Pattern Recognition Letters 000 (2016) 1–7

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Improving pedestrian detection using motion-guided filtering

Yi Wang[a], Sébastien Piérard[b], Song-Zhi Su[c,*], Pierre-Marc Jodoin[a]

[a] Department of Computer Science, University of Sherbrooke, 2500 boul. de l'Université, Sherbrooke, QC, J1K 2R1, Canada
[b] Montefiore Institute, University of Liège, allée de la découverte 10, 4000 Liège, Belgium
[c] School of Information Science and Technology, Xiamen University, Xiamen, Fujian, 361005, China

ARTICLE INFO

Article history:
Available online xxx

MSC:
41A05
41A10
65D05
65D17

Keywords:
Pedestrian detection
Video surveillance
Motion history image
Nonlinear filtering

ABSTRACT

In this letter, we show how a simple motion-guided nonlinear filter can drastically improve the accuracy of several pedestrian detectors. More specifically, we address the problem of how to pre-filter an image so almost any pedestrian detector will see its false detection rate decrease. First, we roughly identify moving pixels by cumulating their temporal gradient into a motion history image (MHI). The MHI is then used in conjunction with a nonlinear filter to filter out background details while leaving untouched foreground moving objects. We also show how a feedback loop as well as a merging procedure between the filtered and the unfiltered frames can further improve results. We tested our method on 26 videos from 6 categories. The results show that for a given miss rate, filtering out background details reduces the false detection rate by a factor of up to 69.6 times. Our method is simple, computationally light, and can be implemented with any pedestrian detector. Code is made publicly available at: https://bitbucket.org/wany1601/pedestriandetection

1. Introduction

Despite the number of publications devoted to pedestrian detection, reliable human-shape detection is still a work in progress. Detecting humans is a difficult task since people may take very different poses, be pictured from different viewpoints, and be occluded by objects or other pedestrians. Also, many background objects have a humanoid shape thus leading to false detections. Objects such as a chair, a fire hydrant, or just a textured area which happens to have the same features than that of a pedestrian are often wrongly associated to pedestrians [11,31]. Also, human detectors are fundamentally ambivalent. A sensitive detector (one with a low decision threshold) will detect most pedestrians but at the same time non-pedestrian background objects. On the other hand, a more conservative detector (one with a higher decision threshold) will have a low false positive rate but will suffer from a large miss rate.

In this letter, instead of proposing new features or an improved pedestrian detection classifier, we focus on the images a pedestrian detector is fed with. We propose a motion-guided nonlinear filter whose goal is to filter out background details while leaving intact everything that is likely to be a pedestrian. To achieve this, we compute a motion history image (MHI) [8] at each frame. Since the content of the MHI is highly correlated with moving objects (and thus pedestrians), we apply a Gaussian filter whose standard deviation is proportional to the content of the MHI. By doing so, fixed background objects are blurred out while areas around moving objects are left untouched. We show that the number of false positives in pre-filtered images is drastically lower than in unfiltered images. The reader shall note that although our filter has been validated with pedestrian detectors, it can also be used in conjunction with other kinds of moving object detectors.

Furthermore, a feedback loop is used to update the MHI. This is done by using the predicted pedestrians to update the background image. Our system also fuses results obtained on the original frames as well as on the filtered frames to decrease even more the false positive rate.

The main contributions of this letter are:

- We propose a simple motion-guided filter which improves by a significant amount the performance of off-shelf pedestrian detectors. The filter is independent of the detector and works on a large variety of surveillance videos.
- The motion-guided filter has two novel characteristics. First, it implements a Gaussian filter whose variance is dynamically adapted to the video (cf. Section 3.2). Second, it benefits from a

* Corresponding author.
E-mail address: ssz@xmu.edu.cn (S.-Z. Su).

http://dx.doi.org/10.1016/j.patrec.2016.11.020
0167-8655/

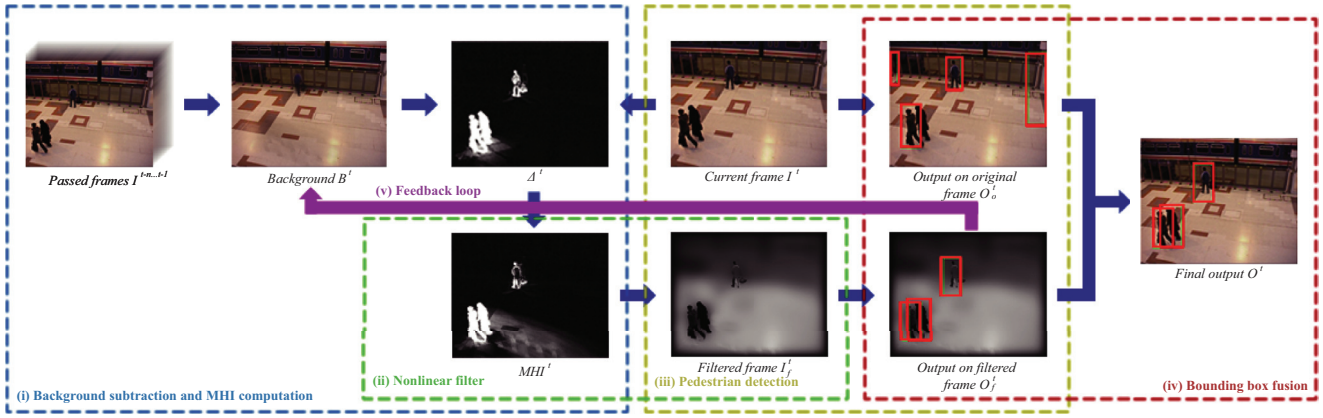

Fig. 1. Pipeline of our method. At each time t, the current frame I^t and the background image B^t are used to update the MHI^t, which is then used to filter the input image. Pedestrians are detected in the filtered image I_f^t and in I^t. The two resulting sets of bounding boxes are intersected. The ones detected in I_f^t are used to update the background image B^t.

feedback loop which takes into account the predicted bounding boxes (cf. Section 3.4).

