Pattern Recognition 44 (2011) 1296–1311

Contents lists available at ScienceDirect

Pattern Recognition

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

Object detection based on a robust and accurate statistical multi-point-pair model

Xinyue Zhao[a,*], Yutaka Satoh[b], Hidenori Takauji[a], Shun'ichi Kaneko[a], Kenji Iwata[b], Ryushi Ozaki[c]

[a] *Graduate School of Information Science and Technology, Hokkaido University, Hokkaido, Japan*
[b] *National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki, Japan*
[c] *Tsukuba University, Tsukuba, Japan*

ARTICLE INFO

Article history:
Received 3 June 2010
Received in revised form
24 November 2010
Accepted 27 November 2010
Available online 17 December 2010

Keywords:
Background subtraction
Object detection
Statistical reach feature (SRF)
Grayscale arranging pairs (GAP)

ABSTRACT

In this paper, we propose a robust and accurate background model, called grayscale arranging pairs (GAP). The model is based on the statistical reach feature (SRF), which is defined as a set of statistical pair-wise features. Using the GAP model, moving objects are successfully detected under a variety of complex environmental conditions. The main concept of the proposed method is the use of multiple point pairs that exhibit a stable statistical intensity relationship as a background model. The intensity difference between pixels of the pair is much more stable than the intensity of a single pixel, especially in varying environments. Our proposed method focuses more on the history of global spatial correlations between pixels than on the history of any given pixel or local spatial correlations. Furthermore, we clarify how to reduce the GAP modeling time and present experimental results comparing GAP with existing object detection methods, demonstrating that superior object detection with higher precision and recall rates is achieved by GAP.

1. Introduction

Background subtraction is one of the typical approaches used in automated surveillance technology, which is aimed at extracting moving objects of interest from video sequences. Typically, as a part of the pre-processing stage for automated activity monitoring, background subtraction is widely used in object tracking [1,2], traffic monitoring [3], behavior recognition [4,5] and unusual event detection [6].

Despite its importance, background subtraction in complex environments is far from being completely mastered. In general, in real-world situations, either indoor or outdoor, variations in illumination cannot be ignored. Outdoor scenes can be affected by sunlight, occasionally leading to global changes caused by the apparent movement of the sun, or to local changes such as shadows and reflections. With indoor scenes, immediate illumination changes can be caused by lights being switched on or off. Furthermore, dynamic background elements, such as small camera jitter, slow moving clouds, and waving leaves, are difficult to deal with. The situation becomes even more difficult in environments with a mixture of these challenges, and examples of these difficulties can be found in the experimental section.

Two approaches have been developed for background subtraction over the past two decades. One uses temporal information and primarily exploits the intensity of a single pixel as a function of time [7–13]. The second approach adds spatial information and primarily considers correlations between pixels in a single frame [14–19]. Temporal methods were initially popular but lost popularity gradually, since they do not consider the spatial correlations between pixels which reduces performance with respect to environmental changes. Current research has concentrated on spatial methods, since they emphasize the correlations in each frame, thereby improving noise resistance. However, most of these methods consider only the correlation of neighboring pixels (especially in an eight-connected neighborhood). Thus, they focus on local spatial information and ignore global spatial information.

To address these issues, we propose a novel background model, called the grayscale arranging pairs (GAP), which is based on previously proposed statistical reach-based models [20–22]. In this paper, however, we consider differences in global spatial intensity. Our objective is to create an accurate and robust background model that is flexible enough to handle different sets of complex conditions. The intensity of a single pixel varies significantly due to background motion (illumination variations, waving leaves, or camera vibration), hence it is difficult to predict its next state. But the problem becomes simple by using point pairs with a stable intensity relationship. For the background model criterion, we choose multiple point pairs whose intensity differences remain stable under complex conditions provided the object being imaged

[*] Corresponding author.
E-mail addresses: zhao@ssc.ssi.ist.hokudai.ac.jp, shuimuzxy@hotmail.com (X. Zhao).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.11.022

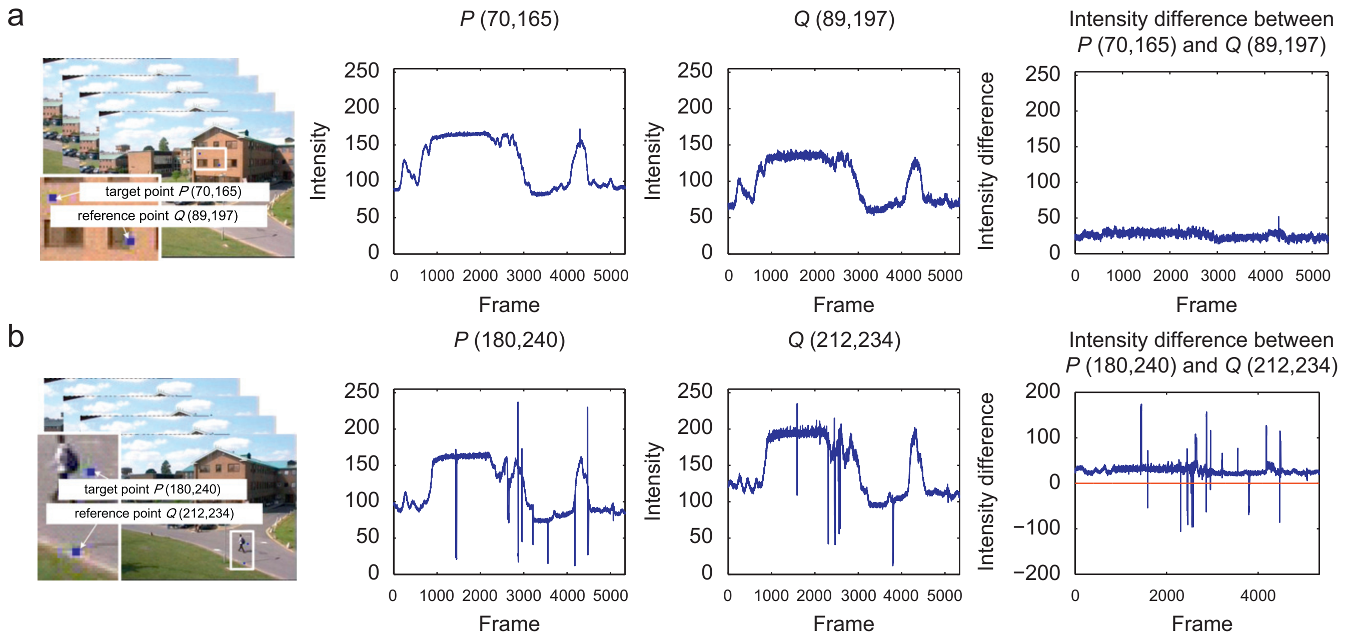


Fig. 1. Intensity difference between two points. Changes in intensity difference between a pair of points are much smaller than those of individual pixels, even under varying illumination. (a) The situation without any moving object and (b) with moving objects.

does not change. These point pairs are not limited to neighboring pairs because background motion occasionally causes neighboring points to share a weak relationship. Fig. 1 shows a stable intensity relationship between one point pair in a scene with varying illumination. Each image point is supported by multiple points, which have consistent pairwise relationships over time. The existence of a foreground point is determined by a substantial deviation from the pre-computed pairwise relations within the support of these multiple points.

The advantages of our proposed method are as follows:

1. Robust to the influence of severe illumination changes and some background motion. Some approaches use background models built by historical judgments of each pixel. In other approaches, although pixel correlations are considered in the modeling, they typically use information between neighboring pixels or pixel blocks; whereas, we adopt the viewpoint that useful pixel correlations are not limited to neighboring pixels or pixel blocks. We select point pairs with a stable intensity difference over the entire image, using both temporal and spatial information, so that illumination variance and some background motion, such as common conditions of swaying leaves, repetitive camera jitter, and slow moving clouds, can be well tolerated.

2. High accuracy. We choose multiple point pairs, rather than a single point pair, since multiple point pairs distributed dispersedly in the spatial domain can eliminate errors caused by a single point pair and provide more robust and credible information for classification. Furthermore, a more robust criterion is used in foreground extraction to ensure that the foreground points are not missed.

3. Fast detection. In the modeling step, we distinguish the cases of positive and negative intensity differences, so we need to store only the signs of the intensity difference between pixel pairs, instead of the intensity value. This requires less memory and reduces the computational load of modeling; in addition, we adopt an accelerated modeling step. During detection, the speed is high enough for real-time implementation, since most of the computation time is spent in comparing the signs between point pairs.

The remainder of our paper is organized as follows: Section 2 gives a brief introduction of previous works; Section 3 describes the background subtraction model; Section 4 presents an accelerated GAP modeling for time reduction; Section 5 presents the experimental results; and Section 6 concludes the main points of the research and future work.

2. Previous works

Since frame differencing was first used for object detection [23], several approaches have been developed in background subtraction. We classify the recent approaches into two categories: time-based methods and space-based methods.

As one of the time-based methods, the mixture of Gaussians (MoG) [8,9] has become well known since it was first proposed. MoG used a mixture of Gaussians model instead of a single Gaussian model [7] that can cope with multiple background objects. However, since using each pixel independently, MoG is very sensitive to sudden changes in illumination. Some other similar approaches based on MoG can also be found in [13]. In [10], Elgammal et al. proposed a kernel density estimation (KDE). Using the probability density functions, the pixel intensity distribution is obtained without input parameters. In other approaches using temporal domain information [11,12], hidden Markov models (HMMs) are applied to model the background.

In space-based methods, Toyama et al. [14] proposed a three-stage Wallflower, which used region and frame-level information, as well as intensity history of pixels. Oliver et al. [15] proposed a spatial approach using eigenspace decomposition. In this method, a mean image is computed from a set of sample images, and the best eigenvectors from a covariance matrix are stored in an eigenvector matrix which is used in the classification step. However, this method suffers from the limitation that the images making up the initial sample are motionless. Seki et al. [17] tried to exploit spatial co-occurrence by making use of $N \times N$ pixel blocks instead of pixel resolution. They computed the temporal average for each block and then got the $N \times N$ covariance matrix with respect to this average, thereby reducing the image dimensions from N^2 to K dimensions by eigenvector transformation. The model is less sensitive to illumination but is only suitable for coarse detection since it makes use of pixel blocks. Sheikh [18] used the joint representation of image pixels in local spatial distribution and color information, and built both background and foreground models as temporal persistence. Monnet et al. [16]built an auto-regressive model in dynamic scenes. Differences in the state space between previous frames and the current frame are considered for detection. In Ref. [19], Heikkilä and Pietikäinen used local binary pattern (LBP) histograms for subtracting the background and detecting moving objects in real time. This method models each pixel as a group of adaptive local binary pattern histograms that are calculated over a circular region around the pixel. The method tolerates general illumination changes and multi-modality of the background well, but is not capable of handling sudden changes such as the switching light problem.

In addition to these two types, a hybrid background model, which consists of both spatial and temporal information, has been used for object detection. However, most approaches put particular emphasis on either the spatial domain or the temporal domain. Liu et al. [24] used an information saliency map (ISM) which is calculated from spatio-temporal volumes to detect both fast and slow moving objects. In another recent approach [25], the temporal variation of intensity and color distributions are analyzed to develop a background model.

