Pattern Recognition 42 (2009) 2897 – 2906

Contents lists available at ScienceDirect

Pattern Recognition

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

Object motion detection using information theoretic spatio-temporal saliency

Chang Liu[a], Pong C. Yuen[a,*], Guoping Qiu[a,b]

[a]Department of Computer Science, Hong Kong Baptist University, Hong Kong
[b]School of Computer Science, University of Nottingham, UK

ARTICLE INFO

Article history:
Received 30 June 2008
Received in revised form 23 January 2009
Accepted 2 February 2009

Keywords:
Moving object detection
Foreground detection

ABSTRACT

This paper proposes to employ the visual saliency for moving object detection via direct analysis from videos. Object saliency is represented by an information saliency map (ISM), which is calculated from spatio-temporal volumes. Both spatial and temporal saliencies are calculated and a dynamic fusion method developed for combination. We use dimensionality reduction and kernel density estimation to develop an efficient information theoretic based procedure for constructing the ISM. The ISM is then used for detecting foreground objects. Three publicly available visual surveillance databases, namely CAVIAR, PETS and OTCBVS-BENCH are selected for evaluation. Experimental results show that the proposed method is robust for both fast and slow moving object detection under illumination changes. The average detection rates are 95.42% and 95.81% while the false detection rates are 2.06% and 2.40% in CAVIAR (INRIA entrance hall and shopping center) dataset and OTCBVS-BENCH database, respectively. The average processing speed is 6.6 fps with frame resolution 320×240 in a typical Pentium IV computer.

1. Introduction

Moving object detection from video is the first and important step in video analysis [1–7]. It has been widely used as low-level tasks of computer vision applications such as target tracking, visual surveillance, human behavior recognition, video retrieval and a pre-stage of MPEG4 image compression. The objective of moving object detection is to locate foreground objects in the scene for further analysis. The captured video can be either from a stationary camera or a moving camera [8]. This paper mainly focuses on the stationary camera scenario which is the case in many applications such as visual surveillance.

Generally speaking, the main challenge of object detection from video is to detect objects with different moving speeds in complex background clusters, and under different illumination changes. Many methods have been developed and reported in the last two decades for object detection from video. These methods can be categorized in three approaches, namely contour-based [9–11], orientation-based [12,13] and distribution-based [2,14–19]. The contour-based approach is able to give a good localization of object contour, but it may not be able to handle fast motion. Optical flow is a type method in orientation-based approach and can accurately detect the motion direction, but it is sensitive to illumination changes. Distribution-based methods model the background based on the intensity distribution and is a popular approach. However, the performance depends on the accuracy in estimating the distribution. A review on these works is given in Section 2.

In order to overcome the limitations on existing methods, this paper proposes a new detection method based on spatio-temporal (ST) information saliency which is calculated from density estimation of pixels both in spatial and temporal domain. Unlike existing methods [5,20–23], our proposed ST model incorporates both spatial and temporal saliencies for moving object detection via direct analysis from videos. A dynamic weighted fusion method is developed to combine the spatial and temporal saliencies. Object saliency is represented by an information saliency map (ISM) which is calculated from ST metrics based on information theory [24]. Dimensionality reduction and kernel density estimation (KDE) are employed to develop an efficient information theoretic based procedure for constructing the ISM. Preliminary versions of our work have been reported in [25].

The rest of this paper is organized as follows. Section 2 will give a brief review on existing work. Section 3 will report the details of our proposed method using ST ISM. Experimental results and the conclusion are given in Sections 4 and 5, respectively.

2. Previous works

Moving object detection from video has been studied for more than two decades and a number of motion detection methods have been proposed in the last decade. We categorize these methods into

* Corresponding author. Tel.: +852 3411 7811; fax: +852 3411 7892.
E-mail addresses: cliu@comp.hkbu.edu.hk (C. Liu), pcyuen@comp.hkbu.edu.hk (P.C. Yuen), qiu@cs.nott.ac.uk (G. Qiu).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.02.002

three approaches, namely distribution-based approach, orientation-based approach and contour-based approach.

In distribution-based approach, background subtraction is the most popular method to detect moving objects. The rationale of this approach is to estimate an appropriate representation (background image model) of the scene based on pixel distribution, so that the object(s) in the current frame can be detected by subtracting the current frame with the background image [26]. Based on this idea, several adaptive background models have been proposed. Stauffer and Grimson [14] developed a method to model pixels as a mixture of Gaussians (MOG) and constructed a model that could be updated on-line. Along this line, other similar methods have been developed [15,16]. However, there is one problem in background modelling methods that it requires long computational time for estimating the background image model. Furthermore, since MOG assumes all pixels are independent and spatial pixel correlation is not considered, the background model based on individual pixels is sensitive to illumination and noise. When the density function is complex, parametric approach may fail. Elgammal et al. [17] proposed a set of Gaussian kernels for modelling the density at pixel level. This model estimates the probability density function (PDF) directly from the data without assumptions of the underlying distributions. Mittal et al. [2] proposed an adaptive kernel density estimation (AKDE) based background subtraction method and introduced a new bandwidth function which was data-dependent for density estimation. Normalized features were used to solve illumination problems. The authors claimed that this method is able to handle mild illumination changes. More recent works on nonparametric background modelling can be found in [18,19]. Generally speaking, moving object detection methods based on pixel distribution require an accurate estimation of the background image. The performance is good if the background image does not change much in a certain period of time.

In orientation-based approach, optical flow is the most widely used method. This approach approximates the object motion by estimating vectors originating or terminating at pixels in image sequences, so it represents the velocity field which warps one image into another high dimensional feature space. Some researchers [12,13] proposed motion detection methods based on optical flow technique, these methods can accurately detect motion in the direction of intensity gradient, but the motion which is tangential to the intensity gradient cannot be well represented by the feature map. Moreover, optical flow based methods also suffer from the illumination problem.

In the contour-based approach, level sets [9], active contours [10] and geodesic active contours [11] have been proposed. These methods can effectively detect moving objects with different sizes and shapes, and claimed to be insensitive to illumination changes [27]. But contour-based methods cannot handle fast moving objects very well and are computationally expensive.

Besides these three approaches for detecting moving object directly from the scene, there is another branch namely visual attention (VA) based approach which is currently applied for image and video understanding. The VA based approach is to determine regions that involuntarily attract our VA. Many VA models [28–32] for still images have been proposed to simulate the cognitive vision mechanism of human beings. Itti et al. [28] proposed a VA model in which three spatial visual feature sets (color, intensity and orientation) are extracted and three saliency maps are built from each feature set. Then a center surround difference filter is applied to each map, and a final saliency map is obtained from combination of these individual maps. This model has then been extensively studied [20] and shown effective in still image analysis. Bruce et al. [30] proposed another VA model for image analysis using local statistics which can be obtained using independent component analysis. Since the temporal information is not considered, these methods did not perform well when applied in video analysis applications, because motion in video is more salient to our vision system than spatial contrast. To solve this problem, Cheng et al. [29] developed another VA model which not only considers intensity and color features, but also motion information. They considered a short video clip as a basic processing unit and obtained the saliency map from both spatial feature and temporal feature. The horizontal slice and the vertical slice were considered independently. However, they did not consider the fusion of spatial features and temporal features. Moreover, their model assumed that illumination could be ignored.

3. Proposed method

This paper proposes a new method to construct an ISM using both spatial and temporal information saliencies. An ST ISM is generated for moving object detection. The ISM is a two-dimensional array and each entry reflects the ST saliency of the corresponding pixel in that video frame. By analyzing the ISM, the detection of moving object(s) with different moving speeds under different illuminations is achieved.

