Nonlinear Background Filter to Improve Pedestrian Detection

Yi Wang[1(✉)], Sébastien Piérard[2], Song-Zhi Su[3], and Pierre-Marc Jodoin[1]

[1] Department of Computer Science, University of Sherbrooke, Sherbrooke, Canada
{yi.wang,Pierre-Marc.Jodoin}@usherbrooke.ca
[2] INTELSIG Laboratory, Montefiore Institute, University of Liège, Liège, Belgium
Sebastien.Pierard@ulg.ac.be
[3] School of Information Science and Technology, Xiamen University, Xiamen, China
ssz@xmu.edu.cn

Abstract. In this paper, we propose a simple nonlinear filter which improves the detection of pedestrians walking in a video. We do so by first cumulating temporal gradient of moving objects into a motion history image (MHI). Then we apply to each frame of the video a motion-guided nonlinear filter whose goal is to smudge out background details while leaving untouched foreground moving objects. The resulting blurry-background image is then fed to a pedestrian detector. Experiments reveal that for a given miss rate, our motion-guided nonlinear filter can decrease the number of false positives per image (FPPI) by a factor of up to 26. Our method is simple, computationally light, and can be applied on a variety of videos to improve the performances of almost any kind of pedestrian detectors.

Keywords: Motion detection · Pedestrian detection · Motion history image · Nonlinear filtering

1 Introduction

Despite the plethora of papers published on the topic of pedestrian detection, detecting human shapes in 2D images is still an open problem. Pedestrian detection comes with fundamental difficulties that even state-of-the-art methods cannot handle properly. By their very nature, pedestrians may have different poses, they may be pictured from arbitrary angles and be occluded by objects or other pedestrians. Another issue that is fundamentally hard to cope with is background objects with a humanoid shape such as an armchair, a lamp post, or a coat rack. Since these objects have roughly the same features than human bodies, they are often wrongly classified as pedestrians [16].

In general, what differentiate pedestrian detection methods are the features they use and/or the classification rule which they implement [10]. The most widely used features are those using histograms of oriented gradients (HOG) [4]. Other commonly-implemented features are local binary patterns (LBP) [3] and

V. Murino et al. (Eds.): ICIAP 2015 Workshops, LNCS 9281, pp. 535–543, 2015.
DOI: 10.1007/978-3-319-23222-5_65

Haar wavelets [15]. When video is available, some methods extract spatio-temporal features such as binary motion labels from background subtraction [9] and motion tracks [13] to describe the motion of pedestrians. Other methods use richer features extracted from specialized hardware like stereo [2] or infrared cameras [8]. Recently, detectors based on deep learning [11] have been proposed. Instead of using man-made features, deep learning methods learn optimal pedestrian features automatically. As for the classifiers, the most widely implemented ones are support vector machines (SVM) and Adaboost [6]. Deformable Parts Model (DPM) [7] has also reported good results.

In this paper, we propose a new way to improve the performances of different pedestrian detectors without changing their design *per se*. Our method is based on two fundamental observations : 1) false positives produced by most pedestrian detectors are located over background objects whose visual features are close to that of a human and 2) by their very nature, pedestrians in surveillance videos are 2D moving blobs. The intuition behind our method is to use the temporal information of the video sequences as a leverage to filter out background objects while leaving untouched moving blobs (and thus pedestrians). By doing so, we get to deteriorate background visual features and thus help pedestrian detectors reduce their false detection rate.

We show that blurring adaptively the video sequences before applying existing pedestrian detection methods improves the results. The amount of blur can vary spatially and temporally, and is a function of the activity. More precisely, the filtered image is such that the graylevel in a pixel results from filtering the input image with a Gaussian filter whose standard deviation is correlated with the probability of motion in that pixel. Since the standard deviation vary spatially, our filter is nonlinear. The probability is estimated based on a Motion History Image (MHI) [5] that we compute by cumulating motion features. Experiments show that the number of false detections in the filtered images is drastically lower than on the original images without impacting much the miss rate.

Our paper comes with three main contributions. First, since our nonlinear filter is independent from the detector, it can be used by almost any pedestrian detection method. Second, our filtering method is robust to MHI inaccuracies which occur when illumination suddenly changes or when a background object moves. And third, since our filter is not optimized for detecting human shapes only, it can be used to detect other kinds of moving objects such as cars or boats.