Besides these traditional background subtraction algorithms, some other methods, which were not originally intended for background subtraction, can detect objects in some situations. For instance, optic flow is feasible for this purpose in motion estimation, such as the estimation of traffic flow. However, in this paper, since we are interested in general background modeling algorithms, and focus more on the silhouettes of moving objects than on their movements, we will do not introduce those methods.

3. GAP background model

In our previous work, a statistical reach feature (SRF) [22] is presented and used in object detection. It achieves good performance under varying illumination conditions, nevertheless, to extend the work to more complex environments, based on it, we propose a novel background model called GAP which focuses on the intensity correlations of pixels in the global distribution.

Our proposed work in this paper has three novel contributions. First, we analyze the intensity stability between point pairs and assume that the intensity difference between the pair is more stable than the intensity of a single point. This is discussed in Section 3.2. Second, unlike previous approaches, we consider the inner relationship of pixel pairs in both the temporal and a global spatial area, and build novel rules for selecting appropriate pixels, as described in Section 3.3. Third, instead of applying a one-sided judging standard, a double-sided judging standard is utilized in foreground extraction to ensure that foreground points are not missed—this is discussed in Section 3.4.

In this section, first we briefly introduce the basics of the SRF algorithm and its existing problems (Section 3.1). We then introduce the concept of the GAP background model (Sections 3.2– 3.5): analysis of properties between point pairs (Section 3.2); modeling of the background (Section 3.3); object detection (Section 3.4); and discussion of influence of background motion and partial illumination changes (Section 3.5).

3.1. SRF algorithm

The SRF operator is defined as a set of statistical pair-wise features, derived by intensity comparison in a local neighborhood. SRF has several properties that favor its usage in background modeling.

Suppose we are given a set of training images $B=\{I_1,\ldots,I_T\}$ for a background sample, where T is the total number of training images. Each image of the image set B has $U\times V$ pixels, which is regarded as the function on the set $\Gamma := \{(u,v)|u=1,\ldots,U, v=1,\ldots,V\}$. In the following, we introduce the classification of background and foreground for every point $P\in\Gamma$, which is called the target point. The classification is performed by a set of selected N points $Q_n\,(1\le n\le N)$, which are called reference points.

The process modeling of SRF can be described as the selection of Q_n. There are three factors that affect the search of Q_n: (1) the absolute value of the intensity difference between P and Q_n must exceed a given threshold W_G. This non-zero threshold plays a critical role—it allows the background model to tolerate noise. (2) Q_n must meet a statistical requirement. The intensity of Q_n remains W_G units larger (or smaller) than that of P in most images. (3) Q_n is searched from the starting point P to the edge of image in N (in Ref. [22], $N=8$) radiation directions. No more than one point is chosen in each direction. Since the intensity of Q_n might be larger or smaller than that of P, two types of Q_n can be selected. SRF defines the sign between P and Q_n, which satisfies $I_t(P)-I_t(Q_n)\ge W_G$ in most images, as $\mathrm{SRF}(P,Q)=1$. It also defines the sign between P and Q_n, which satisfies $I_t(P)-I_t(Q_n)\le -W_G$ in most images, as $\mathrm{SRF}(P,Q)=-1$. Then, comparing the signs in the background model with that in an input image determines the classification of P as a background or foreground pixel, depending on whether the sign between P and Q_n has changed.

Because of the special properties of point pairs, SRF works well in object detection. But three issues must be discussed: First, SRF does not search a sufficient number of Q_n. SRF searches for Q_n in only eight radiation directions, implying that most pixels are not considered. This may lead to an insufficient number of Q_n that causes difficulty in the detection step. Second, the searching way of Q_n is not optimal. SRF searches Q_n in the order of space rather than the order of intensity difference, which leads to an incomplete detection. Actually, the intensity difference between P and Q_n influences the sensitivity of background model. The larger the intensity difference is, the less sensitive the model becomes. Without controlling the magnitude of intensity difference, the intensity difference between P and Q_n searched by SRF tends to be too large. This leads to incomplete detection. Third, the one-sided criterion in the SRF detection step results in false detections. A one-sided criterion means that the two signs of SRF ($\mathrm{SRF}(P,Q)=1$ and $\mathrm{SRF}(P,Q)=-1$) are combined into one binary decision. Since the searching order of SRF is from neighborhood to image edge, the first point in each direction that satisfies the requirements will be stored, irrespective of its sign. In this case, we can control the total number of Q_n but not the number of Q_n with different signs. Thus, the model is not sufficiently robust. For instance, if all chosen Q_n satisfy $\mathrm{SRF}(P,Q)=1$, which means P is brighter than Q_n, then in the case of presence of a moving object with much brighter color, P is misclassified as background.

To solve these problems, we improved SRF both in theory and algorithm. In the next section, we analyze the physical properties of point pairs in theory, which is the basis of the algorithmic improvement.

3.2. Property analysis of point pairs

The effectiveness of SRF is due to the stable intensity relationship between selected point pairs. Although the intensity of a single point may change dramatically over time, the intensity difference between selected point pairs remains stable.

As in Section 3.1, in this paper, P denotes the target point and Q_n denotes the reference point. We define $\mathrm{ref}(P)=\{Q_1,\ldots,Q_N\}$ to denote the reference point set of P. We also define points Q which satisfy the first two factors of SRF as candidate reference points.

They are further filtered through special rules before becoming reference points.

For a single target point P, if its intensity differs from that of another pixel Q_n by more than a given threshold with high probability, then pixel Q_n qualifies as its reference point for background modeling. Thus, the probability of the sign of the difference between $I_t(P)$ and $I_t(Q_n)$ remaining constant at any image I_t is high, even in changing environments, such as those with varying illumination. In other words, if the sign of the difference between $I_t(P)$ and $I_t(Q_n)$ does not change, then no moving object is detected, and pixel P is judged to be background. Conversely, if the sign of the difference between $I_t(P)$ and $I_t(Q_n)$ does change, then a moving object is detected and P is judged to be foreground.

As an example, consider Fig. 1, which shows the intensity difference between two points. In Fig. 1(a), the two points compared are both from the wall of the building, which is stationary; thus, their intensity difference is stable even if the intensity of individual pixels changes over time. In Fig. 1(b), this stable intensity difference condition is disrupted by moving objects. The change in the intensity difference sign (the intensity difference becomes negative) in some frames indicates the presence of moving objects.

From the analysis of stability between point pairs, we can see that selection of point pairs is the key step in this type of algorithm and exerts a strong influence on the final result. In our paper, we utilize the best properties of point pairs and build novel rules for selecting reference points. The steps for choosing point pairs for GAP will be fully described in the next subsection.

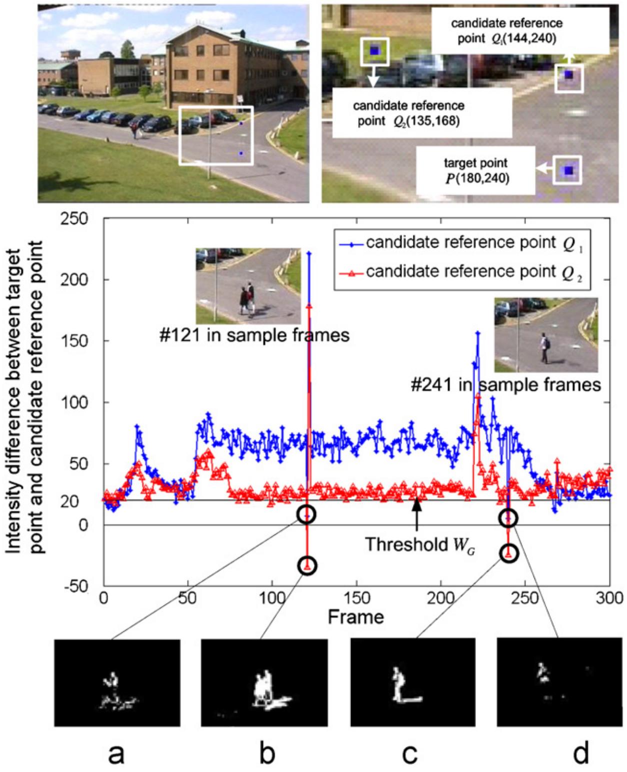


Fig. 2. The image shows that distant pixels with smaller intensity difference have better performance than neighboring pixels. (a) and (d) are partial results obtained by choosing neighboring pixels as high priority. (b) and (c) are the results by selecting pixels with smaller intensity differences (beyond W_G).

3.3. Modeling the background

We build the background model by using intensity differences. Thus, the difference in intensity between reference points and the target point determines the model's sensitivity. The model can be neither over-sensitive nor insensitive; thus, the intensity difference cannot be too small or too large. When the intensity difference is too small, the model is over-sensitive, and noise will be classified as foreground. As explained in the previous subsection, the threshold W_G indicates the ideal capacity of noise tolerance (a detailed discussion of W_G is presented in Section 5.4). Therefore, we avoid over-sensitivity by making the intensity difference beyond W_G. On the other hand, the intensity difference cannot be too large. Otherwise, the model will be insensitive and some foreground will be classified as background. Thus, for each pair of target and reference points, the intensity difference, which satisfies the statistical condition of being beyond W_G, should be as small as possible. For instance, a smaller intensity difference between points P and Q_2 leads to a better performance (shown in Figs. 2(b) and (c)) than that for P and Q_1 (shown in Figs. 2(a) and (d)). The former is more sensitive to identifying all moving objects of interest.

Actually, choosing point pairs is the selection of reference points. Unlike SRF in a spatial order, the proposed method prescribes picking reference points directly from the entire image using an intensity difference order. Although neighboring points have a natural correlation, since we use the intensity difference, rather than the intensity itself, stable point pairs do not always exist in a neighborhood. In many cases, the intensity difference of neighboring points changes significantly from frame to frame. For instance, edges, small objects, or background motion may create neighboring points with no significant relationship. Moreover, selecting sufficient reference points provides ample information about intensity changes, and it is difficult to find sufficient points only in the neighborhood. Therefore, unlike the traditional methods, we do not limit the search range to the neighborhood, but locate suitable reference points from the entire image. This procedure is summarized as follows.

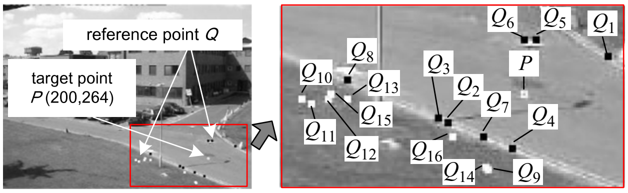


Fig. 3. Distribution of reference points.

1. Search for the candidate reference points. Suppose we are given $M(>N)$ candidate reference points $Q_m (1 \le m \le M)$. Then the sets $X^+(P)$ and $X^-(P)$ are defined as

$$X^+(P) := \{Q | P_r^+(P,Q) \ge W_P\}, \tag{1}$$

$$X^-(P) := \{Q | P_r^-(P,Q) \ge W_P\}. \tag{2}$$

In Eqs. (1) and (2), the probabilities $P_r^+(P,Q)$ and $P_r^-(P,Q)$, which estimate the number of images where the point pair satisfies the intensity requirements, are defined respectively as

$$Pr^+(P,Q) := \frac{\#\{t | I_t(P) - I_t(Q) \ge W_G, t = 1, \ldots, T\}}{T} \tag{3}$$

and

$$Pr^-(P,Q) := \frac{\#\{t | I_t(P) - I_t(Q) \le -W_G, t = 1, \ldots, T\}}{T}, \tag{4}$$

where $\#\{x | f(x)\}$ gives the total number of x satisfying $f(x)$, and T represents the total number of training images. The threshold W_P $(0.5 < W_P < 1)$ is the minimum proportion of the number of images we set.

