Method to improve efficiency of human detection using scalemap

Guntae Bae, Sooyeong Kwak, Hyeran Byun and Daeyong Park

An efficient method is introduced for detecting humans in surveillance video. The method improves the performance of multiscale human detection through the use of a scalemap, and does not require knowledge of the camera parameters or the use of additional devices. A scalemap is a map that links each position in the observed image to the optimal detection scale. The proposed method efficiently reduces the computational costs by estimating the scale of interest and the region of interest based on the scalemap, while maintaining the accuracy of the detection. It is experimentally shown through an experiment that the proposed method can improve both the accuracy and the efficiency of real-world surveillance videos.

Introduction: Human detection is a topic of great interest for many applications such as activity recognition, intelligent surveillance and human–computer interfaces. In surveillance, detecting humans from videos or images is the most fundamental task, and it also critically affects the performance of the overall system. A well-known approach for detecting humans from videos is based on the background subtraction techniques; this approach gives stable and impressive results in static and quiet scenes, but also entails many challenging problems such as shadowing, ghost effects and difficulties responding to changes in illumination. To overcome these problems, pretrained human detectors are widely used in various applications; among them, the histogram of the oriented gradients (HOG) technique is the most popular [1]. However, a number of human detectors use a sliding-window approach, in which a finely sampled multiscale image pyramid is created and the entire image is scanned at each scale to detect humans. To analyse each frame of a video, this approach requires a great deal of computational resources, including both memory and computational operations; thus, it is not appropriate for real-time applications. Various strategies have been suggested to improve the performance of human detection. The first is to parallelise the algorithm, thereby enabling the use of the computing power of a graphics processing unit (GPU). In general, modern GPUs have more cores than central processing units (CPUs) and support running thousands of threads in parallel. However, parallelising the algorithm is not simple, and in some cases is impossible. Another way is to use prior knowledge to improve both the speed and the quality of human detection. However, previous methods based on the use of prior knowledge require additional information. For example, when prior knowledge of the ground plane is used, the additional information required includes camera parameters or depth maps, which are not available for many surveillance cameras that are presently installed.

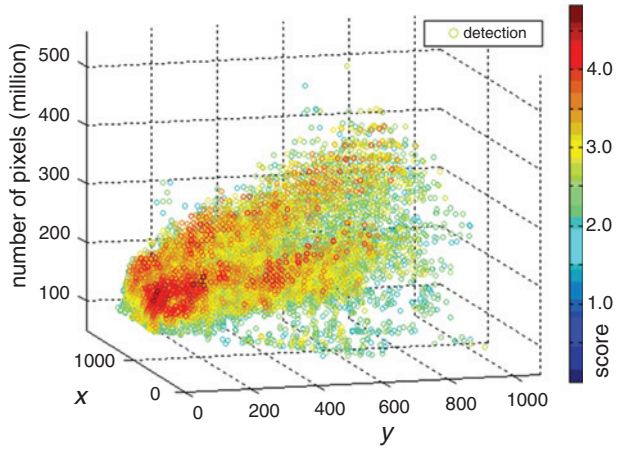

Fig. 1 *Distribution of location and height of detected pedestrians by HOG detector in Town Centre dataset; colour shows its detection score [2]*

Scalemap estimation: A scalemap is a map that links positions in the observed image to the optimal detection scale. The optimal scales at each location are determined by the trained model, under the assumption that the observed ground surface is plane. We used a naïve two-dimensional (2D) pedestrian detector to infer the 3D perspective. Fig. 1 shows a distribution of the detection results on the Town Centre dataset [2] by the baseline HOG detector [1]; the colour of the points represents the classification score, which is a measure of detection confidence. As shown in Fig. 1, the height and location (x, y) of detections in the scene can be correlated by using the plane model

$$h = \beta_0 + \beta_1 y + \beta_2 x \tag{1}$$

where x and y are the coordinates (bottom centre) of the bounding box and h is the height of the bounding box, and β_0, β_1 and β_2 are the model parameters. The model parameters are estimated by a fitting algorithm based on a random sample consensus (RANSAC). We use the classification score as the error weight to minimise the negative effect of the false positives on the accuracy. After the model parameters were estimated, we were able to generate the scalemap by simply dividing the height of the HOG template, H_{template}, by h

$$\text{scalemap}(x,\ y) = \frac{H_{\text{template}}}{\beta_0 + \beta_1 y + \beta_2 x} \tag{2}$$

