Neurocomputing 83 (2012) 121–135

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

An energy model approach to people counting for abnormal crowd behavior detection

Guogang Xiong [a], Jun Cheng [a,b], Xinyu Wu [a,b,*], Yen-Lun Chen [a], Yongsheng Ou [a,b], Yangsheng Xu [a,b]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[b] The Chinese University of Hong Kong, China

ARTICLE INFO

Article history:
Received 2 August 2011
Received in revised form
30 November 2011
Accepted 4 December 2011
Communicated by D. Tao
Available online 26 December 2011

Keywords:
Intelligent surveillance
Image potential energy model
Abnormal events
People counting

ABSTRACT

Abnormal crowd behavior detection plays an important role in surveillance applications. We propose a camera parameter independent and perspective distortion invariant approach to detect two types of abnormal crowd behavior. The two typical abnormal activities are people gathering and running. Since people counting is necessary for detecting the abnormal crowd behavior, we present an potential energy-based model to estimate the number of people in public scenes. Building histograms on the *X*- and *Y*-axes, respectively, we can obtain probability distribution of the foreground object and then define crowd entropy. We define the Crowd Distribution Index by combining the people counting results with crowd entropy to represent the spatial distribution of crowd. We set a threshold on Crowd Distribution Index to detect people gathering. To detect people running, the kinetic energy is determined by computation of optical flow and Crowd Distribution Index. With a threshold, kinetic energy can be used to detect people running. To test the performance of our algorithm, videos of different scenes and different crowd densities are used in the experiments. Without camera calibration and training data, our method can robustly detect abnormal behaviors with low computation load.

1. Introduction

The decreasing costs of video surveillance equipments have resulted in large volumes of video data. However, this excessive volume of information cannot be dealt with enough human operators. On the other hand, a lot of new techniques on image and video analysis have been developed rapidly, such as subspace selection by Tao et al. [1], feature selection by Dalal et al. [2]. At the same time, many useful applications have been tested in experiments or practical systems, such as human gait recognition [3], object tracking [4], scene segmentation [5]. Crowd analysis in computer vision has become a popular research topic. Models able to detect abnormal events within video streams can serve many applications, such as automatic security system and coal mine surveillance. In all cases, automatic anomaly detection can significantly improve the efficiency of video analysis by saving valuable human attention for only the most important events [6]. When an abnormal event is detected, it implies observing areas are at risk or vulnerable, where thefts, vandalism, bomb attacks, or any other dangerous events may occur and the monitoring system can alarm for immediate attention, so that prompt actions can be taken to minimize adverse impact of abnormal events [7,8].

Most traditional approaches on anomaly detection always aim at specific anomalies of a single person or a few moving objects, such as belongings dropping, loitering and crossing over the fence. As only a few people moving in the scenes, these approaches can implement detection and segmentation easily. However, when the environment becomes complicated, these methods are subjected to severe occlusions which makes the tracking, detection and segmentation difficult to be implemented.

Reliable estimation of number of people is an effective way for abnormal crowd behavior detection. An accurate and real-time estimation of pedestrians can provide valuable information to make decisions. It is helpful for the railway staff to optimize services. For another example, at traffic intersections, intelligent walk-signal systems can be designed based on the people-counting results. Obtaining the size and density of a group outside a school or public event can help authorities identify unsafe situations and regulate traffic appropriately. However, when pedestrians in the scenes move in highly irregular motion patterns, which result in severe occlusions, as shown in Fig. 1, the problem becomes challenging because of high computational complexity or non-robustness. A simple, robust and accurate model for people counting is infeasible. This paper presents an effective model to estimate the number of people in simple and

* Corresponding author at: Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China.
E-mail addresses: gg.xiong@siat.ac.cn (G. Xiong), jun.cheng@siat.ac.cn (J. Cheng), xy.wu@siat.ac.cn (X. Wu), yl.chen@siat.ac.cn (Y.-L. Chen), ys.ou@siat.ac.cn (Y. Ou), ysxu@cuhk.edu.hk (Y. Xu).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.12.007

complicated situations. Inspired by gravitational potential energy, we establish the potential energy model to estimate the people count. The proposed method has few assumptions on the scene and can work in real time. It is also found that the approach is robust and can deal with different levels of crowd density.

This paper aims to present a camera parameter independent and perspective distortion invariant model to detect two kinds of anomalies which are the mostly common in public scenes. Fig. 2 shows the typical abnormal scenes. Based on the image potential energy, we define a Crowd Distribution Index (I_{CD}) to determine whether pedestrians gather in a local region, which means events may happen, such as group fighting, tumbling or broiling. Based on the kinetic energy defined in [9], we modify it to detect objects moving with high speed. In general, pedestrian gathering and running are emergency signals indicating some abnormal events happening, and surveillance systems should detect them automatically in time.

The rest of this paper is organized as follows. A summary of the related work is given in Section 2. Our system architecture is described in Section 3. We present the image potential energy model to estimate the number of people in Section 4. In Section 5, the Crowd Distribution Index is defined and the modified definition of kinetic energy is given. In Section 6, we present the experimental results on different video clips. Finally, we summarize the approach and present some clues for future research work.

Fig. 1. Typical crowded scenario. There are always severe occlusions in the public scenes containing groups of people, which make it difficult to track and segment.

2. Related works

In this paper, we summarize related works in the areas of people counting and abnormal crowd behavior detection.

2.1. Related works on people counting

The latest research on people counting can mainly be classified into two categories: map-based methods and detection-based methods [10].

Map-based methods statistically map the number of people to foreground pixels or some other features by training [10]. Some dimension reduction algorithms are used in these approaches, such as [11–13]. Hou et al. [10] find the relationship between the foreground and the crowd density with a neural network. Bozzoli et al. [14] propose a method based on the statistical analysis of the optical flow representing the moving objects in different frames. Kong et al. [15] use a density map to measure the relative size of individuals and estimate a global scale measuring camera orientation. Then the number of pedestrians in the crowds is learned from the labeled training data. Yang et al. [16] adopt a sensor network to count individuals in a crowded scene. Lin et al. [17] use only one single image to estimate the number of people. Firstly, Harr wavelet transform (HWT) is used to extract the feature areas of head-like contour, and then a support vector machine (SVM) is used to classify these featured area as head or not. Finally, the perspective transforming technique is used to estimate the crowd size. To count people, Challa et al. [18] use visual information from a surveillance camera. Using data gathered from single camera, probability models are derived for basic counting scenario. By using the texture analysis and learning, Wu et al. [19] present an automatic method to estimate the people count.

Map-based methods can deal with occlusions and work in real time. However, they always need sufficient training data which may be difficult to obtain and the results depend on the training data quality.

The second category usually deals with occlusions by blob tracking, merging and splitting. Refs. [4,20–23] detect and track individuals in video sequences with some prior knowledge of pedestrians, and count people with the merge-split strategy. Taking advantage of kernel structural information matrices to represent object appearance, Li et al. [4] propose an object tracking framework. Kilambi et al. [20] address the problem of estimating the crowd density in a group. In [20], a heuristic-based and shape-based method is presented for people counting in a group and then each group as a single entity is tracked using a tracker based on an extended Kalman filter. Folgado et al. [21] present a block-based human model for real-time monitoring.

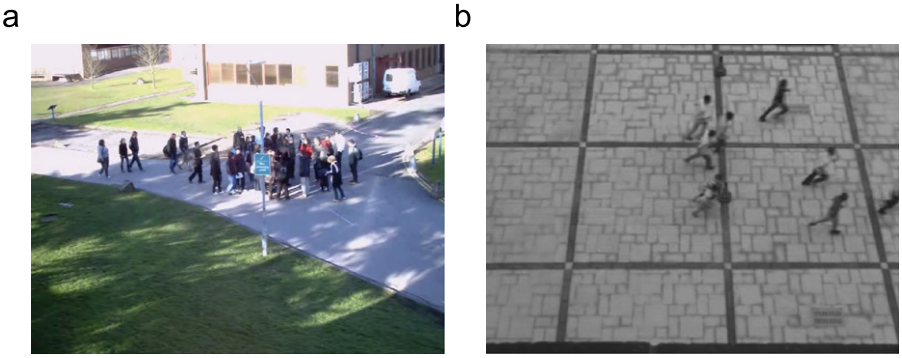

Fig. 2. Typical abnormal scenes. (a) People gathering. (b) People running.

Based on color image processing, Chen and Hsu [22] propose an automatic bi-directional counting method with one color video hung from the ceiling of the gate. By employing Markov random field (MRF), Guo et al. [23] present an algorithm for people counting. Three types of image features which meet the Markov properties are extracted. Then a least-square method is applied to estimate the people count. In related works by Zhang et al. in [24], a combined human segmentation and group tracking method for people counting is presented. In [25] by Zhang et al., a model-specified directional filter (MDF) is used to detect object candidate locations followed by a novel matching process to identify the pedestrian head positions and the number of people can be counted from the number of the heads. Rabaud et al. [26] develop a highly parallel version of Karhunen–Loeve transform (KLT) tracker and cluster a rich set of extended tracked features to identify the number of moving objects in a scene. Zhao et al. [27] segment the foreground blobs with some prior knowledge of human shape in a Bayesian framework. Liu et al. [28] utilize a method for camera auto-calibration based on information gathered by tracking people and present a model based segmentation algorithm which partitions a group of people into individual.

The detection-based method can obtain satisfactory results. However, edge extraction and tracking is a time-consuming process [10]. When the environment is complicated or consisting up to 10 people, the detection-based methods always fail because of high crowd density, occlusions and high computational load. Therefore, these systems are of limited usage in urban environments, which often contain large groups of people with severe occlusions [20].

2.2. Related works on abnormal crowd behavior detection

Abnormal crowd behavior detection can also be divided into two broad families named machine learning-based methods and threshold-based methods.

