Robust Object Detection in Severe Imaging Conditions Using Co-occurrence Background Model

Dong Liang, Shun'ichi Kaneko, Manabu Hashimoto, Kenji Iwata, Xinyue Zhao, and Yutaka Satoh

Abstract

In this study, a spatial-dependent background model for detecting objects is used in severe imaging conditions. It is robust in the cases of sudden illumination fluctuation and burst motion background. More importantly, it is quite sensitive under the cases of underexposure, low-illumination and narrow dynamic range, all of which are very common phenomenon using a surveillance camera. The background model maintains statistical models in the form of multiple pixel pairs with few parameters. Experiments using several challenging datasets (Heavy Fog, PETS-2001, AIST-INDOOR, and a real surveillance application) confirm the robust performance in various imaging conditions.

Index Terms

object detection, narrow dynamic range, low-illumination, background modeling, underexposure

I. Introduction

On the basis of visible spectrum imaging technology, surveillance camera must be the most widely-used imaging sensing device for public security (Hu et al., 2004; Moeslund et al.,

Dong Liang and Shun'ichi Kaneko are with Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan. Email: {liang, kaneko}@ssc.ssi.ist.hokudai.ac.jp

Manabu Hashimoto is with School of Information Science and Technology, Chukyo University, Aichi, Japan.

Kenji Iwata and Yutaka Satoh are with National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan.

Xinyue Zhao is with Department of Mechanical Engineering, Zhejiang University, Zhejiang, China.

2006; Iosifidis et al., 2011). Compared with more sophisticated thermal camera and stereo imaging device (Nalpantidis et al., 2008), its flexibility and low-cost will guarantee a long-term and extensive utilization. Nevertheless, such video camera itself has several weaknesses which influence its imaging quality: first, since the frame rate is fixed (typically 30 fps for NTSC system), the shutter speed cannot delay longer than the frame interval. Thus, under a low-light condition, the sensor would not keep sufficient exposure, making the intensity of signal easily submerge into background noise. Second, surveillance task requires large depth of field (i.e. capture a globally sharp scene), the light-gathering aperture is relatively small (i.e. large F-number of aperture), which also leads to underexposure.

On the other hand, detecting object in severe imaging conditions is an essential task to perform higher-level application, such as video synopsis and indexing (Pritch et al., 2008), person re-identification (Farenzena et al., 2010; Li et al., 2013), and abnormal events detection (Datta et al., 2002). Traditional independent pixel-wise models for object detection assume every pixel as an independent statistic (Kim et al., 2005; Elgammal et al., 2002; Stauffer and Grimson, 1999; Wren et al., 1997) and then subtract the current frame from a statistic background model. This kind of method can handle gradual illumination changes by updating the statistical background models progressively as time goes by. In practice, however, this kind of model update is usually relatively slow to avoid mistakenly integrating foreground elements into the background model, making it difficult to adapt to sudden illumination changes and burst motion. To employ not only an independent statistic of a pixel but also the spatial correlation between pixels, a group of spatial-dependence methods has been proposed (Seki et al., 2003; Toyama et al., 1999; Sheikh and Shah, 2005; Zhao et al., 2011). Seki et al. (Seki et al., 2003) proposed a co-occurrence-based block correlation method, according to which the object can be detected as coarse local blocks. Toyama et al. (Toyama et al., 1999) proposed a three layers algorithm in which Weiner filters were employed, and it used region and frame-level information to verify pixel-wise background model. Nevertheless, this method needs some heuristic scheme so that such a pipeline method can result in a fragile architecture which may suffer from a domino effect, as an error can propagate to the subsequent processing stages, especially under various ill-conditions. Sheikh et al. (Sheikh and Shah, 2005) used the joint representation of image pixels in local spatial distribution (proximal pixels) and colour information to built both background and foreground's kernel density estimation (KDE) models competitively in a decision framework. This frame work is suitable for dynamic background (such as ripple), but can not deal well with illumination change. Zhao et al. (Zhao et al., 2011) proposed a learning-based approach, which aims to find high steady difference between pixels to offset the illumination change. Nevertheless, the presupposed threshold W_G of the intensity difference influences the sensitivity of detection, as the intensity difference increases, the model becomes less sensitive; without optimizing its magnitude, the detecting sensitivity will be far from an ideal level. The above methods can hardly be qualified for robust and accurate object detection under severe imaging conditions.

Recently, we have proposed a basic version of a novel spatial-dependence background model (Liang et al., 2013), called co-occurrence probability-based pixel pairs (CP3), which aims to deal with sudden illumination variation and burst motion background. Its accurate characteristics make it operable under several challenging severe imaging conditions. Compared with our earlier work GAP (Zhao et al., 2011), the proposed method has the following advantages: (1) CP3 employs a unique parametrized statistical model to describe each pixel-pair's co-occurrence rather than a fixed global double-sided threshold for all pixel-pairs in GAP; (2) CP3 derives a self-adaptive threshold for each target pixel to select better-quality supporting pixels rather than a predefined threshold in GAP. Compared with some state-of-the-art independent pixel-wise models or spatial-dependence models, such an accurate background model significantly enhance the robustness of object detection in severe imaging conditions, e.g. foggy scene, low-light and noise, sudden illumination change, and narrow dynamic range, which can be observed in experimental section.

In remainder, Section 2 details CP3 background model; Section 3 details object detection; Section 4 presents the experiments, Section 5 presents an application and some discussions, and Section 6 concludes the main contributions of this work.

II. BACKGROUND MODELING

Fundamental definitions of image data as shown in Fig. 1(a): 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 $\{p_t(u,v)\}_{t=1,2,\ldots,T}$, and $Q(u',v')$ as arbitrary pixel with intensity sequence $\{q_t(u',v')\}_{t=1,2,\ldots,T}$ at location (u',v'). Fig. 1(b) and (c) depict an image sequence and a vertical section of its intensities over time, from which it is clear that the intensity of a pixel have simultaneous variation with its neighbouring pixels as time goes by (i.e. spatio-temporal co-occurrence), especially when sudden illumination variation happens. Note that a spatio-temporal co-occurring pixel pair depends on not only the position and orientation of the camera, but also the irregular geometrical shape and inconsistent relative distance of a physical position pair in the scene.

