2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance

Co-occurrence-based Adaptive Background Model for Robust Object Detection

Dong Liang
Hokkaido University, Japan
liang@ssc.ssi.ist.hokudai.ac.jp

Shun'ichi Kaneko
Hokkaido University, Japan

Manabu Hashimoto
Chukyo University, Japan

Kenji Iwatao
AIST, Japan

Xinyue Zhao
Zhejiang University, China

Yutaka Satoh
AIST, Japan

Abstract

An illumination-invariant background model for detecting objects in dynamic scenes is proposed. It is robust in the cases of sudden illumination fluctuation as well as burst moving background. Unlike previous works, it distinguishes objects from a dynamic background using co-occurrence character between a target pixel and its supporting pixels in the form of multiple pixel pairs. Experiments used several challenging datasets that proved the robust performance of object detection in various environments.

1. Introduction

Object detection suffers from dynamic scenes, especially two types of potentially serious cases: (1) sudden illumination variation, such as outdoor sunlight changes and indoor lights on/off; (2) burst physical motion, such as indoor artificial objects motion including fans, escalators and autodoors. State-of-the-art algorithms [8, 5, 1, 2, 4] can handle gradual illumination changes by updating the statistical background models progressively as time goes by. Generally, this kind of model update is usually relatively slow to avoid mistakenly integrating foreground elements into the background, making it difficult to adapt to sudden illumination change and burst moving background.

In this study, we propose a novel framework to build a background model for object detection, which is brightness-invariant and tolerate burst moving background, named as co-occurrence probability-based pixel pairs (CP3). For modeling background, pixel pairs with high co-occurrence probability in time domain are represented by each other, although the intensity of a single pixel may change dramatically over time. This kind of pixel pairs are selected by using spatio-temporal statistical analyses. Extending our earlier work [9, 3], this paper employs co-occurrence histogram to describe the relationship between pixel pairs and calculates correlation coefficient for measuring the degree of co-occurrence which can deal well with a dynamic background; and introduces a spatial clustering operation to select optimal supporting pixels; then provides a more accurate parameterized detection criterion instead of a fixed double-sided threshold. In remainder, Section 2 details the proposed background model; Section 3 details object detection; Section 4 presents the experiments and Section 4 concludes the main contributions.

2. Background modeling

Fundamental definitions of image data: a training image sequence with a total of T images, each image has $U \times V$ pixel positions. Define P as $target\ pixel$ at location (u, v), and its intensity is denoted as $\{p_t(u, v)\}_{t=1,2,...,T}$, and $Q(u', v')$ as $arbitrary\ pixel$ with intensity sequence $\{q_t(u', v')\}_{t=1,2,...,T}$ at location (u', v'). To analyze the bivariate statistical property of a pixel pair, the co-occurrence probability joint histogram of a pixel pair is defined. The i,jth bin of the joint histogram for an arbitrary pixel pair $(P,\ Q)$ in T training images can be expressed as

$$h_{PQ}(i, j) = \sum_{t=1}^{T} \delta(p_t, q_t, i, j), \tag{1}$$

where $\delta(p_t, q_t, i, j) = 1$ if $(p_t = i) \cap (q_t = j)$ (Kronecker delta). The bins $h_{PQ}(i, j)$ corresponding to $i, j \in [0, L-1]$ represent the co-occurrence probability of $p_t = i$ and $q_t = j$. The joint histogram $\boldsymbol{h}_{PQ}$ can be written compactly as an ordered array, $\boldsymbol{h}_{PQ} = \{h_{PQ}(i, j)\}_{i,j=0}^{L-1}$. We selected a target pixel P located on the "road", and four pixels S, W, G and R from "sky", "wall", "grass" and "road" respectively, as arbitrary pixels Q shown in Fig. 1 (a). The section $h_{PQ}(i, j) > 0$ of co-occurrence probability joint histograms are illustrated in Fig. 1 (b-e), $\boldsymbol{h}_{PS}$, $\boldsymbol{h}_{PW}$, $\boldsymbol{h}_{PG}$ and $\boldsymbol{h}_{PR}$ reveal more and more regular distribution. In Fig. 1 (e), the bins of $\boldsymbol{h}_{PR}$ are parallel to the axis diagonal line. The corresponding intensity sequences of the four histograms shown in Fig. 1 (f-i), the intensity changing in