2 Methodology

As shown in Fig. 1, our method implements a series of operations, which start with a background image that we compute by a moving average. We then compute an MHI which in turn is used by the nonlinear filter. Pedestrians are detected in the filtered image whose background has been smudged out.

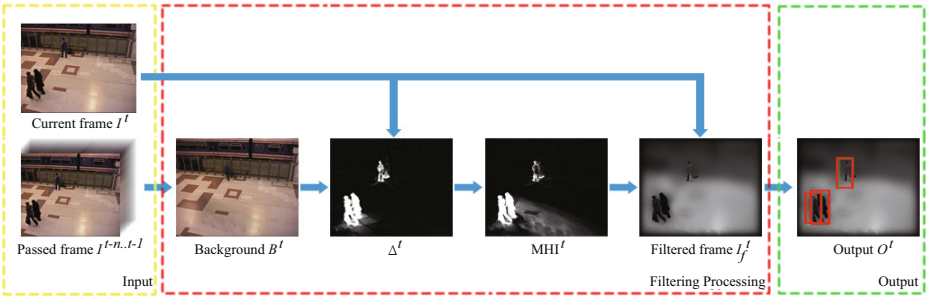


Fig. 1. Pipeline of our method. First, the background image B^t is computed with the past video frames. Then, we cumulate temporal gradients into an MHI. The MHI is then used by the filter to adjust its standard deviation to the amount of activity. Pedestrians are then detected on the filtered image.

2.1 Motion History Image

The first step of our technique is to estimate the background image B^t at each time t. It is updated at every frame with an exponential filter to account for changes in the dynamics of the video:

$$B^{t+1} = \beta I^t + (1 - \beta)B^t, \tag{1}$$

where I^t the input frame, and $\beta \in [0, 1]$ is a forgetting factor.

Based on the estimated background image, we highlight moving objects. We do so by computing a temporal gradient Δ^t

$$\Delta^t = \left| B^t - I^t \right|, \tag{2}$$

where $|\cdot|$ stands for the Euclidean norm in the RGB space. It is a naive estimation of motion at time t.

The Motion History Image MHI^t is obtained by cumulating Δ^t. Our implementation of MHI^t differs from that of Davis *et al.* [5] as we cumulate temporal gradients instead of binary motion maps:

$$\text{MHI}^t = \max(\Delta^t, \alpha\Delta^t + (1 - \alpha)\text{MHI}^{t-1}), \tag{3}$$

where $\alpha \in [0, 1]$ is the updating ratio for MHI. The max operator ensures that the MHI always contains the latest and largest temporal gradient. Without it, Eq.(3) would be a simple exponential moving average unable to grasp short burst of activity caused by fast moving objects.

Computing MHI with Eq.(3) has two main advantages over Davis *et al.*'s method. First, the use of an α ratio allows to adjust the speed at which the MHI is renewed. Second, Eq.(3) requires no detection threshold (we cumulate gradients but not binary motion maps) which is one less parameter to adjust.

(a) (b) (c) (d)

Fig. 2. Example of two filtered images (a, b) and their associate MHI (c, d).

Using MHI instead of a motion detection method to characterize activity has 3 main advantages. First, as opposed to motion detection methods, the MHI summarizes motion information of several frames. In other words, MHI pixel values aggregate the recent motion history that occurred at that location. Thus, MHI comes with a spectrum of values (not just binary values) which allows us to smoothly adjust the filter's standard deviation. Second, aggregating motion information of previous frames adds robustness to the system when a pedestrian stops moving for a short while. Third, MHI helps compensating for camouflage problems. This happens when sections of a moving object have a low temporal gradient. By cumulating gradients in time, we empirically noticed that sections of the moving object with a larger gradient often compensate for another one with a low gradient.

