Tracking Pedestrians Using Local Spatio-Temporal Motion Patterns in Extremely Crowded Scenes

Louis Kratz, *Student Member*, *IEEE*, and Ko Nishino, *Member*, *IEEE*

Abstract—Tracking pedestrians is a vital component of many computer vision applications, including surveillance, scene understanding, and behavior analysis. Videos of crowded scenes present significant challenges to tracking due to the large number of pedestrians and the frequent partial occlusions that they produce. The movement of each pedestrian, however, contributes to the overall crowd motion (i.e., the collective motions of the scene's constituents over the entire video) that exhibits an underlying spatially and temporally varying structured pattern. In this paper, we present a novel Bayesian framework for tracking pedestrians in videos of crowded scenes using a space-time model of the crowd motion. We represent the crowd motion with a collection of hidden Markov models trained on local spatio-temporal motion patterns, i.e., the motion patterns exhibited by pedestrians as they move through local space-time regions of the video. Using this unique representation, we predict the next local spatio-temporal motion pattern a tracked pedestrian will exhibit based on the observed frames of the video. We then use this prediction as a prior for tracking the movement of an individual in videos of extremely crowded scenes. We show that our approach of leveraging the crowd motion enables tracking in videos of complex scenes that present unique difficulty to other approaches.

Index Terms—Tracking, video analysis, crowded scenes, spatio-temporal motion patterns, hidden Markov models.

1 Introduction

Tracking objects or people is a crucial step in video analysis with a wide range of applications including behavior modeling and surveillance. The rising prevalence of video recording technology in crowded areas presents a dire need for tracking methods that can operate on videos containing a large numbers of individuals. Due to the large number of pedestrians in close proximity, crowded areas are at high risk for dangerous activities, including crowd panic, stampedes, and accidents involving a large number of individuals. Extremely crowded scenes, such as that shown in Fig. 1, are one such scenario of high density, cluttered scenes that contain a large number of individuals. The large amount of activity within such scenes is difficult for even human observers to analyze, making extremely crowded scenes perhaps in the most need of automatic video analysis.

Conventional tracking methods typically assume a static background or easily discernible moving objects, and, as a result, are limited to scenes with relatively few constituents. Videos of crowded scenes present significant challenges to these *object-centric* approaches, i.e., those that focus on representing the motion and appearance of an object or individual in the scene. Surveillance videos of extremely crowded scenes exhibit three defining characteristics that make them challenging for such approaches. First, videos of such scenes contain hundreds of pedestrians in each frame, and possibly thousands throughout the video, resulting in frequent partial occlusions. Second, pedestrians may move in any number of different directions depending on their personal goals, neighboring pedestrians, and physical obstacles within the scene. Finally, each individual's appearance may change significantly with their body movements due to the near-field view of the scene.

The various movements of the individuals, however, contribute to the overall *crowd motion*, i.e., the collective motions of all of the scene's constituents over the entire video. Though seemingly chaotic, the crowd motion exhibits a spatially and temporally varying structured pattern due to the natural behavior of pedestrians. While each individual's trajectory through the scene may differ, their instantaneous motions through small areas of the video (i.e., their local spatio-temporal motion pattern) tend to repeat. For example, many of the pedestrians shown in Fig. 1 move from the bottom of the frame toward the top left, even though their full trajectories are dissimilar and spatially separated. The crowd motion may be viewed as a collection of these local spatio-temporal motion patterns, and used as a *scene-centric* (i.e., incorporating information from the entire video) constraint to track individuals.

The local spatio-temporal motion patterns that comprise the crowd motion vary spatially across the frame and temporally over the video. Spatial variations occur when different pedestrians concurrently move in different directions. Temporal variations also occur, as illustrated in Fig. 1 by the intersecting trajectories, when pedestrians in the same area of the frame move in different directions at

• *The authors are with the Department of Computer Science, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104. E-mail: {lak24, kon}@drexel.edu.*

Manuscript received 3 Nov. 2010; revised 24 May 2011; accepted 26 July 2011; published online 11 Aug. 2011.
Recommended for acceptance by B. Schiele.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-11-0842.
Digital Object Identifier no. 10.1109/TPAMI.2011.173.