978-1-4799-0703-8/13/$31.00 ©2013 IEEE

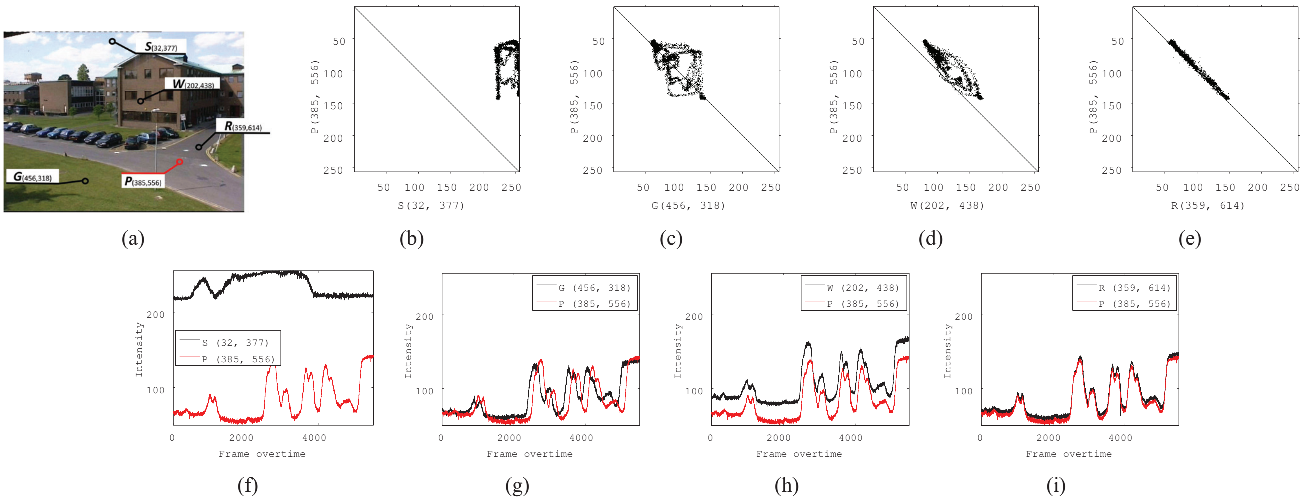


Figure 1. Co-occurrence joint histogram and intensity change

Fig. 1 (c, d) show unexpected phase difference between the two time sequences, which disturb the stable distribution of the joint histograms. In Fig. 1 (e, i), the statistical linearity of the histogram indicates stable simultaneous change of p_t and q_t. As a garyscale/single-channel image, the pixel's intensity changing follows the illumination increment, hence, the statistical linearity of a pixel pair reduces to a stable intensity difference $\mathit{\Delta}(p_t, q_t)$ just as Fig. 1 (e), in which the slope of the regression line approaches to "1". This type of Q pixels can be employed to estimate the intensity of the target pixel. For robust detection, it is necessary to maintain sufficient number of Q as $supporting\ pixels$, and denoted as $\{Q_k^P\}_{k=1,2,\ldots,K}$. $(P, \{Q_k^P\})$ maintains a background model to provide a estimation for P. Once the true intensity of P is far from the background model, P would be regarded as an abnormal-status/foreground-element.

2.1. Measurement of co-occurrence pixel pairs

For a pixel pair $(P,\ Q)$, the one dimensional histograms corresponding to their marginal distributions are,

$$h_P(i) = \sum_{j=0}^{L-1} h_{PQ}(i,j) \qquad (2)$$

The expectations is $\mathcal{E}(p_t) = \frac{1}{T}\sum_{i=0}^{L-1} i h_P(i)$; its variances is $\sigma_{p_t}^2 = \frac{1}{T}\sum_{i=0}^{L-1}[i - \mathcal{E}(p_t)]^2 h_P(i)$. The covariance of a $(P,\ Q)$ pair can be defined as follows:

$$\mathcal{C}_{P,Q} = \frac{1}{T}\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}[i - \mathcal{E}(p_t)][j - \mathcal{E}(q_t)]h_{PQ}(i,j). \qquad (3)$$