3.1. Information theory

From modern attention theory, saliency is the impetus for selective attention. Different attention models may give different definitions of saliency [28]. In this paper, we use the *information* measure as a quantity that reflects saliency [30]. This measure is calculated as self-information in a particular context. Considering a discrete random variable $X \in \{x_1, x_2, \ldots, x_n\}$, suppose an event $X = x_i$ is observed, Shannon's self-information content of this event $I(x_i)$ is defined as follows:

$$I(x_i) = \log_2 1/p(X = x_i) = -\log_2 p(X = x_i) \quad (1)$$

It means that the information content of an event x_i is inversely proportional to the probability of the observation of event x_i. An event that rarely happened contains high information while an event which happens frequently contains low information. The property of information theory shows close relationship with saliency, and information theory can be seen as a channel between saliency and selective attention [30]. In this paper, information saliency and information content are considered as equivalent concepts.

3.2. Computing ISM

We construct the ISM based on spatial and temporal saliencies. For each frame in the video, we compute its ISM which shows the visual saliency of each pixel of the frame.

Fig. 1 shows the block diagram of the proposed method in calculating the ISM. Our method mainly consists of three steps, namely spatial saliency computing, temporal saliency computing and ST saliency computing. Suppose we want to calculate the ISM for the frame Im_0, an ST three-dimensional volume V_Ω is constructed by the current frame Im_0 and its previous $(N-1)$ frames, i.e. $\{Im_1, Im_2, \ldots, Im_{N-1}\}$. The ST volume is then divided into smaller ST sub-volumes with smaller size of $M \times M \times N$. For each sub-volume, a spatial vector set $X' = \{x'_0, x'_1, \ldots, x'_{N'-1}\}$ is constructed by the patch (x'_0) in frame Im_0 and its $N'-1$ spatial neighborhoods. In spatial saliency computing, the spatial saliency $I_S(x'_0)$ and S-weight $P(X'|V)$ (for fusion) are calculated. For the temporal saliency, a temporal vector set with N elements $X = \{x_0, x_1, \ldots, x_{N-1}\}$ is constructed from the sub-volume. The temporal saliency $I_T(x_0)$ and T-weight $P(X|V)$ (for fusion) are calculated. By combining the spatial and temporal saliencies with the S-Weight and T-Weight, the ST saliency for x_0 is then determined. It is noted that x'_0 is the same as x_0. They

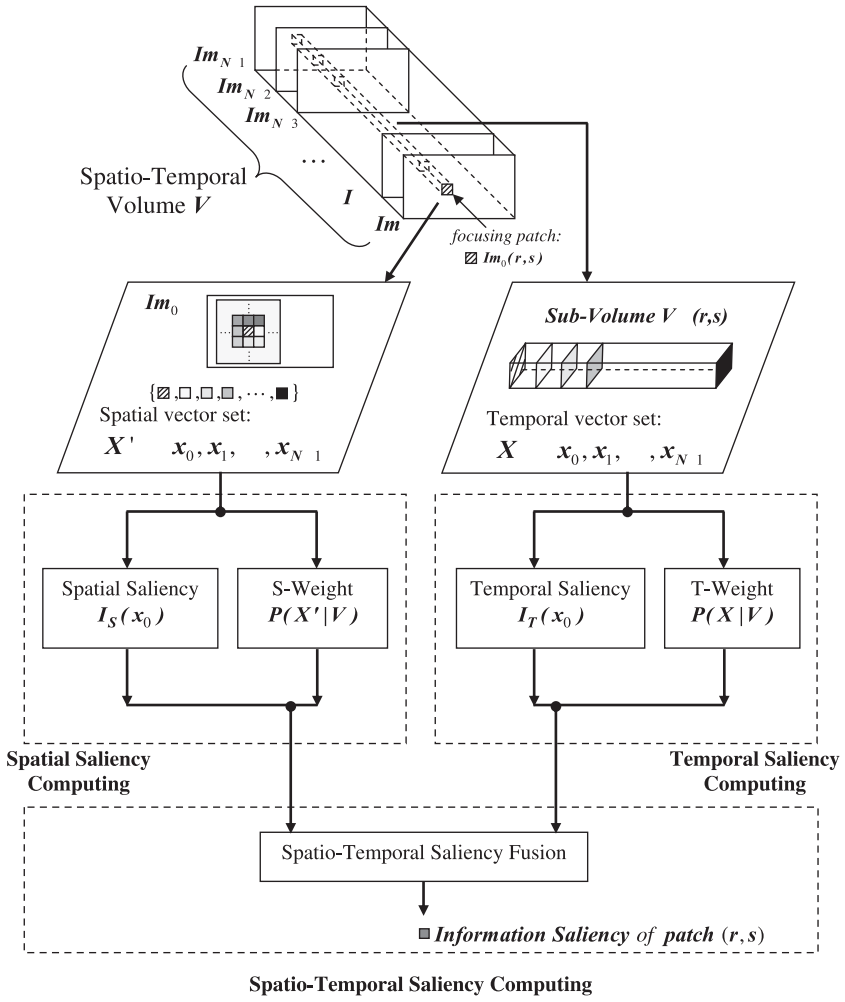

Fig. 1. Flowchart of information saliency map (ISM) computing from one single patch, noted as the *focusing patch*. It mainly contains three parts, namely spatial saliency computing, temporal saliency computing and spatio-temporal saliency computing.

represent thesame vector form of patch $Im_0(r,s)$, called a *focusing patch* in this paper. As is shown in Fig. 1, a focusing patch is temporally the first patch in the sub-volume, and spatially the central patch in a selected context.

The ISM for all the focusing patches are computed in similar methods. Finally, the ISM for the current frame Im_0 can be obtained as follows:

$$I(Im_0)=\begin{pmatrix} I(1,1) & I(1,2) & \dots & I(1,w) \\ I(2,1) & I(2,2) & \dots & I(2,w) \\ \vdots & \vdots & \ddots & \vdots \\ I(h,1) & I(h,2) & \dots & I(h,w) \end{pmatrix} \quad (2)$$

where $I(r,s)$ is the ST information saliency for focusing patch (r,s), $r=\{1,2,\dots,h\}, s=\{1,2,\dots,w\}$, h and w correspond to the number of focusing patches in the saliency map vertically and horizontally. The resolution of $I(Im_0)$ is the same as the original image Im_0.

3.2.1. Computing temporal saliency

This section describes how to compute the temporal saliency of a focusing patch in a sub-volume. Theoretically, based on the entropy equation, the temporal saliency $I_T(r,s)$ is computed by the following equation:

$$\begin{aligned} I_T(x_0) &= I_T(r,s) \\ &= -\log_2[P(Im_0(r,s)|V_\Omega(r,s))] \\ &= -\log_2(P(x_0|X)) \end{aligned} \quad (3)$$

where $V_\Omega(r,s)$ represents the sub-volume constructed at $Im_0(r,s)$, x_0 is the vector form of $Im_0(r,s)$ noted as the focusing patch, as Fig. 1 is shown, $X=\{x_0,x_1,\dots,x_{N-1}\}$. After calculating the saliency of all the focusing patches in the current frame Im_0, the temporal saliency map $I_T(Im_0)$ is then represented by $I_T(r,s)$, $r=\{1,2,\dots,h\}, s=\{1,2,\dots,w\}$, with the same size and structure as in Eq. (2).

The central part in Eq. (3) is to compute the conditional probability $P(x_0|X)$, which can be illustrated as "Given an event that a series of x_i $(i=0,1,\dots,N-1)$ happen, the probability of x_0 to be the first event." In another word, it is equivalent to the probability of chosen x_0 to be the focusing patch in the whole temporal dataset $\{x_0,x_1,\dots,x_{N-1}\}$. In this case, we need to find the density function from the data. Estimating the distribution in high dimension space $X=\{x_0,x_1,\dots,x_{N-1}\}$ is time consuming. To solve this problem, principal component analysis is employed for dimension reduction and a new vector set $Y=\{y_0,y_1,\dots,y_{N-1}\}$ is generated for estimating the probability, where y_i is a $q\times 1$ vector. From Eq. (3), we have

$$I_T(x_0)=-\log_2(P(y_0|Y)) \quad (4)$$

We adopt a non-parametric approach for computing $P(y_0|Y)$ and KDE is employed. KDE gives the exact probabilities regardless of the shape of the population distribution from which the random samples are drawn.