Machine learning-based approaches, such as principal components analysis (PCA), K-means and hidden Markov model (HMM) gain popularity in recent years. Most spatio-temporal representations in these papers assume that the volume contains a dominant, uniform motion patterns, although the motion patterns are exactly non-uniform in many crowded scenes. Kratz et al. [29] construct motion-pattern distributions which capture the variations of local spatio-temporal motion patterns to represent the video volume, then derive a distribution-based HMM, and improve the framework by constructing a coupled HMM. Authors in [6,30] adopt a spatio-temporal MRF model to detect abnormal activities. An unsupervised technique for detecting abnormal behaviors in large video sets is presented in [31]; rather than build explicit model of normal events, their paper compares each event with others to determine their similarity. Using optical flow patterns to estimate the model parameters, Andrade et al. [32] adopt HMMs to represent the flow sequences. Based on a dynamic condition random field (DCRF) model, Yin et al. [7] propose a new spatio-temporal event detection algorithm. By employing PCA for feature selection, Wu et al. [33] adopt SVM to classify human behaviors. By dividing a whole frame into small blocks, Wang et al. [34] use KLT corners to represent moving objects and cluster motion patterns in an unsupervised way. Wu et al. [35] extract the chaotic dynamics of all representative trajectories, and then use a probabilistic model to train these chaotic feature sets.

The machine learning-based methods can be quite effective in the environment where "normal" activity is well-defined and constrained. However, as the number of "normal" observed activity types can easily surpass that of unusual types, defining and modeling the "normal" activity in an unconstrained environment is always difficult. Hence, if the goal is to detect abnormal activities in a long video sequences and to be adaptive to different situations, the machine learning-based approaches are often ineffective [31].

There is also some research on threshold-based methods in video analysis. Once the target value exceeds a preset threshold, the monitor outputs an alert. With multiple local monitors collecting low-level statistics (velocity or direction) and using Bayesian surprise, Xie et al. [36] present a two layer threshold-based approach to detect abnormal events. Chen et al. [37] propose a two-stage hierarchical clustering approach which can group optical flows into crowds and detect abnormal events using the force field model. Employing an unsupervised clustering technique to segment the video into spatial-temporal volumes, Lu et al. [38] utilize spatio-temporal shape and flow correlation to detect abnormal activities. Without learning process and training data, Ihaddadene et al. [39] single out abnormal events by analyzing motion aspect instead of tracking subjects. Using a social force model, Mehran et al. [40] introduce a method to detect abnormal behaviors in crowd scenes. Considering density and direction simultaneously, authors in [41] use the motion heat map to define a region of interest and estimate the entropy of the frames to detect abnormal behaviors. Adopting motion features, Zhong et al. [9] define crowd energy to represent crowd density and detect abnormal activities. Cao et al. [42] combine crowd kinetic energy, motion variation and direction variation for anomaly detection.

Without training data, the threshold-based methods are easy to be implemented. However, these methods always have to deal with occlusions, tracking and segmentation. The threshold is difficult to be determined, always by experience, and the false alarm rate is usually high.

Our approach contributes to the second category of abnormal crowd behavior detection. However, we avoid tracking to avert typical problems such as extensive clutter and dynamic occlusions. Most of the related works only pay attention to motion information, such as optical flow variation, velocity variation, and direction variation. To our knowledge, there is little research on anomaly detection considering density, distribution and motion information simultaneously. Refs. [9,41,42] take crowd density into account. In the above three papers, motion area ratio is adopted to represent the people count. Obviously, the estimation of people count is still inaccurate. In this paper, we adopt a model to obtain accurate estimation of people count which leads to the effective detection of abnormal activities.

3. System overview

The system consists of two procedures: people counting and abnormal crowd behavior detection, as parts A and B shown in Fig. 3. The output of part A is the input of part B.

In the people counting algorithms, we assume that people move on a known ground plane where there are no other moving objects, such as cars or animals. The output of part A are foreground pixels and people count. Similar to most existing people-counting algorithms, we use an adaptive Gaussian mixture modeling (GMM) method to extract the foreground from video sequences, and exclude noise by applying a binary morphology technique. Then an initial model for people counting is discussed. From the pinhole perspective projection, we have two observations. Based on the two observations, we present an image potential energy model to improve the initial model. Finally, we present the method of parameter estimation which is necessary to the image potential energy model.

Part B aims to detect two typical abnormal crowd behaviors: people gathering and running. We design the detection algorithm

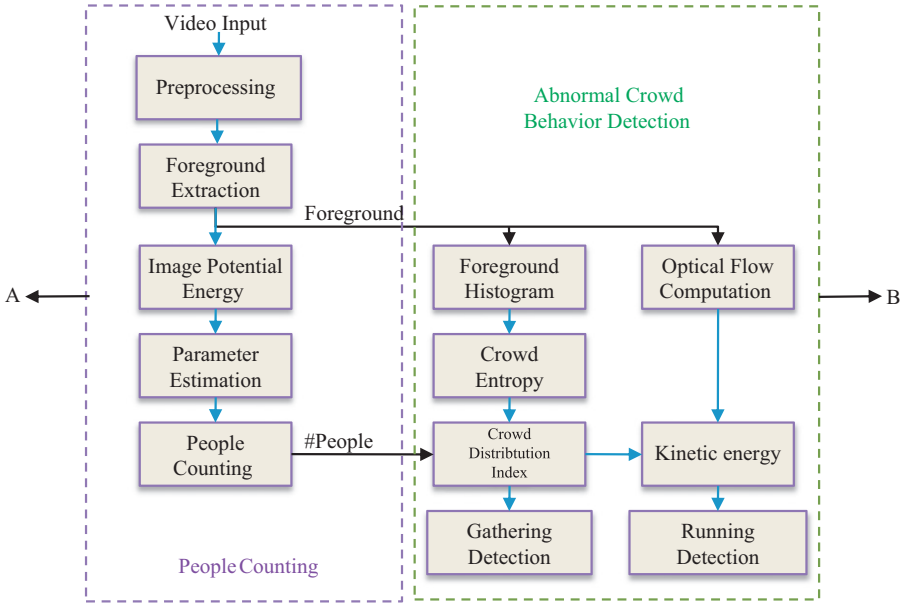

Fig. 3. System architecture.

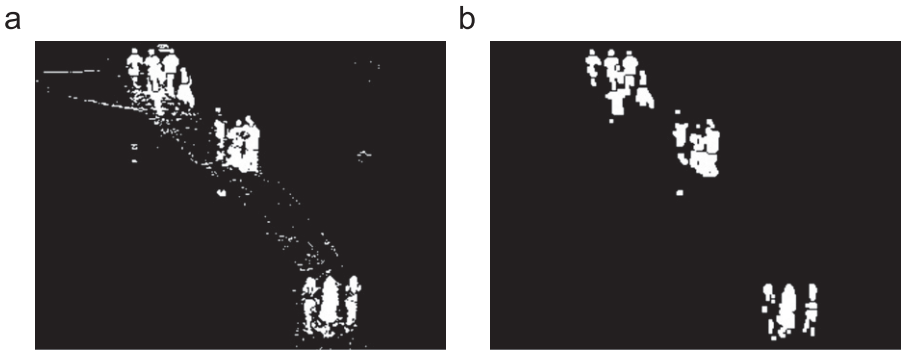

Fig. 4. Foreground extraction. (a) Foreground extracted by the GMM. (b) The foreground is polluted by noise as in (a). (b) Foreground after morphology operation. Morphology operation has improved the foreground greatly.

by the output of part A. Based on the foreground, we build the foreground histograms on *X*- and *Y*-axes, respectively, and then compute the probability distribution. Distribution entropy is defined by the foreground probability distribution. Combining the people count and crowd entropy, we define the Crowd Distribution Index (I_{CD}) to detect pedestrian gathering. To detect people running, based on I_{CD}, we improve the kinetic energy in [9], and the modified kinetic energy model can detect people running effectively.

4. Image potential energy model for people counting

In this section, we introduce the process of foreground extraction, then present an initial model, finally based on the above model, we define the image potential energy model to estimate the people count.

4.1. Foreground extraction

There have been a number of research works in moving object detection. In public scenes, the people moving in irregular motion patterns become foreground, which has been the basis of most existed people counting methods. Foreground detection plays an irreplaceable role in the detection-based people counting method. Although our image potential energy model does not have to track and segment, extracting foreground accurately is also helpful to the effectiveness of our model.

Several background models extract foreground effectively, such as single Gaussian model, non-parametric models, codebook models and mixture Gaussian model. The parameters of GMM evolve with time, so the model is adaptive. In this paper, we employ the mixture Gaussian model to extract foreground. The foreground is extracted as shown in Fig. 4(a).

However, the foreground extracted always accompanies with noise as Fig. 4(a). We exclude these noise by implementing erosion and dilation operations [43]. Fig. 4(b) shows the foreground after morphology operation.

4.2. An initial model

Generally, if the number of foreground pixels and the average number of pixels of one person can be counted, we can obtain the approximate estimation of people count as follows:

$$\#People = \frac{Number\ of\ Foreground\ Pixels}{Number\ of\ One\text{-}person\ Pixels}. \quad (1)$$

Unfortunately, the fact that far objects appear smaller than near ones makes (1) not work well.