In order to measure the independent co-occurrence quantitatively, we utilize Pearson correlation coefficient:

$$\gamma_{(P,\ Q)} = \frac{\mathcal{C}_{P,Q}}{\sigma_{p_t} \cdot \sigma_{q_t}}, \qquad (4)$$

where σ_{p_t} and σ_{q_t} are the standard deviations of P and Q respectively. Fig. 2 shows examples of $\gamma_{(P,\ Q)}$, the black crosses stand for the location of P, and the red coloured area have high correlation coefficient values. In order to accelerate computing, Eq. (4)) can be calculated based on a correlation matrix instead of calculating pixel-by-pixel serial processing. The correlation matrix is the covariance matrix of the standardized random variables $\tilde{p}_t = p_t/\sigma(p_t)$. First, with a total of $M = U \times V$ pixel positions, the image sequence can be arranged progressively as a column vector set $\chi^M = \{\tilde{p}_t(m)\}_{m=1,2,\ldots,M}$. The correlation matrix in the size of $M \times M$ is

$$\Upsilon(\chi^M) = \mathcal{C}(\chi^M, (\chi^M)^T) \qquad (5)$$

where $\mathcal{C}(\cdot)$ is the covariance operation. The correlation matrix is symmetric so that each row and column of the $\Upsilon(\chi^M)$ is an array of $\gamma_{(P,\ Q)}$ for each $P(u,v)$. For speed-up, we modified Eq. (5) using a hierarchical structure of a covariance-matrix: χ^M can be sampled uniformly using a integral sample interval $\mathit{\Lambda}$, the sub-set $\chi^{[M/\mathit{\Lambda}^2]} \subset \chi^M$:

$$\Upsilon(\chi^{[M/\mathit{\Lambda}^2]}) = \mathcal{C}(\chi^{[M/\mathit{\Lambda}^2]}, (\chi^{[M/\mathit{\Lambda}^2]})^T). \qquad (6)$$

In order to cover all the target pixels, we have $\mathit{\Lambda}^2$ hierarchical correlation matrices $\Upsilon(\chi^{[M/\mathit{\Lambda}^2]})$,

$$\chi_\lambda^{[M/\mathit{\Lambda}^2]} = \{\tilde{p}_t(\omega\mathit{\Lambda}^2 + \lambda)\}_{\omega=1,2,\ldots,[M/\mathit{\Lambda}^2]}, \qquad (7)$$

where $\lambda = 1, 2, \ldots, \mathit{\Lambda}^2$.

For each target pixel $P(u, v)$, $U \times V - 1$ number of $\gamma_{(P,\ Q)}$ need to be calculated at different locations (u', v'). Then Q_n corresponding to the highest N components in the array $\gamma_{(P,\ Q(u',v'))}$ can be selected as the candidates of preferred supporting pixels, namely

$$\{Q_n\} = \{Q(u', v') | \gamma_{(P,\ Q)} > \check{\gamma}\},\ n = 1, 2, ..., N, \quad (8)$$

where $\check{\gamma}$ is the lower limit for the co-occurrence pixel pair. In practice, due to sensor noise and encoding noise, any p_t and q_t cannot maintain a full co-occurrence relation. Therefore, the lower limit $\check{\gamma}$ for choosing the high co-occurrence pixel pairs is a key parameter. Our approach to formalization is to assume that, $p_t = p_t{}' + e_1$ and $q_t = q_t{}' + e_2$, where $p_t{}'$ and $q_t{}'$ are the intensities without any noise; e_1 and e_2 are the additive noise independently with each other but with the same density function $\mathcal{N}(0, \sigma_n^2)$. Then we assume $p_t{}'$ and $q_t{}'$ are perfect positive linear correlation with a constant $b = \Delta(p_t{}', q_t{}')$, namely $p_t{}' = q_t{}' + b$, and analyse $\check{\gamma}$ as a statistic for investigating how large degradation is raised by the noise. For the computation of $\gamma_{(P,\ Q)}$, discordance between p_t and q_t can degrade $\check{\gamma}$ value apart from "1". The correlation coefficient $\check{\gamma}$ can be represented by the next expression according to Eq. (4)

$$\begin{aligned} \check{\gamma} &= \frac{\mathcal{C}(p_t{}' + e_1, p_t{}' + e_1 - e_2 - b)}{\sigma_{p_t{}'+e_1} \cdot \sigma_{p_t{}'+e_1-e_2-b}} \\ &= \frac{\sigma^2_{p'_t} + \sigma^2_n}{\sigma_{p'_t+e_1} \cdot \sigma_{p'_t+e_1-e_2-b}} \end{aligned} \quad (9)$$

