Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2014, Article ID 719810, 12 pages
http://dx.doi.org/10.1155/2014/719810

Research Article

Crowd Event Perception Based on Spatiotemporal Weber Field

Zhou Su,[1] Hua Wei,[2] and Sha Wei[3]

[1] *Department of Computer Science, Tsinghua University, Beijing 100086, China*
[2] *Baosight Incorporated Company, Chengdu 610000, China*
[3] *Department of Electronic Engineering, Shanghai Jiaotong University, Shanghai 200241, China*

Correspondence should be addressed to Zhou Su; suhmily@gmail.com

Received 18 September 2013; Accepted 7 November 2013; Published 28 January 2014

Academic Editor: Hang Su

Over the past decade, a wide attention has been paid to the crowd control and management in intelligent video surveillance area. Among the tasks of automatic video-based crowd management, crowd motion modeling is recognized as one of the most critical components, since it lays a crucial foundation for numerous subsequent analyses. However, it still encounters many unsolved challenges due to occlusions among pedestrians, complicated motion patterns in crowded scenarios, and so forth. Addressing these issues, we propose a novel spatiotemporal Weber field, which integrates both appearance characteristics and stimulus of crowd motion patterns, to recognize the large-scale crowd event. On the one hand, crowd motion is recognized as variations of spatiotemporal signal, and we then measure the variation based on Weber law. The result is referred to as spatiotemporal Weber variation feature. On the other hand, motivated by the achievements in crowd dynamics that crowd motion has a close relationship with interaction force, we propose a spatiotemporal Weber force feature to exploit the stimulus of crowd behaviors. Finally, we utilize the latent Dirichlet allocation model to establish the relationship between crowd events and crowd motion patterns. Experiments on PETS2009 and UMN databases demonstrate that our proposed method outperforms the previous methods for the large-scale crowd behavior perception.

1. Introduction

Over the past decades, crowd phenomenon has become an important carrier of economic development and culture exchange along with the steady population growth and worldwide urbanization. Meanwhile, the difficulty in crowd managing is improving rapidly with crowd scale increasing. Examples of large-scale crowd are shown in Figure 1. According to the statistics of *The Guardian*, there have happened more than fifteen fatal crowd accidents that resulted in high casualties within the past twenty years when people lost control during the crowded special events, for example, the stampede in the Cambodia Water Festival and the Love Parade stampede in Germany. Obviously, such terrible event is much easier to control, if we get aware of the abnormal clues and nip the tragedy in the bud before it gets serious. Nowadays, massive applications of the video surveillance system provide a possibility to manage the large-scale crowd and prevent such unfortunate events. Therefore, the perception of the large-scale crowd event has attracted the attention from technical research discipline, especially for the anomalous behaviors in the crowd activities where computer vision algorithms play a growing role.

1.1. Related Work. Vision-based crowd behavior perception is thoroughly studied during the past few years which are categorized into two main philosophies. The first category implements crowd behavior understanding in the appearance perspective. Some works can be regarded as microscopic model, which deals with the crowd as a collection of discrete individuals and extends the methods designed for the individual behavior perception to crowd analysis. The representative works include the methods proposed by Jacques et al. [1] and Pellegrini et al. [2], in which the Voronoi diagram and linear trajectory avoidance are utilized to recognize crowd behavior, respectively. Obviously, it is essential for these algorithms to segment or detect individuals and track isolated pedestrians. A lot of related works have been done on these issues, including the pedestrian detection [3] and crowd tracking [4–7]. However, these methods always perform

(a) (b) (c)

Figure 1: Examples of the large-scale crowd.

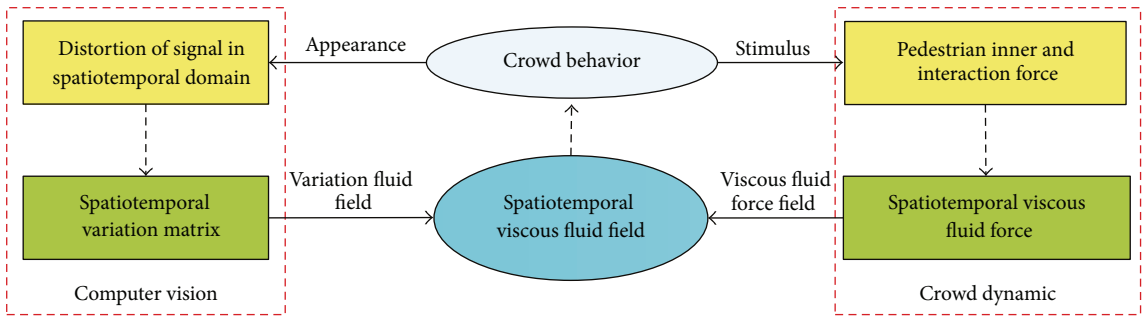

Figure 2: Theoretical framework of crowd behavior perception.

a significant degradation when the density of people in the scene increases. After all, in most cases, it is not necessary to track each pedestrian for crowd event perception, which is time consuming. Another is *macroscopic* model, which deals with the crowd as an entity, without the need to segment each individual. To the best of the authors' knowledge, most of these algorithms capture the low-level motion pattern of crowd behavior via optical flow [8–10], spatiotemporal motion pattern [11–14], or combining both of them [15]. However, most of these methods are sensitive to noise, for example, illumination change, or have a good possibility to fail when crowd motion has sudden changes, which always indicates an irregular crowd flow and is crucial in crowd analysis. An alternative method is to model the crowd motion as a time dependent flow flied [16, 17], since it is observed that pedestrian crowds behave with some striking analogies with the motion of fluids [18, 19]; for example, the footprints of pedestrian crowd in the snow look similar to streamlines of fluids. Although the fluid dynamic model has achieved a success in crowd simulation [20, 21], it still faces a difficulty in estimating velocity of pedestrians for crowd behavior analysis.

Recently, another category of algorithms becomes popular, which analyzes the stimuli, or driven factor, of crowd behavior. These methods lie on an assumption that crowd behavior originates from the interaction of its elementary individuals. Helbing has proposed a social force model [22, 23] to investigate the crowd motion dynamics. According to his work, the crowd behavior is stimulated by a social force field, and the pedestrians will react to energy potentials caused by other pedestrians and static obstacles through a repulsive force. Mehran et al. [24] and Raghavendra et al. [25] proposed to detect the abnormal event of the crowd. However, the model is designed and used for simulation purposes and overall it is a microscopic model. For the application of crowd analysis, it faces the difficulties in estimating the velocity of the pedestrians, especially for the large-scale crowd. Both of the algorithms approximate the *desired force* by averaging the optical flow around the pedestrian, which is not applicable to all cases, since the desired motion is very subjective. Moreover, social force concept aims to model the interaction force between pairwise individuals and thus is inappropriate for dense crowd flow. Nevertheless, it provides us an inspiring perspective to analyze the crowd behavior according to the force model.