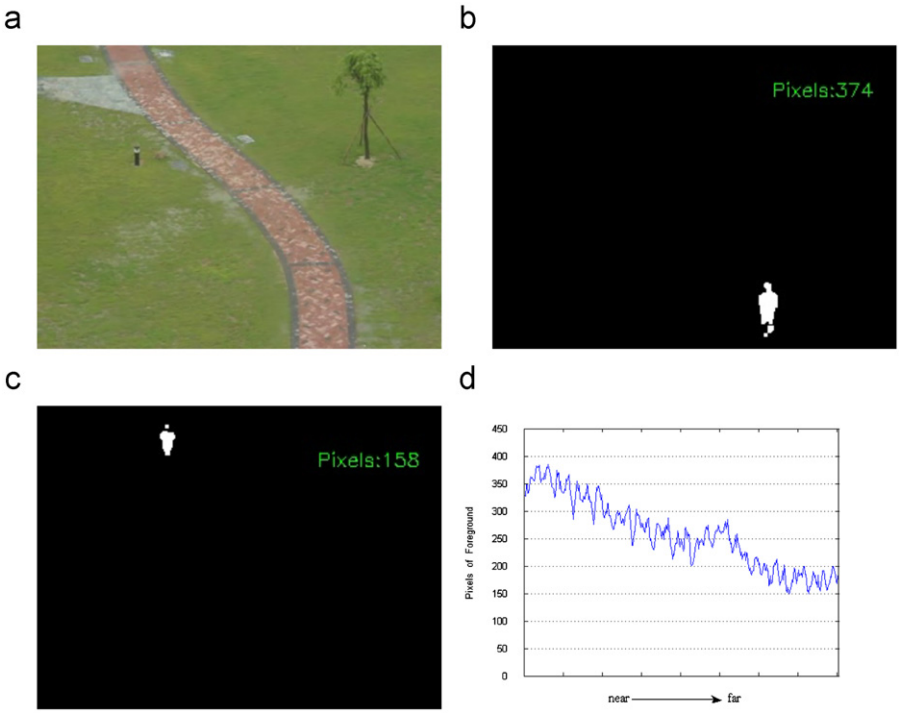

Fig. 5. The reason why the initial model cannot work well. (a) A typical scene in our campus. (b) When the people is closest to the camera, the number of foreground pixels is 374. (c) When the people is farthest to the camera, the number of foreground pixels is 158. (d) When the distance from the camera changes from near to far, the number of foreground pixels changes greatly (from 374 to 158).

As Fig. 5(b) and (c), the foreground pixels of one person with the shortest distance from camera are twice as that with the farthest distance. From Fig. 5(d), we know that when the distance away from the camera changes from near to far, the number of foreground pixels of one person reduces significantly. Based on the above fact, (1) can hardly obtain a satisfactory solution.

However, if we can compensate the defect of the above model by defining a new expression which is invariant when the distance from the camera changes, the modified model can give reasonable estimation. In the following, we propose the image potential energy model to improve the initial model.

4.3. Image potential energy model

Zhong et al. [9] present a model based on the kinetic energy to detect abnormalities in railway stations. Kinetic energy is related to the velocity and motion, and there is another kind of energy called potential energy. Inspired by [9] and gravitational potential energy, we establish the image potential energy model. We consider that potential energy is a kind of energy related to position or distance, for instance, the gravitational potential energy, the elastic potential energy, and the electric potential energy. Knowing the fact that if the object is farther away from the camera, the pixels related to it on the image plane are fewer. We define the image potential energy based on the distance from camera to overcome the shortcomings of the initial model. Depth information could be used to calculate the image potential energy as height information used to calculate the gravitational potential energy. However, depth information is difficult to obtain. Observing from pinhole perspective projection model [44], we obtain two important observations:

(a) The apparent size of an object depends on its distance from the camera.
(b) Supposing the origin of image coordinate system is located at the top-left corner and people walk on a ground plane. The farther people appear closer to the origin on the image plane; that is to say, the Y-axis value is smaller.

Based on the above observations and referring to the formulation of gravitational potential energy, we propose the image potential energy model as follows:

$$E_p = \sum_{i=1}^{X}\sum_{j=1}^{Y} m_{ij} g_{img} (H + Y - y_{ij}) \tag{2}$$

where E_p is the image potential energy, X and Y denote the width and the height of frame, respectively. m_{ij} is the mass of pixel with coordinates (i,j). $m_{ij} \in \{0,1\}$, with 1 denoting the foreground and 0 denoting the static background. Similar to the gravitational potential energy, we let g_{img} be a constant in this paper, $g_{img}=10$; while we can define g_{img} be a larger value in a specific area to detect abnormalities. H is an approximate measurement of the closest distance from camera and H is the only parameter required to be estimated. H is a constant to a fixed camera. y_{ij} is the image coordinate of Y-axis. When the object is farther to the camera, the related pixels are fewer, and y_{ij} is smaller which makes E_p of single pixel larger, so it can compensate the defect of the initial model.

The larger E_p means the people count is larger while smaller E_p implies smaller people count. Let e denote the average E_p of one person and N denote people count. We can estimate the people count as follows:

$$N = \frac{k \cdot E_p}{e} \tag{3}$$

In the above definition, we use k to reflect the effect of occlusions. k is a constant in the definition while it may change in different scenes.

4.4. Estimation of parameter H

The attribute of the image potential energy model is that image potential energy related to the object could be almost invariant to the distance from the object to the camera. The only parameter has to be estimated is H. An appropriate estimation of H can satisfy this requirement.

In order to estimate H effectively, we need a video clip in which there is only one person walking from near to far. There are n frames in the video clip. Let E_{pi} denote the E_p of the ith frame and $\overline{E_p}$ denote the expectation of all frames. We have variance of these frames as follows:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(E_{pi}-\overline{E_p})^2}{n} \tag{4}$$

To minimize σ^2, we can obtain optimal estimation of H. Fig. 6 shows the curve of E_p of one person with different H.

In Fig. 6(b), the variance of curve is the smallest, so 380 is the optimal estimation of H. E_p of one person is approximately equal to 1.1×10^6. E_p with the optimal estimation is not related to the distance and it is helpful to people counting. The image potential energy model compensates the defect of the initial model.

5. Abnormal crowd behavior detection

In this section, we present the foreground histograms on X- and Y-axes, respectively, and then define crowd entropy to represent the spatial distribution of foreground. I_{CD} is defined by crowd entropy and people count (N). In the last part, based on I_{CD}, the modified kinetic energy model which can detect people running is presented.

5.1. Foreground histogram

Entropy is usually a measure of disorder. Here, in order to measure the dispersion of crowd, we refer to the definition of entropy in information theory [45]. We firstly build foreground histograms by projecting the foreground to X- and Y-axes, respectively. The foreground histograms on X- and Y-axes are denoted by $h_x(i)$ and $h_y(j)$:

$$h_x(i) = \{k_i, 0 < i \le n_1\} \tag{5}$$

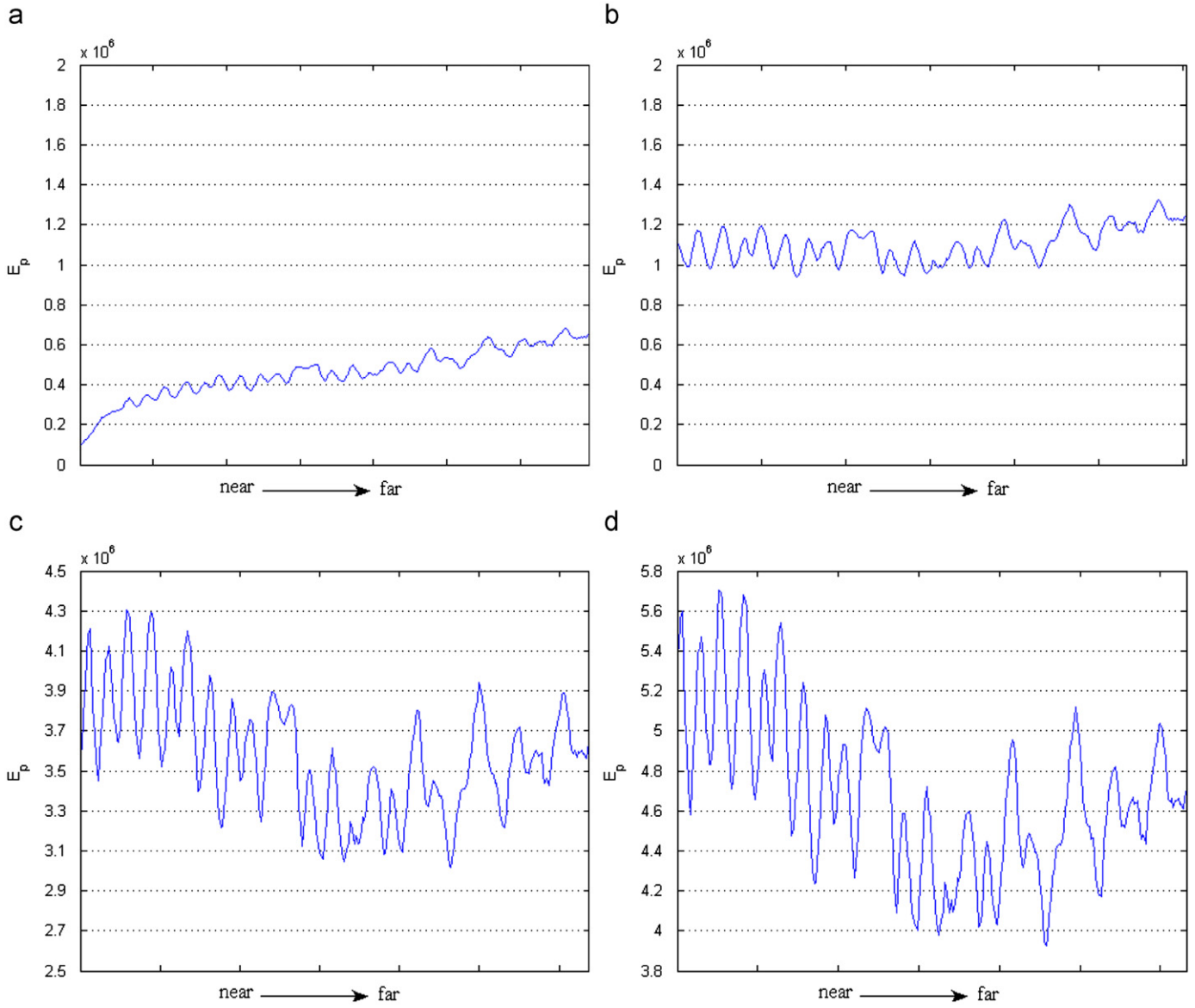

Fig. 6. E_p of one person with different H. (a) $H=100$. (b) $H=380$. (c) $H=1500$. (d) $H=2000$.

$$h_y(j) = \{q_j, 0 < j \le n_2\} \tag{6}$$

where n_1 denotes the number of histogram bins on the horizontal direction (X-axis) and n_2 denotes the number of histogram bins on the vertical direction (Y-axis). k_i implies the number of foreground pixels projected to X-axis at the ith bin and q_j denote the number of foreground pixels projected to the Y-axis at the jth bin for each frame. n_1 and n_2 are determined by the distance from the camera. When the pedestrians in the scenes are farther from the camera, the related pixels on the image plane are fewer, so the bins n_1 and n_2 are bigger. In this paper, we let $n_1=20$, and $n_2=12$.