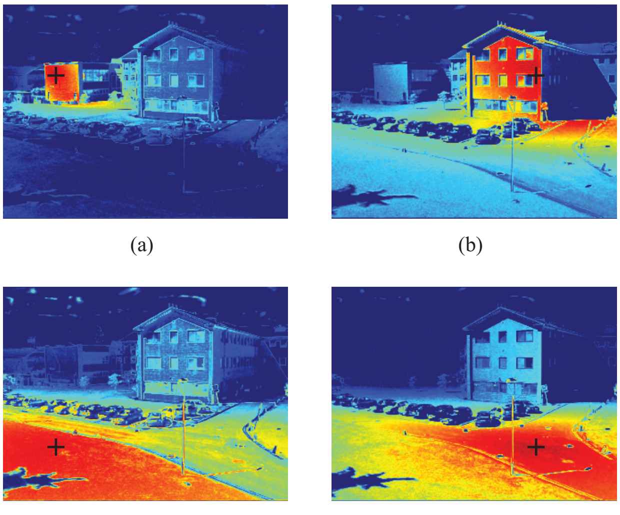

(a)
(b)
(c)
(d)

Figure 2. Diagram of $\gamma_{(P,\ Q)}$ using PETS2001-dataset3.

When $p_t{}'$ is independent with e, Eq. (9) is rewritten as

$$\begin{aligned} \check{\gamma} &= \frac{\sigma^2_{p'_t} + \sigma^2_n}{[(\sigma^2_{p'_t} + \sigma^2_n)(\sigma^2_{p'_t} + 2\sigma^2_n)]^{\frac{1}{2}}} \\ &= (\frac{\sigma^2_{p'_t} + \sigma^2_n}{\sigma^2_{p'_t} + 2\sigma^2_n})^{\frac{1}{2}} = (1 + \frac{\sigma^2_n}{\sigma^2_{p_t}})^{-\frac{1}{2}}, \end{aligned} \quad (10)$$

where σ_n^2 can be determined by the noise level of the image sequence. When the noise level is significantly smaller than the dynamic range of p_t, namely $\sigma^2_{p_t} \gg \sigma^2_n$, Eq. (10) approximate to "1", which reveals that with large-scale intensity variation in training dataset, the noise effect for correlation measurement can be reduced. On the other hand, if the intensity of P keep steady which means $\sigma^2_{p'_t} \to 0$, Eq. (10) will level off to $1/\sqrt{2}$. From the theoretical analysis, the lower limit is determined according to the comprehensive conditions combining with σ_n^2 which can be easily provided by users and a computable $\sigma^2_{p_t}$. Fig. 3 demonstrates that the rules to choose Q_n based on the lower limit $\check{\gamma}$ allow their spatial distributions of Q_n (coloured area) to follow irregular illumination variation patterns, resulting in different numbers of Q_n.

2.2. Background model of pixel pairs

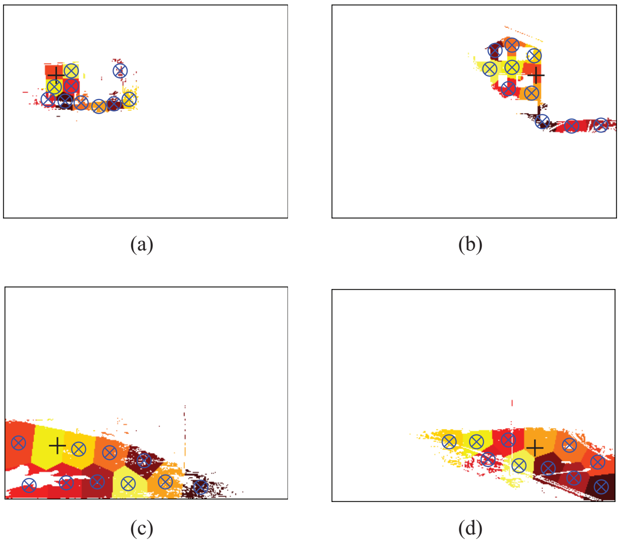

(a)
(b)
(c)
(d)

Figure 3. Selecting Q_k^P using K-means in spatial domain. The black markers is the P pixels in Fig. 2 (a-d). The coloured regions stand for the clusterings. The blue circles are the centres of each clustering, which are selected as Q_k^P. Here, $K = 10$.