A. Co-occurrence character of an arbitrary pixel pair

To further analyse 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, $\boldsymbol{h}_{PQ} = \{h_{PQ}(i,j)\}_{i,j=0}^{L-1}$, where L is the number of discrete intensity. We selected a target pixel P , and four pixels S, W, G and R, as arbitrary pixels shown in Fig. 2 (a). The section $h_{PQ}(i,j) > 0$ of co-occurrence probability joint histograms are illustrated in Fig. 2 (b-e). As shown in Fig. 2 (e), the bins of $\boldsymbol{h}_{PR}$ are parallel to the axis diagonal line, i.e. the slope of the regression line of $\boldsymbol{h}_{PQ}$ approaches to “1”. Then, the statistical linearity of a pixel pair can reduce the bivariate statistic to a univariate statistic of the stable intensity difference $\mathit{\Delta}(p_t, q_t)$. 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 $\{Q_k^P\}_{k=1,2,\ldots,K}$. $(P, \{Q_k^P\})$ denotes co-occurring pixel pairs, which maintain a background model to provide an 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.

B. Background model of co-occurring pixel pairs

For each Q_k^P, it keeps a bivariate difference b with P, $p_t{\sim}\mathcal{N}(q_{t(k)}+b,\sigma_\varepsilon^2)$, 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 rather than a mixture of Gaussian (Stauffer and Grimson, 1999) because we found that a single Gaussian worked better since the selected pixel pairs keep steady differences except for noise, the probability density function (PDF) is estimated as follows,

$$f(\mathit{\Delta}(p_t,q_{t(k)});\hat{b},\hat{\sigma}_\varepsilon)=\frac{1}{\hat{\sigma}_\varepsilon\sqrt{2\pi}}exp\left(-\frac{1}{2}\left(\frac{\mathit{\Delta}(p_t,q_{t(k)})-\hat{b}}{\hat{\sigma}_\varepsilon}\right)^2\right), \tag{2}$$

where the estimation of noise standard deviation $\hat{\sigma}_\varepsilon=\sigma_{p_t-q_{t(k)}}$ and the estimation of difference ε is $\hat{b}=\mathcal{E}[p_t-q_{t(k)}]$.

The above two parameters $\hat{\sigma}_\varepsilon$, $\hat{b}$ are recorded for the following detection procedure. The background model is a look-up table (LUT) consisting of $\{Q_k^P\}\sim[\,u',v',\hat{\sigma}_\varepsilon,\hat{b}\,]$. In the following, we will introduce how to select such kind of high co-occurring supporting pixels for each target pixel.

C. Measurement of co-occurring pixel pairs

For an arbitrary pixel pair $(P,\ Q)$, the one dimensional histograms corresponding to their marginal distributions is,

$$h_P(i)=\sum_{j=0}^{L-1}h_{PQ}(i,j). \tag{3}$$

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

$$\mathcal{C}_{P,Q}=\frac{1}{T}\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}[i-\mathcal{E}(p_t)][j-\mathcal{E}(q_t)]h_{PQ}(i,j). \tag{4}$$

In order to measure the independence, Pearson correlation coefficient is utilized:

$$\gamma_{(P,\ Q)} = \frac{\mathcal{C}_{P,Q}}{\sigma_{p_t} \cdot \sigma_{q_t}}, \tag{5}$$

where σ_{p_t} and σ_{q_t} are the standard deviations of P and Q, respectively. Fig. 3 shows four examples of $\gamma_{(P,\ Q)}$, the black crosses stand for the location of P, and the red coloured area have high correlation coefficient values. The examples provide some characteristics of co-occurring pixel pairs under illumination change: first, the high co-occurring distribute around a target pixel, not only follow the illumination motion, but also relate to the geometrical characteristic of the pixels (position, orientation, shape and relative distance); Second, similar intensity value is not a necessary condition for co-occurring pixel pairs, even a pixel-pair which shows a obvious mean value of intensity difference, is possible to be a qualified co-occurring pixel pair (e.g. the pixel on the road with a low intensity value and the one on the grass with a high intensity value).

Except for illumination change, another 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. Our proposed method employs the spatial-dependence of pixel pairs to keep stable differences regardless of the intensity of a single pixel under any frequency or speed of burst motion. Therefore, a target pixel P can search for the supporting pixels so long as the intensity changes of the pixel pairs are simultaneous. Fig. 4 shows some examples of various burst motion background, and $\gamma_{(P,\ Q)}$ of each target pixel.

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

$$\{Q_n\} = \{Q(u', v') | \gamma_{(P,\ Q)} > \check{\gamma}\},\ n = 1, 2, ..., N, \tag{6}$$

where $\check{\gamma}$ is the lower limit for co-occurring pixel pairs.

D. Adaptive principle for selecting a support pixel

In practice, due to the sensor noise and encoding noise of the image sequence, any p_t and q_t cannot maintain a full co-occurrence relation. Therefore, the lower limit $\check{\gamma}$ for choosing co-occurring pixel pairs is a key factor we tend to solve. 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 $b = \mathit{\Delta}({p_t}', {q_t}')$, 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 $\gamma_{(P,\ Q)}$, dis-concordance 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. (5)

$$\check{\gamma} = \frac{\mathcal{C}({p_t}' + e_1, {p_t}' + e_1 - e_2 - b)}{\sigma_{{p_t}'+e_1} \cdot \sigma_{{p_t}'+e_1-e_2-b}} = \frac{\sigma^2_{p'_t} + \sigma^2_n}{\sigma_{p'_t+e_1} \cdot \sigma_{p'_t+e_1-e_2-b}}. \tag{7}$$

When ${p_t}'$ is independent with e, Eq. (7) is rewritten as

$$\check{\gamma} = \frac{\sigma^2_{p'_t} + \sigma^2_n}{[(\sigma^2_{p'_t} + \sigma^2_n)(\sigma^2_{p'_t} + 2\sigma^2_n)]^{\frac{1}{2}}} = \left(\frac{\sigma^2_{p'_t} + \sigma^2_n}{\sigma^2_{p'_t} + 2\sigma^2_n} \right)^{\frac{1}{2}} = \left(1 + \frac{\sigma^2_n}{\sigma^2_{p_t}} \right)^{-\frac{1}{2}}, \tag{8}$$

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. (8) 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. (8) will level off to $1/\sqrt{2}$, then the candidate supporting pixels can be selected from all the stationary elements of the background. From the theoretical analysis, the lower limit is determined according to the comprehensive conditions combining with a straightforward computable $\sigma^2_{p_t}$, and σ^2_n, which can be steadily implemented by,

$$\sigma_n^2 = \frac{1}{2UV} \sum_u \sum_v [p_1 - p_2 - \frac{1}{UV} \sum_u \sum_v (p_1 - p_2)]^2, \tag{9}$$

where p_1 and p_2 are the intensity values at location (u, v) at the first frame and second frame respectively.

