations (i.e. ADD, BITWISE, etc.).

☆ This paper has been recommended for acceptance by M.T. Sun.
* Corresponding author.
*E-mail address:* zoulian@whu.edu.cn (L. Zou).

http://dx.doi.org/10.1016/j.jvcir.2016.01.010
1047-3203/

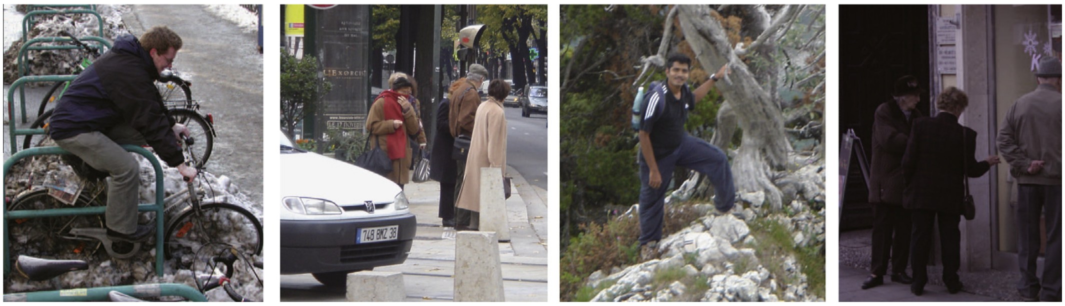

**Fig. 1.** Samples of challenging examples. Including deformations, partial occlusions, and complex backgrounds.

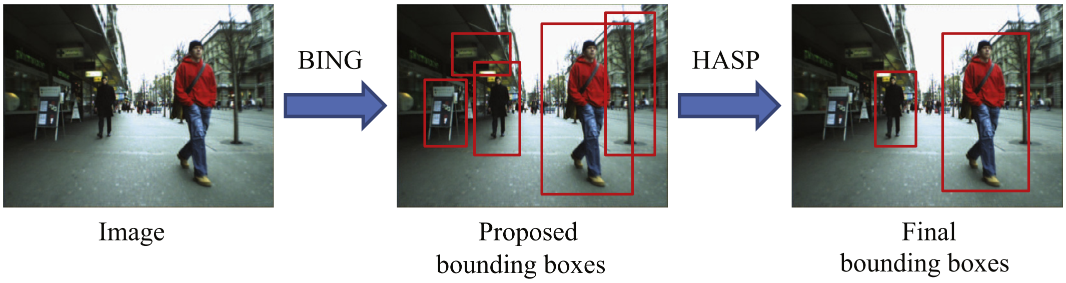

**Fig. 2.** Overview of our approach.

2. We firstly use SWA [18] to segment the objects from an image. Then the segmented image can be encoded to a hash fingerprint [19]. We jointly learn histograms of oriented gradients – adaptive local binary pattern (HOG-ALBP) [17] and SWA-pHash for classification.

Our joint descriptor, denoted by HOG-ALBP-SWA-pHash (HASP), can express textures and geometrical shapes in more details than HOG. It significantly improves the performance obtained by HOG-LBP [7], where there are partial occluded and slight deformable pedestrians, even in complex backgrounds. As a supplementary insight, we point out our joint descriptor improves the detection of small-scale pedestrians. In this paper, we place emphases on feature extraction and our contributions lie in efficiently computing pedestrian features. For simplicity and speed, we apply a stand bootstrapped strategy [21] to train the linear SVM [8,9] throughout the study.

Generally, there are two main environments for detecting: static image and video sequence. Pedestrian detection in static image is a very important prior estimate for people tracking in the video sequences. High accuracy of detection lays the foundation for subsequent studies, such as identification and tacking. It is the cornerstone for advanced people detection and tracking. This paper limits its scope to improve the performance of pedestrian detection in static images without additional tracking cues.

This paper organizes as follows. In Section 2, we review some of state-of-the-art approaches. In Section 3, we describe the proposed descriptor base on the segmentation by weighted aggregation and perceptual hash. The obtained experimental results of our joint descriptor on two datasets are illustrated in Section 4. Finally, Section 5 summarizes the conclusions that can be drawn from the presented research.

## 2. Related work

Ten years ago, Dalal and Triggs [10] proposed the histograms of oriented gradients (HOG) descriptor to obtain the robust human feature, and built a challenging INRIA dataset. Although the problem of pedestrian detection has been well studied in computer vision, it remains a challenging study due to various changes in many factors. Recently, there are a few works for pedestrian detection focusing on developing performance and accelerating computation.

In aspect of algorithm optimization, various studies have acquired good results, together with drawbacks on each other. Hu et al. [11] improved the overall recognition performance for the small-scale pedestrians, although it may not achieve an appropriate result when there is occlusion. Schwartz et al. [12] used partial least squares analysis to reduce the feature dimension to a low level, while computational cost. Chunsheng et al. [13] proposed a effective spatio-temporal based HOG, although the detection would not be effective in the richly-textured environment. Temporal differencing [14] showed its power in the non-occluded side view pedestrians; however, it failed to detect occluded pedestrian and static people from videos. Recently, deep models [29,30] showed their potential capability on pedestrian detection. The performance of unified deep net [29] had surpassed most of shallow models.

As for computation accelerating, a few studies used GPU [15,16] to parallel the integral histogram computation and the anchor sliding. Seung Eun Lee et al. proposed a hardware accelerator method to speed up HOG descriptor extraction, in order to achieve real-time pedestrian recognition with high accuracy on an embedded system. However, these methods did little optimization from the feature itself.

HOG is a good gradient descriptor in pedestrian detection, but it loses the texture information, while LBP descriptor does not. It solves the problem that HOG is vulnerable to the interference of vertical background gradient information. In our past work, We proposed the HOG-ALBP descriptor [17], which could express textures in more details than HOG. It reduced the interference of vertical background gradient cue, and had fine generalization ability. Though HOG-ALBP improves the detection, it has a few false negatives in the case of deformations, partial occlusions, and complex backgrounds (see Fig. 1). In addition, Walk et al. [31] found that LBP could not gain enhancement on some datasets because the changed imaging conditions may blur the texture information. We thought the small-scale pedestrian is another reason.