As the spatial distribution of Q_n follows irregular patterns, we cannot implement any ordinary spatial interpolation approach for selecting high representative Q_k^P from Q_n. To solve this issue, K-means clustering is employed to partition N number of Q_n into K clusters, depending on the nearest clustering centres. With clustering convergence, the pixel that is closest to the k-th cluster centre is selected as a unique Q_k^P. Six demonstrations of the Q_k^P optimization are shown in Fig. 3. Note that, the computational complexity of K-means does not grow linearly with the increasing of K, and the iteration can be convergence within several loops. It is reasonable to assume that selecting more supporting pixels will contribute to a robust result. However, without loss of generality, the number of K for a given video scene is set at 20 in the following experiment. For each Q_k^P, it keeps a bivariate difference with P,

$$p_t \sim \mathcal{N}(q_{t(k)} + b, \sigma_\varepsilon^2), \tag{11}$$

where σ_ε^2 follows a normalized distribution $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. We use this Gaussian function to model the distribution of a pixel pair instead of a mixture of Gaussian [5] because we found that a single Gaussian worked better since the selected pixel pairs keeps steady difference except for noise, the noise standard deviation is estimated as follows,

$$\hat{\sigma}_\varepsilon = \sigma_{p_t - q_{t(k)}}, \tag{12}$$

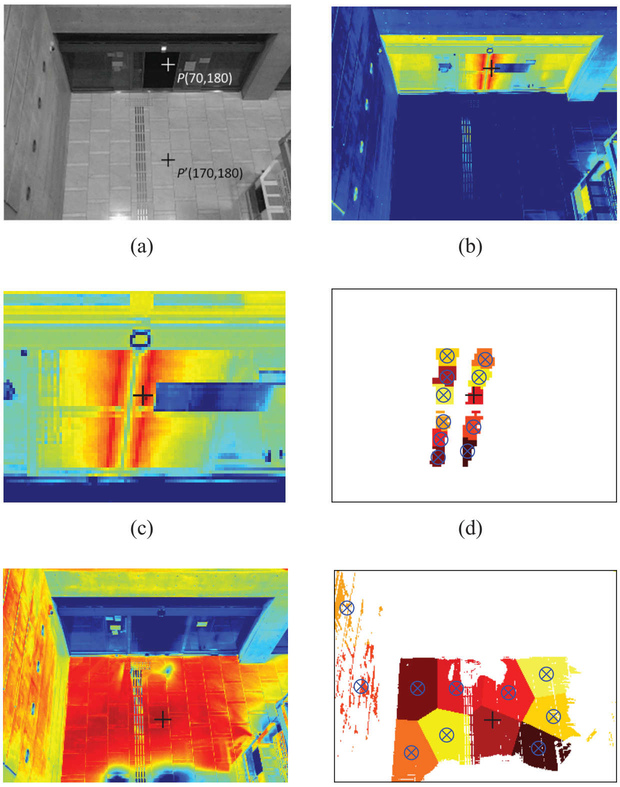
and the estimation of difference b is,

$$\hat{b} = \mathcal{E}[p_t - q_{t(k)}]. \tag{13}$$

The above two parameters $\hat{\sigma}_\varepsilon$, $\hat{b}$ are recorded for the following detection procedure. The background model is a look-up table consisting of $\{Q_k^P\} \sim [\, u', v', \hat{\sigma}_\varepsilon, \hat{b}\,]$.