1.2. Our Proposal. Motivated by the previous work, we proposed a spatiotemporal Weber field which integrates both the appearance and stimulus of crowd behavior. The theoretical framework of the paper is illustrated in Figure 2. In appearance perspective, the moving crowd will disturb the distribution of the background and cause a fluctuation of the signal in both spatial and temporal domains. It motivated us to measure the variation based on which crowd motion pattern can thus be explored. Weber law is an important achievement in psychophysics, which attempts to describe the relationship between the physical magnitudes of stimuli

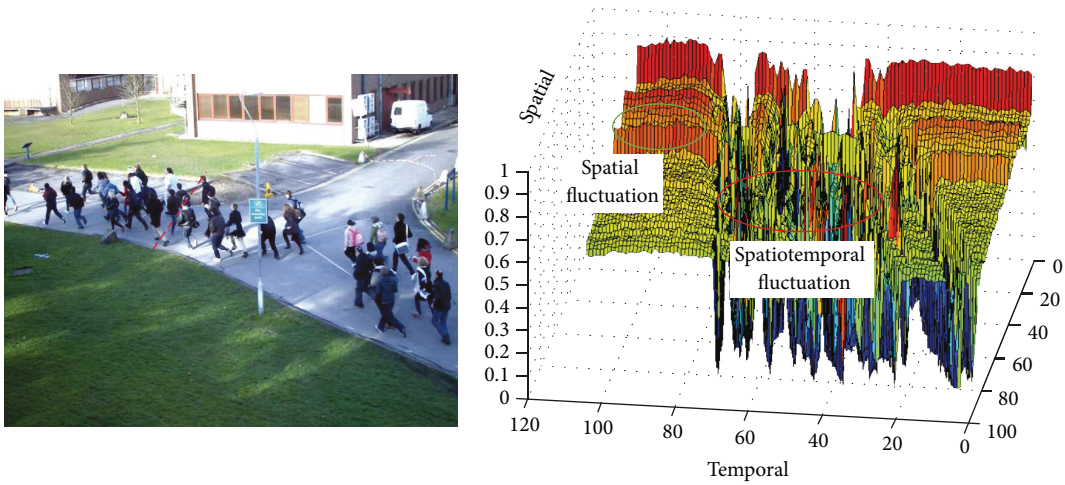

Figure 3: Fluctuation of video signal in spatiotemporal domain. (a) Original crowd image, in which we marked a cross line with red dots to investigate the fluctuation. (b) Fluctuations at the corresponding cross line in spatiotemporal domains, where the green circle indicates variations at background pixels in spatial domain and the red circles indicated fluctuations for moving crowd in both spatial and temporal domains.

and the perceived intensity of the stimuli [26]. The Weber law points out that the threshold that human could discriminate does not only depend on the absolute variation but also on the original stimulus. It has been proved to be an effective tool to measure the variation of the stimulus and has been used as a image texture recognition [27]. In this paper, we propose a *Spatiotemporal Weber Variation Feature* (ST-WVF) for the crowd behavior perception, which adopts the Weber law to measure the variation of the video signal.

According to the research in crowd dynamic, crowd motion is always driven by the stimulus of force. In this paper, we propose a potential function and analyze the intensity of the force field for the Weber variation feature, referred to as *Spatiotemporal Weber Force Feature* (ST-WFF), to explore the stimulus of crowd behavior. The previous spatiotemporal Weber variation and force features thus compose the spatiotemporal Weber field. The contributions of this paper are as follows.

(i) Firstly, we propose a spatiotemporal Weber variation feature to estimate the variation of the video signal and explore crowd motion pattern.

(ii) Secondly, we propose a spatiotemporal Weber force feature, which explores the stimulus of crowd behavior.

(iii) Finally, we propose a crowd behavior perception system, which integrates the previous spatiotemporal Weber variation and force feature with location information and utilizes the latent Dirichlet allocation model to analyze it. Experiments on the UMN database and PETS2009 show that the proposed method can achieve a more desirable result than the conventional methods.

The remainder of this paper is organized as follows. The structure of the system is presented in Section 2. Then the details of the proposed algorithm are discussed in Section 3, including the methods to extract spatiotemporal Weber variation and force feature. In Section 4, we utilize the latent Dirichlet allocation (LDA) model to realize the crowd behavior perception, incorporating with bag of feature algorithm. Experimental results of our system are shown in Section 5. Finally, the conclusion is made in Section 6.

2. System Architecture

In this section, we establish an intelligent vision system that is capable of modeling the large-scale crowd motion pattern and recognizing the crowd behavior. Specifically, motion pattern of crowd behavior is explored in terms of both appearance and stimulus.

In appearance perspective, the moving crowd will disturb the distribution of background and cause a fluctuation of the signal in both spatial and temporal domains. Figure 3 shows a sample of fluctuation resulting from moving crowd in both spatial and temporal domains. Specifically, Figure 3(a) is selected from one sequence of PETS2009, in which the pedestrians enter the scene from the right boundary, run on the paved road, and exit the scene on the left. We marked a red dot line on the pave to demonstrate spatiotemporal fluctuations when pedestrians pass it, and we show the fluctuations in Figure 3(b). The green circle indicates the fluctuation at red dot line on the paved road from the 80th frame to the 120th frame. As is observed in green circle in Figure 3(b), the fluctuation along the temporal direction is small, since all pedestrians have passed the red dot line after the 80th frame and the fluctuation thus resulted from spatial variation in background, for example, strong edges. On the other hand, the fluctuation in the red circle is corresponding to the variation from the 30th frame to the 80th frame, when the pedestrians passed the red dot line.

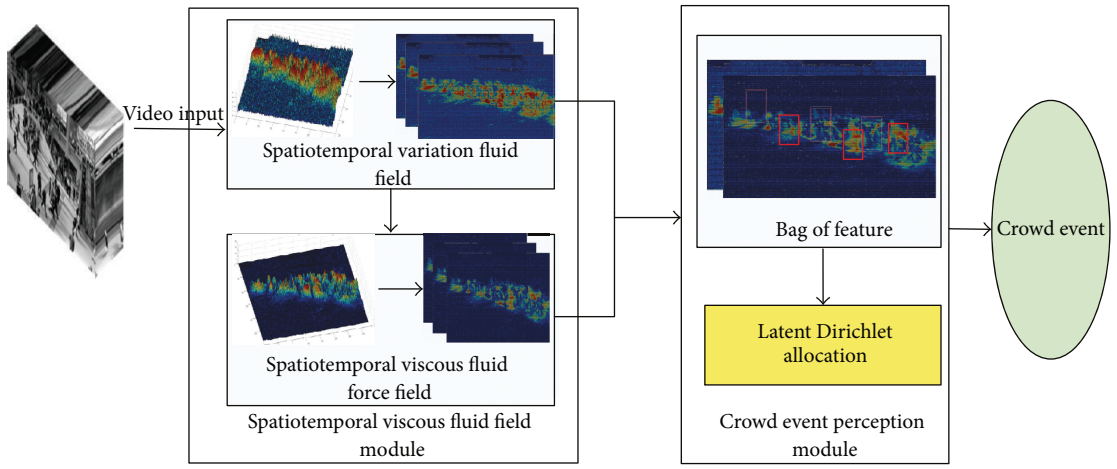

Figure 4: System structure of crowd behavior perception.

The fluctuation thus resulted from the variation in both spatial and temporal domains. From the result, we can observe that the background regions, that is, static pixels, always exhibit uniform distribution with little variation in temporal domain (indicated with the green circle in Figure 3(b)), and regions of moving crowd present drastic fluctuations in spatiotemporal domain (indicated with the red circle in Figure 3(b)).

In stimulus perspective, interaction force in this paper is estimated by investigating the distinctiveness of local motion patterns; that is, the interaction force is relatively small if the pedestrians' motion is identical and large if their motion patterns are distinctive. Force vectors of each pixel construct a *spatiotemporal Weber force field*.