Considering the q-dimensional sample space $Y=\{y_0, y_1, \ldots, y_{N-1}\}$, the multivariate kernel estimator is adopted and defined as:

$$\hat{f}(y)=\frac{1}{N}\sum_{i=0}^{N-1}K_H(y-y_i) \tag{5}$$

where the kernel $K_H(y)=\|H\|^{-1/2}K(H^{-1/2}y)$, H is the bandwidth matrix which specifies the spread of the kernel around sample y_i. In this paper, we use the sample-point estimator [2]:

$$\begin{aligned}\hat{f}(y)&=\frac{1}{N}\sum_{i=0}^{N-1}K_{H(y_i)}(y-y_i)\\&=\frac{1}{N}\sum_{i=0}^{N-1}\|H(y_i)\|^{-1/2}K(H(y_i)^{-1/2}(y-y_i))\end{aligned} \tag{6}$$

This estimator considers the bandwidth matrix as a function $H(y_i)$ of the sample points y_i. So different samples should have kernels with different sizes. $H(y_i)$ is then calculated as follows:

$$H(y_i)=h(y_i)I \tag{7}$$

where $h(y_i)$ is the Euclidean distance from y_i to the k-th nearest point. In order to overcome the illumination effect, we derive a new formulation in Eq. (23) to calculate $H(y_i)$ which will be discussed in Section 3.3.

This method offers two advantages in calculating the bandwidth matrix. First, it avoids the under-smoothness and the over-smoothness problems in data distribution estimation. Second, this is an adaptive method and dependent on statistical data. When data are diverse and far apart, the kernel will be smoother. When data are tightly distributed, the kernel will be sharper. These are good properties for calculating the probabilities in Eq. (4), especially when the data size is small. Gaussian kernel is used in this paper and the density estimator in Eq. (6) becomes Eq. (8) and can be solved.

$$\hat{f}(y)=\frac{1}{(2\pi)^{q/2}N}\sum_{i=0}^{N-1}\left[(h(y_i))^{-q/2}\exp\left(-\frac{1}{2}(y-y_i)^T(h(y_i)^{-1}I)(y-y_i)\right)\right] \tag{8}$$

The temporal saliency of focusing patch x_0 is then generated in the following equation:

$$\begin{aligned}I_T(x_0)&=-\log_2(P(y_0|Y))\\&=-\log_2(\hat{f}(y_0))\end{aligned} \tag{9}$$

$\hat{f}(y_0)$ will be calculated using bandwidth matrix in Eq. (26) (details will be discussed in Section 3.3). We get

$$\begin{aligned}\hat{f}(y_0)=\frac{1}{(2\pi)^{q/2}N}\sum_{i=0}^{N-1}\Bigg[&(D_{KL}(\hat{f}_0\|\hat{f}_i))^{-q/2}\\&\times\exp\left(-\frac{1}{2}(y_0-y_i)^T(D_{KL}(\hat{f}_0\|\hat{f}_i))^{-1}I)(y_0-y_i)\right)\Bigg]\end{aligned} \tag{10}$$

3.2.2. Computing spatial saliency

Spatial saliency computing method is similar with the method in temporal saliency computing. Considering the density of focusing patch x_0 in the spatial vector set X', the spatial saliency of x_0 can be computed using the following equation:

$$\begin{aligned}I_S(x_0)&=I_S(r,s)\\&=-\log_2[P(Im_0(r,s)|B(r,s))]\\&=-\log_2(P(x_0|X'))\end{aligned} \tag{11}$$

where $B(r,s)$ represents the selected spatial context which includes a set of spatial neighboring patches centering at $Im_0(r,s)$, $X'=\{x_0, x'_1, \ldots, x'_{N'-1}\}$ is the spatial vector set with x_0 the current focusing patch. Then principal component analysis is employed for dimension reduction and a new vector set $Y'=\{y_0, y'_1, \ldots, y'_{N'-1}\}$ is generated for estimating the probability, where y'_i is a $q\times 1$ vector. Finally, the spatial saliency of focusing patch x_0 is calculated in the following equation:

$$\begin{aligned}I_S(x_0)=-\log_2\Bigg(\frac{1}{(2\pi)^{q/2}N'}\sum_{i=0}^{N'-1}\Bigg[&(h(y'_i))^{-q/2}\\&\times\exp\left(-\frac{1}{2}(y'-y'_i)^T(h(y'_i)^{-1}I)(y'-y'_i)\right)\Bigg]\Bigg)\end{aligned} \tag{12}$$

A straightforward advantage of using spatial saliency is to detect slow motion. In the case that an object slows down its speed, the object temporal saliency is decreasing to zero. To keep tracking the object saliency, the ST saliency is calculated and the slow motion detection problem can be solved.

3.2.3. ST saliency fusion

After calculating the spatial saliency and temporal saliency of a focusing patch, the next step is to fuse these two saliency maps. The fusion is based on the conditional (probability) information with spatial and temporal contexts [22]. Writing the self information of focusing patch x_0 in the spatial and temporal contexts, we have

$$I(x_0)=-\log_2(P(x_0|X\cup X')) \tag{13}$$

That is, the information saliency of patch x_0 can be obtained from the minus logarithm of probability of x_0 given the conditions of both X and X'. From the property of conditional probability, the following equation can be obtained:

$$I(x_0)=-\log_2\left(\frac{P(x_0, X\cup X')}{P(X\cup X')}\right) \tag{14}$$

where $P(x_0, X\cup X')$ is the joint probability of x_0 and $X\cup X'$. Since x_0 is the intersection of its spatial data set X' and its temporal data set X, in the case that x_0 is chosen to be the focusing patch, it means one existing patch in set X or X' is the focusing patch, so event X (a series of x_i, $i=0,1,\ldots,N-1$) and X' (a series of x'_i, $i=0,1,\ldots,N-1$) will happen. In this case, $P(X\cup X'|x_0)=1$. From the property of conditional probability, $P(x_0, X\cup X')=P(x_0)*P(X\cup X'|x_0)=P(x_0)$. Then we further consider the fact that $P(X\cup X')=P(X)+P(X')-P(X\cap X')=P(X)+P(X')-P(x_0)$, the following equation can be obtained:

$$I(x_0)=-\log_2\left(\frac{P(x_0)}{P(X)+P(X')-P(x_0)}\right) \tag{15}$$

From the total probability theorem, considering the probability space of x_0 is partitioned into its spatial data set X' and its temporal data set X, $P(x_0)=P(X)P(x_0|X)+P(X')P(x_0|X')$, the following equation can be obtained:

$$I(x_0)=-\log_2\left(\frac{P(X)P(x_0|X)+P(X')P(x_0|X')}{P(X)[1-P(x_0|X)]+P(X')[1-P(x_0|X')]}\right) \tag{16}$$

Choosing V as the minimum ST regular hexahedron that contains X and X', the marginal probability $P(X)$ is equal to the joint

