Going with the Flow: Pedestrian Efficiency in Crowded Scenes

Louis Kratz and Ko Nishino

Department of Computer Science
Drexel University, Philadelphia, PA 19104, USA
{lak24,kon}@drexel.edu

Abstract. Video analysis of crowded scenes is challenging due to the complex motion of individual people in the scene. The collective motion of pedestrians form a crowd flow, but individuals often largely deviate from it as they anticipate and react to each other. Deviations from the crowd decreases the pedestrian's *efficiency*: a sociological concept that measures the difference of actual motion from the intended speed and direction. In this paper, we derive a novel method for estimating pedestrian efficiency from videos. We first introduce a novel crowd motion model that encodes the temporal evolution of local motion patterns represented with directional statistics distributions. This model is then used to estimate the intended motion of pedestrians at every space-time location, which enables visual measurement of the pedestrian efficiency. We demonstrate the use of this pedestrian efficiency to detect unusual events and to track individuals in crowded scenes. Experimental results show that the use of pedestrian efficiency leads to state-of-the-art accuracy in these critical applications.

1 Introduction

A key challenge to video analysis of crowded scenes is the complex motion introduced by the intricate interactions between individual pedestrians. The large number of people and their aggregated motion give rise to coherent motion that form the crowd flow. Individuals in the crowd, however, constantly anticipate and react to others surrounding them, causing pauses or changes in direction and speed. These subtle variations of individual motion result in often large deviations from the crowd flow. These deviations are the main source of difficulty for video analysis as they make individual tracking challenging for a microscopic approach and reduces the accuracy of crowd motion models in a macroscopic approach.

Often pedestrians deviating from crowd flow are reacting to an interruption (e.g., someone cutting them off) or congestion. In such cases, the individual avoids collision by deviating from their intended motion. *Efficiency* is a well studied measure in sociology [1] that quantifies the difference between the actual pedestrian motion and his/her intended speed and direction. Helbing et al. [2] define and measure efficiency in physical space (i.e., meters and seconds measured in the 3D world), and show its direct relationship to crowd stability. To our knowledge, despite the possible applications to visual crowd analysis, efficiency has not been addressed by the vision community.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part IV, LNCS 7575, pp. 558–572, 2012.

To compute pedestrian efficiency the intended motion of each individual must be known. Although it is impossible to know each individual's intention, pedestrians form emergent behaviors (e.g., lanes or clusters) that reveal clues to their intended directions. Still [3] notes that emergent behaviors form because it is easier "to follow immediately behind someone who is already moving in your direction." In other words, the emergent behaviors formed by pedestrians suggest where they intend to move. Such behaviors depend on the scene and vary temporally [4], but tend to repeat [2], forming an underlying space-time structure in the collective motion of the crowd. By learning this structure the intended motion of pedestrians can be estimated and used to estimate efficiency.

In this paper, we present a novel method for estimating pedestrian efficiency from videos and use it for video analysis of crowded scenes. Our key insight is that we may estimate the intended motion of individual pedestrians by modeling the crowd motion. First, we introduce a space-time model that captures the latent structure induced by the motion of the crowd. For this, we use a collection of hidden Markov models over directional statistics distributions of optical flow. By training this model on a short video of the scene, we encode the temporally varying multi-modal flows in the image space resulting from the emergent behaviors of the people in the crowd. Second, we use this model to anticipate the motion at each space-time location of the video. These predicted local motions can then be used to estimate the intended motion of individuals passing through each of those space-time regions. We then compare this estimate to the actual motion represented by the instantaneous optical flow to compute pedestrian efficiency over the entire video volume. By doing so, we measure efficiency within the scene without identifying each individual pedestrian.

We use our pedestrian efficiency estimate to robustly detect local and global unusual activities and to dynamically adjust motion priors for tracking individuals in videos of crowded scenes. The experimental results on a number of videos of real-world crowded scenes show that our method enables the accurate computation of pedestrian efficiency which in turn leads to better predictions of scene motions. As a result, the use of pedestrian efficiency achieves state-of-the-art accuracy in these two fundamental tasks in video analysis that are especially challenging in crowded scenes.

2 Related Work

Macroscopic approaches to video analysis of crowded scenes view the crowd as a collection of individuals obeying a set of analytical rules. Moore et al. [5] present a hydrodynamics model, treating each pedestrian as a particle in a fluid. As noted by Still [3], however, emergent behaviors such as lane formations or clustering do not occur in fluids. Particles are affected only by the external forces around them, but pedestrian motion is a result of both external forces and reactions to other pedestrians. Efficiency decreases when pedestrians react to one another, and is inversely related to the deviation from the crowd motion. As such, automatically estimating pedestrian efficiency enables a better understanding of how individuals interact with the crowd, and can be used to more accurately predict their behaviors in the scene.