2. Calculate the mean intensity difference. The mean intensity for every target point $P \in \Gamma$ is defined as $\overline{I(P)} = (1/T)\sum_{t=1}^{T} I_t(P)$, and the mean intensity for every candidate reference point Q_m, which satisfies the requirement $Q_m \in X^+(P)$ or $Q_m \in X^-(P)$, is defined as $\overline{I(Q_m)}$.

3. Search for reference points. Taking sample pixels $Q_1, Q_2 \in X^+(P)$ for example, we define the ordering $<$ such that $Q_1 < Q_2$ if and only if $|\overline{I(P)} - \overline{I(Q_1)}| < |\overline{I(P)} - \overline{I(Q_2)}|$. The sets $X^+(P)$ and $X^-(P)$ are sorted according to this ordering. Then we can give the reference point sets $\text{ref}^+(P), \text{ref}^-(P) \subseteq \Gamma$ associated with each given point $P \in \Gamma$ as

$$\text{ref}^+(P) := \left\{ \text{the first } \frac{N}{2} \text{ elements of } X^+(P) \text{ with respect to} < . \right\}, \tag{5}$$

$$\text{ref}^-(P) := \left\{ \text{the first } \frac{N}{2} \text{ elements of } X^-(P) \text{ with respect to} < . \right\}, \tag{6}$$

where $\text{ref}^+(P) \cup \text{ref}^-(P) = \text{ref}(P)$.

4. Build the background model. We record the sign of the intensity difference between P and Q_n as

$$\text{Mos}(P, Q_n) := \begin{cases} 1, & Q_n \in \text{ref}^+(P), \\ -1, & Q_n \in \text{ref}^-(P). \end{cases} \tag{7}$$

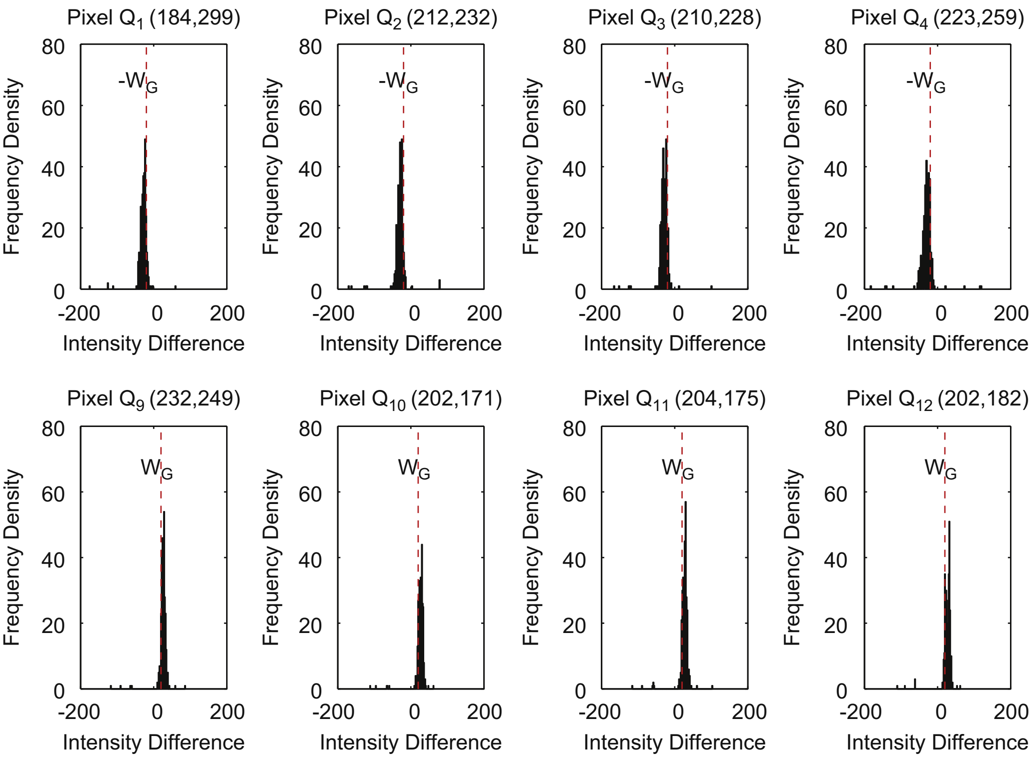
In summary, pseudo-code of the standard version of algorithm for modeling background is given in Algorithm 1.

Algorithm 1. Background modeling (standard version).

```
Input: $T$ Training frames (frame $t$ is denoted by $I_t$); Thresholds $W_G$, $W_P$.
Output: $\text{ref}^+$, $\text{ref}^-$ ($\text{ref}^+(P)$, $\text{ref}^-(P)$ are reference points sets for each pixel $P$).
for each pixel $P$ do                //$P$ is the target point.
    $\text{ref}^+(P) \leftarrow \emptyset$; $\text{ref}^-(P) \leftarrow \emptyset$;        // Initialization.
    Search $Q_m^+$ and $Q_m^-$, where the points $Q \in Q_m^+$ and $Q \in Q_m^-$ satisfy Eqs. (1) and (2) respectively;
                // $Q_m^+$ and $Q_m^-$ are the candidate reference points sets.
    Calculate the sets of the mean intensity difference $D^+, D^-$ between $P$ and each point $Q \in Q_m^+$ and $Q \in Q_m^-$
    in $T$ frames : $D^+(Q) = \frac{1}{T}\sum_{t=1}^{T} |I_t(P) - I_t(Q)|$ $(Q \in Q_m^+)$, $D^-(Q) = \frac{1}{T}\sum_{t=1}^{T} |I_t(P) - I_t(Q)|$ $(Q \in Q_m^-)$;
    Sort $D^+, D^-$ in ascending order respectively;
    Select the first $\frac{N}{2}$ components in $D^+$ and $D^-$ as $\text{ref}^+(P)$ and $\text{ref}^-(P)$ respectively;
return $\text{ref}^+$, $\text{ref}^-$.
```

Fig. 3 shows an example of actual distribution of reference points (we chose $N=16$) and Fig. 4 shows several histograms of the intensity differences between the target point P and selected reference points. As expected, in Fig. 4, the intensity difference histograms between P and Q_n are those concentrated with the smallest magnitude (beyond W_G), demonstrating that the intensity differences between a target point and selected reference points satisfy our requirements: the intensity differences are stable and as small as possible (beyond W_G).

Note that our method is robust to illumination changes and many situations of background motion. The background model is robust to illumination changes due to the stable relationship of


Pixel Q_1 (184,299)

Pixel Q_2 (212,232)

Pixel Q_3 (210,228)

Pixel Q_4 (223,259)

Pixel Q_9 (232,249)

Pixel Q_{10} (202,171)

Pixel Q_{11} (204,175)

Pixel Q_{12} (202,182)

Fig. 4. Histogram of intensity differences $I_t(P) - I_t(Q_n)$ in Fig. 3. In the first row, $Q_n \in \text{ref}^-(P)$; in the second row, $Q_n \in \text{ref}^+(P)$.

Table 1
Testing of selecting reference points.

Different cases	$0<T^-<8$	$T^-=8$	$0<T^+<8$	$T^+=8$	$T^-=T^+=0$
Number of target points from scene of Fig. 10	317	67,477	222	71,745	0
Number of target points from scene of Fig. 11	1252	13,032	19	76,597	154
Number of target points from PETS scene of Fig. 12	148	73,010	0	59,673	0
Number of target points from camera jitter scene of Fig. 12	1176	71,238	162	74,679	0
Number of target points from fog scene of Fig. 12	7	76,604	0	63,005	0

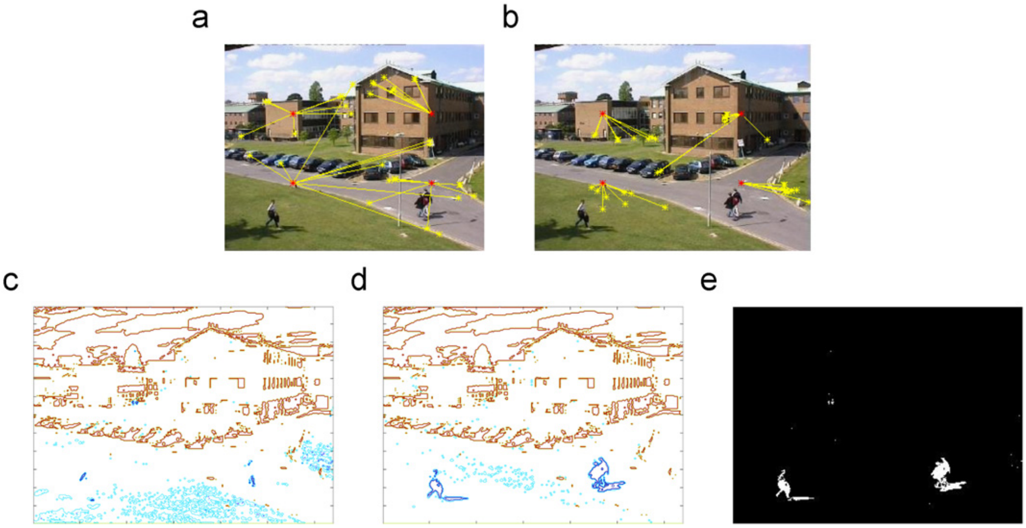


Fig. 5. Description of object detection step. (a) Distribution of point pairs P and $Q_n \in \text{ref}^-(P)$; (b) distribution of point pairs P and $Q_n \in \text{ref}^+(P)$; (c) probability contour maps of $\xi^-(P)$; (d) probability contour maps of $\xi^+(P)$; and (e) the detection result. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

point pairs, which has been explained in the previous subsection. Furthermore, it is also robust to many situations of background motion. From the view point of a single pixel, there are multiple normal background states under the condition of a dynamic background. Suppose a target point P has n background states, $S_1,\ldots,S_n$ (in the increasing order of average intensity), and each state occupies no less than $1-W_P$ proportion of the time. This assumption is quite reasonable, since a general dynamic background always varies among several states repeatedly. We can observe from the procedure of our algorithm that the reference points $\text{ref}^+(P)$ will be darker than but close to S_1, and $\text{ref}^-(P)$ will be brighter than but close to S_n. Therefore, the background model built by these reference points is robust to these background states. Moreover, there is no significant loss of sensitivity to the foreground, since the intensities of reference points are close to the darkest or brightest states. In conclusion, our background model is insensitive to background motion but sensitive to the foreground. A detailed analysis of its robustness to background motion will be presented in Section 3.5.

A sufficient number of reference points provide ample information for detection. It is not easy to provide general results; here we show some examples (the scenes of Figs. 10–12) to test our method of selecting reference points. Although they are examples, they include different scenes and subjects, and are representative to some extent. The detection results for each example are given in the experimental section.

As shown in Table 1, $T^-(T^+)$ defines the real detected number of $Q_n \in \text{ref}^-(P)(Q_n \in \text{ref}^+(P))$ for each P. In our experiment, we choose $N=16$, so our goal is to search eight $Q_n \in \text{ref}^-(P)$ and eight $Q_n \in \text{ref}^+(P)$ corresponding to each P. If $T^-=T^+=0$, the target point P has no reference points Q_n and, hence, cannot be classified; in this case, target points are considered background by default. If T^- or T^+ is less than eight, the judgment reliability is not as high as when T^- or T^+ is eight.