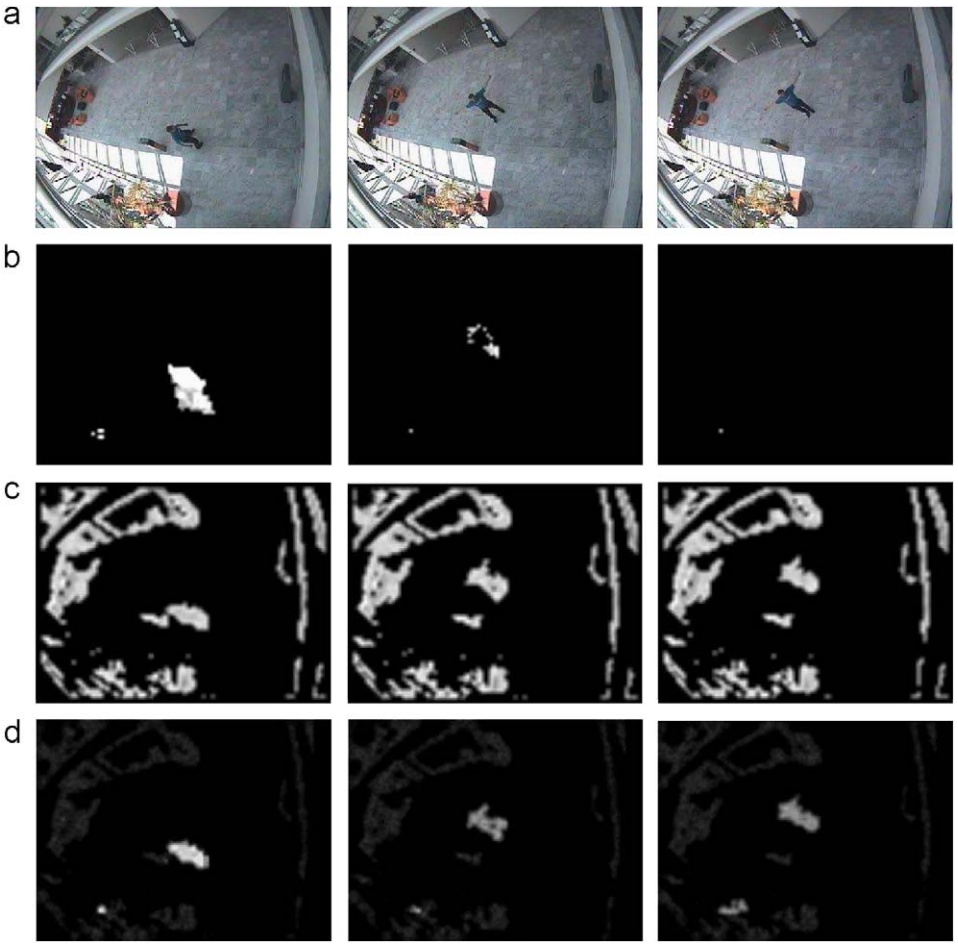

Fig. 2. (a) Original frames, (b) temporal ISM, (c) spatial ISM and (d) spatio-temporal ISM. The three columns are from frame 70, 105 and 140, respectively, in CAVIAR "*Browse* 2" database. This video clip contains two persons, one person walks to the center of the scene, keeps static for several seconds and walks away from the scene; the other person has slow motion near the reception desk.

probability $P(X,V)$ and the marginal probability $P(X')$ is equal to the joint probability $P(X',V)$. Then we get Eq. (17),

$$I(x_0) = -\log_2\left(\frac{P(X,V)P(x_0|X) + P(X',V)P(x_0|X')}{P(X,V)[1-P(x_0|X)] + P(X',V)[1-P(x_0|X')]}\right) \tag{17}$$

Finally, we get Eq. (18) by dividing the numerator and denominator in Eq. (17) by $P(V)$.

$$I(x_0) = -\log_2\left(\frac{P(X|V)P(x_0|X) + P(X'|V)P(x_0|X')}{P(X|V)[1-P(x_0|X)] + P(X'|V)[1-P(x_0|X')]}\right) \tag{18}$$

where $P(x_0|X)$ and $P(x_0|X')$ can be obtained from Eqs. (9) and (12). $P(X|V)$ and $P(X'|V)$ are denoted as T-weight and S-weight as illustrated in Fig. 1. Eq. (18) shows that the ST saliency of x_0 can be calculated from the conditional probability of $P(x_0|X)$, $P(x_0|X')$, S-weight and T-weight.

From Eq. (18), it can be seen that $I(x_0)$ becomes larger when either $P(x_0|X)$ or $P(x_0|X')$ is smaller. This indicates the ST saliency $I(x_0)$ is directly proportional to the spatial saliency $I_S(x_0)$ and the temporal saliency $I_T(x_0)$. This is true in saliency based object motion analysis approaches.

The ST ISM calculated from Eq. (18) represents the saliency of current frame based on spatial and temporal distribution estimation. Since visual saliency $I(x_0)$ is always larger than zero, Eq. (18) satisfies the following condition:

$$\frac{P(X|V)P(x_0|X) + P(X'|V)P(x_0|X')}{P(X|V)[1-P(x_0|X)] + P(X'|V)[1-P(x_0|X')]} < 1$$
$$\Rightarrow \quad P(x_0|X) < 1/2, \quad P(x_0|X') < 1/2 \tag{19}$$

Applying the above condition to Eqs. (3) and (11), it can be obtained that $I_S(x_0) > 1$ and $I_T(x_0) > 1$. This implies that the spatial saliency and temporal saliency will be considered only when their saliency value is larger than 1. Low saliency areas are considered as noisy area and their saliency will not be counted in the overall saliency.

3.3. Illumination effect on ISM

Our proposed ISM is insensitive to illumination effects during object motion detection process. Based on the Lambertian model [33], a frame can be represented by the a product of illumination function $Illu(x,y)$ and reflection function $Ref(x,y)$. Considering the temporal axis, this equation can be represented as follows:

$$f(x,y,t) = Illu(x,y,t) \cdot Ref(x,y,t) \tag{20}$$

Object motion can change the intensity property and also affect the reflection function $Ref(x,y,t)$. If there is no motion information,