Mehran et al. [6] use a social force model but do not measure the full influence of the crowd on the individual. They represent intended velocity using instantaneous optical

Fig. 1. Pedestrian efficiency in videos can be defined by the difference between the intended and actual motions represented with 3D optical flow vectors. Left: The intended direction $\mathbf{u}$ of an individual may be inhibited causing them to move in a different direction $\mathbf{v}$. Right: We measure the difference between these motions with arc-length on the unit sphere, which is inversely proportional to efficiency.

flow, and the average optical flow as the pedestrian's actual velocity. This assumption is not valid in congested scenes: if an area is highly dense and pedestrians are moving slowly, then the averaged (and instantaneous) optical flow has a low velocity and thus their "interaction force" will not reflect the influence of the crowd on the pedestrian's speed. In addition, pedestrians tend to sway when their motion is restricted [7, 8], suggesting that the instantaneous optical flow does not indicate their intended motion. As we show in Sec. 7, by using a model of the crowd motion our method more accurately estimates the intended motion, and can measure efficiency in high-density scenes.

Tracking and anomaly detection methods often degrade when pedestrian motion largely deviates from the crowd motion. Minor, usual deviations appear as noise to anomaly detection, and are often addressed by complex motion descriptors such as distributions of space-time gradients [9] or dynamic textures [10]. Tracking methods designed for crowds [11–13] lose the target when they deviate from the learned model. Other methods based on motion patterns [14] also assume that objects follow dominant flows. Efficiency indicates the severity of the deviation from the learned model, which we can use to detect unusual crowd activities and track pedestrians with a greater robustness to those deviations as we demonstrate.

3 Efficiency

Individuals move through public areas according to their personal goals and with walking speeds they feel comfortable. As shown in the left image in Fig. 1, they have an intended speed and direction, which may be inhibited by surrounding pedestrians. Helbing and Vicsek [1] define the influence of surrounding pedestrians on an individual as the *interaction rate*, and show it is inversely related to efficiency. Rather than computing efficiency for each pedestrian, we estimate efficiency at each space-time pixel location in the video. By doing so, we may analyze the scene without having to detect and track each pedestrian.

Let t denote time and $\mathbf{p} = [x, y]^T$ a 2D pixel location in the video. We denote the intended motion of the pedestrian occupying pixel $\mathbf{p}$ at time t by

$$\mathbf{u}_t(\mathbf{p}) = [\Delta x, \Delta y, \Delta t]^T \ , \tag{1}$$

where Δx, Δy, and Δt is the change (movement) in the horizontal, vertical, and temporal dimensions, respectively, and $|\mathbf{u}_t(\mathbf{p})| = 1$. The 2D optical flow $\tilde{\mathbf{u}}_t(\mathbf{p})$ induced by this indented motion is computed by temporally normalizing this 3D optical flow vector: $\tilde{\mathbf{u}}_t(\mathbf{p}) = [\Delta x/\Delta t, \Delta y/\Delta t]$. Similarly, let $\mathbf{v}_t(\mathbf{p})$ be the 3D instantaneous optical flow observed in the video.

We derive an image-space equivalent of the physical efficiency from Helbing et al. [2]

$$\frac{\tilde{\mathbf{u}}_t(\mathbf{p}) \cdot \tilde{\mathbf{v}}_t(\mathbf{p})}{|\tilde{\mathbf{u}}_t(\mathbf{p})|^2} \,. \tag{2}$$

The bounds of Eq. 2, however, are not well defined. For example, if pedestrians move faster than their intended speed (e.g., in a panic situation), then it is unbounded. As illustrated on the right in Fig. 1, we compute the efficiency using the great-circle distance

$$e_t(\mathbf{p}) = 1 - \frac{\arccos\left(\mathbf{u}_t(\mathbf{p})^T \mathbf{v}_t(\mathbf{p})\right)}{\pi} \tag{3}$$

that is bounded by $[1, 0]$. Since we represent motion using 3D optical flow vectors, Eq. 3 captures both differences in direction (longitudinal variations across the unit sphere) and speed (latitudinal variations). To compute efficiency, however, we need the intended motion $\mathbf{u}_t(\mathbf{p})$. Next, we describe our crowd model which we use to estimate $\mathbf{u}_t(\mathbf{p})$.

4 Directional Statistics Crowd Motion Model

In the absence of other pedestrians, individuals move in straight lines towards their destinations. In higher densities, however, they naturally form organized structures (i.e., emergent behaviors) to utilize the available space and achieve a higher flow [15]. These behaviors vary temporally [4] but tend to repeat [2]. We model this structured crowd motion by training a collection of hidden Markov models (HMMs), one for each spatial location in the frame. Our previous work [12, 9] also use a collection of HMMs but retains appearance information in the form of spatial gradients. In this work, we train the HMMs on directional statistics distributions of optical flow resulting in a more compact and accurate representation. It is worth pointing out that other methods [13, 11] do not retain the temporal dynamics of crowd flow.

As shown in Fig. 2(a), we subsample the video using a regular grid and represent the motion in each sub-volume, or "cuboid." Let ∇I_i be a 3D vector containing the image gradient estimated in the horizontal, vertical, and temporal directions, respectively, and $\{\nabla I_i \,|\, i = 1, \ldots, N\}$ be a set of N space-time gradients within a cuboid. When a cuboid contains motion in a single direction, the space-time gradients lie on a plane orthogonal [16] to the 3D optical flow $\mathbf{q}$. Thus $\mathbf{q}$ can be estimated by solving [16]

$$\left[\frac{1}{N}\sum_{i}^{N} \nabla I_i \nabla I_i^T\right] \mathbf{q} = \mathbf{0} \,. \tag{4}$$

Note that we can use any optical flow estimation algorithm, for instance, those tailored to large displacements [17], if necessary. In this work, we found our gradient-based method sufficient and significantly faster than such dense estimation methods.