The calculated histograms indicate the foreground distribution on the horizontal and vertical directions. Fig. 7 is an example of foreground histograms. The frame size is 320×240.

5.2. Foreground probability distribution

The foreground probability distribution on X- and Y-axes can be directly estimated from $h_x(i)$ and $h_y(j)$:

$$p_x(i) = \frac{h_x(i)}{m}, \quad 0 < i \le n_1, \ i \in N \tag{7}$$

$$p_y(j) = \frac{h_y(j)}{m}, \quad 0 < j \le n_2, \ i \in N \tag{8}$$

where m is the total foreground pixels.

5.3. Crowd entropy

We define crowd entropy ($H(X)$, $H(Y)$) based on the foreground probability:

$$H(X) = \sum_{i=1}^{n_1} p_x(i) \log\left(\frac{1}{p_x(i)}\right), \quad p_x(i) \neq 0 \tag{9}$$

$$H(Y) = \sum_{j=1}^{n_2} p_y(j) \log\left(\frac{1}{p_y(j)}\right), \quad p_y(j) \neq 0 \tag{10}$$

Crowd entropy denotes the dispersion of foreground on horizontal and vertical directions. For example, if the probability is 1 at bin i, then $p(i)=1$, so the entropy of probability distribution is $H = 1 \cdot \log(1) = 0$. If the probability is equally distributed on all bins, then $H = n \cdot 1/n \cdot \log(1/(1/n)) = \log(n)$. Therefore, a distribution with a single sharp peak yields to a low entropy value, whereas a dispersed distribution corresponds to a high entropy value.

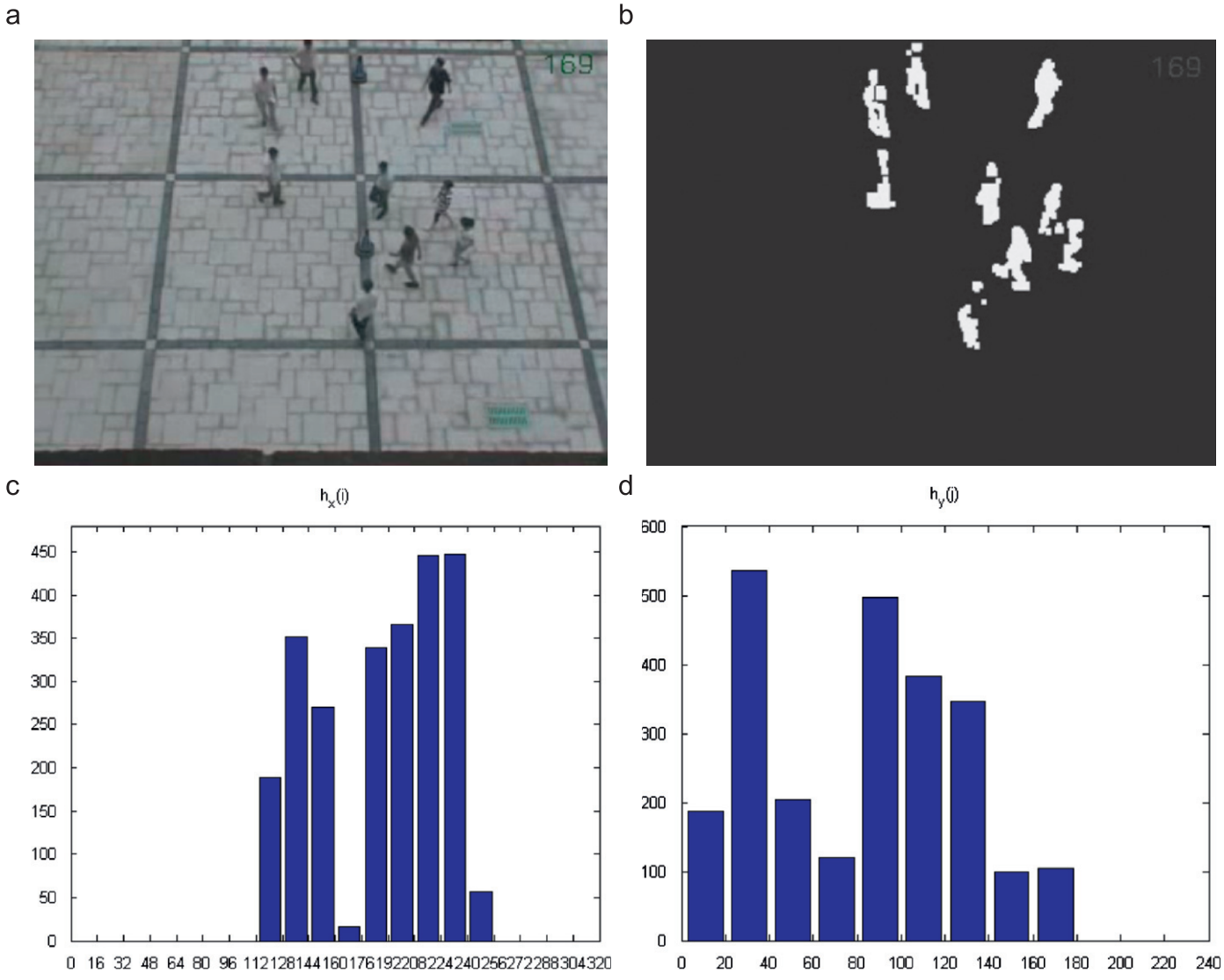

Fig. 7. Foreground histograms. (a) Original frame. (b) Foreground. (c) $h_x(i)$. (d) $h_y(j)$.

5.4. Crowd dispersion

$H(X)$ and $H(Y)$ are the dispersion of foreground on horizontal and vertical directions, respectively. Crowd dispersion (D) is used to reflect the dispersion of one frame globally. D is defined as follows:

$$D = H(X) \cdot H(Y). \tag{11}$$

When there are a few pedestrians moving in the scene, $H(X)$ or $H(Y)$ may equal to 0. However, D will be used to define I_{CD} as denominator, which means D cannot be equal to 0. We modify D as follows:

$$D = \begin{cases} 1 & \text{if } H(X) \cdot H(Y) < 1 \\ H(X) \cdot H(Y) & \text{otherwise} \end{cases} \tag{12}$$

5.5. Crowd distribution Index for people gathering detection

When we obtain the information of N and D, we model people gathering. We assume that larger I_{CD} means people gather in a smaller region. Generally, people count (N) is proportional to I_{CD} and crowd dispersion (D) is inversely proportional to I_{CD}. In experiments, we find out that the following definition works well:

$$I_{CD} = \frac{N^2}{D^3} \tag{13}$$

Eq. (13) means that when people gather in a local region, N is large while D is small, which yields to a large I_{CD} value. Based on a threshold, we can detect pedestrian gathering. However, when the pedestrians move in a horizontal or vertical line (see Fig. 8(a)), D will be very small while N is medium, which makes I_{CD} be a very large value and leads to a false alarm.

For example, in Fig. 8(a), four people walk in a line in the vertical direction, no abnormal activities happen, however, $N^2/D^3 = 16$. In Fig. 8(b), the people gathering in a local region indicate some accidents taking place, but $N^2/D^3 = 9$. So (13) will lead to a false alarm. To avoid this kind of false alarm, we use a piecewise function to modify the definition:

$$I_{CD} = \begin{cases} 1.1N, & D \le 2 \\ \dfrac{N^2}{D^3}, & D > 2 \end{cases} \tag{14}$$

The modified I_{CD} can effectively single out people gathering and is helpful to detect pedestrian running.

When a sudden change of I_{CD} arise, we consider it may be noise and the system should not give an alarm signal. Only when the I_{CD} is greater than its threshold for 10 consecutive frames do the system alarm. This procedure can reduce the false alarm rate effectively.

5.6. Crowd kinetic energy for people running detection

In order to detect people running, we define the kinetic energy. This paper adopts Harris corners as features. Motion vectors are obtained by tracking features of a series of images through the Lucas–Kanade optical flow approach [46]. These motion vectors can be refined by using a mask which represents the crowd region. We generate the mask from the foreground. From the motion vector, we can infer its velocity. Fig. 9 show the result of the optical flow computation.

The kinetic energy of each frame is defined as follows:

$$E_{kn} = I_{CD} \cdot \sum_{i=1}^{m} v_i^2 \tag{15}$$

where E_{kn} is the kinetic energy of the nth frame, and v_i is the velocity. I_{CD} provides not only the information of people count but also crowd distribution. When people gather in a local region, occlusions may occur, which means that motion features tend to

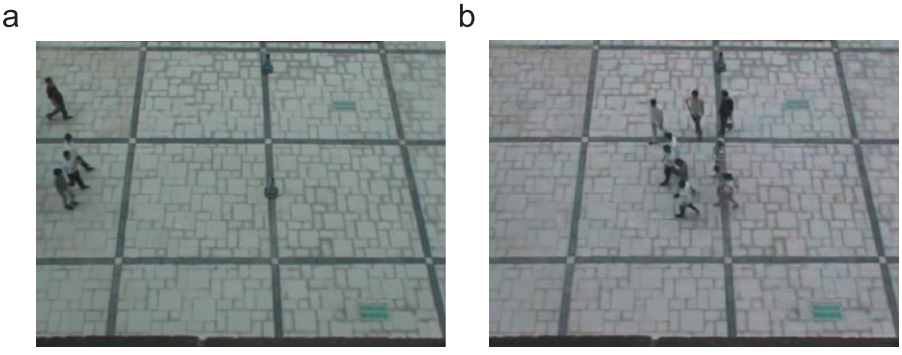

Fig. 8. The reason why we use piecewise function to define I_{CD}. (a) $D = 1.1, N = 4, N^2/D^3 = 16$. (b) $D = 2.22, N = 8, N^2/D^3 = 9$.