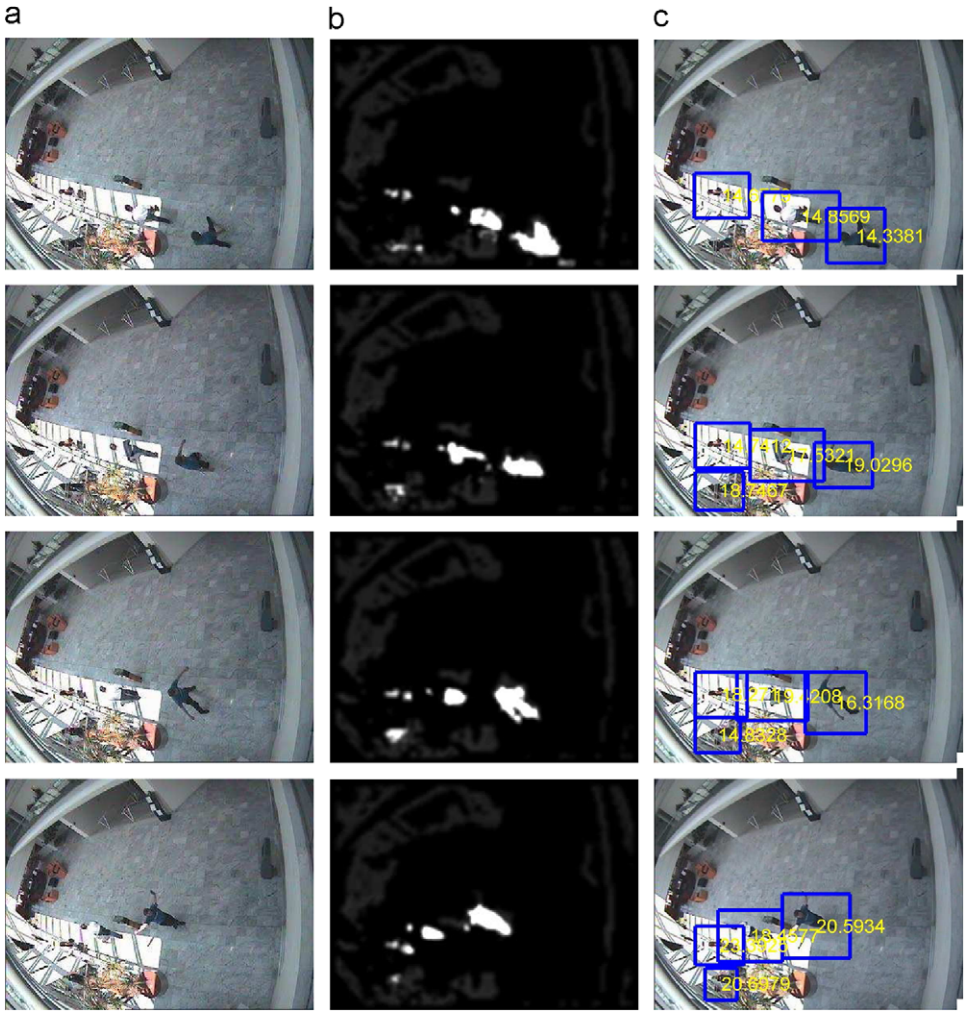

Fig. 3. (a) Original video, (b) spatio-temporal ISM and (c) foreground person detection rectangle result and object information content. From up to down rows: frame 20, 30, 40 and 55 from video "*Browse* 1" CAVIAR database. This video contains four persons moving with different speed, three of them are moving in the illuminated area. The rectangle value is obtained from averaging the spatio-temporal ISM where object motion is detected.

Table 1
Moving object detection results on CAVIAR database.

Database	TP	FP	TG	FAR (%)	DR (%)
Browsing (6 videos)	7021	182	7298	2.53	96.20
Fighting (4 videos)	5346	204	5625	3.68	95.04
Groups_meet (6 videos)	7598	153	7815	1.97	97.22
Leaving_bags (5 videos)	7756	168	8702	2.12	89.13
Rest (4 videos)	5108	151	5322	2.86	96.35
Walking (3 videos)	5380	127	5512	2.31	97.61
ShopCenter (26 videos)	46 723	799	48 753	1.68	95.84
Average	–	–	–	2.06	95.42

TP: true positive; FP: false positive; TG: total ground truth; FAR: false alarm rate, FAR = FP/(TP + FP); DR: detection rate, DR = TP/TG. The number in the bracket represents the total number of video clips in each particular scenario.

the condition of Eq. (21) should be satisfied and Eq. (22) is generated

$$Ref(x,y,t) = Ref(x,y) \tag{21}$$

$$f(x,y,t) = Illu(x,y,t) \cdot Ref(x,y) \tag{22}$$

In this case, reflection function will remain unchanged in the temporal sequence, then $f(x,y,t)$ is directly proportional to the illumination function $Illu(x,y,t)$. Since the ST volume consists of a small number of frames (20 frames in our experiments), it is reasonable to assume that within a short period of time Δt (less than one second in our experiments if we assume 30 frames per second), considering (x_0, y_0) is a sample point, the following equation is satisfied:

$$\lim_{\Delta x\to 0,\Delta y\to 0} \int_{\Delta t} |Illu(x_0+\Delta x, y_0+\Delta y, t) - Illu(x_0,y_0,t)|\, dt = 0 \tag{23}$$

Given the condition of Eq. (21), then

$$\lim_{\Delta x\to 0,\Delta y\to 0} \int_{\Delta t} |\hat{f}'(f(x_0+\Delta x, y_0+\Delta y, t)) - \hat{f}(f(x_0,y_0,t))|\, dt = 0 \tag{24}$$

where $\hat{f}$ is the pixel density function at point (x_0, y_0), $\hat{f}'$ is the pixel density function at point $(x_0+\Delta x, y_0+\Delta y)$. Let $\hat{f}$ be the candidate distribution, to check how close the probability distribution $\hat{f}'$ is to this candidate distribution, Kullback–Leibler divergence D_{KL}, which is a vital concept related to entropy in information theory, is employed, D_{KL} is represented by

$$D_{KL}(\hat{f}\|\hat{f}') = \int_{\Delta t} \hat{f}(f(x_0,y_0,t)) \log \frac{\hat{f}(f(x_0,y_0,t))}{\hat{f}'(f(x+\Delta x, y+\Delta y, t))}\, dt \tag{25}$$

From Eq. (24), $\lim_{\Delta t\to 0} D_{KL}(\hat{f}\|\hat{f}') = 0$ will be satisfied. It shows that if two pixels are close to each other, their density functions will be similar. On the other hand, if the candidate pixel do not follow

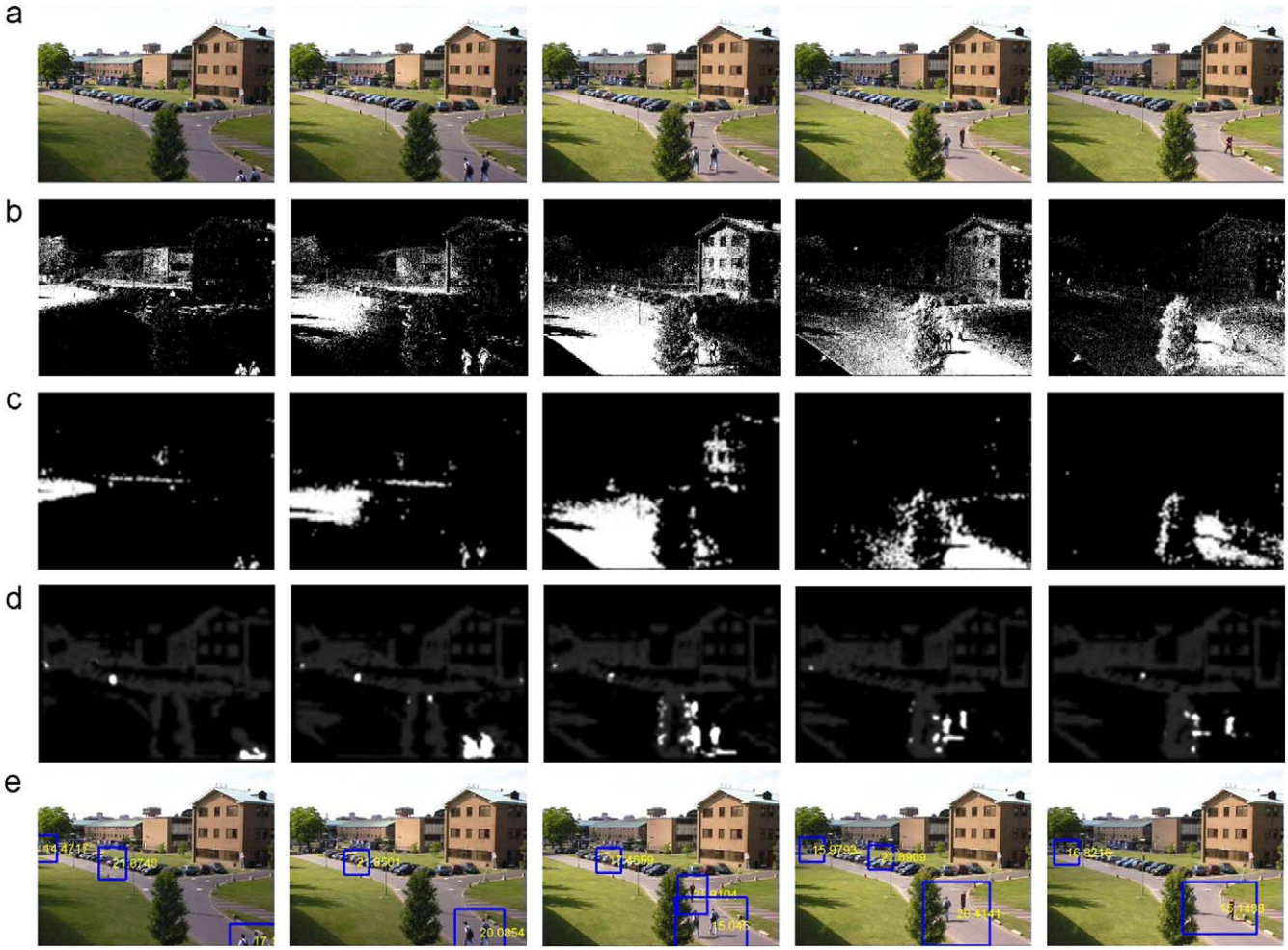

Fig. 4. (a) Original frames, (b) results on MOG [15] ($\alpha = 0.002$ and with fixed number of Gaussian components of $M = 4$), (c) results on AKDE [2] (temporal window size = 300), (d) our proposed ISM and (e) foreground object detection results using the ISM in (d). The five frames are from video "*Dataset* 3_*Testing*_2" in PETS2001 database at frame 4150, 4200, 4250, 4300, 4350. The value in the rectangle correspond to the object information content. In order to simulate fast illumination change, the video is subsumable by five before experiment.