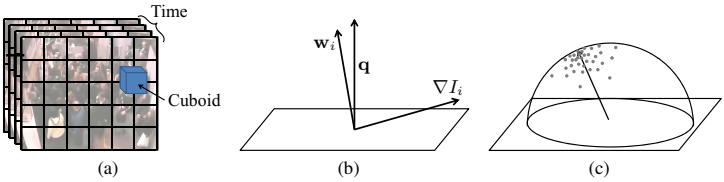


Fig. 2. (a) We subdivide the video into space-time cuboids. (b) The 3D optical flow $\mathbf{q}$ estimated from the cuboid is orthogonal to a plane in spatio-temporal gradient space. A gradient ∇I_i that does not lie on the plane represents uncertainty in the flow, and is orthogonal to another possible flow vector $\mathbf{w}_i$. (c) The set of these possible flow vectors forms a directional distribution on the upper-hemisphere.

Cuboids containing motion in a single direction have gradients that are coplanar, while those containing multiple moving objects have gradients that are not. As illustrated in Fig. 2(b), a space-time gradient ∇I_i that does not lie on the plane suggests motion in another direction $\mathbf{w}_i$ orthogonal to ∇I_i. The vector $\mathbf{w}_i$ is a 3D flow vector

$$\mathbf{w}_i = \frac{\nabla I_i \times \mathbf{q} \times \nabla I_i}{|\nabla I_i \times \mathbf{q} \times \nabla I_i|}\,, \tag{5}$$

where $\times$ is the cross-product.

As shown in Fig. 2(c), the distribution $\{\mathbf{w}_i \,|\, i=1,\ldots,N\}$ exists on the upper hemisphere of $\mathbf{q}$. It's shape characterizes the motion in the cuboid: narrow distributions represent motion in a specific direction, and wide distributions represent motion in multiple directions. A natural representation is the von Mises-Fisher distribution [18]

$$\mathrm{p}(\mathbf{x}) = \frac{1}{c(\kappa)} \exp\left\{\kappa \boldsymbol{\mu}^T \mathbf{x}\right\}, \tag{6}$$

where $\boldsymbol{\mu}$ is the mean direction, $c(\kappa)$ is a normalization constant, and κ is the concentration parameter.

We train an HMM on the von Mises-Fisher distributions observed at each spatial grid location. HMMs are defined by J hidden states, a $J{\times}1$ initial probability vector $\boldsymbol{\pi}$, a $J{\times}J$ transition matrix $\mathbf{A}$, and a set of J emissions densities $\{\mathrm{p}(O|s=j) \,|\, j=1,\ldots,J\}$. In our model, each observation $O=\{\boldsymbol{\mu},\kappa\}$ describes the motion within a specific cuboid. Although κ is not necessary to estimate the intended motion, we include it for tracking in Sec. 6.2. We consider $\boldsymbol{\mu}$ and κ to be statistically independent and define the emission densities analytically

$$\mathrm{p}(O|s=j) = \mathrm{p}(\boldsymbol{\mu}|s=j)\mathrm{p}(\kappa|s=j)\,, \tag{7}$$

where $\mathrm{p}(\kappa|s=j)$ is a Gamma distribution, and $\mathrm{p}(\boldsymbol{\mu}|s=j)$ a von-Mises Fisher distribution (i.e., the conjugate prior on $\boldsymbol{\mu}$ [19]). We train the HMMs on a sample video of the target scene using the Baum-Welch algorithm [20].

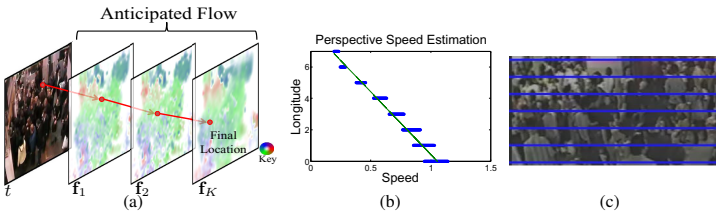


Fig. 3. We estimate the intended direction by advancing each pixel location through a 3D flow field (a) (color indicates speed and direction) that we predict from the HMMs. To estimate the intended speed in scenes captured with a perspective projection, we fit a line to the top 5% of speed measurements (b) at each longitudinal location of the frame (c).

5 Estimation of Intended Motion

Next, we use the trained HMMs to estimate the intended motion at each space-time location in a different video of the same scene. We discuss direction and speed separately, and combine them to compute the intended motion $\mathbf{u}_t(\mathbf{p})$.