The schematic diagram of the proposed system is shown in Figure 4, which consists of feature extraction and feature analysis modules. In this paper, we focus on the module to extract spatiotemporal Weber variation and force feature, which is referred to as spatiotemporal Weber field. For a specific location of input video, we construct a cylindroid spatiotemporal volume with the corresponding pixel as its center. Afterwards, we propose a spatiotemporal Weber variation feature, which adopts the Weber law to measure the variation within the volume. After that, a spatiotemporal Weber force feature is proposed to explore the stimulus of crowd behavior by analyzing the properties of the Weber variation feature. The bag of feature (BOF) algorithm is utilized to estimate the likelihood of the spatiotemporal Weber field. Finally, we utilize the latent Dirichlet allocation model to recognize the crowd behavior.

The proposed spatiotemporal Weber field reveals the crowd motion pattern from the appearance and stimulus aspects, respectively. Furthermore, the abnormal event, which is essentially an eccentric state of the crowd motion, can be regarded as the irregular change of the signal, either in spatial or temporal domain. Meanwhile, there usually exists anomalous stimulus for the abnormal event. Compared with the conventional system for crowd behavior analysis, the system proposed in this paper benefits from the following characteristics.

(i) Firstly, the system extracts the feature directly in the spatiotemporal domain. In this case, it does not depend on individual detecting or tracking, and there is no need for background modeling, which is too complicated to implement in heavily crowded scene. Therefore, the system is more suitable and practical for large-scale crowd analysis.

(ii) Secondly, the system models the crowd behavior as a variation of signal in spatiotemporal domain and measures the change with the Weber law, which has been proved effectively in psychophysics.

(iii) Thirdly, the system explores the stimulus for crowd behavior not only from the appearance of the behavior, which reveals essential motion characteristics.

(iv) Finally, the system utilizes latent Dirichlet allocation to recognize the crowd behavior model, and crowd behavior is recognized effectively.

We detail the proposed algorithm in the following sections.

3. Spatiotemporal Weber Field

3.1. Overview of the Weber Law. Weber law was first proposed in the nineteenth century by the German physiologist Weber and was later formulated quantitatively by Fechner [28], founder of the modern psychophysics. The law reveals that the threshold of a just noticeable difference is a constant proportion of the original stimulus value. This relationship, known since as Weber law, can be expressed as

$$\Delta I = \tau_w I, \tag{1}$$

where ΔI denotes the increment threshold or the just noticeable difference for discrimination, I denotes the initial stimulus intensity, and τ_w represents the Weber fraction. The Weber

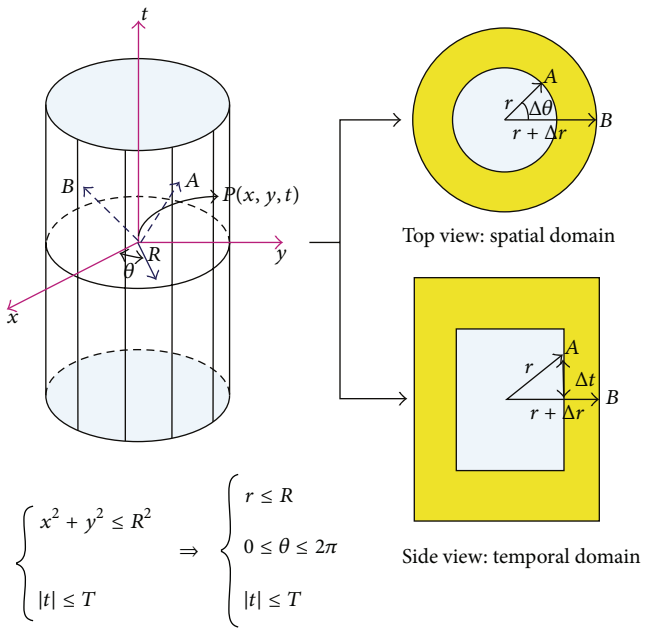

FIGURE 5: Schematic for the calculation of spatiotemporal Weber variation feature. $P(x, y, t)$ is the center of the cylindroid spatiotemporal volume; $A : P_A(r, \theta, t)$ and $B : P_B(r + \Delta r, \theta + \Delta\theta, t + \Delta t)$ are two arbitrary pixels in the volume.

law reveals that the difference threshold for discrimination does not only depend on the absolute variation but more essentially on the relative change. It interprets a law of human beings' cognition and provides a measurement of the variation of the stimulus, which is important for video analysis.

3.2. The Spatiotemporal Weber Variation Feature. Generally speaking, the moving crowd will disturb the distribution of the background and cause a variation both in spatial and temporal domains. Therefore, the behavior can be regarded as a distortion of the signal over the spatiotemporal domain. The Weber law provides an effective way to measure it. For a specific location $P(x, y, t)$ in the image sequence, we construct a cylindroid spatiotemporal volume with radius R and height $2T$ with $P(x, y, t)$ as its center. The volume is denoted as $V(x, y, t)$, as is shown in Figure 5. The analytical expression of the volume is shown in the following:

$$x^2 + y^2 \le R^2, \quad |t| \le T, \tag{2}$$

where R denotes the maximum distance in the spatial domain away from $P(x, y, t)$ and T denotes the largest margin in temporal domain. For the sake of analysis, the expression is transformed into cylindrical coordinates, and the expression is shown as

$$r \le R^2, \quad 0 \le \theta \le 2\pi, \quad |t| \le T. \tag{3}$$

As is shown in Figure 5, $A : P_A(r, \theta, t)$ and $B : P_B(r + \Delta r, \theta + \Delta\theta, t + \Delta t)$ are two arbitrary pixels in the spatiotemporal volume and the corresponding intensity of the pixels is $I_A(r, \theta, t)$ and $I_B(r + \Delta r, \theta + \Delta\theta, t + \Delta t)$. We will calculate the difference of the intensity between the two pixels and integrate the difference with all r in the volume as

$$f(x, y, t) = \int_0^{R-\Delta r} \left(I_B(r + \Delta r, \theta + \Delta\theta, \ldots, t + \Delta t) - I_A(r, \theta, t)\right) dr. \tag{4}$$

In order to exploit the properties of the variation for the signal comprehensively, the parameters Δt and $\Delta\theta$ are changed, with the expression shown in the following:

$$f_{mn}(x, y, t) = \int_0^{R-\Delta r} \left(I_B(r + \Delta r, \theta + m \cdot \Delta\theta, \ldots, t + n \cdot \Delta t) - I_A(r, \theta, t)\right) dr, \tag{5}$$

where $m, n = 1, 2, \ldots, K$. The features with all f_{mn} construct a matrix, as the following shows:

$$F_K = \begin{bmatrix} f_{11} & f_{12} & \cdots & \cdots & f_{1K} \\ f_{21} & f_{22} & \cdots & \cdots & f_{2K} \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ f_{m1} & \cdots & f_{mn} & \cdots & f_{mK} \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ f_{K1} & f_{K2} & \cdots & \cdots & f_{KK} \end{bmatrix}_{K \times K}. \tag{6}$$

The determinant of the feature matrix could reflect the variations in the volume, which is denoted as $\det(F_K)$. If the effective variation or the increment threshold is denoted as f_0,

$$f_0 = \tau_w \cdot I(x, y, t), \tag{7}$$

where τ_w is the Weber fraction and $I(x, y, t)$ is the intensity of the corresponding pixel $P(x, y, t)$. The Weber variation feature for $P(x, y, t)$, which reveals the appearance property of the behavior, could be calculated as the following shows:

$$f_{\text{var}}(x, y, t) = \frac{\det(F_K)}{I(x, y, t)} = \frac{\tau_w \cdot \det(F_K)}{f_0}, \tag{8}$$

where τ_w is the Weber fraction and $\det(F_K)$ is the determinant of the feature matrix F_K.