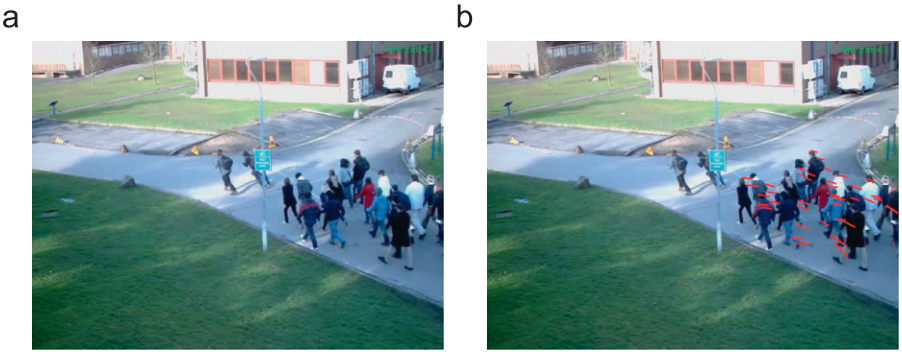

Fig. 9. Optical flow computation. (a) Original frame. (b) Motion vector.

be smaller than that in scene where persons are sparsely scattered. However, in the above situation, I_{CD} will be larger which compensates the influence of occlusions. When a sudden change of E_k arise, it may be noise. To avoid this kind of noise, only when the E_{kn} is greater than its threshold for 10 consecutive frames do the system alarm.

6. Experiments

In this section, we present the experimental results of people counting in the first part while the experimental results of abnormal crowd behavior detection are shown in the second part. The system has been implemented in C++ and tested using AMD 3500+ PC with 994 MHz CPU and 1 GB memory, and it can run in nearly real time (20 fps). We collect a number of video clips with different video size, different levels of crowd density and different scenes to test our model. Some clips are downloaded from the Internet, and others are recorded in our campus.

6.1. Experimental results of people counting

In order to evaluate the effectiveness of our model with low level of crowd density, we install a camera on building B of Shenzhen Institute of Advanced Technology with a topdown view. The video resolution is 320×240. The number of people in the scene is below 12. Fig. 10 shows the typical screen-captures of our results. In this clip, $H=380$, $e=1\,050\,000$ and $k=1.02$.

As shown in Fig. 10, there are severe occlusions between people in some frames, but the estimation is reliable. Fig. 10 shows that our model can cope with severe occlusions. Table 1 shows the accuracy rate of our experiments at low crowd density. Table 1 shows that our model works well when the crowd density is small and the average error is only 1.

Fig. 10 and Table 1 give us a satisfactory result at low level of crowd density. Then the model is tested at high level of crowd density. We download some video clips from the Internet. The resolution of the testing video is 480×360, and we have to estimate H empirically. In this clip, $H=680$, $e=22\,500\,000$, and $k=1.02$. Although H is not estimated by (4), the result is still acceptable. Fig. 13 shows the typical screen-captures of our results.

As shown in Fig. 11, the crowd density is high and there are severe occlusions in the scene, but our image potential energy model can handle this situation. Because the crowd density changes very fast, we cannot make accuracy analysis as Table 1. Ten frames are selected randomly from the result clips to show the effectiveness of our model. The first row denotes the frame index (*FI*). The second row is the actual number of people and the third row is *ER*.

Table 2 shows the effectiveness of our model. For each frame, the error rate is very small. Fig. 12 shows the results of another two clips, where the estimation results are accurate.

6.2. Experimental results of abnormal crowd behavior detection

The people gathering and running events are detected in a static method by comparing I_{CD} and E_{kn} with a specific threshold.

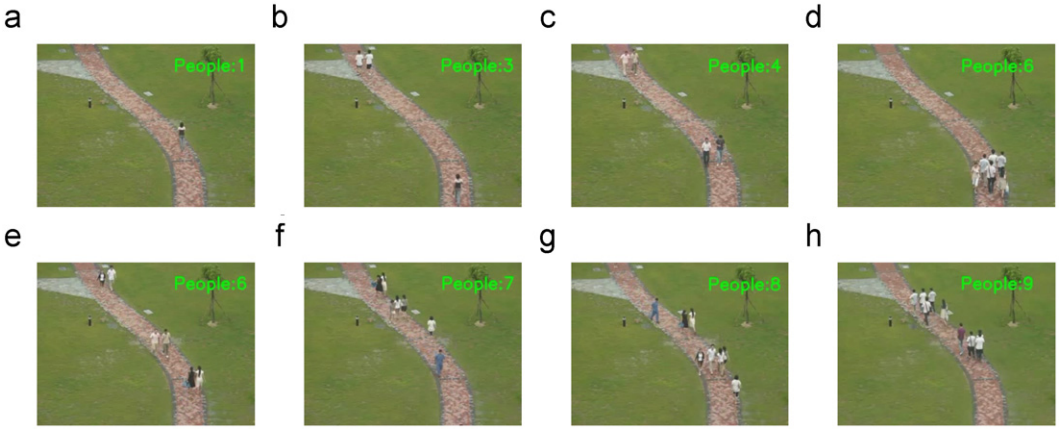

Fig. 10. Typical screen-captures with low level of crowd density.

Table 1
Accuracy rate of our experiments at low crowd density. *ER* represents the experimental result. In the first row, each item is denoted as *M*(*N*), where *M* represents the true number of people and *N* is the frames of special *M*.

ER	1(341) (%)	2(132) (%)	3(221) (%)	4(147) (%)	5(106) (%)	6(181) (%)	7(246) (%)	8(157) (%)	9(164) (%)	12(78) (%)
1	100									
2		96.97	7.69							
3		3.03	90.50							
4			1.81	92.52	10.38					
5				7.48	89.62	0.55				
6						82.88	2.44	4.46		
7						16.57	81.71	27.39		
8							15.85	68.15		
9									20.70	
10									73.81	
11									5.49	
12										64.10
13										35.90

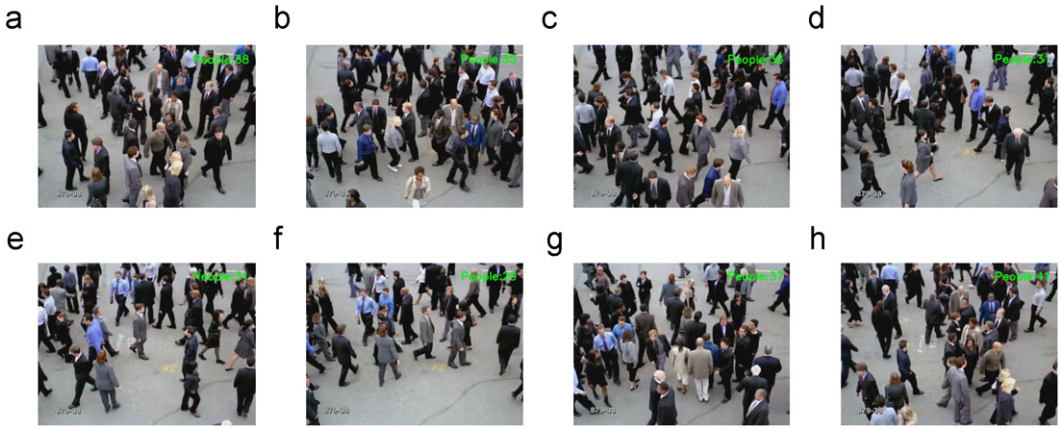

Fig. 11. Typical screen-captures at high crowd density.

Table 2
Experimental results (*ER*) and error (ϵ) at high crowd density.

FI	55	106	131	183	202	217	296	512	636	730	784
#people	37	40	39	42	42	44	39	33	31	29	31
ER	35	38	41	41	40	41	36	33	32	29	30
ϵ (%)	5.41	5.00	5.13	2.38	4.76	6.82	7.69	0.00	3.23	0.00	3.23

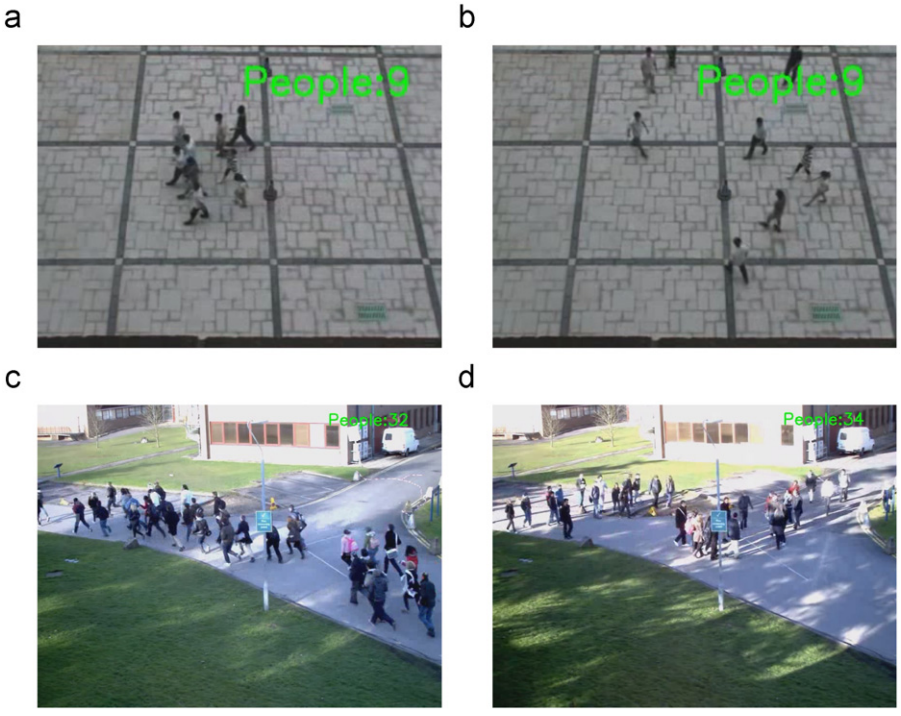

Fig. 12. Results of another two clips.