5.1 Intended Direction

Given the observed video up to time t and an HMM trained at spatial location $\mathbf{p}$, we compute $\mathbf{z}_k(\mathbf{p})$ as a $1 \times J$ vector representing the likelihood of being in state j at time $t+k$

$$\mathbf{z}_k(\mathbf{p}) = \boldsymbol{\alpha}_t \mathbf{A}^k \,, \tag{8}$$

where $\mathbf{A}$ is the state transition matrix from the HMM, and $\boldsymbol{\alpha}_t$ is the scaled forward message from the forwards-backwards algorithm [20]. As $k \to \infty$, Eq. 8 approaches the stationary distribution of the Markov process (if it exists).

We use the set $\{\mathbf{z}_k(\mathbf{p}) | k = 1, \ldots, K\}$ to compute the optical flow after time t. We select K large enough to approach the stationary distribution. Let $\mathbf{f}_k(\mathbf{p})$ be the flow predicted from $\mathbf{z}_k(\mathbf{p})$

$$\mathbf{f}_k(\mathbf{p}) = \sum_{j=1}^{J} \mathbf{z}_{k,j}(\mathbf{p}) \mathbb{E}\left[\mathrm{p}(\boldsymbol{\mu} | s = j)\right] \,, \tag{9}$$

where $\mathrm{p}(\boldsymbol{\mu} | s = j)$ is the emission density from Eq. 7. The resulting flow field (i.e., $\mathbf{f}_k(\mathbf{p})$ for all spatial locations and values of k) represent the anticipated flow of the crowd.

As shown in Fig. 3(a), we estimate the future location of each point $\mathbf{p}$ by advancing it through the anticipated flow field. Let $\tilde{\mathbf{f}}_k(\mathbf{p})$ be the 2D optical flow computed from $\mathbf{f}_k(\mathbf{p})$, and $\hat{\mathbf{p}}_k$ the location of $\mathbf{p}$ at time $k+t$. The next location $\hat{\mathbf{p}}_{k+1}$ is computed by following the predicted flow at the previous point

$$\hat{\mathbf{p}}_{k+1} = \hat{\mathbf{p}}_k + \tilde{\mathbf{f}}_k(\hat{\mathbf{p}}_k) \,. \tag{10}$$

Eq. 10 is initialized with $\hat{\mathbf{p}}_0 = \mathbf{p}$. The final location $\hat{\mathbf{p}}_K$ indicates, according the crowd motion, the intended location of the pedestrian occupying $\mathbf{p}$. The intended direction is the difference of this point from the current location

$$\bar{\mathbf{u}}_t(\mathbf{p}) = \frac{1}{Z}(\hat{\mathbf{p}}_K - \mathbf{p}), \tag{11}$$

where Z is a normalization term such that $|\bar{\mathbf{u}}_t(\mathbf{p})| = 1$.

5.2 Intended Speed

The walking speed of pedestrians has been well studied and is near constant if there is no congestion. Zip's [21] least-effort principle implies that pedestrians minimize metabolic energy when walking at roughly 1.33 meters per second [22], which has been verified in observational studies [23, 24]. For scenes recorded at a distance, we may assume orthographic camera projection and thus a constant intended speed can be estimated for all pedestrians. We approximate the intended speed as the maximum observed speed in the training video. Intuitively, we are identifying the few instances where pedestrians can move freely due to lulls in traffic or less-crowded areas. To address unreliable or erroneous flow estimates, we use Chauvenet's criterion [25] to remove outliers.

For near field views that exhibit perspective distortion, as shown in Fig. 3(b), we estimate the intended speed by observing the relationship between each longitudinal frame location. First, we identify the fastest 5% of speed measurements from each longitudinal frame location. Due to the perspective projection, the speeds across the frame have near-linear relationship. We find a least-squares line fit to the speed measurements to estimate the desired speed over the entire image. Outliers are also removed using Chauvenet's criterion.

Finally, given intended speed $s(\mathbf{p})$ and direction $\bar{\mathbf{u}}_t(\mathbf{p})$, we may compute the intended motion

$$\mathbf{u}_t(\mathbf{p}) = \left[\bar{\mathbf{u}}_t(\mathbf{p})^T, s(\mathbf{p})\right]^T, \tag{12}$$

and normalize such that $|\mathbf{u}_t(\mathbf{p})| = 1$.

6 Applications

The pedestrian efficiency computed for each frame of the video can be used to analyze the scene despite the crowd. In this paper, we demonstrate its use in two critical video analysis tasks that are particularly challenging for crowded scenes: anomaly detection and pedestrian tracking.

6.1 Anomaly Detection

Low pedestrian efficiency is an indicator of unusual activities. Atypical motions decrease efficiency in local areas, and crowd disasters contain people moving irrationally.

We can identify *global* anomalies, i.e., affecting a large portion if not the entire crowd, as frames that have low average efficiency values

$$\bar{e}_t = \frac{1}{|\mathcal{P}|} \sum_{\mathbf{p} \in \mathcal{P}} e_t(\mathbf{p}) \,,$$

where $\mathcal{P}$ is the set of 2D pixel locations.