We illustrate a sample result of the spatiotemporal Weber variation field in Figure 6. The left and right columns are corresponding to the results for PETS2009 and UMN dataset, respectively. The colors pass through blue, green, and red with the amplitude increasing. As is shown in Figures 6(b1), 6(c1), 6(b2), and 6(c2), the positions of walking crowd always have much larger variations in spatiotemporal domain and present red. Oppositely, the positions of background, with smaller variations, always present blue. The motion is indicated distinctly as the simulation results show. Therefore, it can be inferred that the spatiotemporal variation fluid field reflects the crowd motion pattern. Particularly speaking, the abnormal behavior, which is essentially an eccentric state of crowd, is regarded as irregular variations in sequence, either in spatial or temporal domain.

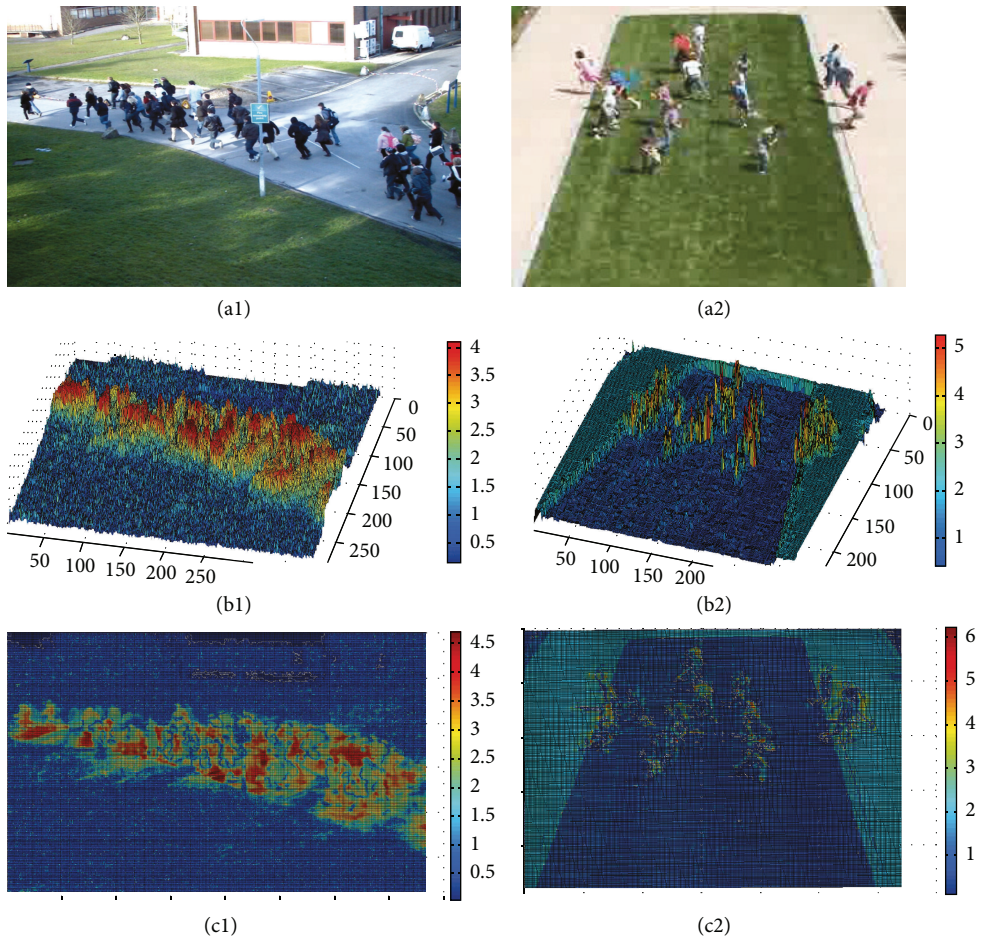

FIGURE 6: Results of spatiotemporal Weber variation feature. (a1) and (a2) are the original images in the sequences. (b1) and (b2) are the corresponding results of the spatiotemporal Weber variation feature. (c1) and (c2) are the top view of spatiotemporal Weber variation feature results.

3.3. The Spatiotemporal Weber Force Feature. The crowd behavior is driven by the social force, including the desired force and interaction force [29]. However, it is still a challenging task to estimate the interaction force. In [29], the interaction force is estimated following the potential between the gas molecules, but it is obviously not applicable for crowd with medium or high density. Mehran et al. [24] calculated the interaction force by subtracting the personal desire force from the acceleration force. However, the work lies on an assumption that the desired velocity of a pedestrian is the average of neighboring optical flow, which is not applicable to all cases, since the desired velocity is quite subjective.

It is observed that the motion of crowd, especially for crowd with medium or high density, presents some striking analogies with fluids [30]. Moreover, it is easily understood that the interaction force between the pedestrians is closely related to motion pattern of the crowd (e.g., the friction between a pedestrian walking in opposite direction among crowds). It is reasonable that the interaction force is highly related to the relative force. In other words, the force is relatively small if the pedestrians' motion is identical, while the force is large if the motion pattern is distinct. Therefore, the interaction force has analogous properties with electromagnetic field. In this paper, the crowd motion pattern is modeled as electromagnetic field lines, and the interaction force, which is driving factor of crowd behavior, is then estimated with the interaction force in the electromagnetic field.

The result of the spatiotemporal Weber variation feature is denoted as $W(x, y, t)$, as the following shows:

$$W(x, y, t) = \{f_{\text{weber}}(x, y, t)\}, \tag{9}$$

where $W(x, y, t)$ is a scalar field constructed by the Weber variation feature of each pixel. According to the related theory of electromagnetic field, the intensity, which reflects the property of the force, has a close relation with the potential of the field, as the following shows:

$$\vec{E} = -\nabla U = -\left(\frac{\partial U}{\partial x}\vec{i} + \frac{\partial U}{\partial y}\vec{j} + \frac{\partial U}{\partial z}\vec{k}\right), \tag{10}$$

where $\vec{E}$ denotes the intensity of the field and U denotes the potential of the field. The equation reveals that the more dense