2.3. Moving background case

A typical motion pattern in backgrounds is *burst motion*. This motion pattern can be described as a moving part of the background following regular directions but with an irregularly scheduled occurrence; hence, the speed and frequency can not be directly predicted. In the case of moving background, applying independent pixel-wise methods (such as GMM [5] or Codebook [2]) only employ pixel's history to continuously update background, and the fixed learning rate make these background models sensitive to burst motion. However, the moving regions covering several pixels in the same frames also present co-occurrence. Our proposed method employ the spatial-dependence of pixel pairs to keep stable differences regardless of the intensity of a single pixel under any frequency and speed of burst motion. Therefore, a target pixel P can search for the supporting pixels Q_k^P if the intensity changes of the pixel pairs are simultaneous. Fig. 4 shows a typical example of a target pixel repeatedly passed by a moving automated door. The supporting pixels with high co-occurrence are located along its vertical direction meeting the simultaneity of burst motion shown in Fig. 4 (b-d), but not around a uniform neighborhood. In contrast, Fig. 4 (e, f) show the case of a target pixel in a static area of the same background.



(a) (b)

(c) (d)

(e) (f)

Figure 4. Examples using AIST-INDOOR dataset. (a) Location of $P(70, 180)$ and $p'(170, 180)$. (b) $\gamma_{(P,\ Q)}$ of $P(70, 180)$. (c) Partial enlarged drawing of (b). (d) The supporting pixels of (c). (e) $\gamma_{(P,\ Q)}$ of $P'(170, 180)$. (f) The supporting pixels of (e).

3. Object detection

The proposed background model converts the object detection problem into a competitive binary classification problem by comparing the pairs $(P, \{Q_k^P\}_{k=1,2,...,K})$ in turn. For each pixel pair $(P,\ Q_k^P)$, the binary function $\beta(Q_k^P)$ for discriminating the normal/abnormal state between P and Q_k^P can be estimated as the following condition according to Eq. (11):

$$\beta(Q_k^P) = \begin{cases} 1 & if\ ||(p - q_k) - \hat{b})|| < C \cdot \hat{\sigma}_\varepsilon \\ 0 & otherwise \end{cases} \tag{14}$$

where p and q_k are the intensity value of P and Q_k^P in the current frame respectively, and C is a constant. Note that Eq. (14) use a bivariate normal distribution of the pixel pair is differ from traditional single Gaussian *pdf*-based identification function; In a single Gaussian *pdf*-based method, an ideal threshold should be changed following the latest intensity variation. For example, the standard deviation should be larger when the illumination fluctuate become more intense. In our proposed version, the stable difference of a pixel pair provides a normalized observation so that $\hat{\sigma}_\varepsilon$ is only related to the noise acting on each pixel. Therefore, we do not need an adjustable C to adapt to its changes caused by illumination changes or background motion. After identifying the normal/abnormal state of the pixel pair, K bits of $\beta(Q_k^P)$ are produced for the following decisions of each P. To classify whether P is a foreground pixel, the probability $\xi(P)$ of the background is defined as,

$$\xi(P) = \frac{1}{K}\sum_{k=1}^{K}\beta(Q_k^P). \tag{15}$$

Target pixel P in the input image is considered as a foreground pixel only if both $\xi(P) < PF$, where PF is a global threshold that can be adjusted to achieve the desired result. Otherwise, pixel P is considered as a background pixel.