As the spatial distribution of Q_n follows irregular patterns, we cannot implement any ordinary spatial interpolation approach for selecting 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 closest to the k-th cluster centre is selected as a unique Q_k^P.

E. Speed-up version

For convenient computation, Eq. (5) 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 $\chi^M = \{\tilde{p}_t(m)\}_{m=1,2,...,M}$. The correlation matrix in the size of $M \times M$ is

$$\varUpsilon(\chi^M) = \mathcal{C}(\chi^M, (\chi^M)^T), \tag{10}$$

where $\mathcal{C}(\cdot)$ is the covariance operation. The correlation matrix is symmetric so that each row and column of the $\varUpsilon(\chi^M)$ is an array of $\gamma_{(P,\ Q)}$ for each $P(u, v)$. The main issue is the cost of computation of all the potentials, since they are combinatorially as many as $U \times V$. When we want a speed-up version, we typically only consider a sparse number of well-separated locations, i.e. we modified Eq. (10) using a hierarchical structure of a covariance-matrix χ^M, which can be sampled uniformly using an integral sample interval $\varLambda$, the sub-set $\chi^{[M/\varLambda^2]} \subset \chi^M$:

$$\varUpsilon(\chi^{[M/\varLambda^2]}) = \mathcal{C}(\chi^{[M/\varLambda^2]}, (\chi^{[M/\varLambda^2]})^T). \tag{11}$$

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

$$\chi_\lambda^{[M/\varLambda^2]} = \{\tilde{p}_t(\omega\varLambda^2 + \lambda)\}_{\omega=1,2,...,[M/\varLambda^2]}, \tag{12}$$

where $\lambda = 1, 2, ..., \varLambda^2$.

III. Object detection

The proposed background model converts an object detection problem into a competitive binary classification problem by comparing the pairs $(P, \{Q_k^P\}_{k=1,2,...,K})$ in turn:

$$\xi(P) = \frac{1}{K}\sum_{k=1}^{K}\beta(Q_k^P), \quad where \quad \beta(Q_k^P) = \begin{cases} 1 & if\ |(p-q_k)-\hat{b})| < C\cdot\hat{\sigma}_\varepsilon \\ 0 & otherwise \end{cases}, \qquad (13)$$

where p and q_k are the intensity values of P and Q_k^P in the current frame respectively, and C is a constant. 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 according to Eq. (13). Target pixel P in the input image is considered as a foreground pixel only if $\xi(P) < pf$, where pf is a probability threshold of foreground that can be adjusted to achieve the desired result. Otherwise, pixel P is considered as a background pixel. That is,

$$\xi(P|pf) = \begin{cases} foreground & if \quad \xi(P) < pf \\ background & otherwise \end{cases}. \qquad (14)$$

Note that Eq. (13) uses a bivariate normal distribution of a pixel pair is different 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 becomes 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. The constant C can be set from 1.0 to 3.0 in order to contain approximately an area of 68-99 % of its probability density function. In the following experiments, we set $C = 2.5$. Considering computational complexity, the procedure used to calculate $\xi(P)$ for every target pixel, is performed by a LUT for calculating $\beta(Q_k^P)$ along with bit counting operations for calculating $\xi(P)$, both of them are quite efficient to implement on any conventional hardware.

IV. Experimental results

To evaluate the performance of the proposed method, we tested it on video datasets including a variety of severe imaging conditions. We compared our algorithm with three methods: (1) GMM (Stauffer and Grimson, 2000), a standardized method among independent pixel-wise models; (2) Sheikh's KDE (Sheikh and Shah, 2005), a representative method among spatial-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 (Zhao et al., 2011), which is a predecessor and has a homologous methodology with CP3. The parameters for GMM were set as defaults in OpenCV tool; for Sheikh's KDE were set according to the author's recommendations with the size of model [26, 26, 26, 21, 31]; and in GAP $W_G = 20, W_P = 0.9, W_H = 0.3$.

For quantitative analysis, the three information retrieval measurements, $Precision$, $Recall$ and $F - measure$ were utilized,

$$Precision = \frac{TP}{TP + FP}, \tag{15}$$

and

$$Recall = \frac{TP}{TP + FN}, \tag{16}$$

where TP, FP and FN stand for the number of true positive pixels, false positive pixels and false negative pixels, represent the number of pixels which are correctly classified as foreground, the number of pixels which are incorrectly classified as foreground and the number of pixels which are incorrectly classified as background, respectively. The $precision$ ration (also called positive predictive value) is the fraction of detected pixels which belong to the foreground, which can show the detection noise level; The $recall$ ration (also known as sensitivity) is the fraction of object's completeness after detection. $F - measure$ is a weighted harmonic mean of the $Precision$ and $Recall$ to compute a score,

$$F = \frac{2Precision \cdot Recall}{Precision + Recall}. \tag{17}$$

A. Parameters discussion

In our proposed model CP3, there are two important parameters. One is the number of supporting pixels K in background modeling step, and the other is probability of foreground pf in detection step. We synthetically investigate the relationship among $F-measure$, K, and pf, using PETS2001-dataset3-camera1 (Raw data: ftp://ftp.cs.rdg.ac.uk/pub/), shown in Fig. 5. The highest $F-measure$ at around $pf = 0.4$ to $pf = 0.7$. The larger K is, the more steady $F-measure$ will be provided. On the other hand, from the results of Fig. 5, it is reasonable to assume that selecting more supporting pixels will contribute to a robust result. However, without loss of generality and saving computing time, the number of K for a given video scene is set at $K = 20$ and $pf = 0.5$ in the following experiments.

B. Experiments on datasets

First, we use a dataset of traffic sequence with heavy fog (Raw data: http://i21www.ira.uka.de/image_sequences/). In which there is only gradually varied illumination but no burst motion background. The only difficulty is that the heavy fog compresses the dynamic range of the scene. The detection results are shown in Fig. 6. In the point of view of sensitivity for detecting object, GAP method is the weakest one, because of the fixed threshold during the training and testing phase.

Second, we use PETS2001-dataset3-camera1 (Ground truth: http://limu.ait.kyushu-u.ac.jp/en/dataset/) to test outdoor severe illumination fluctuation (Fig. 7). The 300 frames ground truth data allows us to do a long-term quantitative test as shown in Fig. 8 (a-c). The sudden partial illumination variations in this scene can be clearly represented as average intensity change shown in Fig. 8 (d), after 150 frames, it became a low-light phase with a sudden illumination change. CP3 has an obviously higher $Precision$, $Recall$ and $F-measure$ than any other methods. Even under low-light and sudden illumination changes phase, the method is still relatively steady.