The threshold estimation is necessary for each crowd scene because the threshold varies depending on the camera position.

In order to evaluate the power of our algorithm, we choose four video clips to show the results. Each clip consists of two kinds of situation: normal and abnormal. The first two video clips are obtained by a stationary camera installed on building B of Shenzhen Institute of Advanced Technology in a topdown view. Figs. 13 and 14 show the results and analysis. The video resolution is 320×240. Under this circumstance, the threshold of I_{CD} is 6 and the threshold of E_k is 180.

In the first clip shown in Fig. 13, people are gathering from different directions and then walk together in one direction; our algorithm detects the anomaly in time. Fig. 13(b) shows the curve of I_{CD}, where the abnormal gathering occurs at the 218th frame. As there is no people running in the scene, curve of E_k is ignored.

In the second clip shown in Fig. 14, people walk in a low speed, and then begin to run suddenly. The abnormal running is detected at the 500th frame. Fig. 14(b) shows the curve of E_k. No people gather in this clip, so the curve of I_{CD} is not shown here. In Fig. 14(b), we can see a sharp peak which is actually noise

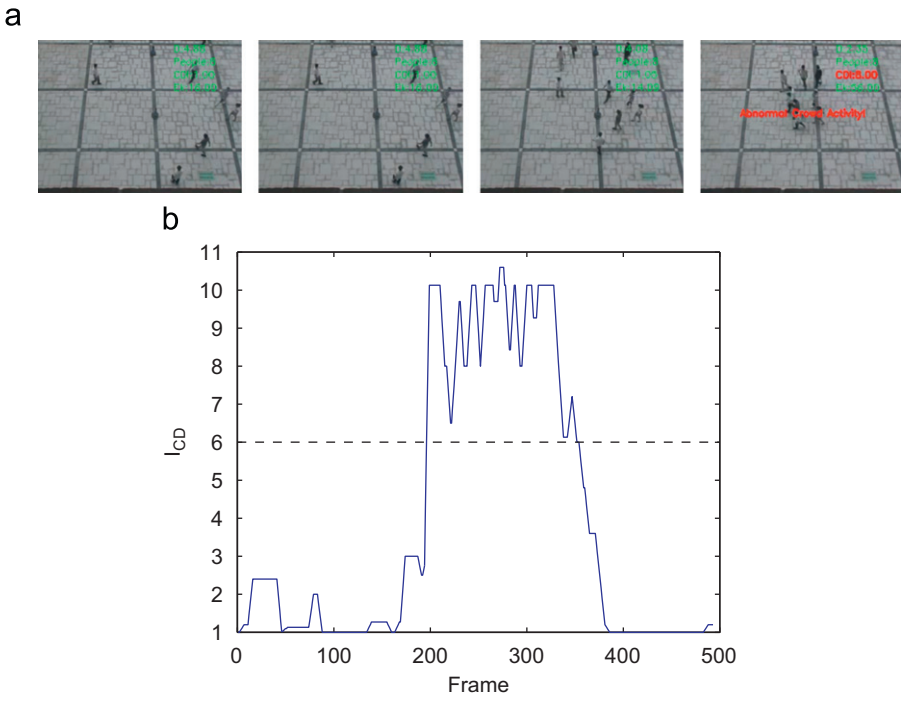

Fig. 13. Pedestrian gathering. (a) Typical screen-captures of the video clip. (b) I_{CD} curve.

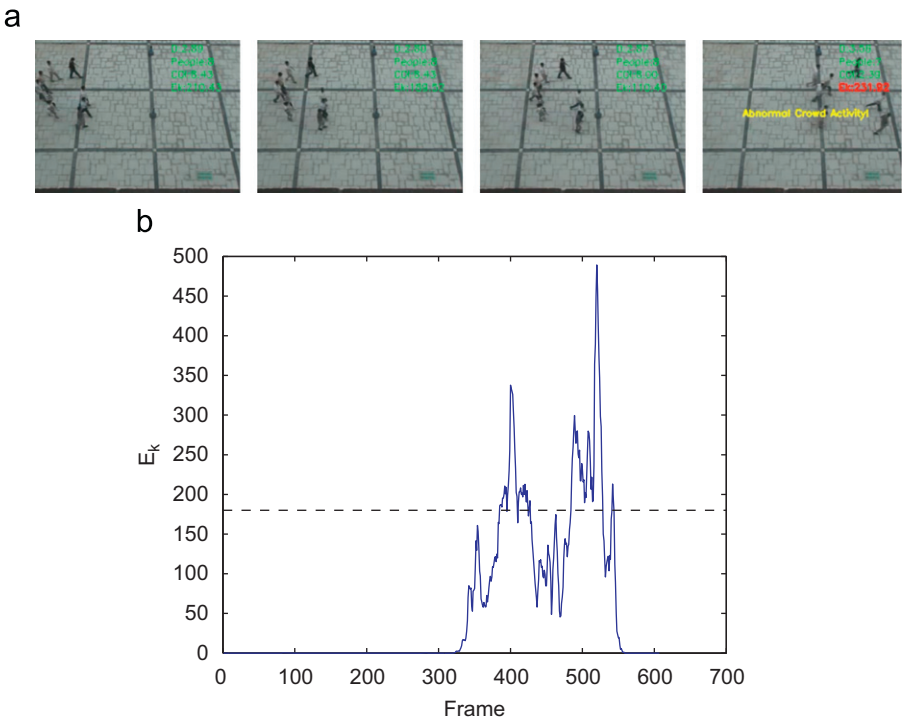

Fig. 14. Pedestrian running. (a) Typical screen-captures of the video clip. (b) E_k curve.

occurring at the 400th frame; however, it only lasts six frames, and our system avoids false alarm successfully.

The crowd density of the above two video clips is low and the two typical abnormal events take place separately in different clips. Another two video clips downloaded from the Internet are with high crowd density, and two typical anomalies occur in the same clip. The frame resolution is 768×576. Under this circumstance, the threshold of I_{CD} is 12 and the threshold of E_k is 40 000. Figs. 15 and 16 show the experimental results.

In the third clip shown in Fig. 15, we can see the crowd density is high and occlusions are severe. However, our algorithm detects the pedestrian gathering in a local region at the 100th frame effectively. In Fig. 15(a), I_{CD} is red, which indicates that people gathering is taking place. Fig. 15(b) presents the curve of I_{CD}. In

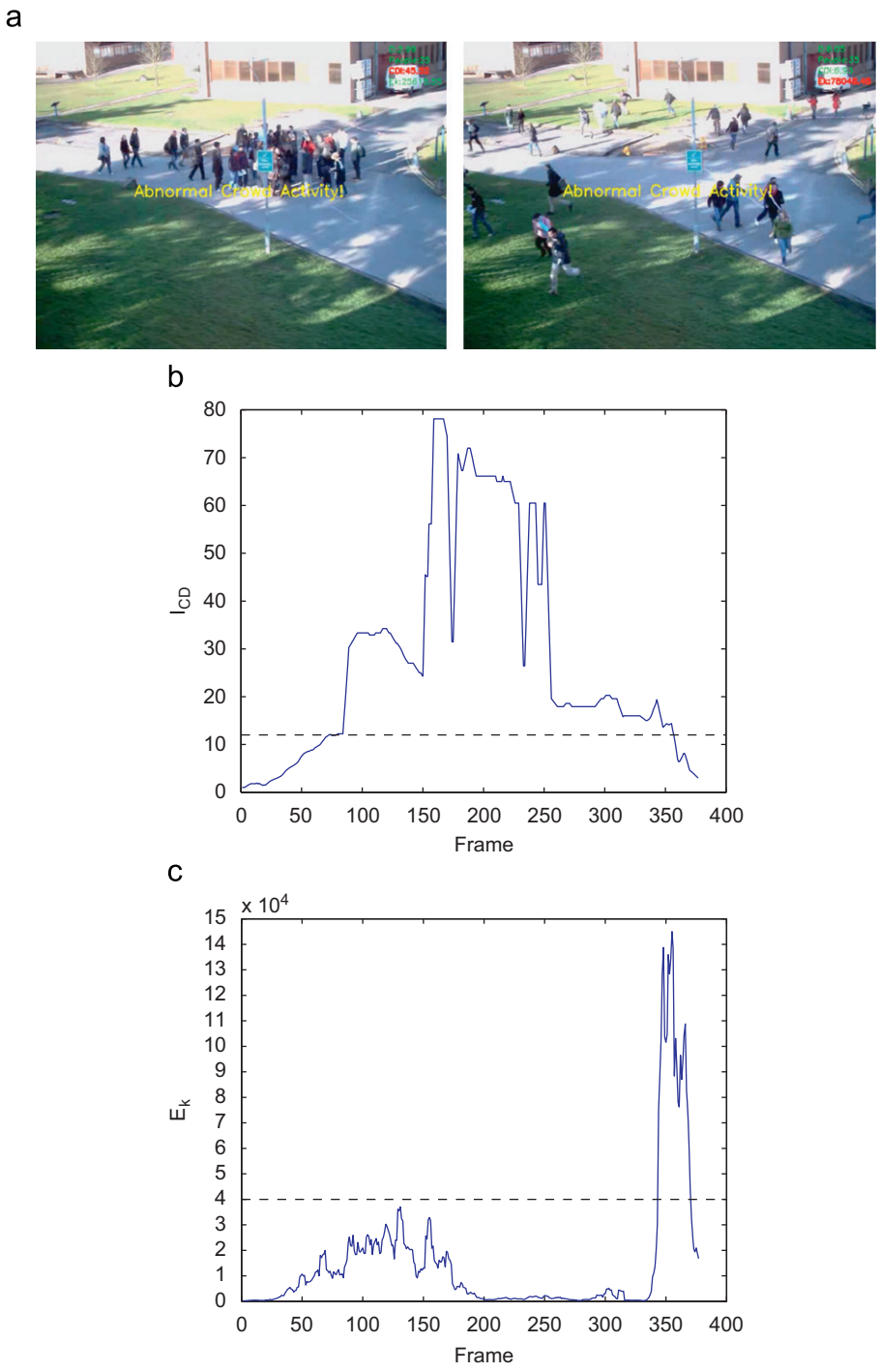

Fig. 15. Gathering firstly and then running. (a) Typical screen-captures of the video clip. (b) I_{CD} curve. (c) E_k curve. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the same clip, after gathering, people start to run suddenly. Fig. 15(c) shows the curve of E_k, where running is detected at the 355th frame.