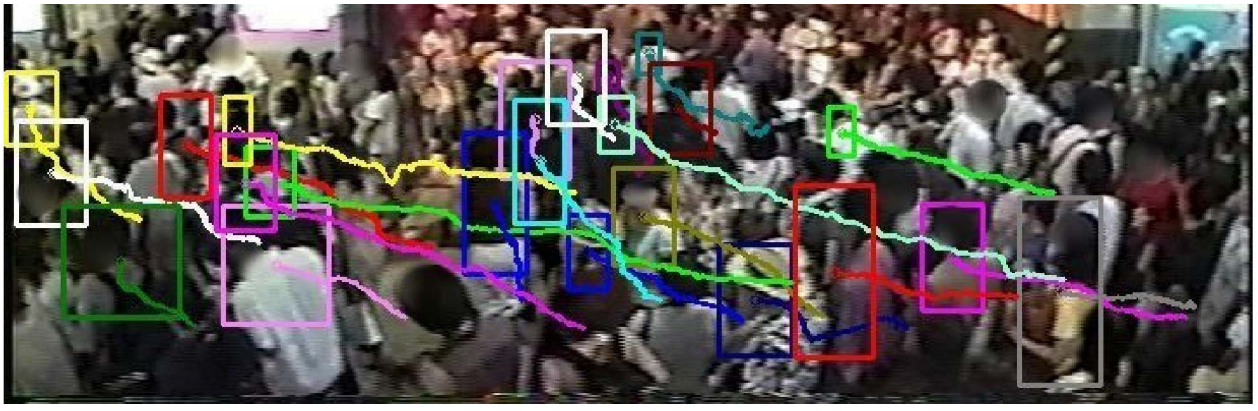

Fig. 1. Extremely crowded scenes, such as the one shown here from a subway station during rush hour, contain hundreds of pedestrians in each frame moving in any number of directions and exhibiting frequent partial occlusions. These characteristics make tracking individuals in videos of such scenes extremely difficult. The movements of pedestrians, however, collectively contribute to the overall crowd motion, whose structure varies spatially across the frame, as illustrated by the pedestrians moving toward different exits, and temporally over the video, as illustrated by the intersecting trajectories caused by pedestrians moving in different directions. We track pedestrians in videos of extremely crowded scenes by leveraging the spatially and temporally varying structure in the crowd motion.

different times of the video. Temporally sequential local spatio-temporal motion patterns are strongly related as they are either exhibited by the same pedestrian or indicate a transition between two pedestrians, both of which contribute to the overall crowd motion. Characteristic spatial and temporal variations recur frequently due to the large number of pedestrians moving through the scene, and are captured by the fixed-view surveillance systems that continuously record the scene. Thus, the crowd motion (i.e., the typical space-time dynamics of the local spatio-temporal motion patterns comprising the collective motion within the scene) may be learned from a video of the scene and used to predict the motion of individuals moving through local regions.

In this paper, we track individual pedestrians in videos of extremely crowded scenes by leveraging a model of the spatially and temporally varying structure in the crowd motion. Our key idea is to exploit the dynamic latent structured pattern of crowd motion in the scene to predict the movement of individuals. Specifically, we leverage the crowd motion as a prior in a Bayesian tracking framework. First, we model the crowd motion as a statistical model of local spatio-temporal motion patterns that changes spatially across the frame and temporally over the video. By doing so, we model the time-varying dynamics of the crowd motion that occurs at and across different spatial locations of the video. Next, we use this temporal model to *predict* the local spatio-temporal motion patterns of the individual being tracked. Finally, we use the predicted local spatio-temporal motion pattern to hypothesize the parameters of a particle filter to track the target. By using a space-time model of the crowd motion, we are able to accurately track individuals in crowded scenes that contain pedestrians whose motions vary across the frame and over time.

Fig. 2 depicts the outline of our approach for tracking pedestrians in videos of extremely crowded scenes. First, as shown in Fig. 2a, we divide a training video into spatio-temporal subvolumes, or "cuboids," defined by a regular grid, and model the motion within each using a local spatio-temporal motion pattern. We then represent the crowd motion with a collection of hidden Markov models (HMMs), one for each spatial location, trained on the local spatio-temporal motion patterns. The hidden states of the HMMs encode the multiple possible motions that can occur