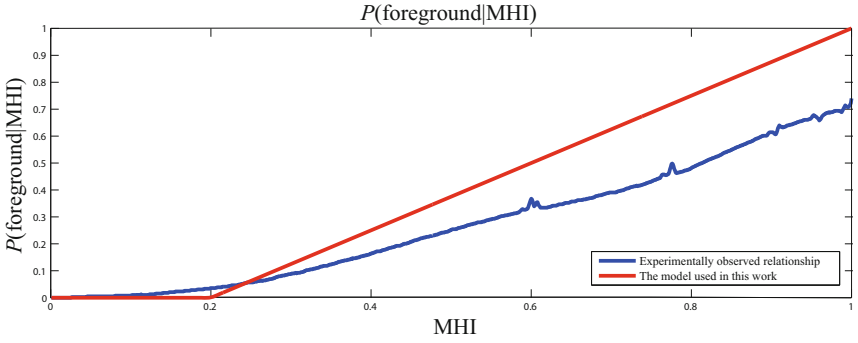


Fig. 3. The blue plot shows the empirical relationship $P(\text{foreground}|\text{MHI})$ obtained after processing all 54 videos from the changedetection.net dataset. As can be seen, the posterior probability and MHI values are almost linearly correlated. The red curve shows the model used in this work.

2.2 Motion-guided Filtering

As mentioned previously, background objects with humanoid shapes often lead to false detections. To decrease the false positive rate, we propose a nonlinear

motion-guided filter whose goal is to blur out background details. In order to do so, we need to characterize the probability of a region to be associated to a moving blob, *i.e.* $P(\text{foreground}^t|\text{MHI}^t)$. This formulation is based on the working assumption that MHI^t is a discrete variable.

In order to define the probabilistic model between foreground pixels and MHI^t values, we conducted an experiment on the changedetection.net 2014 dataset, the largest video dataset with pixel-accurate groundtruth. We took all 54 videos, computed an MHI for each frame and empirically computed $P(\text{foreground}|\text{MHI})$ by counting foreground and background pixels for each MHI value. This experiment resulted into the graphic of Fig. 3. As can be seen, pixels with MHI values smaller than 0.2 hardly correlate to any motion while those with values above 0.2 have a linear relationship with the probability of the foreground. Considering the noise of the data and motion detection errors, $P(\text{foreground}|\text{MHI})$ can never reach 1 in practice. However, for our foreground model we set $P(\text{foreground}|\text{MHI}) \in [0, 1]$:

$$P(\text{foreground}|\text{MHI}) = \max(0,\ 1.25(\text{MHI} - 0.2)), \tag{4}$$

In order for our filter to adapt to the content of the video (*i.e.* to filter out background areas while leaving untouched moving blobs), we use a Gaussian filter $\mathcal{G}(0, \sigma)$ whose standard deviation is a function of the posterior probability estimated at position (x, y). More specifically:

$$I_f^t = I^t * \mathcal{G}(0, \sigma^t), \tag{5}$$

$$\sigma_{x,y}^t = \sigma_{max} \times (1 - P_{x,y}(\text{foreground}|\text{MHI})), \tag{6}$$

where σ_{max} is set as $\frac{1}{5}$ of the height of the image and $P_{x,y}$ is the posterior probability at position (x, y). The reason that σ_{max} is a function of the image size is to account for high-resolution images whose background details are much larger pixel-wise. We also observed that a Gaussian filter with a standard deviation of σ_{max} eliminates background details in such a decisive manner that detecting pedestrians in those areas is very unlikely.

It should be stressed that the proposed filter aims at improving pedestrian detectors with a low risk of deteriorating their results. In particular, it is a well known fact that background subtraction techniques are highly sensitive to changes in the lighting conditions, to hard shadows, camera jitter and background motion [14]. Under these conditions, the temporal gradient Δ^t contains false positives which lead to an overestimation of activity within the MHI. That being said, more activity within the MHI leads to less filtering and thus performances close to the ones obtained by the detector alone.


(a) small pedestrians


(b) top-down


(c) baseline

Fig. 4. Examples of video sequences from the three categories.

3 Experiments and Results

3.1 Setup

We evaluated our method on 13 videos with pedestrians from 3 datasets, namely Caviar [1], changedetection.net [14] and the CUHK Square dataset [12]. The 13 videos were separated into 3 categories:

1. the **small pedestrians** category contains pedestrians with an height of at most 50 pixels;
2. the **top-down** category shows pedestrians filmed by a camera looking downward. Pedestrians in those videos all suffer from perspective distortion;
3. the **baseline** category contains videos showing easily detectable pedestrians.