2. Related work

As of today, top performing pedestrian detectors mostly rely on sophisticated features or discriminative classifiers [15,21,22]. At test time, these classifiers output a score indicating how confident they are that a pedestrian is located in the currently-scanned window. What differentiates most pedestrian detectors are the features and the classifiers they use. Although histogram of oriented gradients (HOG) is probably the most frequently-used feature [4,16], local binary patterns (LBP) [6] and Haar-like features [33] have also been shown effective. Since pedestrians are usually moving, several methods use spatio-temporal features such as binary motion labels [18,26] and tracking [19,28]. Other methods use richer features based on specialized hardware such as stereo [1,20] and infrared features [10,12,34]. A trend recently emerged with deep learning where features are learned instead of being handcrafted [24].

The most common classifiers used for pedestrian detection are support vector machines (SVM) [23], AdaBoost [14], Hough forests [13], and deep learning methods such as convolutional neural networks (CNN) [24].

Motion detection is also used for pedestrian detection, [32] uses Gaussian mixture model (GMM) in luma space and temporal saliency map obtained by background subtraction to extract semantic information, which is then used to adjust the pixel-wise learning rate adaptively. In [35], a video is split into spatio-temporal texture patches, in which dynamic texture is extracted. In the end, a conventional GMM is used to separate foreground motion from background image. With an advanced conditional random field model, [17] combined multiple motion and visual saliency induced features, such as shape, foreground/background color models, and visual saliency, to extract the foreground objects in videos. However, all these methods are only focused on motion detection but never extended to pedestrian detection.

3. Proposed method

As shown in Fig. 1, our method is a 5-step procedure made of: (i) a background subtraction and MHI computation (Section 3.1), (ii) a nonlinear filter (Section 3.2), (iii) pedestrian detection, (iv) bounding boxes fusion (Section 3.3), and (v) a feedback loop (Section 3.4).

3.1. Motion history image (MHI)

The first step of our method is to identify where moving objects (and thus pedestrians) roughly are. This information will later on be used to filter out background details. Motion is characterized with a temporal gradient:

$$\Delta_{x,y}^t = |B_{x,y}^{t-1} - I_{x,y}^t|, \tag{1}$$

where (x, y) denotes the coordinates of a pixel, $|\cdot|$ is the Euclidean norm in the RGB space, I^t is the video frame at time t, and B^{t-1} the background image at time $t-1$. Since in this step, the goal is to roughly detect the moving objects, B^t is updated with a running average:

$$B_{x,y}^t = \beta_{x,y}^t I_{x,y}^t + (1 - \beta_{x,y}^t) B_{x,y}^{t-1}, \tag{2}$$

where $\beta_{x,y} \in [0, 1]$ is the updating ratio which may be fixed or, as will be shown in Section 3.4, adjusted according to a feedback loop. The initial background B^0 is obtained following a temporal median filter on the first 200 frames of the video.

Once the temporal gradient Δ^t has been computed, we cumulate it into an MHI as follows:

$$\mathrm{MHI}_{x,y}^t = \max(\Delta_{x,y}^t, \alpha \Delta_{x,y}^t + (1 - \alpha)\mathrm{MHI}_{x,y}^{t-1}), \tag{3}$$

where $\alpha \in [0, 1]$ is the MHI updating ratio. MHI^0 is initialized with zero values. The max operator ensures the MHI always contains the latest and largest temporal gradients. In this case, Eq. (3) can grasp short bursts of activity caused by fast moving objects. As for the values of α and $\beta_{x,y}$, please refer to Section 3.4 and 4 for how we fix on it.

Note that Eq. (3) differs from the original MHI implementation by Davis [8]. First, the use of an α ratio allows to adjust the speed at which the MHI is renewed in time. Second, since we directly cumulate the gradient instead of binary motion maps, there is no detection threshold and thus one less parameter to tune.

Examples of MHI are shown in Fig. 2(a) and (e). As can be seen, MHI aggregates layers of motion so a pixel value is a function of the recent activity at that position. MHI values are also strongly correlated with the presence of foreground moving objects: the larger a grayscale value is at a given pixel, the more probable a moving object is at that position. As opposed to background subtraction which produces binary maps, MHI contains a much richer set of information, especially in low-contrasted areas. In fact, MHI helps compensating for camouflage problems which happens when sections of a moving object have a low temporal gradient. By cumulating gradients in time, it is likely that a section of the moving object with a larger gradient will eventually compensate for another.

Please cite this article as: Y. Wang et al., Improving pedestrian detection using motion-guided filtering, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.11.020