4. Experimental results

To evaluate the performance of the proposed method, we tested it on video datasets including a variety of environments. The number of Q_k^P is $K = 20$ and $\sigma_n^2 = 100$ in the training stage; two thresholds were set as $C = 2.5$ and PF=0.5 respectively in the detection stage. We compared our algorithm with three methods: (1) GMM [6], a standardized method among independent pixel-wise models; (2) Sheikh's KDE [4], a representative method among spatially dependent models, which is different from the original KDE that it employs KDE over the joint domain (location) and range (intensity) representation of image pixels; (3) GAP [9]. The parameters for GMM were set as defaults in OpenCV tool; for Sheikh's KDE were set according to the author's recommendations in Sheikh's KDE with the size of model [26, 26, 26, 21, 31]; and in GAP $W_G = 20, W_P = 0.9, W_H = 0.3$. First, we use PETS2001-dataset3-camera1 to test outdoor severe illumination fluctuation (Fig. 6). The sudden partial illumination variations in this scene can be clearly represented as average intensity change shown in Fig. 5 (d). In Fig. 5 (a-c), CP3 has an obviously higher $Recall$ and F-measure than other methods even under sudden illumination changes. During the frames from 150-200, GAP and Sheikh's KDE methods show clearly decreasing performance of $Recall$, that because the test video comes into a darker phase after the frame 150, and the dynamic range of the intensity is compressed, as shown in Fig. 5 (d). The second dataset for testing indoor environment is AIST dataset. It contains several indoor extreme conditions: low contrast illumination, lights sudden on-off and an auto-door rapid open-shut. The average $Precision$, $Recall$ and F-measure are shown in Table 1. In, Fig. 7, compared with other approach, CP3 is insensitive to varying illumination and robust to reciprocating motion of the auto door. The third dataset is "Wallflower", which is introduced in the work of Toyama et al. [7]. This dataset consists of seven video sequences, each of which addresses a special canonical background subtraction problem shown in Fig. 8. Our method masters the illumination changes and background fluctuation well. Using speed-up background model, both the memory cost and time consumption is $O(TM^2\Lambda^{-2})$ that the hierarchical covariance-matrix reduces Λ^2 computational complexity. On a computer with a Intel Xeon 3.0 GHz processor and 16 GB RAM, the optimized C++ implementation costs 15.73s for 50 training frames with a size of 1024×1024, and the detection implementation can process about 40 fps, which is sufficiently fast for real-time detection.

Table 1. Mean precision, recall, and F-measure for each method.

Methods	Precision	Recall	F-measure
GMM	0.402	0.290	0.323
Sheikh's KDE	0.374	0.517	0.306
GAP	0.912	0.575	0.703
CP3	0.922	0.780	0.845

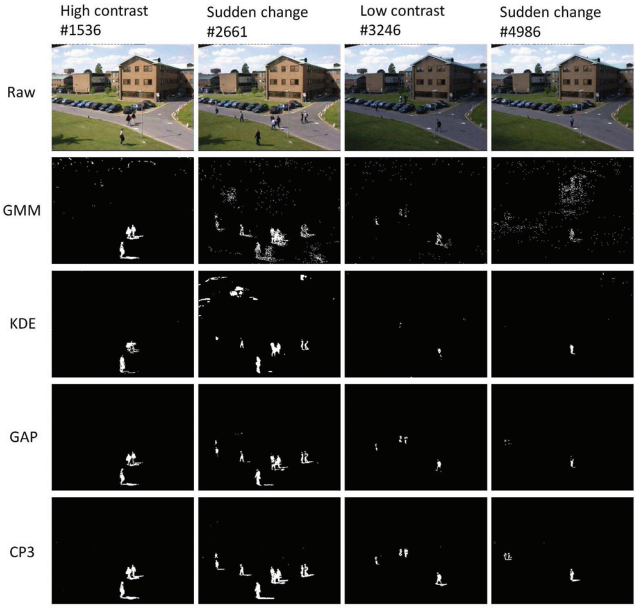
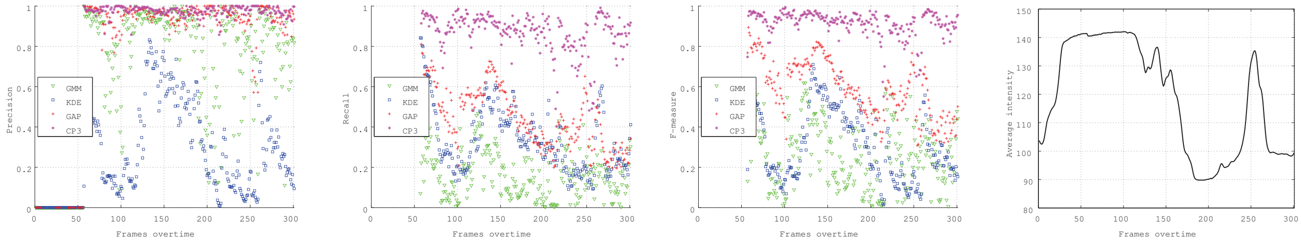


Figure 6. Qualitative analysis in PETS2001.