Scale of interest (SOI) and region of interest (ROI) estimations: To minimise the unnecessary computations in practical multiscale human detection, a scalemap can be used in two steps. The first step is a SOI estimation, which reduces the possible scale range of the image pyramid. Without a scalemap, all possible scales need to be considered, and this can cause false positives. However, a scalemap can be used as useful prior knowledge, provided that the video is acquired in static cameras. The SOI is determined by finding the minimum and maximum values in the scale space. By restricting the range of the scale to be generated in the image pyramid, we effectively reduce unnecessary computations and wasted memory in every step of human detection. The second step further minimises computations through ROI estimation on each scale of the SOI by using the following equation:

$$\text{ROI}(s_i,\ \theta) = \left\{(x,\ y) | s_{i-\theta} < \text{scalemap}(x,\ y) < s_{i+\theta}\right\} \tag{3}$$

where s_i is the observed image on the ith scale of the SOI and θ is an overlapping scale parameter that can control the efficiency and the accuracy of the algorithm.

Experimental results: For a quantitative evaluation of the proposed method, we used public datasets that were captured in indoor and outdoor scenes. Table 1 shows the composition of the datasets and the number of human detections by using the baseline HOG detector. All experiments were performed on a desktop computer that was equipped with an Intel Core i5 760 CPU (2.8 GHz) and 16 GB of random access memory (RAM). The operating system was a 64-bit version of Microsoft Windows 7.

Table 1: Characteristics of datasets

Scene	Dataset	Resolution	Frames	Detections
1	Town Centre [2]	1920 × 1080	4500	78 028
2	PETS2009 [3]	1536 × 1152	795	820
3	CAVIAR [4]	384 × 288	1462	712

Table 2 shows the plane model parameters estimated by RANSAC for each dataset. By using the plane model, Fig. 2 shows the estimated scalemap that corresponds to the perspective perceived by humans; the intensity depicted in the Figure was multiplied by 128 from the original value, and the yellow region represents the mask for the unnecessary region in which the human size is smaller than 32 × 64 pixels. The blue line represents the same scale value (interval: 0.2).

Table 2: Estimated model parameters

Scene	β_0	β_1	β_2
1	109.195	0.246	0.012
2	33.598	0.297	0.011
3	23.162	0.510	0.025

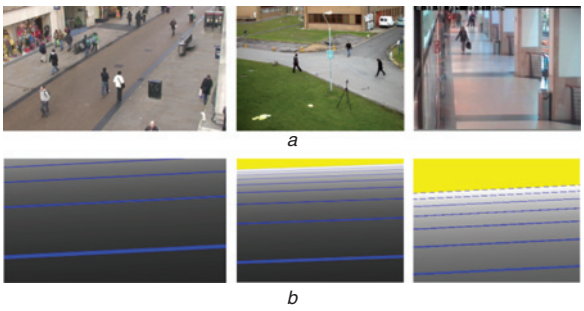

Fig. 2 *Result of scalemap estimation*

a View of dataset
b Estimated scalemap

To evaluate the effectiveness of our method, we first compared it with the baseline HOG detector in terms of the number of pixels to be processed and the processing time. As shown in Fig. 3, by selecting the SOI and ROI, the proposed method efficiently reduced the number of pixels by 24.5% of the original pixels on average. As a result, the required time of the proposed method is ~27.2% that of the baseline detector, while achieving approximately three times its accuracy. The false positive per image (FPPI) measure with PASCAL also shows the improvements. Each result is specified in Table 3.