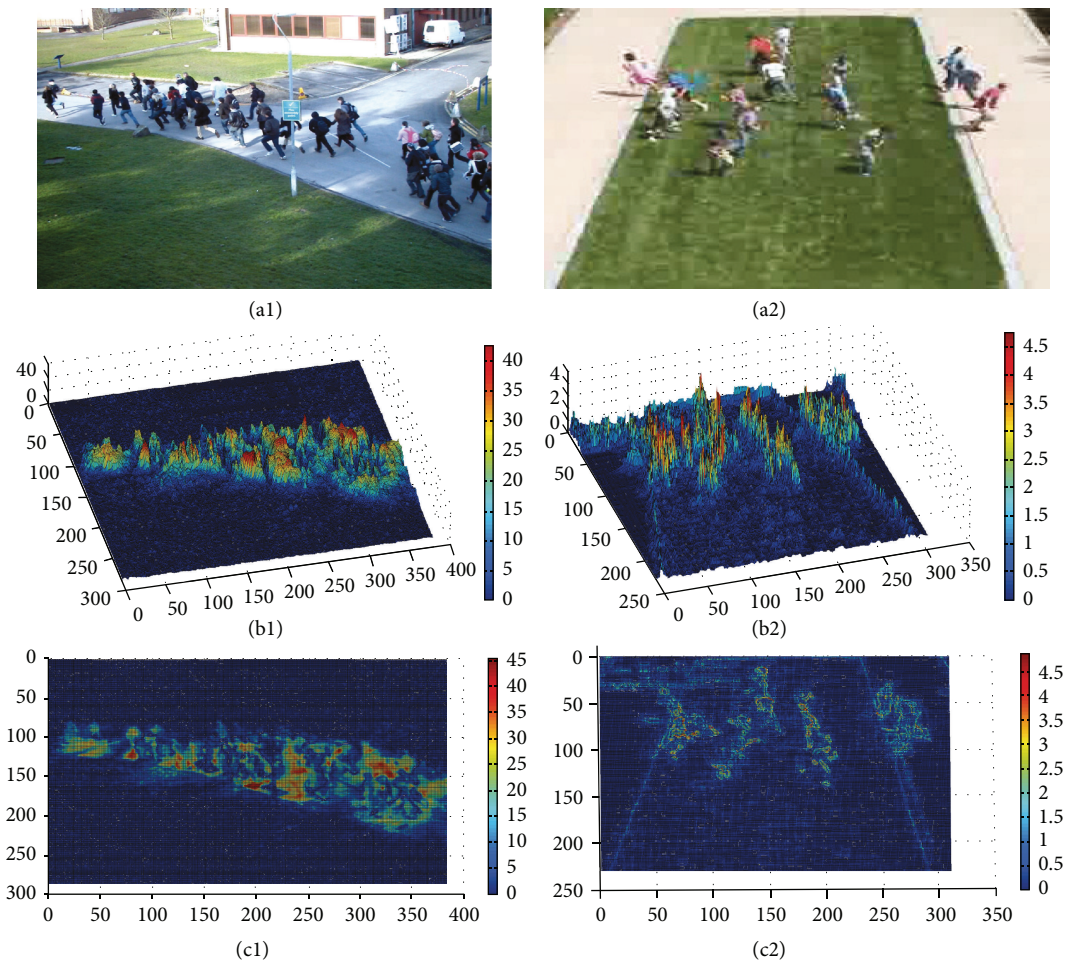

FIGURE 7: Results of spatiotemporal Weber variation feature. (a1) and (a2) are the original images in the sequences. (b1) and (b2) are the corresponding results of the spatiotemporal Weber force feature. (c1) and (c2) are the top view of spatiotemporal Weber force feature.

the isopotential line, the stronger the field intensity. In this paper, the intensity of the Weber variation feature field can be calculated as

$$\vec{E}(x,y,t) = -\nabla U = -\left(\frac{\partial W}{\partial x}\vec{i} + \frac{\partial W}{\partial y}\vec{j} + \frac{\partial W}{\partial t}\vec{k}\right), \quad (11)$$

where $\vec{E}$ denotes the intensity of the Weber variation feature field $W(x, y, t)$. In this paper, we concentrate on the amplitude of the intensity and the force field can be calculated as the following shows:

$$F(x,y,t) = \left|\vec{E}\right| = \sqrt{\left(\frac{\partial W}{\partial x}\right)^2 + \left(\frac{\partial W}{\partial y}\right)^2 + \left(\frac{\partial W}{\partial t}\right)^2}. \quad (12)$$

The force field is the driving factor of the variation feature and reveals the stimulus of the crowd behavior. The simulation result of force field is shown in Figure 7. The result is also described in JET color-map, ranging from the colors blue, green to red. Different from the previous figures, the value of the force is constantly positive in this paper, so the blue indicates the positions with smaller amplitude and with the amplitude increasing, the color is shifting to red. As the results show, the background or the positions with little variation of motion present blue color and indicate smaller amplitude in force field. However, the positions with significant movement variation present red and with a larger amplitude force field, which indicate a greater possibility for the change of motion pattern or the evidence of irregular behavior.

4. Crowd Event Perception Based on the Spatiotemporal Weber Field

The spatiotemporal Weber variation and force features provide a picture of the local crowd activities, but the features do not capture the relationship between their occurrences. In other words, the discrete values are not a clear evidence of crowd behavior or abnormal event, and the crowd behavior cannot be recognized robust merely depending on the approach in the previous sections. Therefore, we adopt the latent Dirichlet allocation (LDA) model to establish the relation between the proposed features and the crowd behavior, which is proved effectively in video activity perception [24, 31].

4.1. Spatiotemporal Weber Word Generation. LDA is a hierarchical Bayesian model, which has gained great success in language processing [32]. To use LDA, we partition a video sequence into blocks with size $X \times Y \times Z$, and the properties of each block are treated as words for word-document analysis. We aim to infer the distribution of word cooccurrence and thus recognize the crowd behavior.

By combining the spatiotemporal variation feature in (9) and the corresponding force feature in (12) together, we get a *spatiotemporal Weber* feature $\boldsymbol{\Psi}(x, y, t)$, which is denoted by

$$\boldsymbol{\Psi} \triangleq (W, F), \tag{13}$$

where W and F are the spatiotemporal variation and force feature, respectively.

Furthermore, we utilize a multivariate normal distribution to model all the feature vectors $\boldsymbol{\Psi}^i(x, y, t)$ in block Blk_i:

$$\boldsymbol{\Psi}^i \sim \mathcal{N}\left(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i\right), \tag{14}$$

where $\boldsymbol{\mu}$ is the J-dimensional mean vector $\boldsymbol{\mu} = (\mu_W, \mu_F)^T$ with μ_W and μ_F being the mean of the variation and force feature, respectively; and $\boldsymbol{\Sigma}^i$ is a 2×2 covariance matrix $\boldsymbol{\Sigma}^i = \text{diag}(\sigma_W^2, \sigma_F^2)$, because we assume that each dimension of $\boldsymbol{\Psi}^i$ is independent. Furthermore, $\boldsymbol{\mu}^i$ and variation $\boldsymbol{\Sigma}^i$ can be calculated as

$$\boldsymbol{\mu}^i = \frac{1}{XYZ}\sum_{t=1}^{Z}\sum_{y=1}^{Y}\sum_{x=1}^{X}\boldsymbol{\Psi}^i(x, y, t),$$

$$\boldsymbol{\Sigma}^i = \frac{1}{XYZ}\sum_{t=1}^{Z}\sum_{y=1}^{Y}\sum_{x=1}^{X}\left[\boldsymbol{\Psi}^i(x, y, t) - \boldsymbol{\mu}^i\right]\left[\boldsymbol{\Psi}^i(x, y, t) - \boldsymbol{\mu}^i\right]^T. \tag{15}$$

Afterwards, we use the BoF method [33] to construct a codebook including K visual words. The descriptors $\{(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i)\}$ for block $\{\text{Blk}^i\}$ are partitioned into K clusters by minimizing a cost function, which is a sum of pairwise distance between the descriptors within the same cluster c_k:

$$D\left(\boldsymbol{\Psi}; c_k\right) = \sum_{k=1}^{K}\sum_{\boldsymbol{\Psi}^i, \boldsymbol{\Psi}^j \in c_k}\text{dist}\left(\boldsymbol{\Psi}^i, \boldsymbol{\Psi}^j\right). \tag{16}$$

In this paper, we employ the Bhattacharyya distance [34] to measure the distance between $\boldsymbol{\Psi}^i$ and $\boldsymbol{\Psi}^j$:

$$\text{dist}\left(\boldsymbol{\Psi}^i, \boldsymbol{\Psi}^j\right) = \frac{1}{8}\left(\boldsymbol{\mu}^i - \boldsymbol{\mu}^j\right)^T\left(\frac{\boldsymbol{\Sigma}^i + \boldsymbol{\Sigma}^j}{2}\right)^{-1}\left(\boldsymbol{\mu}^i - \boldsymbol{\mu}^j\right) + \frac{1}{2}\ln\left(\frac{\det\left(\boldsymbol{\Sigma}^i + \boldsymbol{\Sigma}^j\right)}{2\sqrt{\det\boldsymbol{\Sigma}^i \cdot \det\boldsymbol{\Sigma}^j}}\right), \tag{17}$$

where $(\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i)$ and $(\boldsymbol{\mu}^j, \boldsymbol{\Sigma}^j)$ are the normal distribution parameters for $\boldsymbol{\Psi}^i$ and $\boldsymbol{\Psi}^i$, respectively. Note that the first term is related to the distance of the mean vector, and the second term takes into account the variance distance.