We set the thresholds W_G at 20 and W_P at 0.9 in all the examples in Table 1. In most examples in Table 1 (row 1, 3, 4, and 5), we did not find any target points those have no reference point. The number of target points with less than eight reference points is small. The data in these examples indicates that there are sufficient reference points for classification. However, one may claim that whether sufficient reference points can be searched depends strongly on the image content. Therefore, in images with low contrast the result would be different. Fig. 11 (row 2 of Table 1) is one example of a scene with low contrast and low texture. The number of target points with no reference points increases to 154. However, this number of pixels does not influence the overall situation, since it represents only a small percentage of all pixels. Since our method executes at the pixel-level, even if all of these pixels are misjudged, only some dispersed foreground pixels are undetected, which has little influence on the detection of entire objects.

3.4. Object detection

The proposed background subtraction process transforms the problem into a competitive binary classification problem [26]. In the context of object detection with a fixed camera, the objective is to add a binary label to each pixel in the input image. The most common technique of performing foreground detection is to take the difference between the input frame and background model [27,28]. We incorporate this technique into our model as well.

For an input image J, consider a target point P and its reference point $Q_n \in \mathrm{ref}(P)$. Their intensity relationship is

$$\mathrm{Ins}(P,Q_n) := \begin{cases} 1, & J(P) \geq J(Q_n), \\ -1, & J(P) < J(Q_n). \end{cases} \tag{8}$$

Then, comparing the sign of the intensity difference $\mathrm{Mos}(P,Q_n)$ (defined in Eq. (7)), determined by the model with the real sign of the intensity difference $\mathrm{Ins}(P,Q_n)$ (defined in Eq. (8)) and calculating the probability of the correct number of pairs, we find

$$\xi^+(P) = \frac{\#\{Q_n | \mathrm{Mos}(P,Q_n) = \mathrm{Ins}(P,Q_n) = 1\}}{\#\{Q_n | Q_n \in \mathrm{ref}^+(P)\}}, \tag{9}$$

$$\xi^-(P) = \frac{\#\{Q_n | \mathrm{Mos}(P,Q_n) = \mathrm{Ins}(P,Q_n) = -1\}}{\#\{Q_n | Q_n \in \mathrm{ref}^-(P)\}}. \tag{10}$$

For the point pair P and $Q_n \in \mathrm{ref}^-(P)$, and P and $Q_n \in \mathrm{ref}^+(P)$, we show their real distributions by picking four sample target points in Figs. 5(a) and (b), respectively. Red pixels are sample target points, which are connected to their corresponding yellow reference points. Using the pair points with different signs, the probability contour maps are shown in Figs. 5(c) and (d), where blue pixels represent higher foreground probability areas and red pixels represent lower foreground probability areas. Fig. 5(c) shows the contour map of probability $\xi^-(P)$, which helps detect light colored moving objects; Fig. 5(d) shows the contour map of $\xi^+(P)$, which assists in detecting dark colored moving objects.

Unlike SRF, we use a double-sided criterion. $\xi^+(P)$ and $\xi^-(P)$ are estimated with separate judgment systems, so that moving objects of darker or lighter color can be detected correctly. Pixel P in the input image is considered a background pixel only if both $\xi^+(P) > W_H$ and $\xi^-(P) > W_H$ (W_H is a global threshold that can be adjusted to achieve the desired result). Otherwise, pixel P is considered a foreground pixel, as shown in Fig. 5(e). The pseudo-code of the algorithm for detecting objects is presented in Algorithm 2.

3.5. Influence of background motion, and partial illumination changes

Our proposed method works well in complex environments. It is robust to illumination changes, and many situations containing background fluctuation.

For instance, in the general case of swaying leaves, when leaves swing in the wind, they change their location by moving back and forth, so pixels on the swaying leaves (especially those near the edge of leaves) sometimes represent a part of the leaf or a part of a road (or other background). Without loss of generality, we assume that the intensity of the road is lower than that of the leaves. In our background modeling, we find that reference points $Q_n \in \mathrm{ref}^-(P)$ are brighter than the leaves, and reference points $Q_n \in \mathrm{ref}^+(P)$ are darker than the road. Therefore, in both situations, the pixels would be classified as background correctly, as long as the probabilities of both situations are no less than $1-W_P$. Note $1-W_P$ is relatively small, for example, 0.1. This means the pixels on swaying leaves with probability from 10% to 90% can be classified as background correctly. Most pixels near the edge of leaves can be contained in such a probability range. Thus, neither the leaves nor the road will be classified as foreground. But note that, in some extreme situations, for instance, when the swing suddenly becomes severe (since this condition has not been learned or the proportion of images with this condition is less than $1-W_P$ in the training images), the influence of swaying leaves cannot be easily removed.

When there is a small amount of camera jitter, the vibration replaces pixels of fixed positions with neighboring pixels. Objects are smooth and the pixel intensities in the middle of objects change very little. Since the camera vibration is very small and neighboring pixels in a single object have similar intensities, pixels in the middle of each object are not significantly influenced by a small amount of camera jitter. Only pixels near the edges of objects are influenced and they alternate between two objects. So each pixel has two states: object one and object two. Thus, the problem is the same as with swaying leaves, and the pixels can generally be correctly classified as background.

Another frequent problem of background modeling is partial illumination changes, such as shadows. In some cases (shadows of stationary objects and those of regular moving objects), the

Algorithm 2. Object detection.

```
Input: Testing frame $J$; Threshold $W_H$; ref$^+$ and ref$^-$.
Output: Classification result of $J$.
for each pixel $P$ in $J$ do
    $Nall^+ \leftarrow |\mathrm{ref}^+(P)|; Nall^- \leftarrow |\mathrm{ref}^-(P)|; Ncnt^+ \leftarrow 0; Ncnt^- \leftarrow 0;$    // Initialization.
    for each component $Q \in \mathrm{ref}^+(P)$ do
        if $J(P) \geq J(Q)$ then
            $Ncnt^+ \leftarrow Ncnt^+ + 1;$
    $\xi^+(P) \leftarrow \frac{Ncnt^+}{Nall^+};$
    for each component $Q \in \mathrm{ref}^-(P)$ do
        if $J(P) < J(Q)$ then
            $Ncnt^- \leftarrow Ncnt^- + 1;$
    $\xi^-(P) \leftarrow \frac{Ncnt^-}{Nall^-};$
    if $\xi^+(P) > W_H$ and $\xi^-(P) > W_H$ then    // Classification results.
        $P$ is classified as background;
    else
        $P$ is classified as foreground.
```

problem can be handled by statistical point selection. In our proposed method, the absolute values of intensity differences between reference points and target points are larger than W_G, furthermore, as small as possible. This means the reference points are pixels with the most stable intensity difference with their corresponding target points. Obviously, the relationship of the pair of points in the same illumination environment is more stable than that in different illumination environments. In the case of partial illumination variation, the reference points are usually inside the same objects or objects have a similar illumination environment to that of their corresponding target points. For instance, the shadow of a house occasionally appears. The intensity difference between a pair of points that are both in the shadow is stable, while that between one point in the shadow and another in the sunshine is unstable (occasionally large or small). Thus, a reference point in the same shadow as the target point is preferred. The partial illumination changes do not change the stable relationship between the chosen points. Cast shadows are correctly classified as background, as long as the proportion of images under this shadow condition is not less than $1-W_P$ in the training images.

4. Accelerated modeling for GAP

In this section, we present an accelerated modeling of GAP. In the method presented thus far, every pixel of each image is processed by searching for reference points, which is an extremely computation-intensive process. To solve this problem, we adopt two steps: searching range reduction and spatial sampling steps. In the searching range reduction step, for each target point, the searching range of reference points is reduced; in the spatial sampling step, instead of searching reference points for each target point, the same reference points can be shared among target points with similar conditions. Adopting these two steps accelerates the GAP modeling.

4.1. Searching range reduction

The searching range reduction (SRR) step reduces the cost of calculation by constraining the search for reference points. In Section 3, we search for reference points through the whole image, which incurs a large amount of calculation. Therefore, in this step, we narrow the search for reference points to a small range of points, without missing any reference points. Section 3 showed that reference points have a constant intensity difference with the corresponding target points in most frames, so the mean intensity differences among frames are not too small or too large. In other words, the range of mean intensity differences is limited. Next, we discuss how to determine this range.

To simplify the explanation, consider reference points $Q \in \text{ref}^-(P)$. The situation is the same for $Q \in \text{ref}^+(P)$. In most frames (a fraction greater than W_P), pair points P and Q satisfy the known intensity relationship as $I_t(P)-I_t(Q) \leq -W_G$. On the other hand, in the remaining frames (a fraction smaller than $1-W_P$), since the range of intensity is from 0 to 255 (here we use 8-bit grayscale images), the extreme situation of intensity is $I_t(P)=255$ and $I_t(Q)=0$. In this case, the relationship between P and Q satisfies $I_t(P)-I_t(Q) \leq 255$. From the above assumption, we can make the conclusion that the mean intensity difference $(\overline{I(P)}-\overline{I(Q)})$ is smaller than $-W_G \cdot W_P+255 \cdot (1-W_P)$ in the case of $Q \in \text{ref}^-(P)$. Based on the same principle, in the case of $Q \in \text{ref}^+(P)$, the mean intensity difference $(\overline{I(P)}-\overline{I(Q)})$ is larger than $W_G \cdot W_P+(-255)(1-W_P)$.

For instance, supposing W_G and W_P are set at 20 and 0.9, respectively, in the case of $Q \in \text{ref}^-(P)$, the mean intensity difference is smaller than 7.5. If we take $\overline{I(P)}=128$ for example, we only search points that satisfy the requirement of $\overline{I(Q)} > 120.5$. Obviously, instead of searching points Q, whose mean intensities are in the range from 0 to 255, only points with a mean intensity in the range from 120.5 to 255 are involved. And, in the case of $Q \in \text{ref}^+(P)$, only points that satisfy $\overline{I(Q)} < 135.5$ are searched. Thus, nearly half of the searching range has been reduced, which consequently leads to the reduction of total searching time.

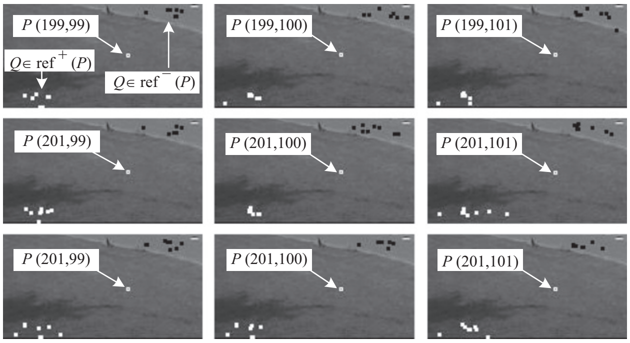


Fig. 6. Distribution of reference points for proximal target points.

4.2. Spatial sampling

The spatial sampling (SS) step reduces the computation by having several target points share the same set of reference points. Only a portion of the target points, called sample target points, need to search reference points, while others can share these same reference points. From the natural high correlation of neighboring points, it is known that pixel intensities in the neighborhood remain similar over time. In this case, the reference points of neighboring points are likely to have a similar distribution, as shown in Fig. 6. Thus, for several target points, it is unnecessary to execute a complete search for reference points, since reference points of neighboring target points can be used.