Eq. (21), $f(x,y,t)$ will not only dependent on illumination function $Illu(x,y,t)$, but also object reflection function $Ref(x,y,t)$. Pixel density functions will be different. This is a necessary condition to identify if adjacent pixels are under illumination.

To calculate the ST ISM in model Eq. (18), the conditional probability of $P(X|V)$ and $P(X'|V)$ are calculated by KDE, using $D_{KL} \cdot I$ as the bandwidth matrix

$$H(y_i) = D_{KL}(\hat{f}\|\hat{f}_i) \cdot I \tag{26}$$

From Eq. (26), illuminated data in V will be tightly distributed because of a sharp kernel, their probabilities become larger while lowering the information saliency. From Eq. (18), the overall saliency will change a bit with the influence of illumination.

4. Experimental results

We have applied our ST ISM to the detection of moving foreground objects in real video data. The experimental results are divided into two parts. First, we evaluate the performance of our proposed method using two datasets in CAVIAR [34], namely INRIA entrance hall and shopping mall front view. Second, the proposed method is compared with existing methods using CAVIAR INRIA entrance hall [34], PETS2001 [35] and OTCBVS-BENCH [36] datasets. Two existing methods, namely MOG [15] and AKDE [2], are selected for comparison. In all experiments, we set the number of patches in each sub-volume to be $N = 20$ and the number of patches in each spatial context to be $N' = 25$. RGB images are converted into gray level image. Patch resolution is set to be 4×4, so the dimension of the original feature vector is $\langle 16 \times 1 \rangle$. q is chosen to be 4.

4.1. Evaluation of the proposed method

CAVIAR database [34] consists of three datasets, namely INRIA entrance hall, shopping mall front view and shopping mall corridor view. The INRIA entrance hall dataset has six types of events, namely "*Browsing*", "*Fighting*", "*Groups_meeting*", "*Leaving_bags*", "*Rest*" and "*Walking*", totally 28 video sequences. These video sequences are captured from inclined look-down camera with a wide angle. The bottom left region of the video is under severe illumination condition. The shopping mall front view dataset consists of 26 video clips. This database is selected to evaluate our method because of significant illumination variations. Furthermore, people outside the shops have a shadow, which is challenging for human detection.

Fig. 2 illustrates the moving foreground object detection process using the ST ISM when object's state changes from moving to stationary. Fig. 2(a) shows three frames from the video "*Browse* 2" in

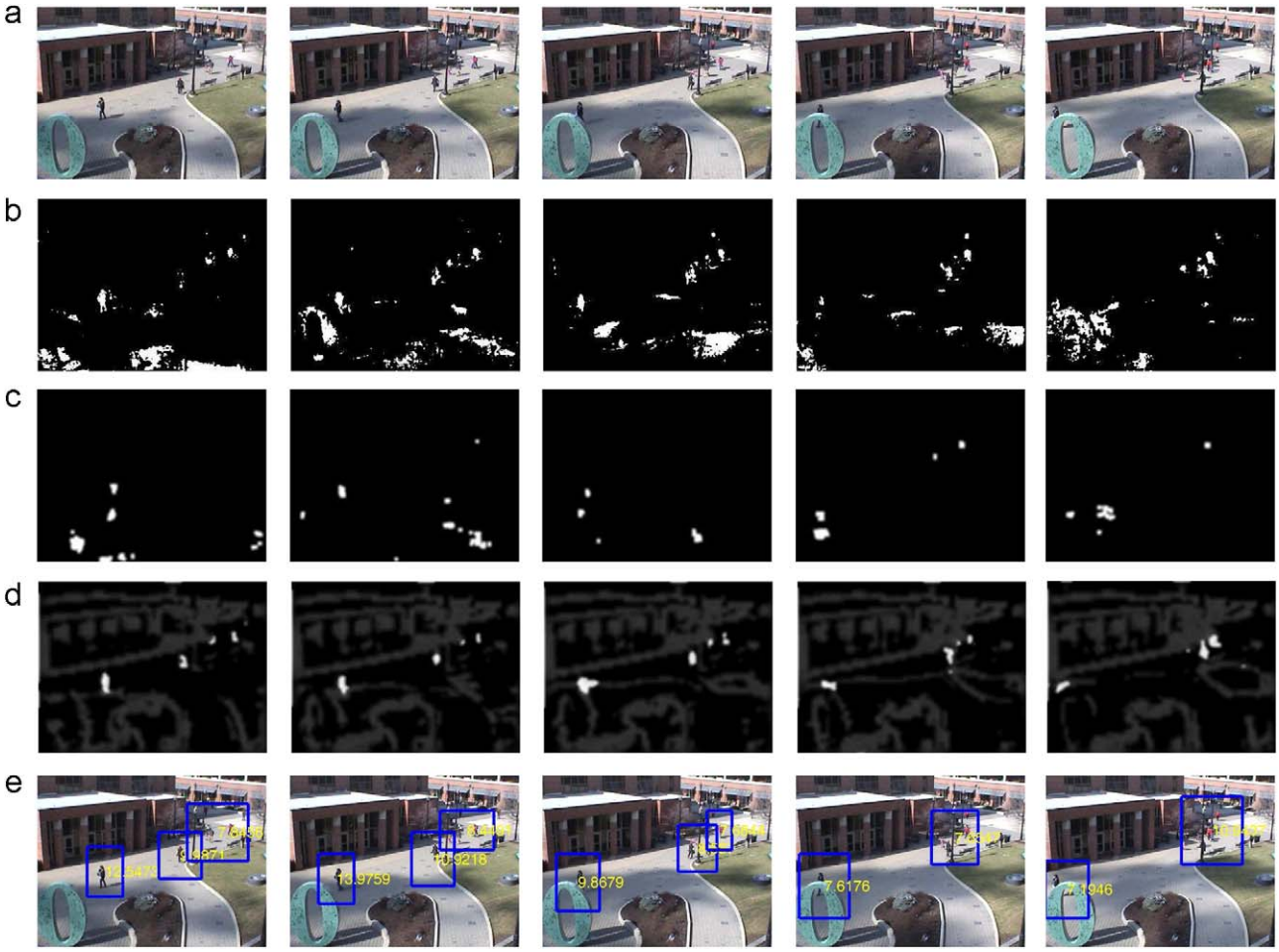

Fig. 5. (a) Original frames, (b) results on MOG [15] ($\alpha = 0.002$ and with fixed number of Gaussian components of $M = 4$), (c) results on AKDE [2] (temporal window size = 300), (d) our proposed ISM and (e) foreground object detection results using the ISM in (d). The five frames are from video "*OTCBVS-BENCH-3b*" in *OTCBVS-BENCH* database at frame 110, 160, 200, 240 and 280. There are illumination changes during these frames caused by moving clouds. The value in the rectangle correspond to the object information content.

the CAVIAR dataset. When the person walks slowly and then stops moving, the temporal saliency of the person region becomes smaller as illustrated in Fig. 2(b). When the person stops moving, the temporal ISM shows that area of the person has very low saliency. However, the spatial saliency would not be affected by the object motion speed as shown in Fig. 2(c). It can be seen that the person does not lose his spatial saliency at different speeds. Fig. 2(d) shows the ST ISM, which combines the spatial ISM and temporal ISM naturally by Eq. (18). It is shown that the areas covering the person show a high degree of saliency no matter the person is moving or not.