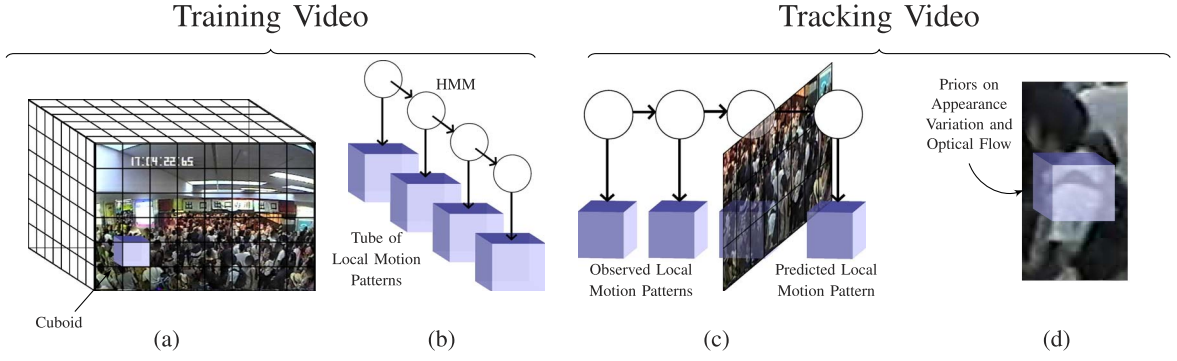

Fig. 2. An overview of our model of the crowd motion for tracking in videos of extremely crowded scenes. We divide the video into spatio-temporal subvolumes, or cuboids, defined by a regular grid (a) and compute the local spatio-temporal motion pattern within each. We then train a hidden Markov model (b) on the local spatio-temporal motion patterns at each spatial location, or tube, of the training video to represent the spatially and temporally varying crowd motion. Using the HMM and previously observed frames of a separate tracking video of the same scene, we predict (c) the local spatio-temporal motion pattern that describes how a target moves through the video. Finally, we use this predicted local spatio-temporal motion pattern to hypothesize a set of priors (d) on the motion and appearance variation of individuals that we wish to track.

at each spatial location. The transition probabilities of the HMMs encode the time-varying dynamics of the crowd motion. By training a collection of HMMs, we encode the spatially and temporally varying motions of pedestrians that comprise the entire crowd motion. To accurately represent the motion within the video, we avoid quantization by training novel HMMs whose observations are the local spatio-temporal motion patterns themselves. As illustrated in Fig. 2c, we then use this model of the crowd motion to predict local spatio-temporal motion patterns in a video of the same scene containing the pedestrian that we wish to track. We use this prediction to hypothesize a full distribution of the optical flow for the cuboids the target currently occupies. In addition, we hypothesize the amount of appearance variation the target will exhibit within the cuboids. This is in contrast to our previous work [16], which did not dynamically adjust for the pedestrian's changing appearance. We use these predictions as a prior on a particle filter to track individuals. We demonstrate the ability of our method to track pedestrians in videos of extremely crowded scenes that present significant difficulty to conventional methods. We show that our approach accurately predicts the optical flow and appearance variations that a target will exhibit during tracking. By modeling the crowd motion as a spatially and temporally varying structured pattern, we achieve more accurate tracking results compared to previous models that only use a fixed model or disregard temporal information.

2 Tracking in Crowded Scenes

Previous work on tracking are mostly object centric, based on the modeling of the motion and appearance of the target individual. Recent work extend these methods to handle crowded scenes by mainly focusing on deriving a robust appearance model of each individual. A representative approach by Zhao et al. [29] tracks multiple pedestrians in videos of crowds by modeling the camera, human shapes, and the background's appearance. Though improved accuracy can be achieved by a more descriptive model of each tracked target, the large amount of partial occlusions and significant variations in the movement of pedestrians in videos of dense crowds still present significant challenges to such approaches.

Other object-centric techniques, such as the work by Betke et al. [5], Li et al. [20], or Wu and Nevatia [28], use object-detection to create short trajectories and then, based on a global model of all of the objects within the scene, merge the trajectories into the final tracking results. In contrast, our approach uses global information regarding the entire crowd motion *prior* to tracking the individual pedestrians and thus may operate online. In addition, these approaches assume that the underlying detection is reliable. Videos of extremely crowded scenes may contain over a hundred pedestrians in a single frame, and possibly thousands throughout the video, resulting in significant partial occlusions well beyond those found in typical crowded scenes. As such, they present challenges to detection-based methods since frequently the entire individual is not visible and correlating detections between frames becomes difficult, if not impossible.