Certain conditions must be met for a pixel to skip the search for reference points. First, the points it can share reference points which should be in the neighborhood. Second, obviously not all the neighboring points can share reference points, so we use an additional condition. Consider one point P, whose reference points are unknown, and its neighboring point P' whose reference points are known (P' is the selected sample target point, and the first pixel in the image is always considered a sample target point). These two points must have the smallest mean intensity difference in the eight-connected neighborhood. Then, comparing their intensities in each frame allows us to define their statistical intensity difference as

$$L = \#\{t \| I_t(P')-I_t(P)| < W_S, 1 \leq t \leq T\}/T, \tag{11}$$

where W_S is an intensity difference threshold—a small W_S leads to a small risk. At this point, whether pixel P has the need to search for reference points or not can be estimated using L. If $L > W_B$ (W_B is another given threshold; $0 < W_B < 1$), then P and P' can share the same reference points. Otherwise, a complete search for reference points for P must be executed.

In this subsection, two additional parameters W_S and W_B are used. The selection of these two parameters will be discussed in Section 5.2.

The pseudo-code of the accelerated version of background modeling for our algorithm is shown in Algorithm 3.

Algorithm 3. Background modeling (accelerated version).

```
Input: $T$ Training frames (frame $t$ is denoted by $I_t$); Thresholds $W_B$, $W_S$, $W_G$, $W_P$.
Output: $\text{ref}^+$, $\text{ref}^-$.
$W \leftarrow W_G \cdot W_P + (-255)(1-W_P)$; $S \leftarrow \emptyset$;    // $S$ is the pixel set containing the points which have already found their reference points.
Sort all the pixels as $\{Q_1, Q_2, \ldots Q_U\}$ in the ascending order of mean intensity $\overline{I(Q_u)}$ among $T$ frames;
    // $U$ is the number of pixels of frame, and $u \in \{1, \ldots, U\}$.
for each pixel $P$ do
    $\text{ref}^+(P) \leftarrow \emptyset$; $\text{ref}^-(P) \leftarrow \emptyset$; $Cnt^+ \leftarrow 0$; $Cnt^- \leftarrow 0$;
    $P' = \operatorname{argmin}_{\hat{P} \in Nbr(P)} |\overline{I(\hat{P})} - \overline{I(P)}|$;    // $Nbr(P)$ is the neighboring points set of $P$.
    if $P' \in S$ and $L > W_B$ then    // $L$ is calculated by Eq. (11).
        $\text{ref}^+(P) \leftarrow \text{ref}^+(P')$; $\text{ref}^-(P) \leftarrow \text{ref}^-(P')$;
        Add $P$ into $S$;
        break;
    for $u = U : 1$ do
        if $\overline{I(P)} - \overline{I(Q_u)} > W$ then
            if $Q_u$ satisfies Eq. (1) then
                Add $Q_u$ into $\text{ref}^+(P)$;
                $Cnt^+ \leftarrow Cnt^+ + 1$;
                if $Cnt^+ = \frac{N}{2}$
                    break;
    for $u = 1 : U$ do
        if $\overline{I(P)} - \overline{I(Q_u)} < -W$ then
            if $Q_u$ satisfies Eq. (2) then
                Add $Q_u$ into $\text{ref}^-(P)$;
                $Cnt^- \leftarrow Cnt^- + 1$;
                if $Cnt^- = \frac{N}{2}$ then
                    break;
    Add $P$ into $S$;
Return $\text{ref}^+$, $\text{ref}^-$.
```

5. Experimental results

To evaluate the performance of our proposed method, we applied it to object detection with real video data, in both indoor and outdoor environments. The experiment is divided into two parts. In part one, we tested our results with several challenging video sequences under different complex conditions and analyzed our method by comparing the results with those of four existing approaches: adapted MoG [8], Co-occurrence [17], Sheikh's Bayesian model [18], and SRF [22]. For part two of our experiment, we evaluated the performance of accelerated modeling of GAP under different parameter settings. The images used in the experiments are 320 × 240 pixels (in the database of Toyama [14], the image resolution is 120 × 160), and RGB images are converted into gray level images. Note that we did not apply any post-processing in our results, so some dispersed noise may appear. However, that noise can be easily removed by simple post-processing, such as a morphological operation. The dynamic results, shown in Figs. 10–12, can be found in our website [29].

5.1. Evaluation under different complex conditions

Generally, in most papers, authors like to use their own test sequences. This makes comparing the different approaches rather difficult. In this paper, we demonstrate the performance of our method with various complex video data, which are often used in comparison.

Toyama et al. in their work [14] provided a challenging dataset which contains seven video sequences, addressing specific background subtraction problems. We chose this dataset as one of our experimental subjects. Dataset 3 in the PETS2001 database consists of two camera scenes and each contains a sequence of 5336 frames [30]. These are both outdoor scenes captured under severe illumination conditions—one features moving clouds while the other contains the difficult problem of swaying trees. Another database is an indoor scene provided by the National Institute of Advanced Industrial Science and Technology (AIST) in Japan. It comprises 4890 frames under more abrupt and severe illumination changes caused by the switching electric lights on and off. We also selected two traffic video sequences from the website maintained by KOGS/IAKS Universität Karlsruhe [31]. The first sequence contains 336 frames in an environment of constant illumination, but with heavy fog. The second one containing 1733 frames is under small camera vibration. Excepting the fog database, 300 training frames with an average interval we picked to build our background model. For the fog database, we picked 100 training frames. Since the exposure time of moving objects on a pixel is short, the existence of moving objects does not affect the statistical point choice. Thus, there is no need to initialize the training images without foreground objects. The parameters in the following methods are set to the recommended values by the references respectively.

Fig. 10 shows five frames from camera one of dataset 3 in the PETS2001 database with rapid changes in illumination. This outdoor scene contains a sequence with groups of people and has significant lighting variation. The slow moving clouds are another

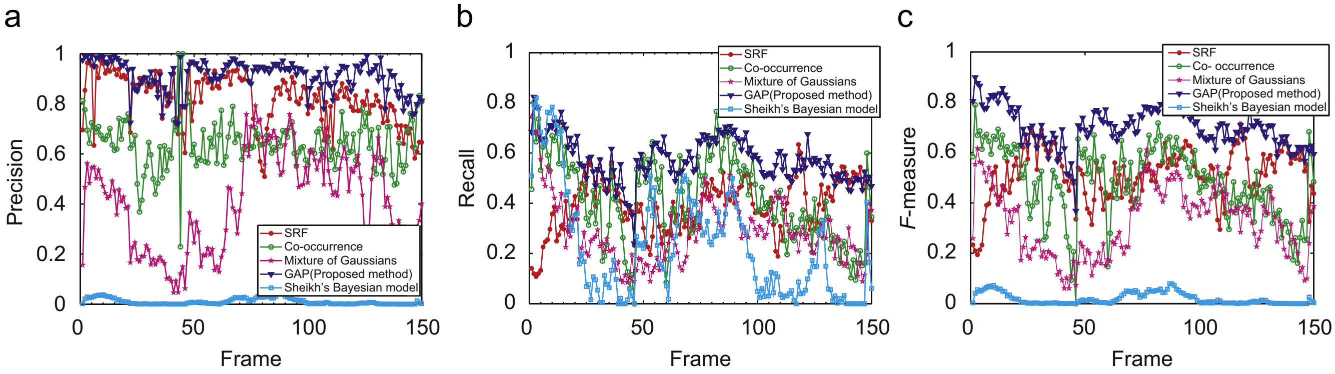


Fig. 7. Comparison between Co-occurrence, SRF, Mixture of Gaussians, Sheikh's Bayesian model, and GAP. (a) shows the precision comparison result, (b) shows the recall comparison result, and (c) shows the *F*-measure comparison result.

Table 2
Mean precision, recall, and *F*-measure values for each method.

Methods	Precision	Recall	*F*-measure
SRF	0.8191	0.3894	0.5152
Co-occurrence	0.6432	0.4143	0.4828
Mixture of Gaussians	0.4020	0.2896	0.3233
Sheikh's Bayesian model	0.0104	0.2288	0.0196
GAP (proposed method)	0.9120	0.5751	0.7026

challenge in this scene. Row (b) in Fig. 10 shows the result of the Co-occurrence method with a dimension of $K=3$ and $L=6$ neighboring points. Moving people are roughly detected by blocks, which should only be used for approximate detection. Row (c) shows the results of adopting the SRF method (in which, $T_P=10$, $T_B=0.75$, $T_R=0.9$). We find that some moving people are not detected, and some are incompletely extracted. Moreover, there is a large amount of noise in crowded scenes. Row (d) gives the result from the adapted MoG method with five Gaussian components and a learning rate of 0.01. This method does not properly model the background or correctly detect the moving people. The results in Row (e) are given by Sheikh's Bayesian model, where the size of matrix H is $[26,26,26,31,41]$. This is a recent method that performs very well in dynamic scenes; however, this method is very sensitive to illumination. Large levels of noises are distributed in the area of changing illumination. Row (f) shows the results of the GAP method (in which $W_G=20$, $W_P=0.9$, $W_H=0.3$), and post-processing, such as morphological processing, was not used in our results. Compared with the other four approaches, our background model appears more robust and accurate, our results have less noise, and the moving people are more clearly detected even from a distance (Row (f), second column).

Fig. 11 shows another typical experimental result from an indoor scene containing more rapid and severe illumination changes caused by switching lights on and off. Highlights on the floor are another detection problem. We did not use Sheikh's Bayesian model for comparison, since the frames in this database are gray level images, and Sheikh's Bayesian model becomes inaccurate if color information is lost. The adapted MoG method loses the person in a suddenly dark environment (Row (d), second column), while SRF detects the persons in all frames, but suffers from significant noise. Although the Co-occurrence algorithm provides an outline of the moving person with the least noise, it only gives a rough outline of the person and loses the person in some situations (Row (b), second column). Thus, under extreme variation of illumination, these methods typically cannot provide either accurate detection or robustness with respect to noise.

Three more experimental results of our proposed method are shown in Fig. 12. The images shown in Row (b) of Fig. 12 are the results from camera two of dataset 3 in PETS2001. In addition to the challenge of varying illumination, the swaying tree in this scene makes detection more difficult. Also, shadows cast by trees and partial illumination variations caused by clouds are obvious in this scenario. It can be seen that our proposed method detects the moving person accurately, even when the person passes through the shadow cast by trees (Row (b), first column). Moreover, note that, in the second column of Row (b) in Fig. 12, in the middle-right part of the image, a fast moving bird has also been detected successfully. In our results, although when swings of branches suddenly become severe, noises come to appear on the edge of leaves, this is because such extreme situations have not been learned in the training images. Nerveless, in general situations, such as when the swings are not so severe, our detection results are good. Row (c) in Fig. 12 shows sample frames from KOGS/IAKS Universität Karlsruhe under different conditions. Frames in the first two columns are the cases suffering small camera jitter and slightly swaying leaves. The challenge in the cases of last two columns is the heavy fog, which causes low visibility. Our proposed method successfully detects most of the moving cars in the two sequences. The undetected cars are those that stopped in the street for a long time, and which are treated as background.