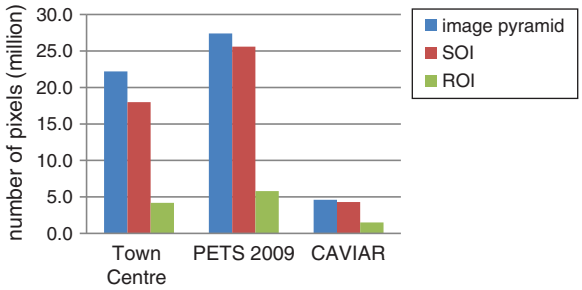

Fig. 3 *Decrease in number of processed pixels by SOI and ROI selections*

Table 3: Performance of proposed method

Scene	Baseline detector [1]			Proposed method		
	FPPI	Time	Speed	FPPI	Time	Speed
1	4.40	5654	0.17	0.40	1423	0.70
2	4.31	7178	0.13	2.55	1509	0.66
3	3.20	1121	0.82	2.09	362	2.76

Next, we compared the performance of the proposed method against two reference algorithms [5, 6]; one represents the state-of-the-art for detection accuracy and the other for processing time. The former, the discriminatively trained part-based model (DPM), uses structured HOG detectors to find a deformable object robustly [5]. The latter, the fastest pedestrian detector in the west (FPDW), is known as the best choice for detecting medium-scale pedestrians [6]. In this experiment, we used the codes supplied by the authors, and implemented our own method using multi-threading and optimisation. As shown in Table 4, our method effectively improved both detection accuracy and processing time.

Table 4: Comparison with state-of-the-art algorithms

Scene	DPM [5]			FPDW [6]			Proposed methods		
	FPPI	Recall	Time (ms)	FPPI	Recall	Time (ms)	FPPI	Recall	Time (ms)
1	2.56	0.96	36 743	2.66	0.70	19 171	0.45	0.55	374
2	0.57	0.55	31 311	1.13	0.55	363	2.24	0.80	329
3	0.36	0.55	2896	0.70	0.67	117	0.68	0.71	99

Conclusion: In this Letter, we have proposed a novel method to improve the efficiency of a human detection algorithm in surveillance video without requiring additional information such as camera parameters, stereo imagery or depth measurements. The proposed method generated the scalemap automatically by estimating a perspective of the scene based on the unreliable initial detection results. The scalemap reduces the detection area in the image pyramid by estimating the SOI and ROI. We have demonstrated that our method improves both detection accuracy and processing time with public datasets. An advantage of our method is that it can be practically applied to existing systems without adding an overhead.

Acknowledgment: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2013R1A2A1A01015870).

17 December 2013
doi: 10.1049/el.2013.3588
One or more of the Figures in this Letter are available in colour online.

Guntae Bae, Hyeran Byun and Daeyong Park (*Department of Computer Science, Yonsei University, 134, Shinchon-Dong, Seodaemun-gu, Seoul 120-749, Republic of Korea*)

E-mail: hrbyun@yonsei.ac.kr

Sooyeong Kwak (*Hanbat National University, San 16-1, Duckmyoung-Dong, Daejeon 305-719, Republic of Korea*)

References

1 Dalal, B., and Triggs, B.: 'Histogram of oriented gradients for human detection'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, San Diego, CA, USA, June 2005, pp. 888–893
2 Benfold, B., and Reid, I.: 'Stable multi-target tracking in real-time surveillance video'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, June 2011, pp. 3457–3464
3 Computational Vision Group in Reading University: 'PETS 2009 benchmark data' http://www.cvg.rdg.ac.uk/PETS2009/a.html accessed December 2013
4 Fisher, B.: 'CAVIAR test case scenarios' http://www.homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/ accessed December 2013
5 Felzenszwalb, P.F., Girshick, R.B., McAllester, D., and Ramanan, D.: 'Object detection with discriminatively trained part based models', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **32**, (9), pp. 1627–1645
6 Dollar, P., Wojek, C., Schiele, B., and Perona, P.: 'Pedestrian detection: an evaluation of the state of the art', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, **34**, (4), pp. 743–761