The third dataset for testing indoor environment is AIST-INDOOR dataset (http://ssc-lab.com/~liang/CP3_project/AIST_INDOOR_DATASET.rar). It contains several indoor extreme conditions: low contrast illumination, lights sudden on-off and an auto-door rapid open-shut. The detection results are shown in Fig. 9. Compared with other approaches, CP3 is insensitive to sudden illumination and robust to reciprocating motion of the auto door. Note that, when in the low-contrast frame $\#1129$, the object and the background also have low-contrast between each other, rather than a easier case $\#0042$.

The average $Precision$, $Recall$ and $F-measure$ of the above three experiments are shown in Table I.

V. Applications and discussions

We have already integrated CP3 method to an off-line supermarket shopper analysis system (Etchuya et al., 2013) as its first step for person detection, to make a quite time-consuming transformation of coordinates to be more practicable, i.e. only transform the region of interest (ROI), rather than the whole scene. In this application, the dataset is a high resolution (1024×1536) surveillance video, where an optimized implementation of CP3 algorithm for object detection can process about 20 fps. The runtime is measured on a computer with a Intel Xeon 3.0 GHz processor. Some detection samples are shown in Fig. 10. In this supermarket scene, there is a large-area glass window facing the camera's direction, so that the scene not only has light on/off, but also have sunlight change, under which CP3 method can work well. At present, our method needs to model the background based on an off-line framework. Therefore, two ways are available for detecting object. One is periodical off-line model a scene and then do on-line detection alternately; another way is, implementing both an off-line modeling and an off-line detecting, just as such kind of supermarket shopper analysis system. So one possible further work can be implementing CP3 to model the background on-line for person and object detection.

VI. Conclusions

In conclusion, CP3 performs robust detection under severe imaging conditions. It determines stable co-occurring pixel pairs instead of building the parametrized/non-parametrized 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, but provides an accurate detection criterion even under weak illumination.

ACKNOWLEDGEMENTS

The authors would like to thank Professor Takayuki Tanaka for his helpful comments. Dong Liang would like to thank China Scholarship Council (CSC) for its continual support. Xinyue Zhao is supported by Zhejiang Provincial Natural Science Foundation of China (LQ13F030003).

REFERENCES

Datta, A., Shah, M., da Vitoria Lobo, N., 2002. Person-on-person violence detection in video data. In: 16th International Conference on Pattern Recognition. Vol. 1. pp. 433–438.

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

Etchuya, T., Nara, H., Kaneko, S., Li, Y., Miyoshi, M., Fujiyoshi, H., Shishido, K., 2013. Integration of image and id-pos in iszot for behavior analysis of shoppers. In: 2013 International Symposium on Optomechatronic Technologies (ISOT). pp. 1–9.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M., 2010. Person re-identification by symmetry-driven accumulation of local features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 2360–2367.

Hu, W., Tan, T., Wang, L., Maybank, S., 2004. A survey on visual surveillance of object motion and behaviors. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 34 (3), 334–352.

Iosifidis, A., Mouroutsos, S. G., Gasteratos, A., 2011. A hybrid static/active video surveillance system. International Journal of Optomechatronics 5 (1), 80–95.

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

Li, W., Wu, Y., Mukunoki, M., Minoh, M., 2013. Coupled metric learning for single-shot versus single-shot person reidentification. Optical Engineering 52 (2), 027203.

Liang, D., Kaneko, S., Hashimoto, M., Iwata, K., Zhao, X., Satoh, Y., 2013. Co-occurrence-based adaptive background model for robust object detection. In: 10th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 401–406.

Moeslund, T. B., Hilton, A., Krüger, V., 2006. A survey of advances in vision-based human motion capture and analysis. Computer vision and image understanding 104 (2), 90–126.

Nalpantidis, L., Sirakoulis, G. C., Gasteratos, A., 2008. Review of stereo vision algorithms: From software to hardware. International Journal of Optomechatronics 2 (4), 435–462.

Pritch, Y., Rav-Acha, A., Peleg, S., 2008. Nonchronological video synopsis and indexing. IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (11), 1971–1984.

Seki, M., Wada, T., Fujiwara, H., Sumi, K., 2003. Background subtraction based on cooccurrence of image variations. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2. IEEE, pp. 65–72.

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

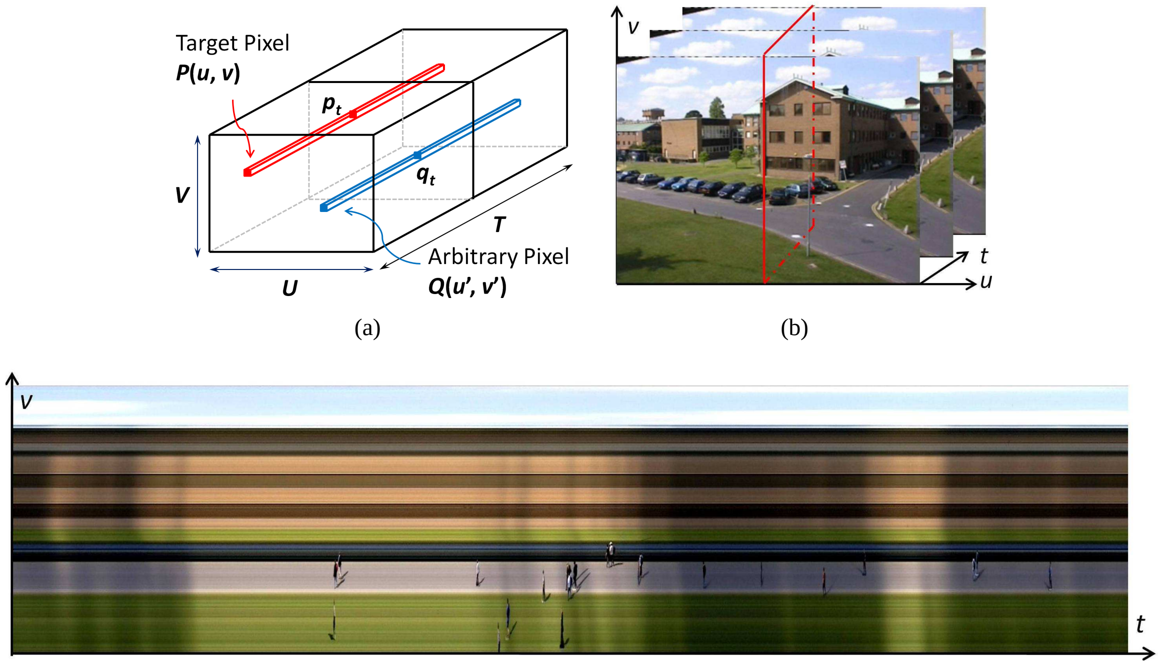
Stauffer, C., Grimson, W. E. L., 1999. Adaptive background mixture models for real-time tracking. In: IEEE Conference on Computer Vision and Pattern Recognition. Vol. 2.