Another dataset, which is introduced in the work of Toyama et al. [14], is also used to test our method. This dataset consists of seven video sequences, each of which addresses a specific canonical background subtraction problem. Our results are shown in Fig. 13. Compared with the results in [14], our method masters the illumination changes and background fluctuation well, especially the case of switching lights, which cannot be handled by most of the other methods. "Camouflage" is a special case. It is difficult to avoid the false detection of wall pixels, because a large area of the moving object covers most of the image textures, leaving only the low texture of wall; the reference points of wall pixels are easily covered by foreground, which changes the original stable intensity relationship. However, in many cases of interest, it is unusual for the foreground to occupy such a large proportion of pixels.

For the PETS2001 database, ground truth data are available [32], so we can form a qualitative analysis by comparing adaptive MoG, Co-occurrence, Sheikh's Bayesian model, and SRF with the GAP method using 150 sample images. The sample images contain different conditions (such as a group of people and illumination changes), which would result in an unstable result. Therefore, to analyze the temporal stability of method, three information retrieval measurements, recall, precision and *F*-measure, are utilized [33]. Precision can be seen as a measure of exactness or fidelity, whereas recall is a measure of completeness. They are

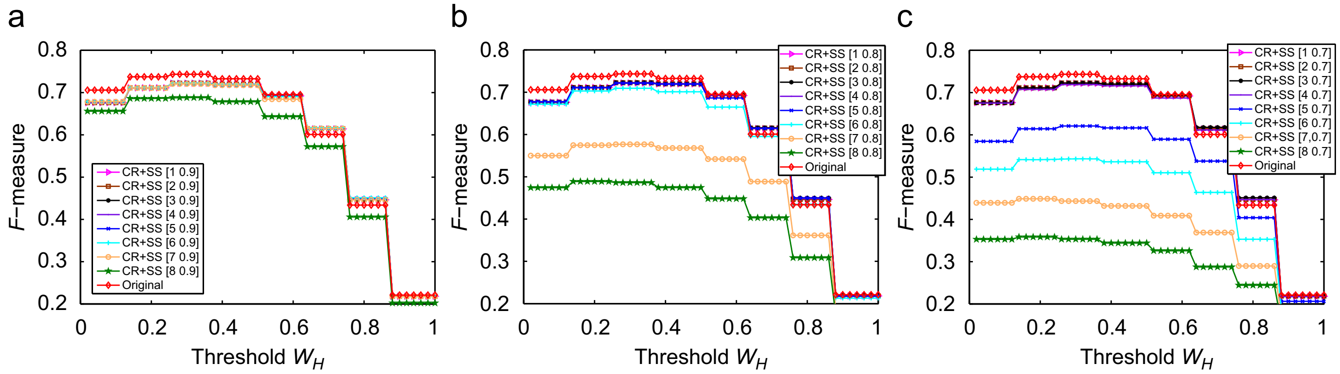


Fig. 8. *F*-measure for different parameter values. While one parameter W_S was varied, the other one W_B kept fixed.

Table 3
Experimental results of accelerated modeling under different threshold settings for W_S and W_B.

Input conditions	[W_S ,W_B]	Number of sample target points	Degradation of *F*-measure (%)	Reduction of computation time (%)
SRR only	N.A.	76,800	2.852	59.520
SRR and SS	[1 0.9]	76,528	2.852	73.099
SRR and SS	[2 0.9]	71,040	2.852	73.372
SRR and SS	[3 0.9]	63,513	2.812	78.400
SRR and SS	[4 0.9]	47,515	2.812	82.060
SRR and SS	[5 0.9]	38,797	2.825	84.547
SRR and SS	[6 0.9]	32,318	2.825	86.906
SRR and SS	[7 0.9]	27,249	3.081	88.672
SRR and SS	[8 0.9]	23,354	7.440	89.452
SRR and SS	[1 0.8]	76,083	2.852	73.396
SRR and SS	[2 0.8]	64,463	2.798	76.531
SRR and SS	[3 0.8]	49,396	2.758	81.125
SRR and SS	[4 0.8]	38,196	2.825	84.012
SRR and SS	[5 0.8]	36,715	2.960	87.440
SRR and SS	[6 0.8]	24,879	4.561	89.509
SRR and SS	[7 0.8]	20,741	22.400	91.047
SRR and SS	[8 0.8]	17,423	32.212	92.093
SRR and SS	[1 0.7]	75,237	2.852	73.578
SRR and SS	[2 0.7]	58,213	2.785	77.717
SRR and SS	[3 0.7]	42,119	2.785	82.866
SRR and SS	[4 0.7]	31,464	3.283	86.771
SRR and SS	[5 0.7]	24,661	16.454	89.519
SRR and SS	[6 0.7]	19,900	26.853	91.108
SRR and SS	[7 0.7]	16,193	39.620	92.360
SRR and SS	[8 0.7]	13,282	51.769	93.800

defined as

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}, \tag{12}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}, \tag{13}$$

The *F*-measure considers both the precision and the recall in computing the score, which can be interpreted as a weighted harmonic mean of the precision and recall. The formula of the *F*-measure is

$$F = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{14}$$

A comparison of precision, recall and *F*-measure is shown in Fig. 7. We also recorded the mean precision, recall and *F*-measure values of each method in Table 2. Clearly, the detection accuracy for precision, recall and *F*-measure is consistently higher using our method than with adapted MoG, SRF, Sheikh's Bayesian model or the Co-occurrence approach. This indicates that our method has good temporal stability.

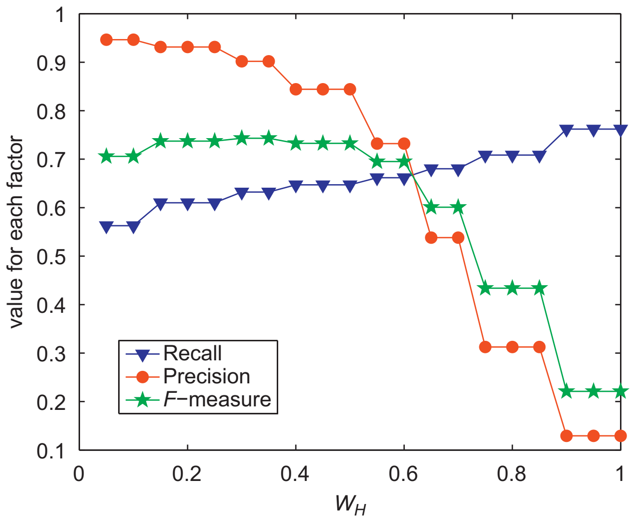


Fig. 9. The discussion of W_H.

5.2. Analysis of accelerated modeling

In a separate experiment, we examined the results from the accelerated modeling method, wherein we used 300 frames from camera one of dataset 3 in the PETS2001 for background modeling. Here, the *F*-measure is used to measure the error rate.

The performance of our method was tested under two input conditions: with SRR and SS steps and with only a SRR step. The SRR step is parameterless. In the SS step, there are two parameters, W_B and W_S, which are used for determining whether two target points can share the same reference points. As introduced in Section 4.2, theoretically speaking, two target points can share the same reference points as long as their intensities are similar statistically. Here, the similarity means that their intensity difference is below W_S. And the statistical condition is that the probability of being similar is beyond W_B. Therefore, in practical applications, it is obvious that W_S cannot be set too large and W_B cannot be set too small. Thus, here, we chose parameter W_B as 0.9, 0.8 and 0.7, independently. For each value of W_B, W_S is set from one to eight.

Figs. 8(a)–(c) compare the *F*-measure value of the original modeling method with those of accelerated modeling at different thresholds. Table 3 shows the performance comparisons expressed by the reduction rates of computation time and degradation rates of *F*-measure. From the first row of Table 3, we can see that by using

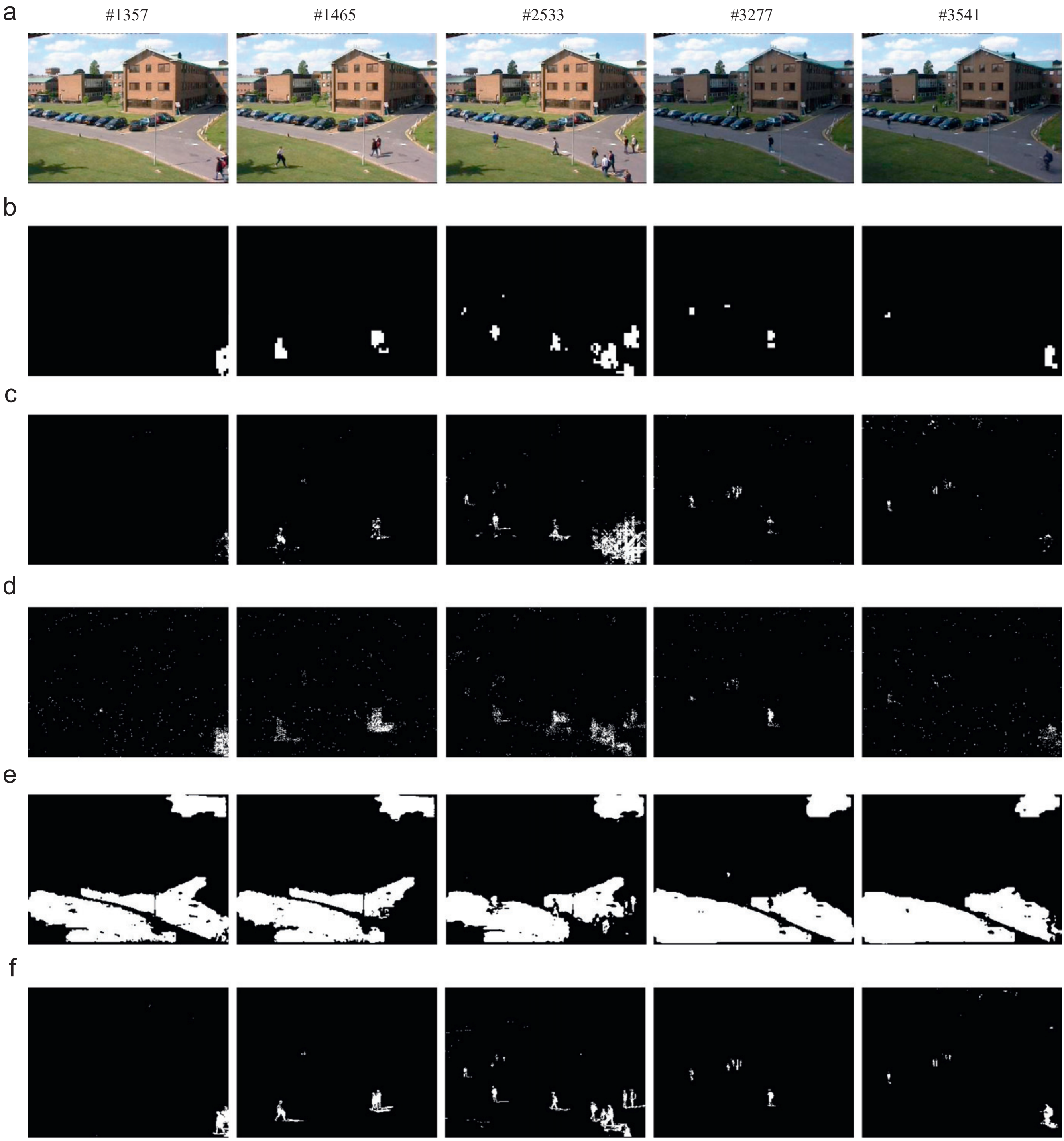


Fig. 10. Results of applying several object detection methods to PETS2001 (camera one). (a) The original images; (b) the results of Co-occurrence; (c) the results of SRF; (d) the results of adapted MoG; (e) the results of Sheikh's Bayesian model; (f) the results of GAP (our proposed method).

only the SRR step, more than half of computation time (59.520%) has been reduced without much change in the *F*-measure (2.852% of degradation). The other rows of Table 3 show that, in the SS step, smaller values of W_B and larger values of W_S decrease the number of sample target pixels, resulting in a reduction in computation time. When $W_S \leq 7$, and $W_B \geq 0.9$, the degradation of *F*-measure is quite small (smaller than 2.9%), while the computational time can be reduced dramatically (up to 88%). When W_B is decreased to 0.8 or 0.7, although the degradation of *F*-measure is dramatic, the reduction of computational time is quite small compared with that when $W_B = 0.9$. Thus, there is no need to set a very small W_B. To reduce the computation time with little degradation of the *F*-measure, we suggest $W_S \leq 7$ and $W_B \geq 0.9$, which are consistent with our empirical study with different videos.