Furthermore, the visual word w_k is the center of cluster c_k, and we thus construct a codebook $\mathscr{C} \doteq \{w_1, w_2, \ldots, w_k, \ldots, w_K\}$. The features $\boldsymbol{\Psi}^i$ are then quantized to w^* with the codebook:

$$w^* = \arg\min_{w_k \in \mathscr{C}}\left\|\boldsymbol{\mu}^i - w_k\right\|, \tag{18}$$

where $\boldsymbol{\mu}^i$ is the mean vector of $\boldsymbol{\Psi}^i$.

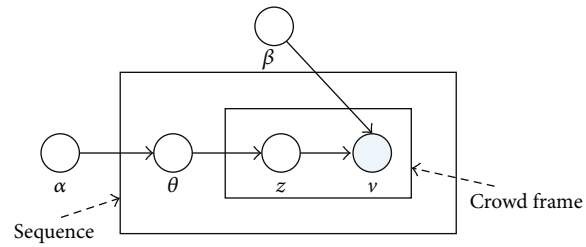

Figure 8: LDA probabilistic graphical representation.

4.2. Crowd Behavior Recognition Based on LDA. We then learn the distribution of cooccurrence for visual words with LDA model and infer the crowd behavior of video. The basic idea of LDA model is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [32]. In this case, the inference problem for a document is to compute the posterior distribution given the hidden variables. In this paper, the whole video sequence is treated as a corpus; the sequence is uniformly divided into nonoverlapping short clips $\mathbf{v}$, and the video clips are treated as documents. Each document is modeled as a mixture of N topics $\mathbf{z}$, with a joint distribution parameter $\boldsymbol{\theta}$. Moreover, the corpus has two Dirichlet prior parameters: $\boldsymbol{\alpha}$ determines the per-document topic distributions, and $\boldsymbol{\beta}$ is corresponding to the topic-word distribution. The corpus-level parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated by maximizing the marginal log-likelihood of the data during the learning procedure. The probabilistic graphical representation is illustrated in Figure 8.

Based on these descriptions, we calculate the posterior distribution for each video clip:

$$p\left(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \frac{p\left(\boldsymbol{\theta}, \mathbf{z}, \mathbf{v} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}\right)}{p\left(\mathbf{v} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}\right)}. \tag{19}$$

A video clip is then classified into various crowd behaviors based on its particular topic mixture $(\boldsymbol{\theta}^*, \mathbf{z}^*)$ by variational Bayes inference algorithm [32]. Crowd behavior recognition is implemented by calculating the pairwise similarity between the video clip and the training data, which is measured by JS divergence [35].

5. Experiments and Discussion

Data and Parameters. In this paper, the approach is tested on some publicly available datasets of crowd videos. Seq. 1 are the videos from PETS2009, which contain videos of crowd walking, evacuation (rapid dispersion), local dispersion, and crowd gathering/splitting. We select more than 5000

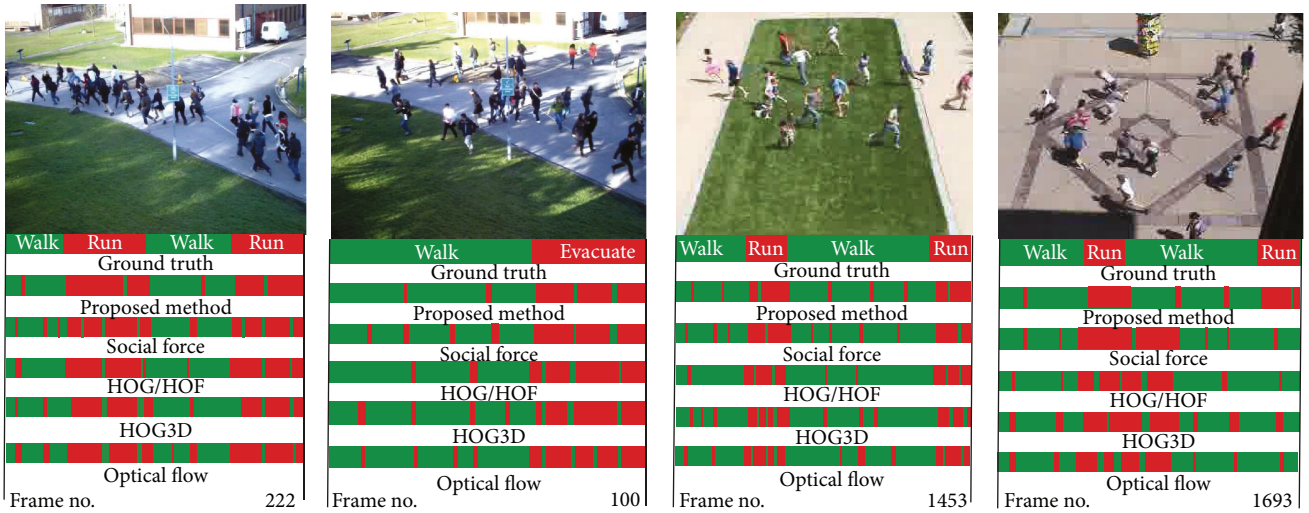

FIGURE 9: The qualitative results of the crowd behavior perception for four sample videos. (a) are sample results from Seq. 1, and (b) are sample results from Seq. 2. The bars are utilized to represent the labels of each frame for the videos in the sample video, and different colors represent the different corresponding crowd behaviors. The bars indicate the ground truth, the detection results based on the proposed spatiotemporal Weber field, social force model [24], HOG/HOF [36], HOG3D [37], and pure optical flow [8], respectively.

frames from 30 sequences in the dataset and manually label every frame as ground truth. Seq. 2 are crowd videos from University of Minnesota (UMN), which contain videos of 11 different scenarios of escape events. The videos are shot in 3 different scenes, including both indoor and outdoor. We randomly select 40% frames for each scenario for the parameter optimizing during the learning procedure and use the other frames for testing.

In order to explore the variation properties of the videos, the parameters in (5) are set to $M = P = N = 3$. For construction of visual words, videos are partitioned into blocks with size $8 \times 8 \times 5$, and $K = 64$ visual words are extracted from the properties of the blocks with BoF method to construct the codebook. In LDA model, we use $L = 32$ latent topics for learning and recognition. We adopt the proposed method to identify the crowd event as well as the start and end of the event. Furthermore, the results are compared with the ground truth, the detection results based on social force model [24], HOG/HOF [36], HOG3D [37], and optical flow [8]. For comparison, we use the same videos and parameter setting for model learning. Visual words are extracted with the other corresponding features in the same spatiotemporal patches, and we also create a codebook from them.