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

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

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

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



(a)

(b)

(c)

Fig. 1. (a) Definition of CP3 elementary unit. (b) Using pixels along this line to create visible 2D spatial-temporal image (c). It is clear that the intensity of a pixel have simultaneous variation with its neighbouring pixels as time goes by, especially when sudden illumination variation happens. (Dataset: PET2001-dataset3-camera1)

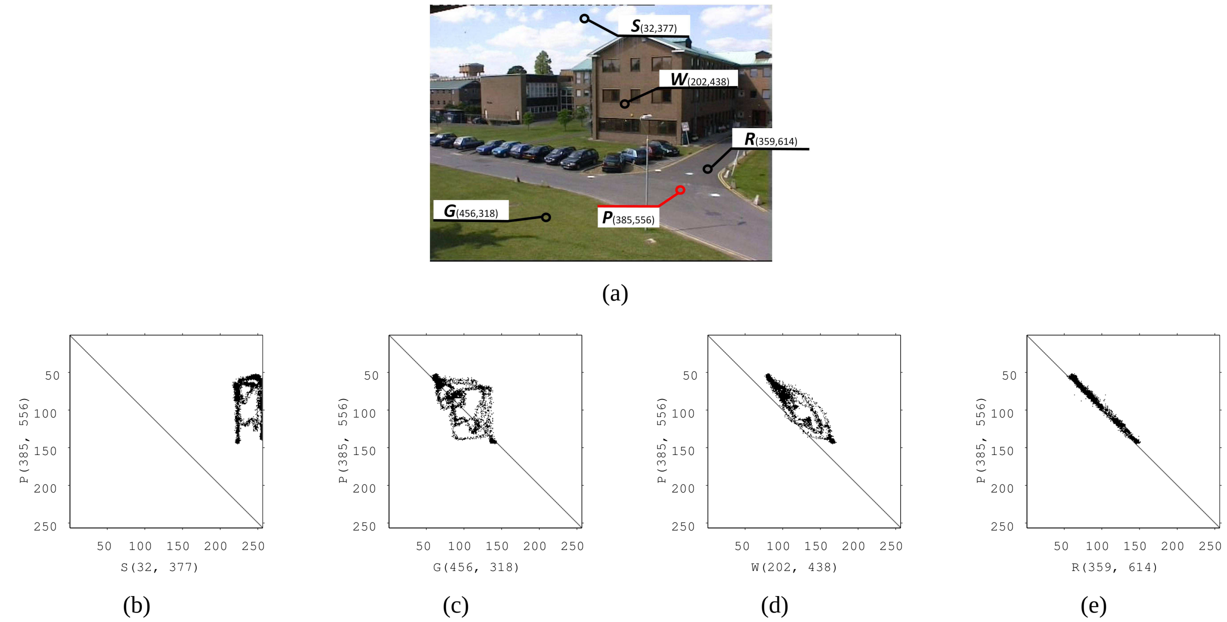


Fig. 2. (a) One target pixel P and four arbitrary pixels S, G, W, and R. (b) - (e) The sections $h_{PQ}(i,j) > 0$ of four co-occurrence probability joint histograms $\boldsymbol{h}_{PS}$, $\boldsymbol{h}_{PG}$, $\boldsymbol{h}_{PW}$, and $\boldsymbol{h}_{PR}$. In (e), the bins of $\boldsymbol{h}_{PR}$ are parallel to the axis diagonal line, implying high co-occurrence.

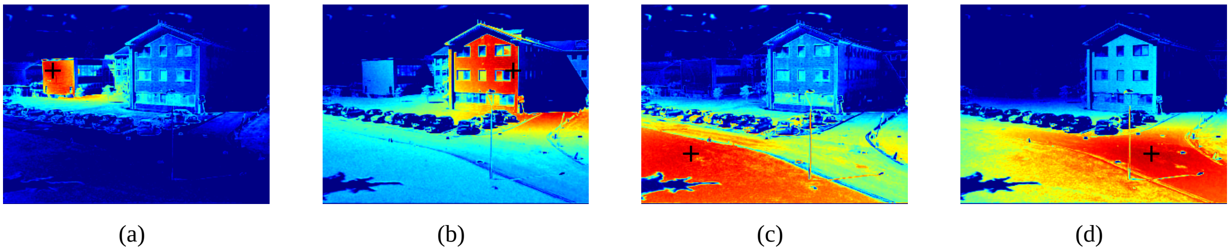


Fig. 3. Correlation coefficients $\gamma_{(P,\ Q)}$ using PETS2001-dataset3-camera1 dataset. The black crosses stand for the locations of P, and the red coloured area have high correlation coefficient values.

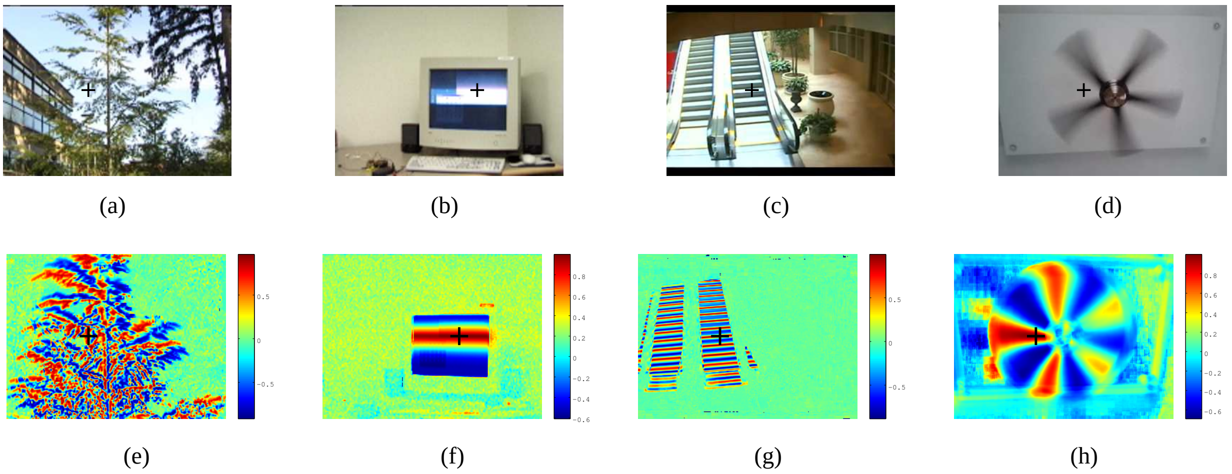


Fig. 4. Examples of various burst motion. (a) Tree swing. (b) Dynamic horizontal lines of a displayer. (c) Auto-induction escalator. (d) Speed-adjustable fan. (e) - (h) Correlation coefficients $\gamma_{(P,\ Q)}$ values of a selected pixel in (a) - (d). The black crosses stand for the locations of P, and the red coloured area have high correlation coefficient values.