5.3. Time and computational complexity

There are two components of computation time: (1) background modeling for training images is performed off-line and (2) objects of interest are detected on-line. The computer used for the testing is Intel (R) 3.06 GHz with 2.99 GB RAM. The algorithm is implemented in MATLAB (Version 7).

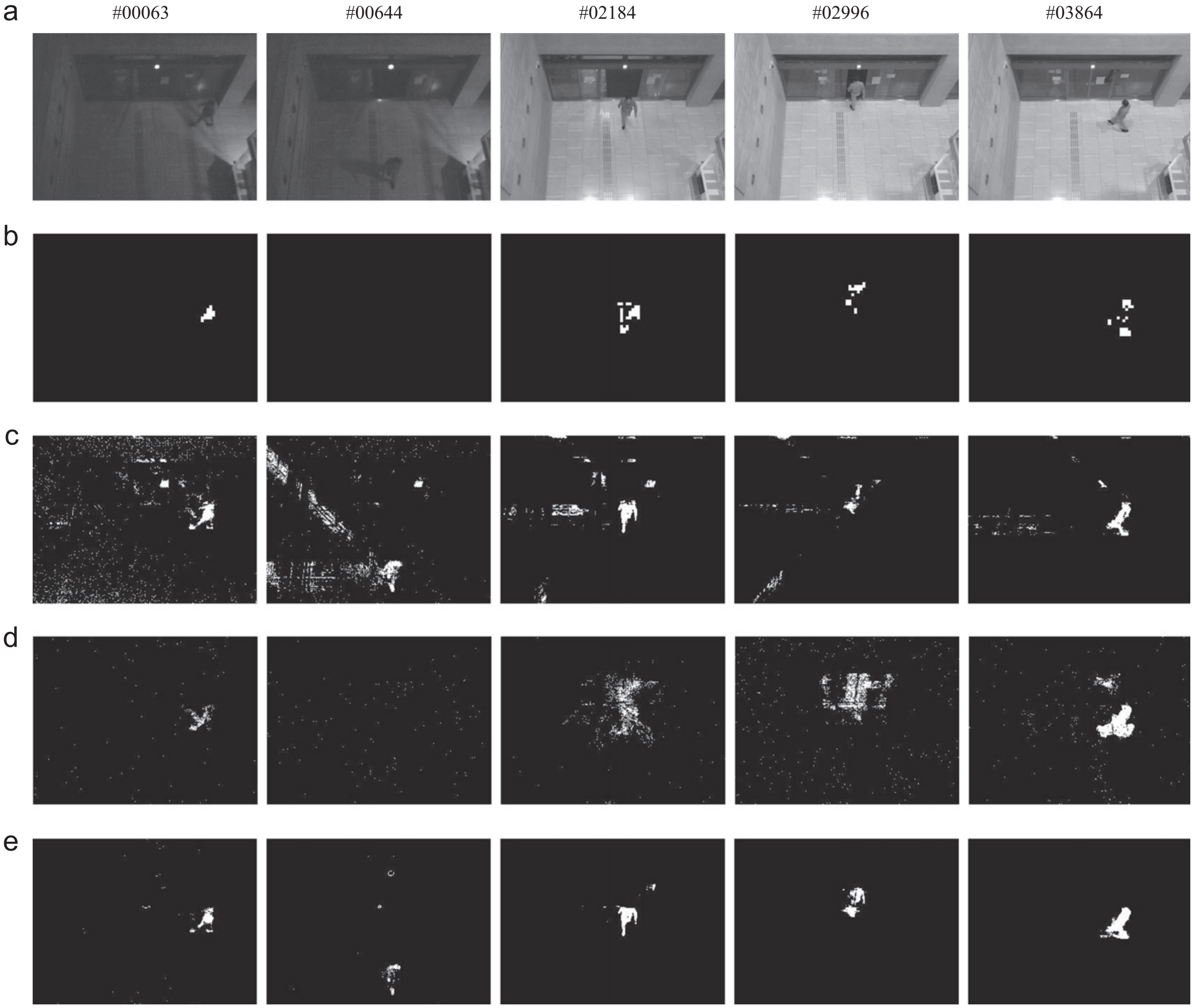


Fig. 11. Results of applying the various detection methods to an indoor scene. (a) The original images; (b) the results of Co-occurrence; (c) the results of SRF; (d) the results of adapted MoG; (e) the results of GAP (our proposed method).

In background modeling, the worst case is calculating all of the pixels in the training images for searching reference points of a single pixel. In this case, all the n pixels in all T training images participate in the calculation. Thus, in terms of computational complexity, for training data of length T, and image pixels of number n, in the worst case, the time complexity for background modeling is $O(T \cdot n^2)$. The running time of the original background modeling in our experiment is 192 s per frame, and the time is reduced to 25 s per frame with the accelerated step ($W_B = 0.9$, $W_S = 6$). Since our algorithm is implemented in MATLAB, the system will become much more efficient with the implementation of C programming language.

In terms of the execution speed of object detection, the method is very efficient. Since it requires only the computation of comparing the sign of the intensity difference, the computational complexity is linear for the number of image pixels: $O(n)$. In our experiment, the average computation time for one frame takes 18 ms, which is sufficiently fast for real-time applications. With the optimization of the implementation, the system's efficiency could be much further improved.

5.4. Parameter discussion

The proposed method mainly involves three significant parameters: W_G, W_P and W_H. Next, we discuss how to select suitable values.

As discussed in Section 3.3, W_G represents the ideal value of environmental fluctuation. The more severe the environmental fluctuation (e.g., dramatic illumination changes), the larger W_G should be. Otherwise, more foreground pixels will be incorrectly classified as background or more noises will be classified as foreground. This threshold has been set according to different environmental changes and represents the ideal value of environmental fluctuation. However, tuning the value to be appropriate for different environments is difficult for users. Based on the empirical study, we offered an interval [10, 30] of W_G for reference. A value lower than 10 may make the model too sensitive to small changes, thus causing the false detection of background. On the other hand, a value higher than 30 will make foreground not easy to be detected. Also, it may lead to an insufficient number of points. Consequently, in most of our databases, we defined $W_G = 20$ as a suitable value.

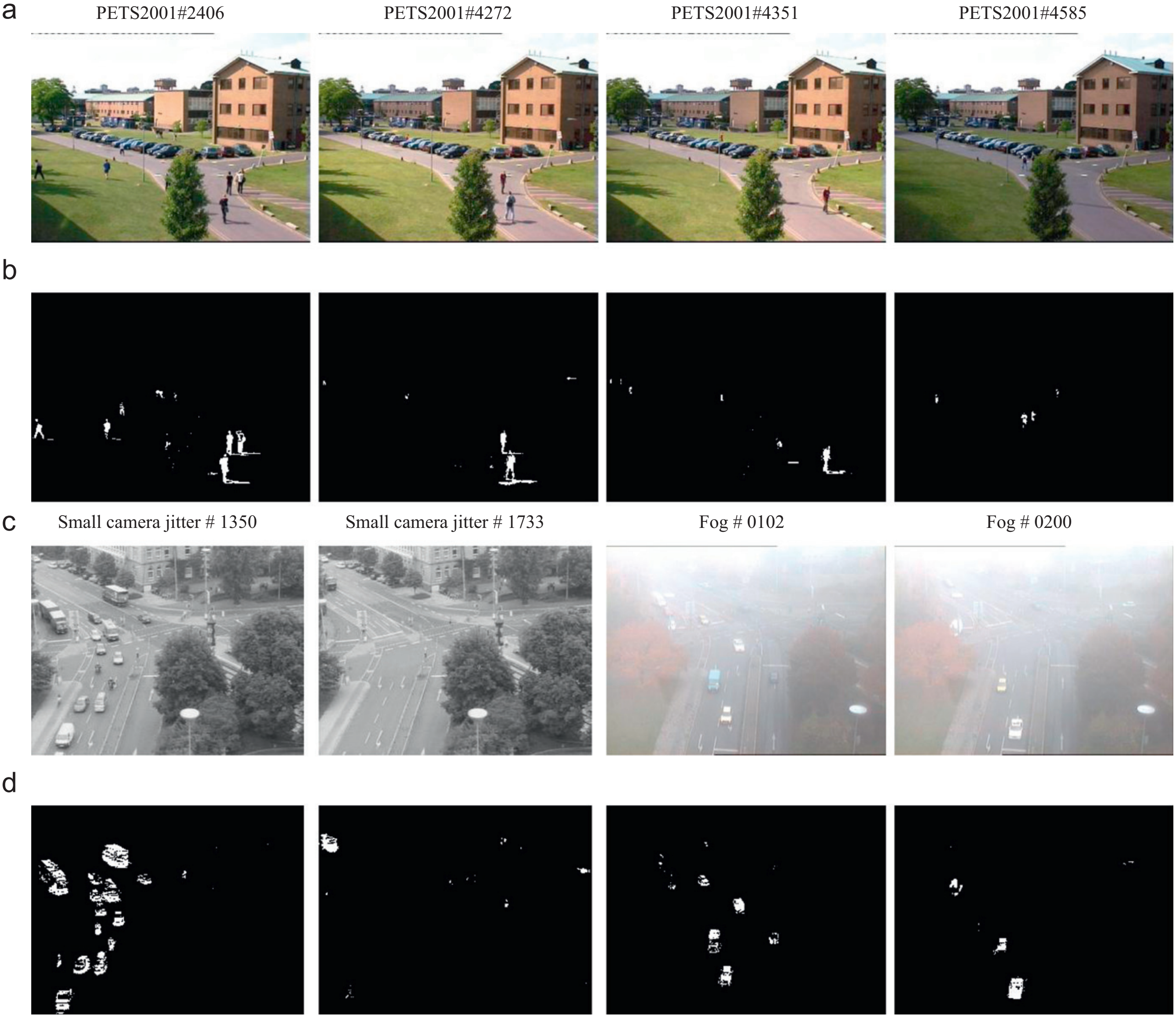


Fig. 12. Results of our proposed methods with three other challenging databases. (a) The original images of PETS2001 (camera two) database; (b) the background subtraction results of (a) using GAP (our proposed method); (c) the original images of database from KOGS/IAKS Universität Karlsruhe. (d) the background subtraction results of (c) using GAP (our proposed method).