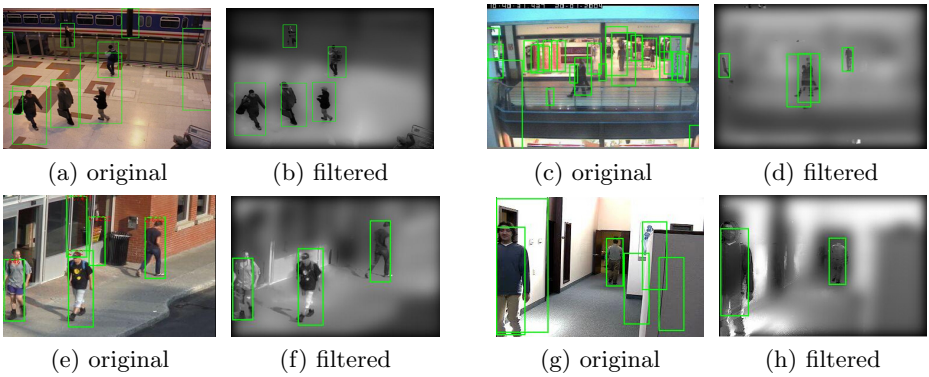
For evaluation, we extract a subset of N (uniformly spaced) frames per video, where N is the number of frames in the shortest video of the category. All the frames selected in a category are considered as a unique set. This selection allows one to avoid biasing the results in favor of longer videos (their lengths vary from 400 to 7200 frames). Snapshots from the 3 categories are shown in Fig. 4.

We tested our method with 3 widely-implemented pedestrian detectors, namely **HOG+SVM** from Dalal and Triggs [4], the **C4** contour-based method by Wu *et al.* [16], and **DPM** by Felzenszwalb *et al.* [7]. For every video, the background image B^1 was initialized with a temporal median filter over the first 200 frames. The MHI update ratio α and the background update ratio β were set to 0.8 and 0.016 after some empirical validation.

3.2 Results

Fig. 5 shows pedestrian detection results with and without our motion-guided nonlinear filter. As one can see, in every case the number of false detections is much smaller in the filtered images than in the original images. Since the number of true detections is the same, the use of a filtered image significantly improves accuracy.

We also evaluated the performance by drawing miss rate *vs* FPPI curves, as shown in Fig. 6. Again here, the motion-guided nonlinear filter drastically



(a) original (b) filtered (c) original (d) filtered

(e) original (f) filtered (g) original (h) filtered

Fig. 5. Results obtained on original (*i.e.* without our technique) and filtered (*i.e.* with it) images with the detectors HOG+SVM (1st row), C4(2nd row left) and DPM (2nd row right).

reduces the FPPI for all 3 detectors. The reader shall notice that results from our filtering method have a miss rate 1.1 to 1.9 times larger (the end of the dotted lines are slightly higher than the full lines) which is nonetheless negligible compared with the FPPI reduction which is between 2.4 and 16.5 times smaller. It is thus clear that our nonlinear filter can be used to improve the performance of existing pedestrian detectors on different kinds of videos.

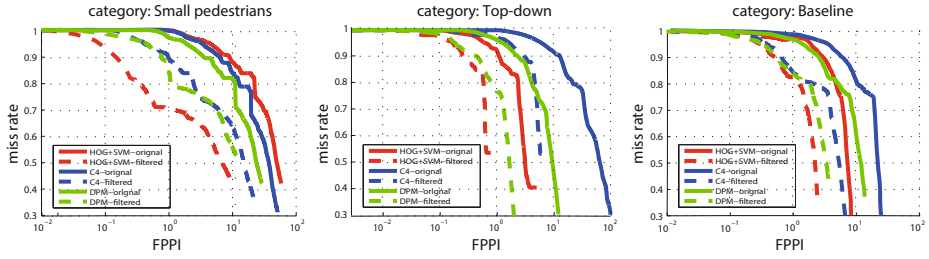


Fig. 6. Miss rate - FPPI curves for all 3 categories. Curves with different colors correspond to different detectors. Solid curves are for the original results without our pre-processing filtering, whereas the dashed ones are for the results with our filtering.

In Table 1 and Table 2, we further illustrated the quantitative analysis of our results. In Table 1, we selected the last point of each curve in Fig. 6 and compared the miss rate and FPPI value before and after applying nonlinear filtering on each detector with the same set of parameters. As can be seen, the FPPI decreases on average by a factor of 5.8 for each detector while the miss rate increases on average by a factor of only 1.3.

Table 1. Factor by which the miss rate increases and the FPPI decreases after applying our nonlinear filter.