We may also detect *local* anomalies, such as individuals moving against the crowd flow. Such offenders will exhibit low efficiency (since the training data lacks their intended motion) and decrease the efficiency in their immediate vicinity (as surrounding pedestrians must avoid them). We identify local unusual events as space-time regions with low efficiency. Since many scenes naturally contain low efficiency (a congested train station, for example), we normalize the efficiency $\tilde{e}_t(\mathbf{p}) = \frac{e_t(\mathbf{p})}{Z(\mathbf{p})}$, where $Z(\mathbf{p})$ is the average efficiency at spatial location $\mathbf{p}$ of the training data.

We identify the space-time locations with low efficiency using a space-time Markov random field. Details are omitted for limited space, but this can be achieved with binary latent variables indicating whether the scene point exhibits usual activities or not. The latent variables can be computed through energy minimization of an error function consisting of a data term that returns the efficiency value if the scene point contains unusual activity together with an Ising model smoothing term. This energy minimization can be efficiently solved with graph-cuts [26, 27].

6.2 Tracking

Efficiency indicates how much an individual is conforming to the flow of the crowd. As such, we may use it as a dynamic prior on the individual's motion to probabilistically track pedestrians in crowded scenes.

Let $\mathbf{x}_t$ be the 2D pixel location at time t of a pedestrian being tracked. Object-centric methods [28, 29] assume pedestrians exhibit smooth motion and impose (often first order) stochastic dynamics to update the location

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{h}_t + \boldsymbol{\epsilon} \,, \tag{13}$$

where $\mathbf{h}_t$ is a 2D flow vector and $\boldsymbol{\epsilon}$ is (typically Gaussian) noise. Crowd methods [12, 13, 11] use a learned model of the crowd

$$\mathbf{x}_{t+1} = \mathbf{x}_t + c(\mathbf{x}_t) + \boldsymbol{\epsilon} \,, \tag{14}$$

where $c(\mathbf{x}_t)$ is the flow of the crowd at location $\mathbf{x}_t$. Using our model, $c(\mathbf{x}_t)$ is the predicted von Mises-Fisher distribution ($\boldsymbol{\mu}$ and κ) from the HMM at location $\mathbf{x}_t$.

Macroscopic approaches assume the crowd motion model yields an accurate prediction, and do not perform well when pedestrians deviate from the crowd (i.e., areas of low efficiency). Microscopic (object-centric) approaches that rely on individual motion models, such as a linear model, struggle in areas without visible backgrounds (often high efficiency). We use pedestrian efficiency as an indicator of how much to trust the crowd motion model and dynamically weight the two motion models

$$\mathbf{x}_{t+1} = \mathbf{x}_t + e_t(\mathbf{x}_t) c_t(\mathbf{x}_t) + [1 - e_t(\mathbf{x}_t)] \, \mathbf{h}_t + \boldsymbol{\epsilon} \,. \tag{15}$$

(a) (b) (c) (d) (e) (f)

Fig. 4. Frames from six videos on which we evaluate our method. The concourse (a) [12], street (b) [6], and sidewalk (c) [30] scenes contain pedestrians moving in many different directions. The platform (d), escalator (f) (both from [31]), and intersection (e) [30] contain more obvious emergent behaviors such as lane formation.

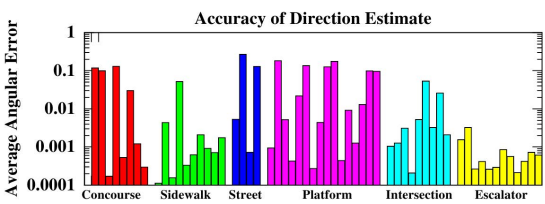
For the individual's motion $\mathbf{h}_t$ we use the expected vector of a von Mises-Fisher distribution fitted to the previous flow observations. Intuitively, we are switching between the crowd motion model and a simple individual motion model that maintains the momentum at that location based on the pedestrian efficiency; when the pedestrian efficiency is high go with the crowd flow and otherwise let the individual maintain its own previous motion. Our final state-transition density is a von Mises-Fisher distribution computed by weighting the expected directions and variances.

7 Experimental Results

Fig. 4 shows frames from six videos of crowded scenes that we use to evaluate our method. For each scene, we train the HMMs on a sample video sequence, and use them to compute the efficiency in a video of the same scene recorded at a different time. The concourse (4(a) from [12]), sidewalk (4(c) from [30]), and street (4(b) from [6]) scenes have few physical obstacles and contain many interactions. The platform (d) and escalator (f) (both from [31]) scenes contain low efficiency due to bottlenecks. The intersection ((e) from [30]) contains pedestrians avoiding each other as they intersect in the middle of the frame. Many of the videos are available from the respective authors.

Fig. 5 shows examples of pedestrians moving inefficiently. The left most example shows an individual changing direction due to congestion. His intended direction is to the left, and efficiency drops when moving around other pedestrians. The middle example shows pedestrians avoiding an oncoming individual (video from [6]). Their intended direction is vertical, and efficiency decreases as they move to the side. The

Fig. 5. Low efficiency (red=low efficiency, blue=high) due to congestion (left), pedestrians avoiding an individual (middle), and a lack of motion (right). The yellow solid arrow is the intended motion, and the green dashed arrow depicts the actual motion (optical flow).