Recognition Results. We demonstrate some sample recognition results in Figure 9. Color bars are utilized to represent the labels of each frame for videos in the sequence, and different crowd behaviors are indicated with different colors. As a comparison, results based on social force, HOG/HOF, HOG3D, and optical flow are also shown in Figure 9. It is observed that our proposed method outperforms the conventional methods, because we exploit both the appearance and driven factor of the behavior. The social force model gains a comparable result in abnormal event detection with our proposed method, such as crowd splitting and evacuation. However, it results in some false detection for normal crowd behavior recognition, for example, crowd walking, because the social force is not obvious in such cases. HOG/HOF has balanced results in both normal and abnormal events perception, but it faces a difficulty if the motion is not obvious such as crowd gathering. The performance of HOG3D is not as desirable as the previous ones, because the spatial feature is "overweight" for this descriptor, and the spatiotemporal volume construction for HOG3D analysis also leads to a *lag* in detection. Optical flow fails to exploit the spatial characteristic of the behavior, and the performance is also inferior to our proposed method. Overall, these results demonstrate that crowd behavior can be recognized effectively and accurately based on our proposed spatiotemporal Weber field, because it exploits both the appearance and driven factor of the behavior.

Quantitative Evaluations. In order to evaluate our method quantitatively, we denote the behaviors, which are labeled red in Figure 9, as positive events, that is, run, gather, split, evacuate, and so forth. The positive events are usually more important in practical crowd analysis. The behaviors, which are labeled as green in Figure 9, are denoted as negative events, that is, walk, wait, and so forth.

We measured performance of each algorithm in terms of precision, recall, and F_{score}:

$$\begin{aligned} \text{precision} &= \frac{\text{tp}}{\text{tp} + \text{fp}}, \\ \text{recall} &= \frac{\text{tp}}{\text{tp} + \text{fn}}, \\ F_{\text{score}} &= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \end{aligned} \tag{20}$$

Table 1: Comparison of precision and recall for different algorithms.

Index	Run (PETS209)		Run (UMN)		Gather		Split		Evacuation		Opposite flow	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Proposed method	0.887	0.902	0.847	0.902	0.825	0.883	0.912	0.882	0.903	0.878	0.879	0.854
Social force [24]	0.812	0.887	0.792	0.891	0.767	0.895	0.807	0.879	0.781	0.882	0.817	0.832
HOG/HOF [36]	0.892	0.919	0.832	0.878	0.817	0.836	0.923	0.813	0.915	0.867	0.833	0.781
HOG3D [37]	0.803	0.786	0.812	0.793	0.803	0.784	0.827	0.832	0.831	0.852	0.821	0.771
Optical flow [8]	0.836	0.868	0.803	0.822	0.809	0.812	0.903	0.792	0.822	0.841	0.812	0.769

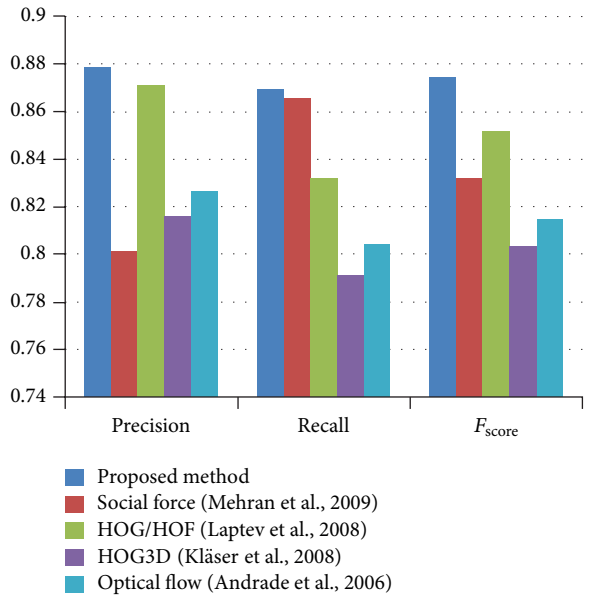

Figure 10: Comparison of different algorithms by averaging precision, recall, and F_{score} for all crowd behaviors.

where tp is the true positive, or correct detection for positive events, fp is the false positive, or error detection of positive events, and fn is the false negative, or missing detection of positive events.

The results of criteria for each crowd behavior are illustrated in Table 1. The results show that, compared with social force model, our proposed model has a much better performance in terms of precision, since the social force is not obvious for normal behaviors. HOG/HOF has a good performance in terms of precision, but it misses some detections for positive events. The main reason is that it fails to exploit the interaction between pedestrian, which is a crucial factor such as opposite flow, and crowd gathering. HOG3D and optical flow perform inferiorly to other algorithms, because the spatial feature is too much emphasized in HOG3D, whereas the optical flow fails to utilize the spatial information. The overall evaluation for all the behaviors is demonstrated in Figure 10. It is observed that our proposed method performs good in terms of precision and recall and thus has a much better result in F_{score}.

Moreover, by changing the decision parameter of LDA, we illustrate the receiver operating characteristic (ROC) curves (true positive rate versus false positive rate) in Figure 11. Figures 11(a) and 11(b) are the results of our proposed and other comparative algorithms for Seq. 1 and Seq. 2, respectively.

We use area under ROC curve (AUC) and accuracy (ACC) as the metrics to evaluate the performance. A larger AUC indicates a better performance in robustness of recognition, and ACC, which indicates the effectiveness of the perception, is defined as

$$\mathrm{ACC} = \frac{\mathrm{tp} + \mathrm{tn}}{\mathrm{tp} + \mathrm{fp} + \mathrm{tn} + \mathrm{fn}}, \tag{21}$$

where fn is false negative, or error recognition of negative events.

The results are reported in Table 2, which demonstrates that our proposed method outperforms other algorithms in terms of both AUC and ACC. Social force model has a comparable AUC index, but the ACC is much lower. It is due to the difficulty in estimating social force for normal behavior, which leads to some error detections. Our method gains a much better performance by integrating both the appearance and driven factors of the crowd behavior, comprehensively.

Table 2: Comparison for AUC and ACC of different algorithms.

	Seq. 1		Seq. 2		Average	
	AUR	ACC	AUR	ACC	AUR	ACC
Proposed method	0.873	91.3%	0.929	89.7%	0.893	90.5%
Social force [24]	0.857	81.7%	0.912	80.2%	0.880	80.9%
HOG/HOF [36]	0.785	86.6%	0.875	84.5%	0.821	85.2%
HOG3D [37]	0.728	77.2%	0.771	73.3%	0.747	74.0%
Optical flow [8]	0.719	78.9%	0.789	76.2%	0.741	77.9%