As shown in Fig. 16, at the 44th frame, our algorithm detects abnormal activity happening, while I_{CD} and E_k are both red, which means pedestrian gathering and running occur simultaneously.

Finally, we compare our model with other typical approaches. Table 3 shows the details. The machine learning-based method can be quite effective in the environment where "normal" activity is well-defined and detect more kinds of anomalies. However, the high computational load makes the system cannot work in real time. Without training data, the threshold-based methods are easy to be implemented and always run in real time. In this paper, our system can effectively single out two typical anomalies and run in real time. Comparing to other threshold-based approaches, our model considers more features and can deal with higher crowd density with lower false alarm rate.

7. Conclusions and future work

The main characteristics of the image potential energy model was that the image potential energy related to the object was invariant to the distance from the object to the camera.

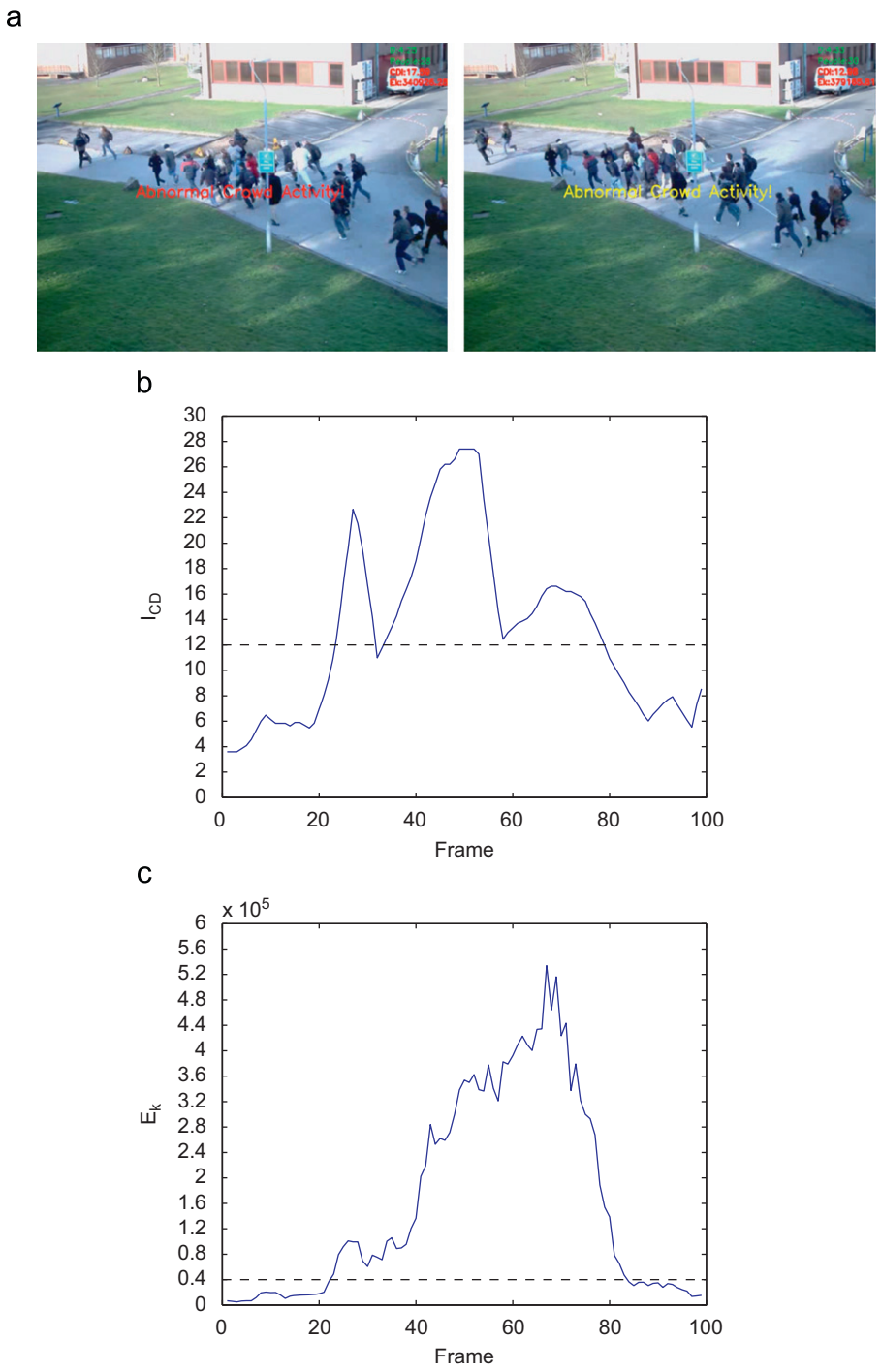

Fig. 16. Gathering and running occur at the same time. (a) Typical screen-captures of the video clip. (b) I_{CD} curve. (c) E_k curve. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Depending on the energy invariance feature, the average image potential energy of each person was obtained from typical video clips, and the image potential energy of all people divided by the average potential energy of one person was the people count. In this paper, we presented a method to detect two typical abnormal activities: people gathering and running. Building histograms on the *X*- and *Y*-axes, respectively, we could obtain probability distribution of the foreground object and then defined as crowd entropy. We defined the Crowd Distribution Index by combining the people counting results with crowd entropy to represent the spatial distribution of crowd. Crowd Distribution Index was thresholded to detect people gathering. To detect people running, the kinetic energy was determined by computation of optical flow and Crowd Distribution Index. Kinetic energy was also thresholded to detect people running. To test the performance of our algorithm, videos of different scenes and different crowd density were used in the experiments. Without camera calibration and training data, our method could robustly detect abnormal behaviors with low computation load.

However, this model is not perfect. To achieve the best performance, we need specific video clips to estimate *H*. If we can obtain the depth information as in [47], we will establish a

Table 3
Comparison with other models.

Reference papers	Methods	Level of crowd density	Features	Real time?	Anomalies
Kratz et al. [19]	Machine-learning	High	Spatio-temporal gradients	No	Irregular directions, individuals obstructing traffic and so on
Kim et al. [20]	Machine-learning	Low	Distribution of optical flow	No	Loitering, belonging dropping and so on
Zhong et al. [16]	Threshold-based	Low	Motion feature	Yes	Fighting, pedestrian running
This paper	Threshold-based	Medium	People count, velocity, dispersion of crowd	Yes	Pedestrian gathering, pedestrian running

perfect model theoretically. The shortcoming of the threshold-based approach is also obvious: the appropriate estimation of threshold is necessary and sensitive to the effectiveness. The threshold always has to be determined by experience. In the future work, we will combine the threshold-based methods and machine-learning-based methods together to detect more kinds of abnormal behaviors, such as belonging dropping, loitering, crossing over the fence. Some appropriate modeling approaches and hardware with high computational power such as graphics processing unit (GPU) will be included in our future work.

Acknowledgment

The work described in this paper is partially supported by the Nature Science Foundation of China (61005012), Shenzhen/Hongkong Innovation Circle Project (ZYB200907070024A), the Key Laboratory of Robotics and Intelligent System of Guangdong Province (2009A060800016), Introduced Innovative R&D Team of Guangdong Province (201001D0104648280), and the grant from Shenzhen public science and technology (SY200806300R1A). The authors would like to thank Ruiqing Fu, Lei Zhang, Ke Xu, and Long Han for their valuable contribution to this paper. The authors also would like to thank professor Dezhen Song from Texas A&M university for his suggestions and revision.