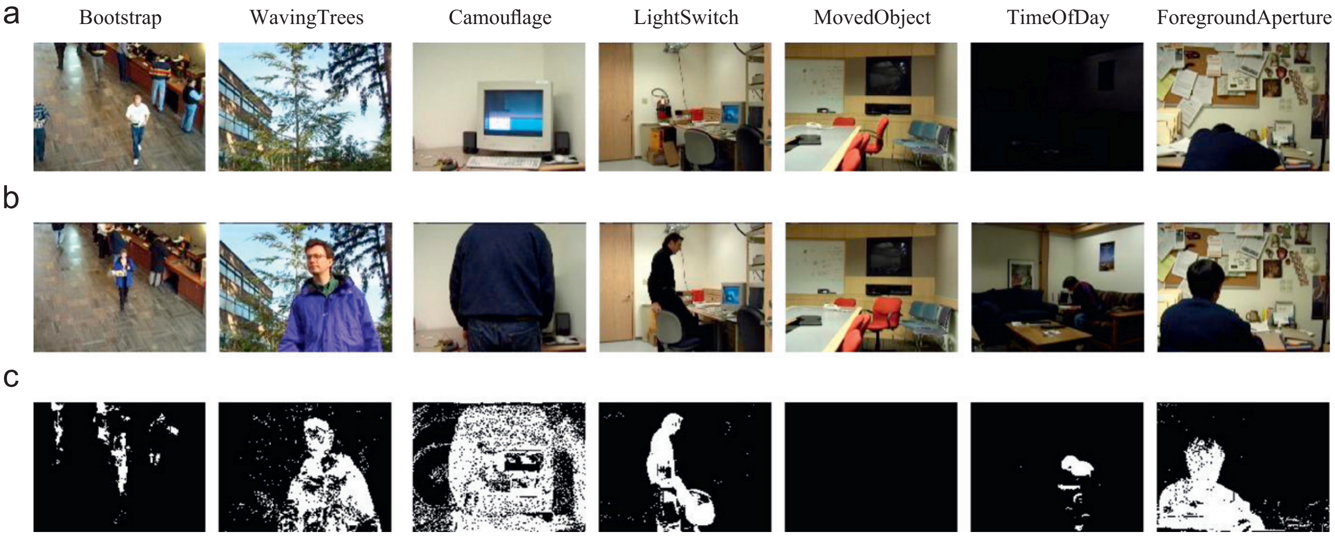


Fig. 13. Detection results of our method for the test sequences presented in Ref. [14]. (a) The first frames; (b) the test frames; and (c) the results of GAP (our proposed method).

W_P is the proportion threshold of the number of required images. Generally, a large W_P works well in a complex environment, but if it is too large, it is hard to get a sufficient number of points, which will lead to a false detection. Experimental results show that setting $W_P = 0.9$ is already efficient in different complex environments. So in this paper, we set $W_P = 0.9$.

W_H is a trade-off parameter. Fig. 9 shows that a smaller value of W_H leads to a higher performance of precision, while larger value leads to higher performance of recall. As can be seen in this figure, the *F*-measure shows that in the range from 0.2 to 0.5, W_H leads to the best performance considering both the precision and the recall. Thus, we selected $W_H = 0.3$ as a suitable value in this paper.

6. Conclusion and future works

In this paper, we proposed a novel background modeling method called the grayscale arranging pairs (GAP) for object detection in scenes with different complex conditions. There are a number of innovations in this work. First, we analyzed the stability of intensity between point pairs. The intensity difference showed better stability than the intensity, even in complex environments. Then, we built novel selection rules to choose appropriate point pairs that maintain a stable intensity difference during changes in the global spatial domain. We also used a double-sided judging standard to reduce false detection. Finally, we presented the results of experiments, comparing GAP to other object detection methods, which indicted superior results for the GAP method.

In practice, the adaptation of the background model is also important in real-time detection, which we will demonstrate in our future work. We plan to adjust the detection criterion to accommodate real environmental changes using an updated model. Furthermore, we will use our method in other fields such as tracking and behavior recognition.

References

[1] A. Yilmaz, O. Javed, M. Shah, Object tracking: a survey, ACM Computing Surveys (CSUR) 38 (4) (2006) 12.

[2] G.A. Robert, R.W. Lance, Multiple target tracking with lazy background subtraction and connected components analysis, Machine Vision and Applications 20 (2) (2009) 93–101.

[3] S. Cheung, C. Kamath, Robust background subtraction with foreground validation for urban traffic video, EURASIP Journal on Applied Signal Processing 2005 (14) (2005) 2330–2340.

[4] S. Ali, A. Basharat, M. Shah, Chaotic invariants for human action recognition, in: IEEE International Conference on Computer Vision (ICCV), 2007, pp. 1–8.

[5] I. Haritaoglu, D. Harwood, L.S. Davis, W4: real-time surveillance of people and their activities, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 809–830.

[6] A. Datta, M. Shah, N.D.V. Lobo, Person-on-person violence detection in video data, in: IEEE International Conference on Pattern Recognition (ICPR), 2002, pp. 11–15.

[7] C.R. Wren, A. Azarbayejani, T. Darrell, A.P. Pentland, Pfinder: real-time tracking of the human body, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 780–785.

[8] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real-time tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, 1999, pp. 246–252.

[9] C. Stauffer, W.E.L. Grimson, Learning patterns of activity using realtime tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 747–757.

[10] A. Elgammal, R. Duraiswami, D. Harwood, L.S. Davis, Background and foreground modeling using nonparametric kernel density estimation for visual surveillance, Proceedings of the IEEE 90 (2002) 1151–1163.

[11] J. Rittscher, J. Kato, S. Joga, A. Blake, A probabilistic background model for tracking, in: European Conference on Computer Vision (ECCV), 2000, pp. 336–350.

[12] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, J. Buhmann, Topology free hidden Markov models: application to background modeling, in: IEEE International Conference on Computer Vision (ICCV), 2001, pp. 294–301.

[13] Z. Zivkovic, Improved adaptive Gaussian mixture model for background subtraction, in: IEEE International Conference on Pattern Recognition (ICPR), 2004.

[14] K. Toyama, J. Krumm, B. Brumitt, B. Meyers, Wallflower: principles and practice of background maintenance, IEEE International Conference on Computer Vision (ICCV), vol. 1, 1999, pp. 255–261.

[15] N.M. Oliver, B. Rosario, A.P. Pentland, A Bayesian computer vision system for modeling human interactions, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 831–843.

[16] A. Monnet, A. Mittal, N. Paragios, V. Ramesh, Background modeling and subtraction of dynamic scenes, IEEE International Conference on Computer Vision (ICCV), vol. 2, 2003, pp. 1305–1312.

[17] M. Seki, T. Wada, H. Fujiwara, K. Sumi, Background subtraction based on cooccurrence of image variations, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2003, pp. 65–72.

[18] Y. Sheikh, M. Shah, Bayesian modeling of dynamic scences for object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (11) (2005) 1778–1792.

[19] M. Heikkilä, M. Pietikäinen, A texture-based method for modeling the background and detecting moving objects, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (4) (2006) 657–662.

[20] Y. Satoh, C. Wang, H. Tanahashi, Y. Niwa, K. Yamamoto, Robust object detection for intelligent surveillance systems based on radial reach correlation, IEEE/RSJ International Conference on Intelligent Robots and Systems, vol. 1, 2003, pp. 224–229.

[21] N. Wajima, S. Takahashi, M. Itoh, Y. Satoh, S. Kaneko, Robust object detection based on radial reach filter for mobile robots, in: SICE-ICASE International Joint Conference, 2006, pp. 1828–1831.

[22] K. Iwata, Y. Satoh, R. Ozaki, K. Sakaue, Robust background subtraction based on statistical reach feature method, The IEICE Transactions on Information and Systems (8) (2009) 1251–1259 (in Japanese).

[23] R. Jain, H. Nagel, On the analysis of accumulative difference pictures from image sequences of real world scenes, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1 (2) (1979) 206–213.

[24] C. Liu, P.C. Yuen, G.P. Qiu, Object motion detection using information theoretic spatio-temporal saliency, Pattern Recognition 42 (11) (2009) 2897–2906.

[25] T. Ko, S. Soatto, D. Estrin, Background subtraction on distributions, in: European Conference on Computer Vision (ECCV), 2008, pp. 276–289.

[26] S.Y. Elhabian, K.M. El-Sayed, S.H. Ahmed, Moving object detection in spatial domain using background removal techniques-state-of-art, Recent Patents on Computer Science 1 (1) (2008) 32–54.

[27] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, Detecting moving objects, ghosts and shadows in video streams, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (10) (2003) 1337–1342.

[28] J.C. Nascimento, J.S. Marques, Performance evaluation of object detection algorithms for video surveillance, IEEE Transactions on Multimedia 8 (4) (2006) 761–774.

[29] ⟨http://www.ssc-lab.com/research_contents/gap.html⟩.

[30] ⟨http://ftp.pets.rdg.ac.uk/pub/⟩.

[31] ⟨http://i21www.ira.uka.de/image_sequences/⟩.

[32] ⟨http://limu.ait.kyushuu.ac.jp/en/dataset/⟩.

[33] J. Makhoul, F. Kubala, R. Schwartz, R. Weischedel, Performance measures for information extraction, in: DARPA Broadcast News Workshop, 1999, pp. 249–252.

Xinyue Zhao received the M.E. degrees in Mechanical Engineering from Zhejiang University, China, in 2008. She is studying for the Ph.D. course at Graduate School of Information Science and Technology, Hokkaido University, Japan. Her research interests include computer vision and image processing.

Yutaka Satoh received Ph.D. degree in Systems Engineering from Hokkaido University, Japan, in 2001. He is a research scientist of National Institute of Advanced Industrial Science and Technology(AIST), Japan, from 2004. He is also an Associate Professor (Cooperative Graduate School Program) of the University of Tsukuba, Japan. His research interest includes machine vision systems, mobile robots and robust pattern recognition algorithms.

Hidenori Takauji received the Ph.D. degree in Systems and Information Engineering from Hokkaido University, Japan, in 2006. He is currently a post-doctoral fellow of GCOE Program in Hokkaido University. His research interests include robotic vision, image sensing and robotic intelligence.

Shun'ichi Kaneko received the B.S. degree in Precision Engineering and the M.S. degree in Information Engineering from Hokkaido University, Japan, in 1978 and 1980, respectively, and then the Ph.D. degree in System Engineering from the University of Tokyo, Japan, in 1990. He had been a research assistant of the Department of Computer Science since 1980 to 1991, an associate Professor of the Department of Electronic engineering since 1991 to 1995, and an associate Professor of the Department of Bio-application and Systems Engineering since 1996 to 1996, in Tokyo University of Agriculture and Technology, Japan. He is currently a Professor at the Graduate School of Information Science and Technology, Hokkaido University, Japan. His research interests include machine and robot vision, image sensing and understanding, and robust image registration.

Kenji Iwata received his Ph.D. degree in Engineering from Gifu University, Japan, in 2002. He is a research scientist of National Institute of Advanced Industrial Science and Technology (AIST), Japan, from 2005. His research interest includes computer vision and middleware for its systems.

Ryushi Ozaki is a student of Graduate School of Systems and Information Engineering, the University of Tsukuba, Japan. He is interested in computer vision.