Average Error

Scene	Ours	Mehran et al. [6]
Concourse	0.042	0.458
Street	0.086	0.860
Sidewalk	0.006	0.260
Platform	0.051	0.729
Intersection	0.010	0.207
Escalator	0.001	0.258

Fig. 6. The accuracy of our estimate of future directions for a number of pedestrians (left), along with averages compared with Mehran et al. [6] (right)

pedestrians in the right most example are standing still and exhibit lower efficiency than those moving in the lower left of the image.

Since it is impossible to know a pedestrian's intentions, we cannot directly measure the accuracy of our estimated intended motion. We can, however, assume that pedestrians move in their intended direction over time. Let $\{\hat{\mathbf{x}}_t | t = 1, \ldots, T\}$ be a sequence of ground-truth tracking locations for a specific pedestrian. We measure the error

$$\frac{1}{T}\sum_{t=1}^{T} \arccos\left(\frac{\bar{\mathbf{u}}_t(\hat{\mathbf{x}}_t)^T\left[\hat{\mathbf{x}}_{t+w} - \hat{\mathbf{x}}_t\right]}{|\hat{\mathbf{x}}_{t+w} - \hat{\mathbf{x}}_t|}\right), \tag{16}$$

where $\bar{\mathbf{u}}_t$ is the estimated intended direction from Eq. 11, and w is a window size that depends on the subject (typically the duration the subject is in the scene).

The left graph in Fig. 6 shows the estimation error for a number of subjects from different scenes. For almost all of the subjects the estimation error is below 0.1 (about $6°$). None of the error rates exceed 0.2 which is small given the resolution of the video. The theoretical maximum error is π, and thus at most the error is $0.2/\pi \approx 6\%$. The right table in Fig. 6 shows the average error for all scenes, and the error using the optical flow for the intended motion as suggested by Mehran et al. [6]. Scenes with less structure, such as the concourse and street, have higher errors due to the larger number of directions that pedestrians move. Compared with Mehran et al. [6], our method achieves consistently lower errors.

7.1 Anomaly Detection

First, we detect global anomalies as frames with low average efficiency on the University of Minnesota Crowd Dataset [32]. The dataset contains a number of usual and unusual video segments from 3 different scenes. For each scene, we train the HMMs on a usual sequence, and estimate efficiency on the remaining sequences. A frame is considered unusual if its average efficiency is below a specific threshold that is selected empirically. Fig. 7(a) shows visualizations of the efficiency for usual (top) and unusual activities (bottom) for the first scene. The pedestrians in the unusual frame (bottom) exhibit lower efficiency than those in the usual frame (top).

Fig. 7. The efficiency on frames from the UMN data set (a) is high in usual scenes (top) and low in unusual scenes (bottom). Pedestrians that move against the crowd exhibit low efficiency (b). We detect such anomalies (c) with higher accuracy than our previous method [9] (d) (top) and Mahadevan et al. [10] (bottom). The color indicates the detection results: blue are true positives, red are false negatives, and pink are false positives.

The left graph in Fig. 8 shows the average efficiency plotted over time for a specific scene in the UMN data set. The red and green points are the average efficiency from clips of usual and unusual activities, respectively. The average efficiency drops during all six clips of unusual activities. We vary the threshold to compute an ROC curve. The area under the ROC curve was 0.92, which compares favorably with 0.96 in [6] and 0.99 in [33]. Our slightly poorer performance is due to the higher efficiency at the beginning and end of each unusual sequence (where pedestrians are moving normally) as shown in the left graph in Fig. 8.

We evaluate our local anomaly detection method on the UCSD Anomaly Detection Dataset [34] from [10] and videos of two train station scenes from [9]. We measure detection accuracy by the average of the true positive rates and true negative rates. The UCSD data set provides ground truth for some sequences. We hand-labeled the ground-truth for the remaining sequences and those of the train station.

Fig. 7 shows example frames of local anomalies detected in both datasets. The intended motion of pedestrians moving against the crowd cannot be determined, and thus such individuals exhibit low efficiency as shown in Fig. 7(b). We successfully detect such pedestrians as shown in Fig. 7(c). As shown in Fig. 7(d), efficiency is less sensitive to minor deviations than our previous method [9] and that of Mahadevan et al. [10].

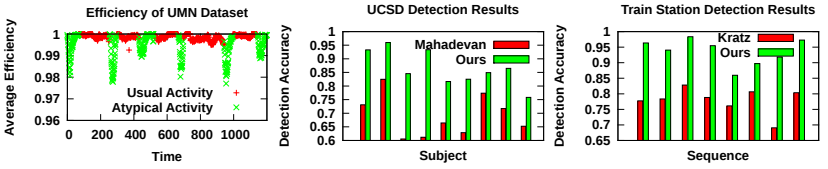


Fig. 8. Left: Efficiency drops when crowds in the UMN data set enter unusual states, as shown by the green points in the graph. Middle: Accuracy of local anomaly detection for 9 sequences in the UCSD Crowd Dataset [34] compared with [10]. Right: Accuracy of 8 sequences from two train station scenes compared with our previous method [9]. Using efficiency achieves higher accuracy for all sequences compared with other approaches.