References

[1] D. Tao, X. Li, X. Wu, S. Maybank, Geometric mean for subspace selection, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2008) 260–274.
[2] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR'05: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, 2005, pp. 20–25.
[3] D. Tao, X. Li, X. Wu, S. Maybank, General tensor discriminant analysis and gabor features for gait recognition, IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007) 1700–1715.
[4] X. Li, W. Hu, H. Wang, Z. Zhang, Robust object tracking using a spatial pyramid heat kernel structural information representation, Neurocomputing 73 (16–18) (2010) 3179–3190.
[5] Y. Yuan, Y. Pang, J. Pan, X. Li, Scene segmentation based on IPCA for visual surveillance, Neurocomputing 72 (10–12) (2009) 2450–2454.
[6] J. Kim, K. Grauman, Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates, in: CVPR'09: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Florida, USA, 2009, pp. 2921–2928.
[7] J. Yin, D. Hu, Q. Yang, Spatio-temporal event detection using dynamic conditional random fields, in: IJCAI'09: Proceedings of 21st International Joint Conference on Artificial Intelligence, vol. 9, Los Angeles, USA, 2009, pp. 1321–1327.
[8] E. Carmona, M. Rincón, M. Bachiller, J. Martínez-Cantos, R. Martínez-Tomás, J. Mira, On the effect of feedback in multilevel representation spaces for visual surveillance tasks, Neurocomputing 72 (4–6) (2009) 916–927.
[9] Z. Zhong, W. Ye, S. Wang, M. Yang, Y. Xu, Crowd energy and feature analysis, in: ICIT'07: IEEE International Conference on Integration Technology, Shenzhen, China, 2007, pp. 144–150.
[10] Y. Hou, G. Pang, Automated people counting at a mass site, in: ICAL'08: IEEE International Conference on Automation and Logistics, Qingdao, 2008, pp. 464–469.
[11] W. Bian, D. Tao, Max–min distance analysis by using sequential SDP relaxation for dimension reduction, IEEE Trans. Pattern Anal. Mach. Intell. 33 (5) (2011) 1037–1050.
[12] N. Guan, D. Tao, Z. Luo, B. Yuan, Non-negative patch alignment framework, IEEE Trans. Neural Networks 22 (8) (2011) 1218–1230.
[13] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (6) (1933) 417–441.
[14] M. Bozzoli, L. Cinque, E. Sangineto, A statistical method for people counting in crowded environments, in: ICIAP'07: the 14th International Conference on Image Analysis and Processing, Modena, 2007, pp. 506–511.
[15] D. Kong, D. Gray, H. Tao, A viewpoint invariant approach for crowd counting, Pattern Recognition 3 (2006) 1187–1190.
[16] D. Yang, H. González-Baños, L. Guibas, Counting people in crowds with a real-time network of simple image sensors, in: The 9th IEEE International Conference on Computer Vision, Nice, France, 2003, pp. 122–129.
[17] S. Lin, J. Chen, H. Chao, Estimation of number of people in crowded scenes using perspective transformation, IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. 31 (6) (2001) 645–654.
[18] S. Challa, K. Aboura, K. Ravikanth, S. Deshpande, Estimating the number of people in buildings using visual information, in: IDC'07: Information, Decision and Control, Adelaide, Australia, 2007, pp. 124–129.
[19] X. Wu, G. Liang, K. Lee, Y. Xu, Crowd density estimation using texture analysis and learning, in: ROBIO'06: IEEE International Conference on Robotics and Biomimetics, Kunming, China, 2006, pp. 214–219.
[20] P. Kilambi, E. Ribnick, A. Joshi, O. Masoud, N. Papanikolopoulos, Estimating pedestrian counts in groups, Comput. Vision Image Understanding 110 (1) (2008) 43–59.
[21] E. Folgado, M. Rincón, E. Carmona, M. Bachiller, A block-based model for monitoring of human activity, Neurocomputing 74 (8) (2010) 1283–1289.
[22] T. Chen, An automatic bi-directional passing-people counting method based on color image processing, in: The 37th IEEE International Carnahan Conference on Security Technology, Taiwan, 2003, pp. 200–207.
[23] J. Guo, X. Wu, T. Cao, S. Yu, Y. Xu, Crowd density estimation via Markov random field (MRF), in: WCICA'10: The 8th World Congress on Intelligent Control and Automation, Jinan, 2010, pp. 258–263.
[24] E. Zhang, F. Chen, A fast and robust people counting method in video surveillance, in: 2007 International Conference on Computational Intelligence and Security, Harbin, 2007, pp. 339–343.
[25] X. Zhang, G. Sexton, Automatic human head location for pedestrian counting, in: The Sixth International Conference on Image Processing and Its Applications, vol. 2, Dublin, Ireland, 1997, pp. 535–540.
[26] V. Rabaud, S. Belongie, Counting crowded moving objects, in: CVPR'06: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, New York, 2006, pp. 705–711.
[27] T. Zhao, R. Nevatia, Tracking multiple humans in crowded environment, in: CVPR'04: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, Washington, DC, USA, 2004, pp. 406–413.
[28] X. Liu, P. Tu, J. Rittscher, A. Perera, N. Krahnstoever, Detecting and counting people in surveillance applications, in: AVSS'05: IEEE Conference on Advanced Video and Signal Based Surveillance, Genova, 2005, pp. 306–311.
[29] L. Kratz, K. Nishino, Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models, in: CVPR'09: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, FL, USA, 2009, pp. 1446–1453.
[30] Y. Benezeth, P. Jodoin, V. Saligrama, C. Rosenberger, Abnormal events detection based on spatio-temporal co-occurrences, in: CVPR'09: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, FL, USA, 2009, pp. 2458–2465.
[31] H. Zhong, J. Shi, M. Visontai, Detecting unusual activity in video, in: CVPR'04: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, Washington, DC, USA, 2004, pp. 819–826.
[32] E. Andrade, R. Fisher, S. Blunsden, Detection of emergency events in crowded scenes, in: The Institution of Engineering and Technology Conference on Crime and Security, 2006, London, UK, 2006, pp. 528–533.

[33] X. Wu, Y. Ou, H. Qian, Y. Xu, A detection system for human abnormal behavior, in: IROS'05: IEEE/RSJ International Conference on Intelligent Robots and Systems, Alberta, Canada, 2005, pp. 1204–1208.
[34] S. Wang, Z. Miao, Anomaly detection in crowd scene, in: ICSP'10: 10th IEEE International Conference on Signal Processing, Beijing, China, 2010, pp. 1220–1223.
[35] S. Wu, B. Moore, M. Shah, Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes, in: CVPR'10: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Florida, USA, 2010, pp. 2054–2060.
[36] J.S. Xie, L. Guo, Y.B. Chen, Z.L., The detection of unusual events in video based on Bayesian surprise model, in: ICISE'10: the 2nd International Conference on Information Science and Engineering, Hangzhou, China, 2010, pp. 4784–4788.
[37] D. Chen, P. Huang, Dynamic human crowd modeling and its application to anomalous events detection, in: ICME'10: IEEE International Conference on Multimedia and Expo, 2010, Singapore, 2010, pp. 1582–1587.
[38] L. Yong, H. Dongjian, Video-based detection of abnormal behavior in the examination room, in: IFITA'10: International Forum on Information Technology and Applications, vol. 3, Kunming, China, 2010, pp. 295–298.
[39] N. Ihaddadene, C. Djeraba, Real-time crowd motion analysis, in: ICPR'08: the 19th International Conference on Pattern Recognition, FL, USA, 2008, pp. 1–4.
[40] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: CVPR'09: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, FL, USA, 2009, pp. 935–942.
[41] M. Sharif, N. Ihaddadene, C. Djeraba, Crowd behaviour monitoring on the escalator exits, in: ICCIT'08: the 11th International Conference on Computer and Information Technology, Venice, Italy, 2008, pp. 194–200.
[42] T. Cao, X. Wu, J. Guo, S. Yu, Y. Xu, Abnormal crowd motion analysis, in: ROBIO'09: IEEE International Conference on Robotics and Biomimetics, Guilin, China, 2009, pp. 1709–1714.
[43] R. Gonzalez, R. Woods, Digital Image Processing, Prentice Hall, NJ, 2002.
[44] D. Forsyth, J. Ponce, Computer Vision: A Modern Approach, vol. 54, Prentice Hall, 2002.
[45] L. Brillouin, Science and Information Theory, Dover Publications, 2004.
[46] B. Lucas, T. Kanade, et al., An iterative image registration technique with an application to stereo vision. in: IJCAI'81: International Joint Conference on Artificial Intelligence, Vancouver, Canada, vol. 3, 1981, pp. 674–679.
[47] G. Zhang, J. Jia, T. Wong, H. Bao, Recovering consistent video depth maps via bundle optimization, in: CVPR'08: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 2008, pp. 1–8.

Guogang Xiong received Bachelor of Engineering from Chongqing University, China, in 2009. He is currently a postgraduate student of the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests include computer vision and source location.

Jun Cheng received Bachelor of Engineering, Bachelor of Finance and Master of Engineering from the University of Science and Technology of China in 1999, 2002 respectively. His PhD degree was awarded at the Chinese University of Hong Kong in 2006. Currently he is with the Shenzhen Institutes of Advanced Integration Technology, Chinese Academy of Sciences, as a Professor and Director of the Laboratory for Human Machine Control. His research interests include computer vision, robotics, machine intelligence, and control.

Xinyu Wu is now an associate professor at Shenzhen Institutes of Advanced Technology, and an associate director of Center for Intelligent and Biomimetic systems. He received his BE and ME degrees from the Department of Automation, University of Science and Technology of China in 2001 and 2004, respectively. His PhD degree was awarded at the Chinese University of Hong Kong in 2008. He has published nearly 50 papers and a monograph. His research interests include computer vision, robotics, and intelligent system.

Yen-Lun Chen received the BS and MS degrees from the Department of Electrical Engineering at National Taiwan University, Taipei, Taiwan, and the PhD degree in Electrical and Computer Engineering from the Ohio State University, Columbus, Ohio, USA. Currently she is an Assistant Research Fellow at Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. Her research interests include machine learning, pattern recognition, computer vision, and their applications in the area of robotics and multimedia signal processing.

Yongsheng Ou is a Professor at Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. He holds a BSc degree in Mechanical and Electrical Engineering (Beijing University of Aeronautics and Astronautics, 1995) and an MSc degree in Electrical Engineering (Institute of Automation, Chinese Academy of Sciences, 1998). Ou received a PhD degree in Automation and Computer-Aided Engineering from the Chinese University of Hong Kong in 2004. He is a coauthor of the monograph on Control of Single Wheel Robots (Springer, 2005). His research interests include learning control by demonstration, computer vision, control of biped robots and control of distributed parameter systems with applications to fusion reactors.

Yangsheng Xu is Professor of Automation and Computer-aided Engineering at the Chinese University of Hong Kong where he has worked since 1997. He received BSE and MSE from Zhejiang University in China, and PhD from the University of Pennsylvania in US in the area of Robotics. His current research interests are in the areas of robotics, intelligent systems, and electric vehicles.