5. Conclusions

In conclusion, CP3 performs robust detections under extreme environments. It determines stable co-occurrence pixel pairs instead of building the parameterized/nonparametrized model for a single pixel. These pixel pairs are adaptive to capture structural background motion and cope with local and global illumination changes. As a spatial-dependence method, CP3 does not predefine any local operator, subspace or block, and it provides an accurate detection criterion even under weak illumination.



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

Figure 5. (a) $Precision$, (b) $Recall$ and (c) F-measure of CP3, GAP, Sheikh's KDE and GMM of PETS2001 dataset3 camera1.

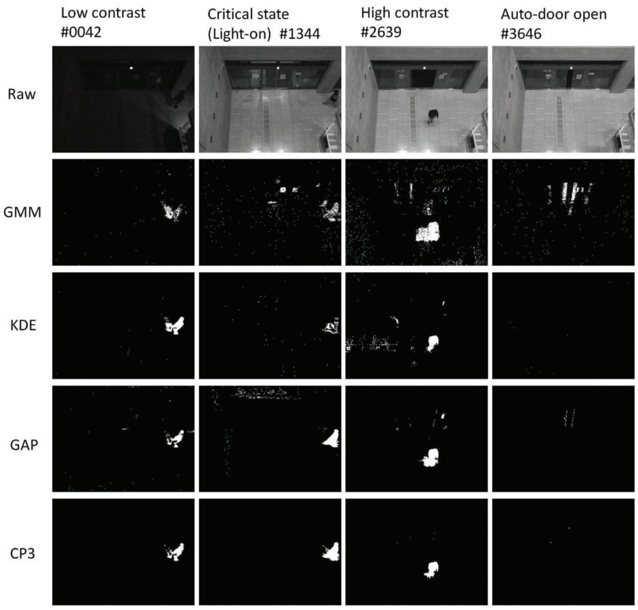


Figure 7. Qualitative analysis in AIST dataset.

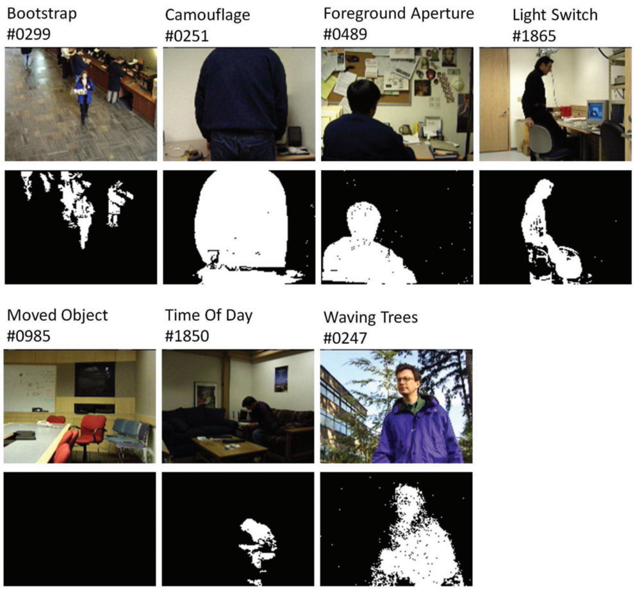


Figure 8. CP3 results of Wallflower dataset.

Acknowledgements

The authors would like to thank Professor Takayuki Tanaka for his helpful comments. This project is partly supported by Zhejiang Provincial Natural Science Foundation of China (LQ13F030003).

References

[1] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151–1163, 2002.

[2] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis. Real-time foreground–background segmentation using codebook model. *Real-time imaging*, 11(3):172–185, 2005.

[3] D. Liang, S. Kaneko, M. Hashimoto, K. Iwata, and X. Zhao. Statistical spatial multi-pixel-pair model for object detection. In *Optomechatronic Technologies (ISOT), 2012 International Symposium on*, pages 1–6, 2012.

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

[5] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Conference on Computer Vision and Pattern Recognition.*, volume 2, 1999.

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

[7] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 255–261, 1999.

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

[9] X. Zhao, Y. Satoh, H. Takauji, S. Kaneko, K. Iwata, and R. Ozaki. Object detection based on a robust and accurate statistical multi-point-pair model. *Pattern Recognition*, 44(6):1296–1311, 2011.