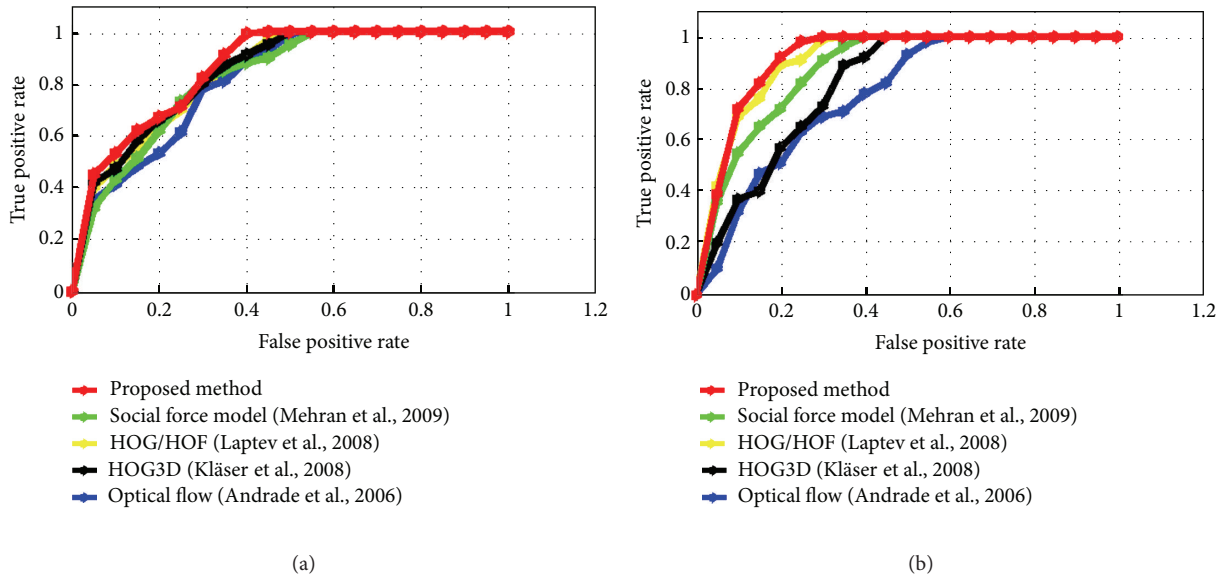

Figure 11: The ROC curves for the crowd behavior perception results based on our proposed spatiotemporal Weber field, social force model [24], HOG/HOF [36], HOG3D [37], and optical flow [8], respectively. Figure 9(a) are the results for Seq. 1; Figures 2 and 3 are the results for Seq. 2, respectively.

6. Conclusion

This paper proposes a novel spatiotemporal Weber field to recognize the large-scale crowd event, which is an attractive topic in the area of intelligent video analysis. The motion of the crowd is modeled as a variation of signal in spatiotemporal domain, and we propose a spatiotemporal Weber variation feature to measure the change, which adopts the Weber law. Afterwards, the paper proposes a potential function and analyzes the intensity of the force for the Weber variation feature to exploit the stimulus of the crowd behavior. Finally, the authors utilize the latent Dirichlet allocation model to recognize the crowd behavior, combined with the bag of feature algorithm. Overall, the paper exploits the characteristics of the behavior from both appearance and stimulus perspectives. The experiments show that the proposed method is effective and robust for the large-scale crowd event perception. Additionally, the proposed method does not base on the premise that the background should be extracted perfectly or individual tracking, which makes it more suitable for the large-scale crowd behavior perception.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] J. C. S. Jacques Jr., A. Braun, J. Soldera, S. R. Musse, and C. R. Jung, "Understanding people motion in video sequences using Voronoi diagrams," *Pattern Analysis and Applications*, vol. 10, no. 4, pp. 321–332, 2007.

[2] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: modeling social behavior for multi-target tracking," in *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV '09)*, pp. 261–268, Kyoto, Japan, October 2009.

[3] M. J. Jones and D. Snow, "Pedestrian detection using boosted features over many frames," in *Proceedings of the 19th International Conference on Pattern Recognition (ICPR '08)*, pp. 1–4, Tampa, Fla, USA, December 2008.

[4] S. Ali and M. Shah, "Floor fields for tracking in high density crowd scenes," in *European Conference on Computer Vision*, Lecture Notes in Computer Science, pp. 1–14, Springer, Berlin, Germany, 2008.

[5] M. Rodriguez, S. Ali, and T. Kanade, "Tracking in unstructured crowded scenes," in *Proceedings of the IEEE 12th International Conference on Computer Vision*, pp. 1389–1396, Kyoto, Japan, October 2009.

[6] L. Kratz and K. Nishino, "Tracking with local spatio-temporal motion patterns in extremely crowded scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern*

Recognition (CVPR '10), pp. 693–700, San Francisco, Calif, USA, June 2010.

[7] L. Kratz and K. Nishino, "Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 987–1002, 2012.

[8] E. L. Andrade, S. Blunsden, and R. B. Fisher, "Modelling crowd scenes for event detection," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, vol. 1, pp. 175–178, Hong Kong, China, August 2006.

[9] Y. Yang, J. Liu, and M. Shah, "Video scene understanding using multi-scale analysis," in *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV '09)*, pp. 1669–1676, Kyoto, Japan, October 2009.

[10] B. Solmaz, B. E. Moore, and M. Shah, "Identifying behaviors in crowd scenes using stability analysis for dynamical systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 2064–2070, 2012.

[11] M. Hu, S. Ali, and M. Shah, "Detecting global motion patterns in complex videos," in *Proceedings of the 19th International Conference on Pattern Recognition (ICPR '08)*, pp. 1–5, Tampa, Fla, USA, December 2008.

[12] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1446–1453, Miami, Fla, USA, June 2009.

[13] Y. Ke, R. Sukthankar, and M. Hebert, "Event detection in crowded videos," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–8, Rio de Janeiro, Brazil, October 2007.

[14] I. Saleemi, L. Hartung, and M. Shah, "Scene understanding by statistical modeling of motion patterns," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2069–2076, San Francisco, Calif, USA, June 2010.

[15] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, "Data-driven crowd analysis in videos," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 1235–1242, Barcelona, Spain, November 2011.

[16] S. Ali and M. Shah, "A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–6, Minneapolis, Minn, USA, June 2007.

[17] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2054–2060, San Francisco, Calif, USA, June 2010.

[18] D. Helbing, "A fluid dynamic model for the movement of pedestrians," *Complex Systems*, vol. 6, pp. 391–415, 1992.

[19] R. L. Hughes, "A continuum theory for the flow of pedestrians," *Transportation Research B*, vol. 36, no. 6, pp. 507–535, 2002.

[20] P. Allain, N. Courty, and T. Corpetti, "Crowd flow characterization with optimal control theory," in *Computer Vision*, Lecture Notes in Computer Science, 2010.

[21] A. Treuille, S. Cooper, and Z. Popovic, "Continuum crowds," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 1160–1168, 2006.

[22] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, no. 5, pp. 4282–4286, 1995.

[23] D. Helbing, I. Farkas, and T. Vicsek, "Simulating dynamical features of escape panic," *Nature*, vol. 407, no. 6803, pp. 487–490, 2000.

[24] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 935–942, Miami, Fla, USA, June 2009.

[25] R. Raghavendra, A. Del Bue, M. Cristani, and V. Murino, "Optimizing interaction force for global anomaly detection in crowded scenes," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV '11)*, pp. 136–143, Barcelona, Spain, November 2011.

[26] A. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall Information and System Sciences Series, Prentice Hall, New York, NY, USA, 1989.

[27] J. Chen, S. Shan, C. He et al., "WLD: a robust local image descriptor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1705–1720, 2010.

[28] G. T. Fechner, "An important psychophysical basic law and its relationship with the estimation of star sizes," *Elemente der Psychophysik*, vol. 31, 1860.

[29] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, no. 5, pp. 4282–4286, 1995.

[30] D. Helbing, "A fluid dynamic model for the movement of pedestrians," *Complex Systems*, vol. 6, pp. 391–415, 1998.

[31] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 539–555, 2009.

[32] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[33] H. Jégou, M. Douze, and C. Schmid, "Packing bag-of-features," in *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV '09)*, pp. 2357–2364, Kyoto, Japan, October 2009.

[34] E. Choi and C. Lee, "Feature extraction based on the Bhattacharyya distance," *Pattern Recognition*, vol. 36, no. 8, pp. 1703–1709, 2003.

[35] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, NY, USA, 1991.

[36] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, Anchorage, Alaska, USA, June 2008.

[37] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *Proceedings of the 19th British Machine Vision Conference*, pp. 995–1004, September 2008.