Table 1. Tracking errors averaged over multiple subjects for the different scenes using estimated pedestrian efficiency compared with our previous crowd motion model approach [12] and that of Rodriguez et al. [11]. Using efficiency achieves lowest error on almost all scenes. On the concourse scene we achieve comparable results to our previous method [12].

	Concourse	Street	Platform	Escalator	Intersection	Sidewalk
Ours	8.7	**3.3**	**2.8**	**8.5**	**3.1**	**7.4**
Kratz and Nishino [12]	**6.8**	47.6	17.3	24.8	3.56	9.9
Rodriguez et al. [11]	24.7	14.8	29.9	60.4	25.9	11.9

The middle graph in Fig. 8 shows the detection accuracy of our method on 9 sequences compared with that of Mahadevan et al. [10], and the right graph in Fig. 8 compares the results on 8 sequences with our previous method [9]. We use the results of Mahadevan et al. [10] posted on the web for comparison. The use of pedestrian efficiency achieves consistently higher accuracy in both cases.

7.2 Tracking

We quantitatively evaluated our tracking method using hand-labeled ground truth of targets. Given a ground-truth location $\hat{\mathbf{x}}_t$ and tracking result $\mathbf{x}_t$, the tracking error $|\hat{\mathbf{x}}_t - \mathbf{x}_t|$ is averaged over all frames $\{t=1,\ldots,T\}$. Table 1 shows the tracking errors (average over multiple subjects for each sequence) using the estimated pedestrian efficiency compared with our previous method [12] and that of Rodriguez et al. [11]. Using pedestrian efficiency achieves superior results on all scenes but one, and significantly lower errors on the platform, escalator, and street scenes where pedestrians move with lower efficiency due to higher density.

Pedestrians that deviate from the flow of the crowd present challenges to tracking. Since such pedestrians naturally have low efficiency, our method is able to reliably track them by gracefully switching to simple individual motion models as defined in Eq. 15. The left most four images in Fig. 9 shows two tracking results using our method and just the crowd motion model (Eq. 14). In both cases, the pedestrian is moving against the crowd: the first is moving left to right, and the second is moving towards the bottom of the frame. As shown in green, the crowd model assumes pedestrians are moving with the crowd, drifts, and loses the target. Our method, shown in red, is able to compensate for the anomaly and accurately track the targets. The middle graph in Fig. 9 shows the tracking errors for 16 anomalous targets using both methods. Using pedestrian efficiency achieves a consistently lower error.

The right graph in Fig. 9 shows the ratio of our tracking error to the tracking error using just the crowd model for different subjects. High ratios (i.e., 1) indicate that our method performs similar to using just the crowd motion model, while a low ratio indicates improvement by our method. The downward trend of the points show the advantage of using pedestrian efficiency: our method vastly improves tracking in crowds when pedestrians are moving inefficiently, and performs similarly to crowd motion models when pedestrians are moving with the flow.

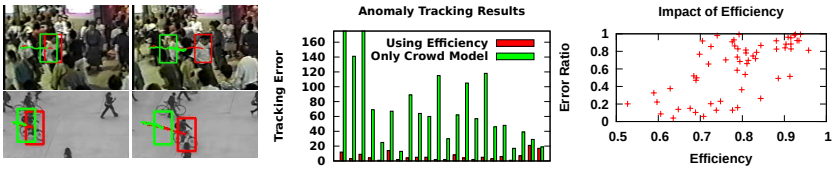


Fig. 9. Left: Tracking results of pedestrians deviating from the general crowd flow in the concourse (top row) and UCSD dataset (bottom row) using our method (red) and just a crowd motion model (green). The crowd motion model assumes that the pedestrians are moving with the crowd causing the tracker to drift (left column) or lose the target (right column). Middle: Since such anomalous pedestrians naturally have low efficiency, our method achieves a lower tracking error for all the tested subjects. Right: Pedestrians moving inefficiently have a low ratio (close to 0) and the downward trend indicates crowd motion models are only accurate when pedestrians are moving efficiently.

8 Conclusion

In this paper, we introduced the use of pedestrian efficiency for video analysis of crowded scenes. We showed that the pedestrian efficiency can be computed from a video without detecting and tracking individuals. The computed pedestrian efficiency can be used to reliably identify global and local anomalous activities, and robustly track individuals through crowded scenes regardless of whether they are conforming to the crowd flow or not. The experimental results show that the computation and use of pedestrian efficiency can indeed enable more reliable video analysis of crowded scenes. We believe that measuring efficiency is but the first step to recognizing the impact of individuality on crowds, and provides new means to further study the complex interactions between pedestrians in videos of crowded scenes.

Acknowledgments. This work was supported in part by National Science Foundation grants IIS-0746717 and Nippon Telegraph and Telephone Corporation. The authors thank Nippon Telegraph and Telephone Corporation for providing the train station videos.