Another typical experimental result from CAVIAR INRIA entrance hall is shown in Fig. 3. Fig. 3(a) shows the original video "*Browse* 1" at frame 20, 30, 40 and 55, while the ISM and the detection results are shown in Fig. 3(b) and (c), respectively. The rectangles show the detected foreground objects and the values indicate the average information saliency value of each region. The difficulty is to detect the three moving persons in the illuminated area. The lighting changes from time to time in this region, which makes the corresponding background very unstable. Moreover, when persons are passing this area, their appearance will change greatly because of the strong illumination reflection. The proposed ST ISM gives good results, including the two persons with very slow motion on the left. This is because illumination effect shows much lower ST saliency than motion effect.

For the CAVIAR INRIA entrance hall and shopping mall front view datasets, ground truth data are available so that we can make quantitative analysis of our proposed method. The detection rate (DR) and the false DR of each video sequence are recorded and tabulated in Table 1. Our performance evaluation method is the same as the one in [37], where true positive (TP), false positive (FP), false negative (FN) and total ground truth (TG) are used. Let R_{GT} and R_D be the ground truth rectangular region and detected region, respectively. The detected region is considered as the TP if

$$\frac{(R_D \cap R_{GT})}{R_D} \geqslant Th \tag{27}$$

where "$\cap$" is the overlapped area between two regions and Th is a pre-defined threshold (90% in our experiments). Otherwise, the detected region is classified as FP.

As presented in Table 1, the average DR using our proposed method is 95.42% while the average false alarm rate (FAR) is 2.06%. The experiment results for the *Leaving_bags* scenario is not very good because small bags show much lower saliency, and sometimes the bags are left on the ground for a long time that they are considered to be part of the background. If only moving people are considered in this scenario, the DR is 96.92%.

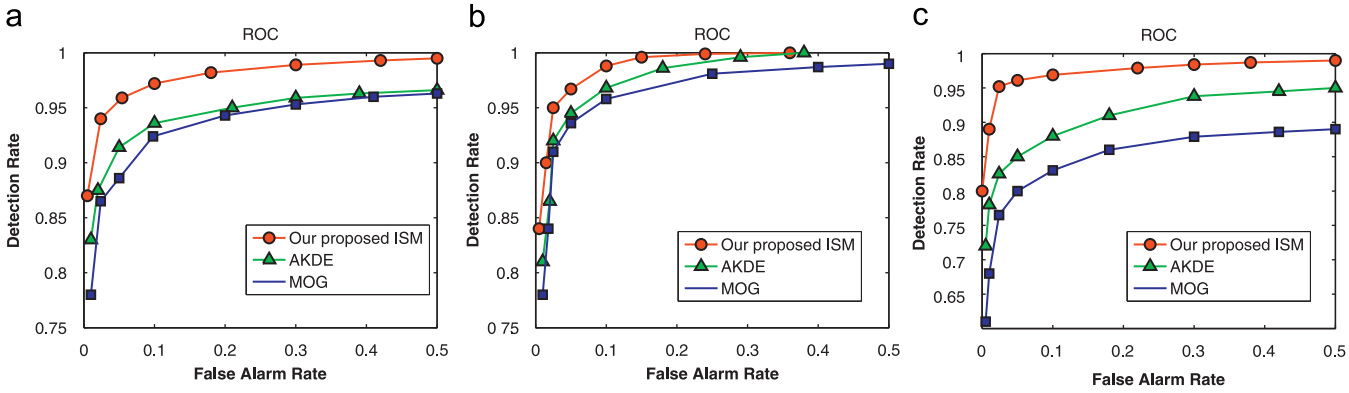

Fig. 6. Comparison between the proposed method with MOG [15] and AKDE [2]. (a) ROC curve in CAVIAR INRIA entrance hall dataset, (b) ROC curve in CAVIAR shopping center front view dataset and (c) ROC curve in OTCBVS-BENCH (1b, 2b, 3b) database.

4.2. Comparing the proposed method with existing methods

This section compares the proposed method with two existing methods, namely MOG [15] and AKDE [2] using the INRIA entrance hall dataset in CAVIAR [34], PETS2001 [35] and OTCBVS-BENCH [36].

The improved adaptive MOG model [15] is used to construct the PDF for each pixel independently and pixel-level background subtraction is performed to find the regions of interest. Experimental results show that MOG is able to successfully model background that has regular variations. However, MOG does not perform well when there are severe illumination changes in a short period of time. Furthermore, when an object is with relatively slow motion, MOG may mis-classify that region(s) as background and update the background accordingly. These situations can be illustrated using the example in Fig. 4. This video clip from PETS2001 [35] is under fast illumination change when the sunlight is blocked by a piece of cloud. Another challenge in this video is that the tree is waving in the present of wind. For implementation, we set $\alpha = 0.002$ and with fixed number of gaussian components of $M=4$ in [15]. From Fig. 4(b), it can be seen that MOG cannot model the background well and does not detect the moving persons correctly, and the person in the center of the frame who moves slowly is also missed. The AKDE [2] shows good performance in dynamic scenes modelling, including ocean waves, tree motions and mild illumination change, but this method does not perform well in severe illumination cases. This can be illustrated in Fig. 4(c), for the temporal window size is set to be 300. The results using our proposed method is shown in Fig. 4(d) in which all the moving objects with different speeds are detected.

Another comparison of our proposed method with existing methods was performed on the OTCBVS-BENCH [36] database. This database (1b, 2b, 3b video set) contains rapid changes of illumination video, caused by moving clouds under the strong sunlight. Also, the moving persons are small. Fig. 5 shows the results using MOG, AKDE and our proposed method of five representative frames from a video. It can be seen that our proposed method can correctly detect slow object motions under severe illumination variations while MOG and AKDE methods give a relatively high false detection error. The ROC curves for these methods are also recorded and plotted in Fig. 6. It can be seen that the proposed method outperforms MOG and AKDE.

5. Conclusions

A novel moving object detection method based on information theory and ST ISM have been developed and reported in this paper. An ISM is introduced to represent each frame in video and is estimated through direct analysis of video frames. It is shown that the moving object detection is feasible by using the ISM. Two publicly available databases have been selected to evaluate the proposed method. The DR and FAR on CAVIAR datasets are 95.42% and 2.06%, respectively, while the DR and FAR on OTCBVS-BENCH datasets are 95.81% and 2.40%, respectively. Comparison with two popular methods, namely MOG and AKDE, are also reported. Experimental results show that the proposed method is robust to illumination changes and no prior knowledge of the scene is required. Moreover, ISM not only provides the saliency of each pixel for object detection, but also gives additional higher level object saliency information which can be used as one of the cue for event recognition.

Our future work will be concentrated on exploring object saliency correlation between successive frames in a multi-dimensional space. We will make use of the ISM together with other cues for human activity recognition and event understanding.

Acknowledgments

This project was partially supported by the Faculty Research Grant of Hong Kong Baptist University and NSFC-GuangDong research Grant U0835005. The authors would like to thank the EC Funded CAVIAR project/IST 2001 37540 for the contribution of the CAVIAR dataset, IEEE International Workshops on Performance Evaluation of Tracking and Surveillance for the contribution of the PETS2001 dataset and IEEE OTCBVS WS Series Bench for the contribution of the OTCBVS-BENCH dataset.