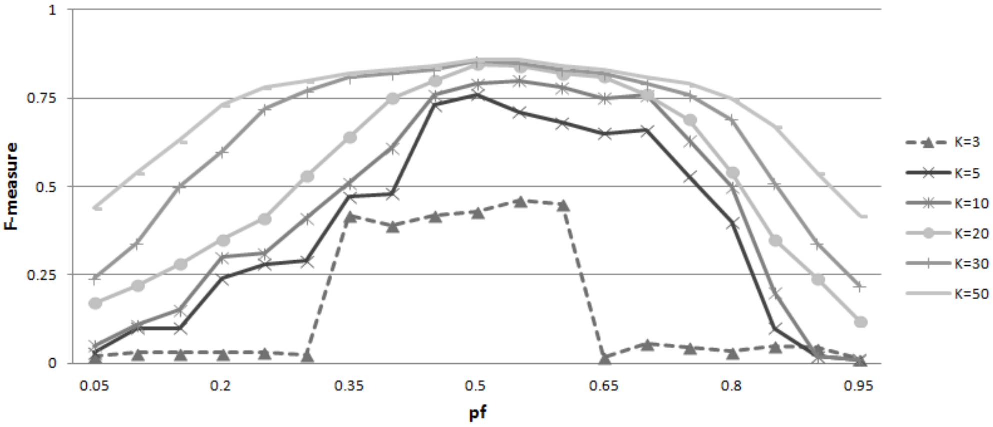


Fig. 5. The relationship among $F\!-\!measure$, K, and pf, using PETS2001-dataset3-camera1 dataset. The highest $F\!-\!measure$ at around $pf = 0.4$ to $pf = 0.7$. The larger K is, the more steady $F - measure$ will be provided.

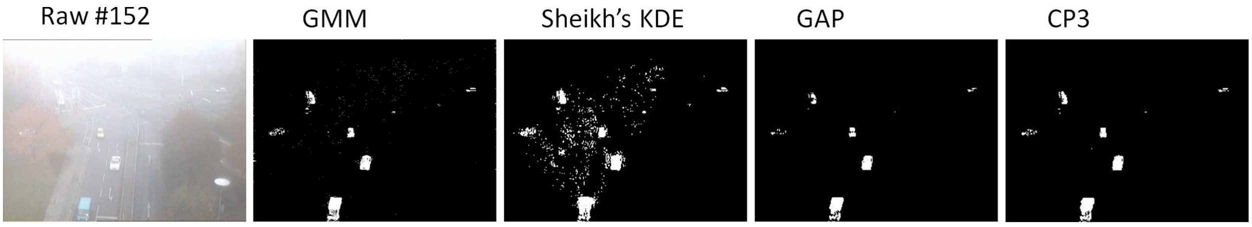


Fig. 6. Qualitative comparison of GMM, Sheikh'KDE, GAP, and proposed CP3 method using a dataset of traffic sequence with heavy fog. The difficulty is that the heavy fog compresses the dynamic range of the scene.

TABLE I

Mean $precision$, $recall$, and $F - measure$ of GMM, Sheikh's KDE, GAP, and proposed CP3 method using Heavy fog, PATS2001, and AIST-INDOOR datasets.

Methods	Quantitative evaluation	Heavy fog	PATS2001	AIST-INDOOR	Total
GMM	$Precision$	0.614	0.816	0.402	0.611
	$Recall$	0.747	0.311	0.290	0.422
	$F - measure$	0.674	0.450	0.323	0.482
Sheikh's KDE	$Precision$	0.439	0.390	0.374	0.401
	$Recall$	0.763	0.464	0.517	0.327
	$F - measure$	0.557	0.424	0.306	0.429
GAP	$Precision$	0.847	0.905	0.912	0.888
	$Recall$	0.605	0.539	0.575	0.573
	$F - measure$	0.706	0.676	0.703	0.695
Proposed CP3	$Precision$	0.862	0.918	0.922	0.901
	$Recall$	0.795	0.836	0.780	0.804
	$F - measure$	0.827	0.875	0.845	0.849