References

1. Helbing, D., Vicsek, T.: Optimal Self-Organization. New Journal of Physics 13 (1999)
2. Helbing, D., Moln, P., Farkas, I.J., Bolay, K.: Self-Organizing Pedestrian Movement. Environment and Planning B: Planning and Design 28, 361–383 (2001)
3. Still, K.: Crowd Dynamics. PhD thesis, University of Warwick (2000)
4. Schadschneider, A., Klingsch, W., Kluepfel, H., Kretz, T., Rogsch, C., Seyfried, A.: Evacuation Dynamics: Empirical Results, Modeling and Applications. In: Encyclopedia of Complexity and Systems Science, pp. 3142–3176 (2009)
5. Moore, B.E., Ali, S., Mehran, R., Shah, M.: Visual Crowd Surveillance Through a Hydrodynamics Lens. Comm. of ACM 54, 64–73 (2011)
6. Mehran, R., Oyama, A., Shah, M.: Abnormal Crowd Behavior Detection using Social Force Model. In: Proc. of IEEE CVPR (2009)

7. Krausz, B., Bauckhage, C.: Analyzing Pedestrian Behavior in Crowds for Automatic Detection of Congestions. In: Proc. of IEEE Workshop on MSVLC (2011)
8. Hoogendoorn, S.P., Daamen, W.: Pedestrian Behavior at Bottlenecks. Transportation Science 39 (2005)
9. Kratz, L., Nishino, K.: Anomaly Detection in Extremely Crowded Scenes Using Spatio-Temporal Motion Pattern Models. In: Proc. of IEEE CVPR, pp. 1446–1453 (2009)
10. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly Detection in Crowded Scenes. In: Proc. of IEEE CVPR, pp. 1975–1981 (2010)
11. Rodriguez, M., Ali, S., Kanade, T.: Tracking in Unstructured Crowded Scenes. In: Proc. of IEEE ICCV (2009)
12. Kratz, L., Nishino, K.: Tracking Pedestrians using Local Spatio-Temporal Motion Patterns in Extremely Crowded Scenes. IEEE TPAMI 34, 987–1002 (2012)
13. Ali, S., Shah, M.: Floor Fields for Tracking in High Density Crowd Scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 1–14. Springer, Heidelberg (2008)
14. Yu, Q., Medioni, G.: Motion pattern interpretation and detection for tracking moving vehicles in airborne video. In: Proc. of IEEE CVPR, pp. 2671–2678 (2009)
15. Kretz, T., Grunebohm, A., Kaufman, M., Mazur, F., Schreckenberg, M.: Experimental Study of Pedestrian Counterflow in a Corridor. JSTAT 2006, P10001 (2006)
16. Wright, J., Pless, R.: Analysis of Persistent Motion Patterns Using the 3D Structure Tensor. In: IEEE WACV, pp. 14–19 (2005)
17. Brox, T., Malik, J.: Large Displacement Optical Flow: Descriptor Matching in Variational Motion Estimation. IEEE TPAMI 33, 500–513 (2011)
18. Mardia, K.V., Jupp, P.: Directional Statistics. John Wiley and Sons Ltd. (1999)
19. Mardia, A., El-Atoum, S.: Bayesian Inference for The Von Mises-Fisher Distribution Miscellanea. Biometrika 63, 203–206 (1976)
20. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2007)
21. Zipf, G.: Human Behavior and the Principle of Least Effort. Addison-Wesley Press (1949)
22. Guy, S., Chhugani, J., Curtis, S., Dubey, P., Lin, M., Manocha, D.: PLEdestrians: A Least-Effort Approach to Crowd Simulation. In: Proc. of ACM/EG SCA, pp. 119–128 (2010)
23. Teknomo, K.: Microscopic Pedestrian Flow Characteristics: Development of an Image Processing Data Collection and Simulation Model. PhD thesis, Tohoku University (2002)
24. Henderson, L.F.: The Statistics of Crowd Fluids. Nature 229 (1971)
25. Chauvenet, W.: In: A Manual of Spherical and Practical Astronomy, 5th edn., pp. 474–566. Adamant Media Corporation (1891)
26. Boykov, Y., Kolmogorov, V.: An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. IEEE TPAMI 26, 1124–1137 (2004)
27. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C.: A Comparative Study of Energy Minimization Methods for Markov Random Fields. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 16–29. Springer, Heidelberg (2006)
28. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-Based Probabilistic Tracking. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part I. LNCS, vol. 2350, pp. 661–675. Springer, Heidelberg (2002)
29. Isard, M., Blake, A.: CONDENSATION-Conditional Density Propagation for Visual Tracking. IJCV 29, 5–28 (1998)
30. Ali, S., Shah, M.: A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis. In: Proc. of IEEE CVPR, pp. 1–6 (2007)

31. Cheriyadat, A., Radke, R.: Detecting Dominant Motions in Dense Crowds. IEEE Journal of Selected Topics in Signal Processing 2, 568–581 (2008)
32. University of Minnesota: Unusual Crowd Activity Dataset (2006), http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi
33. Raghavendra, R., Bue, A.D., Cristani, M., Murino, V.: Optimizing Interaction Force for Global Anomaly Detection in Crowded Scenes. In: Proc. of IEEE ICCV, pp. 136–143 (2011)
34. University of California San Diego: Anomaly Detection Dataset (2010), http://www.svcl.ucsd.edu/projects/anomaly/