References

[1] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V.K. Papastathis, M.G. Strintzis, Knowledge-assisted semantic video object detection, IEEE Transactions on Circuits and Systems for Video Technology 15 (10) (2005) 1210–1224.
[2] A. Mittal, N. Paragios, Motion-based background subtraction using adaptive kernel density estimation, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, pp. 302–309.
[3] A. Bugeau, P. Perez, Detection and segmentation of moving objects in highly dynamic scenes, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007.
[4] G. Zhang, J. Jia, W. Xiong, T.-T. Wong, P.-A. Heng, H. Bao, Moving object extraction with a hand-held camera, in: IEEE International Conference on Computer Vision, 2007.
[5] Z. Yin, R. Collins, Belief propagation in a 3d spatio-temporal MRF for moving object detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007.
[6] M. Heikkila, M. Pietikainen, 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.
[7] M. Andriluka, S. Roth, B. Schiele, People-tracking-by-detection and people-detection-by-tracking, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008.

[8] K.A. Patwardhan, G. Sapiro, V. Morellas, Robust foreground detection in video using pixel layers, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (4) (2008) 746–751.
[9] T. Brox, A. Bruhn, J. Weickert, Variational motion segmentation with level sets, in: European Conference on Computer Vision, 2006, pp. 471–483.
[10] M. Yokoyama, T. Poggio, A contour-based moving object detection and tracking, in: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 271–276.
[11] W. Fang, K.L. Chan, Using statistical shape priors in geodesic active contours for robust object detection, in: International Conference on Pattern Recognition, 2006, pp. 304–307.
[12] A.A. Stocker, An improved 2d optical flow sensor for motion segmentation, Proceedings of IEEE International Symposium on Circuits and Systems 2 (2002) 332–335.
[13] S.P.N. Singh, P.J. Csonka, K.J. Waldron, Optical flow aided motion estimation for legged locomotion, in: IEEE International Conference on Intelligent Robots and Systems, 2006, pp. 1738–1743.
[14] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real-time tracking, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1999.
[15] Z. Zivkovic, Efficient adaptive density estimation per image pixel for the task of background subtraction, Pattern Recognition Letters 27 (7) (2006) 773–780.
[16] T. Yu, C. Zhang, M. Cohen, Y. Rui, Y. Wu, Monocular video foreground/background segmentation by tracking spatial-color gaussian mixture models, in: IEEE Workshop on Motion and Video Computing, 2007.
[17] A. Elgammal, D. Harwood, L. Davis, Non-parametric model for background subtraction, in: Proceedings of the 6th European Conference on Computer Vision, 2000, pp. 751–767.
[18] Y. Liu, H. Yao, W. Gao, X. Chen, D. Zhao, Nonparametric background generation, in: International Conference on Pattern Recognition, 2006, pp. 916–919.
[19] T. Parag, A. Elgammal, A. Mittal, A framework for feature selection for background subtraction, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 1916–1923.
[20] S. Li, M.-C. Lee, An efficient spatiotemporal attention model and its application to shot matching, IEEE Transactions on Circuits and Systems for Video Technology 17 (10) (2007) 1383–1387.
[21] Y. Zhai, M. Shah, Visual attention detection in video sequences using spatiotemporal cues, in: ACM International Conference on Multimedia, October 2006.
[22] G. Qiu, X. Gu, Z. Chen, Q. Chen, C. Wang, An information theoretic model of spatiotemporal visual saliency, in: International Conference on Multimedia & Expo, 2007, pp. 1806–1809.
[23] C. Guo, Q. Ma, L. Zhang, Spatio-temporal saliency detection using the phase spectrum of quaternion fourier transform, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008.
[24] Z. Cernekova, I. Pitas, C. Nikou, Information theory-based shot cut/fade detection and video summarization, IEEE Transactions on Circuits and Systems for Video Technology 16 (1) (2006) 82–91.
[25] C. Liu, P.C. Yuen, G.P. Qiu, Generating novel information salient maps for foreground object detection in video, Proceedings of International Congress on Image and Signal Processing 4 (2008) 196–200.
[26] J. Sun, W. Zhang, X. Tang, H. Shum, Background cut, in: European Conference on Computer Vision, 2006, pp. 628–641.
[27] V. Ferrari, T. Tuytelaars, L.V. Gool, Object detection by contour segment networks, in: European Conference on Computer Vision, 2006, pp. 14–28.
[28] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11) (1998) 1254–1259.
[29] W.H. Cheng, W.T. Chu, J.L. Wu, A visual attention based region-of-interest determination framework for video sequences, IEICE Transaction on Information and Systems 88 (7) (2005) 1578–1586.
[30] N.D.B. Bruce, J.K. Tsotsos, Saliency based on information maximization, in: Advances in Neural Information Processing Systems, 2006, pp. 155–162.
[31] J. Han, K.N. Ngan, M. Li, H.-J. Zhang, Unsupervised extraction of visual attention objects in color images, IEEE Transactions on Circuits and Systems for Video Technology 16 (1) (2006) 141–145.
[32] T. Liu, J. Sun, N.-N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007.
[33] R.C. Gonzalez, R.E. Woods, Digital Image Processing, second ed., Prentice-Hall, Englewood-Cliffs, NJ, 2003.
[34] ⟨http://homepages.inf.ed.ac.uk/rbf/caviar/⟩.
[35] ⟨http://ftp.pets.rdg.ac.uk/pub/⟩.
[36] J.W. Davis, V. Sharma, Background-subtraction using contour-based fusion of thermal and visible imagery, Computer Vision and Image Understanding 106 (2–3) (2007) 162–182.
[37] F. Bashir, F. Porikli, Performance evaluation of object detection and tracking systems, in: Proceedings of 9th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, 2006, pp. 7–14.

About the Author—CHANG LIU received the B.S. degree in Information System from Nankai University, China, in 2003. He received the M.S. degree in Applied Mathematics from Sun Yat-Sen University, China, in 2006. He is now pursuing the Ph.D. degree in Computer Science in Hong Kong Baptist University. His current research interests include pattern recognition, computer vision and machine learning.

About the Author—PONG C. YUEN received his B.Sc. degree in Electronic Engineering with first class honours in 1989 from City Polytechnic of Hong Kong, and his Ph.D. degree in Electrical and Electronic Engineering in 1993 from The University of Hong Kong. He joined the Department of Computer Science, Hong Kong Baptist University in 1993 as an Assistant Professor and currently is a Professor.
Dr. Yuen was a recipient of the University Fellowship to visit The University of Sydney in 1996. He was associated with the Laboratory of Imaging Science and Engineering, Department of Electrical Engineering. In 1998, Dr. Yuen spent a 6-month sabbatical leave in The University of Maryland Institute for Advanced Computer Studies (UMIACS), University of Maryland at college park. He was associated with the Computer Vision Laboratory, CFAR. From June 2005 to January 2006, he was a visiting professor in GRAVIR laboratory (GRAphics, VIsion and Robotics) of INRIA Rhone Alpes, France. He was associated with PRIMA Group. Dr. Yuen was the director of Croucher Advanced Study Institute (ASI) on biometric authentication in 2004 and the director of Croucher ASI on Biometric Security and Privacy in 2007. Dr. Yuen has been actively involved in many international conferences as an organizing committee and/or technical program committee member such as FG and ICB. Recently, he was the track co-chair of International Conference on Pattern Recognition 2006. Currently, Dr. Yuen is an editorial board member of Pattern Recognition. Dr. Yuen's current research interests include human face processing and recognition, biometric security and privacy, context modelling and learning for human activity recognition.

About the Author—GUOPING QIU received the B.Sc. degree in Electronic Measurement and Instrumentation from the University of Electronic Science and Technology of China in 1984 and the Ph.D. degree in Electrical and Electronic Engineering from the University of Central Lancashire, Preston, UK, in 1993. He is currently a Reader in the School of Computer Science, University of Nottingham, UK. He has research interests in the broad area of computational visual information processing and has published widely in this area. More about his research can be found in: http://www.viplab.cs.nott.ac.uk/.