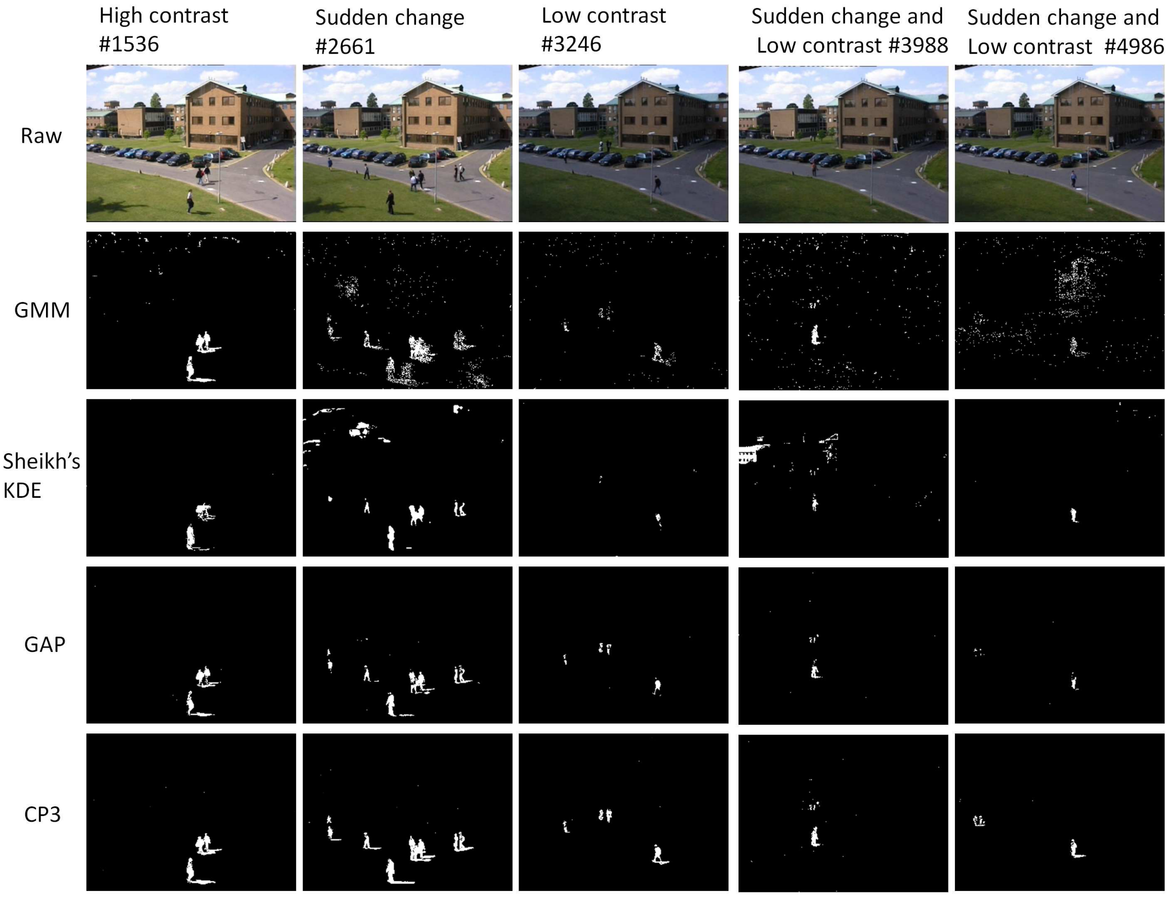
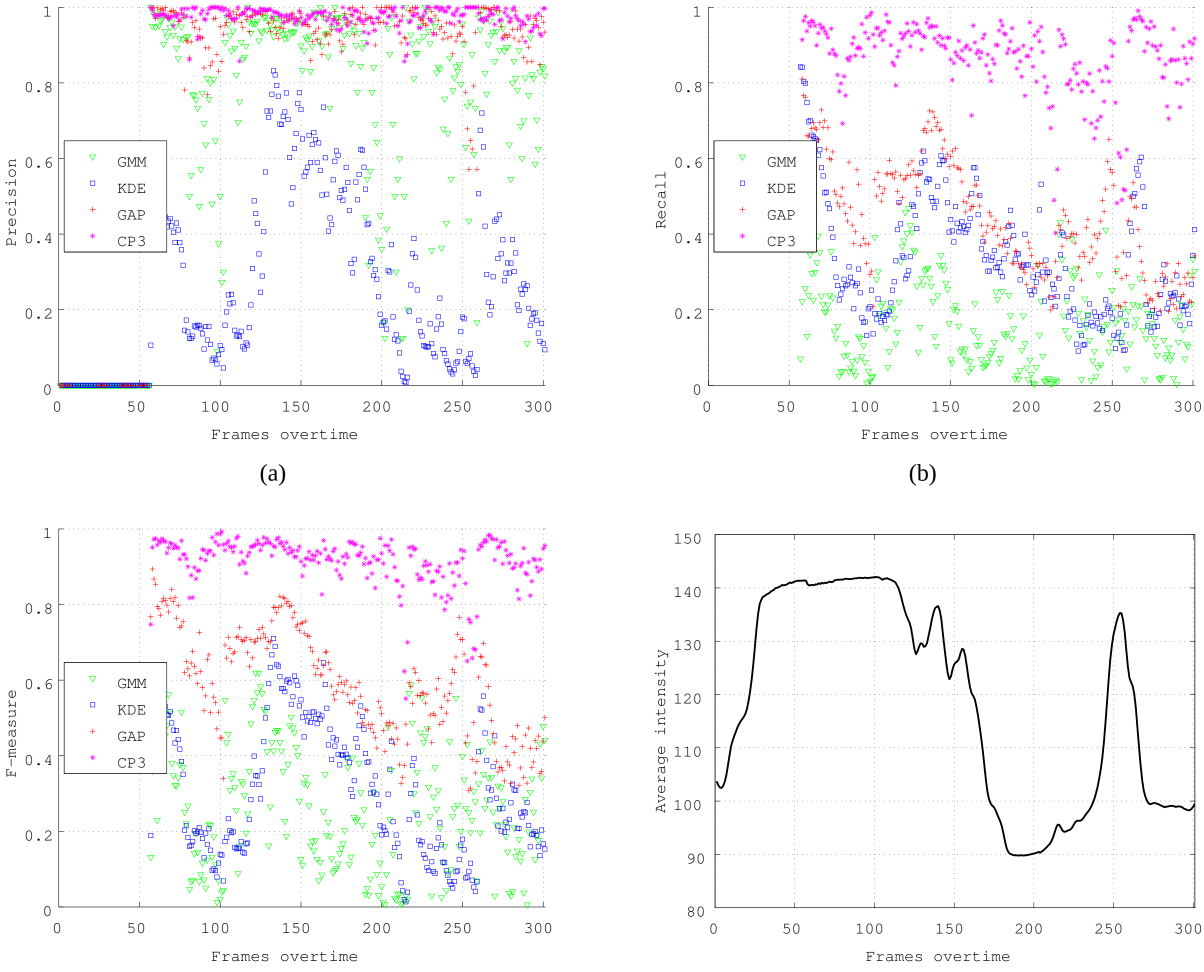


Fig. 7. Qualitative comparison of GMM, Sheikh'KDE, GAP, and proposed CP3 using PETS2001-dataset3-camera1 dataset with outdoor severe illumination fluctuation.



(a)

(b)

(c)

(d)

Fig. 8. (a) *Precision*, (b) *Recall* and (c) $F - measure$ of CP3, GAP, Sheikh's KDE and GMM using PETS2001-dataset3-camera1 dataset. (d) Average intensity over time of 300 testing frames. Even under low-light and sudden illumination changes phase, CP3 is still relatively steady.