Detector	**Small pedestrians**		**Top-down**		**Baseline**	
	miss rate increase	FPPI decrease	miss rate increase	FPPI decrease	miss rate increase	FPPI decrease
HOG+SVM	1.1	6.5	1.3	6.4	1.4	3.5
C4	1.2	2.4	1.9	16.5	1.1	3.6
DPM	1.3	2.5	1.5	6.8	1.2	3.8

Table 2. The FPPI value for a fixed miss rate of 0.7.

Detector		Small pedestrians	Top-down	Baseline
HOG+SVM	Original	**10.4**	2.6	5.9
	Filtered	**0.4**	0.6	1.6
C4	Original	6.6	35.3	18.5
	Filtered	2.0	5.2	3.8
DPM	Original	**4.0**	7.1	8.3
	Filtered	**1.9**	1.3	2.3

Table 2 compares the FPPI with and without our nonlinear motion-guided filter for a fix miss rate of 0.7. The table shows that our method reduces the FPPI values by a factor between 2.1 and 26 (see the bold values in the table).

Experiments were conducted on a 2.8GHz Intel Core2 Quad computer with a Matlab implementation. On average, it takes approximately 0.03 secs to update the background, compute the MHI and filter an image. This time does not depend on the number of pedestrians. Considering that it takes 0.06 sec to detect a pedestrian with HOG+SVM, 0.1 sec with C4 and 1.8 sec with DPM, our preprocessing (filtering) method does not bring a major processing overhead.

4 Conclusion

In this paper, we proposed a novel nonlinear filtering method which can be combined with almost any existing pedestrian detector. The method extracts motion features and cumulates it into an MHI. The MHI is then used to determine the standard deviation of a Gaussian filter which is used to nonlinearly filter video frames. Our method is easy to implement, fast and robust on different kinds of challenging circumstances. A drawback of this method is that if a pedestrian is a stationary from the first images, it would be integrated into background. We test our system with 3 widely used detectors on 3 categories of videos. The experiment results show that our nonlinear filter significantly decreases the false positive rate while keeping almost unchanged the miss rate.

References

1. http://homepages.inf.ed.ac.uk/rbf/caviar/
2. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Pedestrian detection at 100 frames per second. In: IEEE-C-CVPR, pp. 2903–2910 (2012)
3. Chen, G., Ding, Y., Xiao, J., Han, T.X.: Detection evolution with multi-order contextual co-occurrence. In: IEEE-C-CVPR, pp. 1798–1805 (2013)
4. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)
5. Davis, J.W.: Hierarchical motion history images for recognizing human motion. In: IEEE-W-Det. and Rec. of Eve. in Vid., pp. 39–46 (2001)
6. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE-T-PAMI **34**(4), 743–761 (2012)
7. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE-T-PAMI (2010)
8. Fernández-Caballero, A., Castillo, J., Serrano-Cuerda, J., Maldonado-Bascón, S.: Real-time human segmentation in infrared videos. Exp. Sys. with App. **38**(3), 2577–2584 (2011)
9. Lin, Z., Davis, L.S.: Shape-based human detection and segmentation via hierarchical part-template matching. IEEE-T-PAMI **32**(4), 604–618 (2010)
10. Ouyang, W., Wang, X.: Single-pedestrian detection aided by multi-pedestrian detection. In: IEEE-C-CVPR, pp. 3198–3205 (2013)
11. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: IEEE-C-CVPR (2013)
12. Wang, M., Li, W., Wang, X.: Transferring a generic pedestrian detector towards specific scenes. In: IEEE-C-CVPR, pp. 3274–3281 (2012)
13. Wang, X., Wang, M., Li, W.: Scene-specific pedestrian detection for static video surveillance. IEEE-T-PAMI **36**(2), 361–374 (2014)
14. Wang, Y., Jodoin, P.M., Porikli, F., Konrad, J., Benezeth, Y., Ishwar, P.: Cdnet 2014: an expanded change detection benchmark dataset. In: IEEE-W-CVPRW, pp. 393–400 (2014)
15. Wojek, C., Schiele, B.: A performance evaluation of single and multi-feature people detection. In: Pat. Rec., pp. 82–91 (2008)
16. Wu, J., Rehg, J.M.: Centrist: A visual descriptor for scene categorization. IEEE-T-PAMI **33**(8), 1489–1501 (2011)