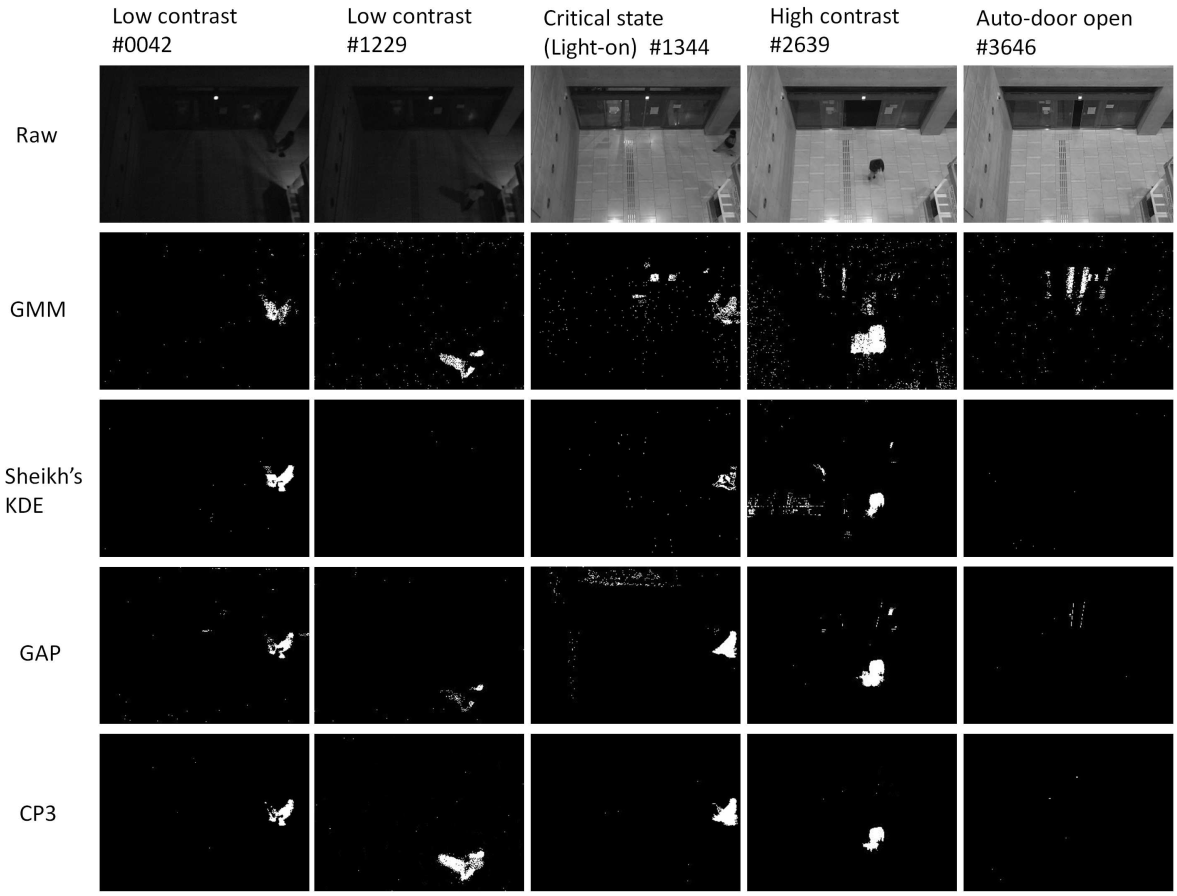


Fig. 9. Qualitative comparison of GMM, Sheikh'KDE, GAP, and proposed CP3 using AIST-INDOOR dataset. It contains several indoor extreme conditions: low contrast illumination, lights sudden on-off and an auto-door rapid open-shut.

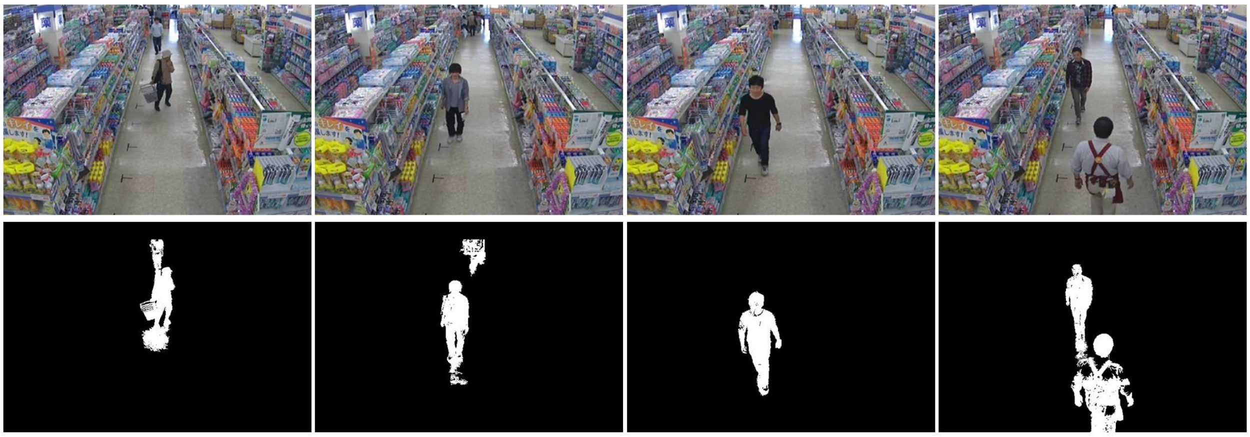


Fig. 10. Integrating CP3 method to an off-line supermarket shopper analysis system as its first step for person detection. In this supermarket scene, there is a large-area glass window facing the camera's direction, so that the scene not only has light on/off, but also have sunlight change.

NOMENCLATURE

$\beta(\cdot)$	binary function
$\hat{b}$	estimation of difference
C	constant of a Gaussian function
$\mathcal{C}$	covariance operation
$\varDelta(\cdot)$	intensity difference
e	intensity of noise
$\mathcal{E}(\cdot)$	mathematical expectation
γ	Pearson correlation coefficient
$\check{\gamma}$	lower limit of γ
$\varUpsilon$	correlation matrix
$\boldsymbol{h}(\cdot)$	joint histogram of intensity
$h(\cdot)$	a bin of the joint histogram of intensity
K	number of supporting pixels
L	number of discrete intensity level
$\varLambda$	integral sample interval
P	target pixel
p_t	intensity of a target pixel
p	current intensity of a target pixel
pf	probability threshold of foreground
Q	arbitrary pixel
q	current intensity of a supporting pixel
q_t	intensity of an arbitrary pixel
Q_k^P	a supporting pixel
$\hat{\sigma}_\varepsilon$	estimation of standard deviation of a co-occurring pixel pair
σ_n^2	variance of noise
T	total number of images
t	frame number
$\xi(\cdot)$	probability function
χ^M	column vector