Techniques that track multiple objects, such as that of Hue et al. [10], detect interest points within each frame to describe the objects within the scene. Tracking is performed concurrently on all of the objects by establishing correspondences between the points in different frames. As noted by Khan et al. [13], a single point may be shared between multiple targets and present ambiguity to the tracker. Shared points are often the result of motion boundaries or clutter, both of which occur frequently in videos of extremely crowded scenes. Rather than associating points between frames, our approach predicts the motion and appearance variation a single target will exhibit between frames. Khan et al. [12] model the interaction among the interest points to improve the tracking of each object. Due to the large number of pedestrians within extremely crowded scenes, modeling the interaction among and the separate motion of individual objects is intractable. By using a scene-centric model of the crowd motion we can track an individual without explicitly modeling each moving object.

Other work, such as that by Brostow and Cipolla [8] or Sugimura et al. [26], detects individual moving entities by associating similar feature trajectories. They assume that the subjects move in distinct directions, and thus disregard possible local motion inconsistencies between different body parts. As noted by Sugimura et al. [26], such inconsistencies cause a single pedestrian to be detected as multiple targets or multiple individuals that move similarly to be detected as the same target. Extremely crowded scenes, especially when captured in relatively near-field views, as is often the case in video surveillance, necessitate a method that captures the multiple motions of a single target and also discerns different pedestrians with similar movements.

Ali and Shah [2] track pedestrians in videos of crowded scenes captured at a distance using a number of "floor fields" that describe how a pedestrian should move based on scene-wide constraints. Their sink seeking and boundary floor fields, which describe the exit points and motion-boundary areas of the frame, respectively, impose a single likely direction at each spatial location of the video. In contrast, we model the crowd motion as a temporally evolving system that allows for any number of likely movements in each space-time location of the video. Though their dynamic floor field allows for some variation with the moving pedestrian, it is computed using frames after the current tracking position. Our approach uses the previously observed frames of the video to predict the motion of the target, allowing it to operate online. We compare the results of our method to that of Ali and Shah in Section 6.

Rodriguez et al. [24] use a topical model to represent motion in different directions at each spatial location. They represent a local motion in the form of a quantized optical flow vector at each spatial location of the video. This approach imposes a fixed number of motion directions at each spatial location, but disregards the temporal relationship between sequentially occurring local motions that videos of crowded scenes naturally exhibit. We explicitly encode this relationship with an HMM at each spatial location of the video. In addition, Rodriguez et al. quantize the optical flow vectors into 10 possible directions, while our approach uses a full distribution of optical flow for a

more robust and descriptive representation of the motion of tracked targets. Such a coarse quantization limits tracking in crowded scenes to only a few directions. A finer quantization results in diminishing returns (as discussed by Rodriguez et al. [24]). We directly compare the results of our method to that of Rodriguez et al. [24] in Section 6.

The relationship between motion patterns has been used to improve the tracking of motion boundaries. Nestares and Fleet [21] assume that spatially neighboring motion patterns are similar. They increase the continuity of the motion boundary tracking from Black and Fleet [7] by using motion patterns from the local area surrounding the tracked region. In videos of extremely crowded scenes, however, adjacent pedestrians may move in completely different directions, and thus the neighboring local motion patterns that they exhibit are dissimilar. In contrast, we track pedestrians by modeling the temporal relationship between sequentially occurring local spatio-temporal motion patterns.

3 Spatio-Temporal Model of the Crowd Motion

We begin by representing the crowd motion as a statistical model of spatially and temporally varying local spatio-temporal motion patterns. Following our previous work [15], we model the motion occurring in local regions using a local spatio-temporal motion pattern, and then learn the spatially and temporally varying crowd motion by training a collection of HMMs on the local spatio-temporal motion patterns that comprise the video. This is in sharp contrast to previous bag-of-word models [24] that disregard the temporal relationship between sequentially occurring local motions.

We use a local spatio-temporal motion pattern to represent the motion occurring in local space-time regions of the video. As shown in Fig. 2a, the video is subdivided into nonoverlapping spatio-temporal subvolumes, or "cuboids," defined by a regular grid. We define the local spatio-temporal motion pattern in each cuboid by a 3D Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of spatio-temporal gradients. The parameters of the local spatio-temporal motion pattern are

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i}^{N} \nabla I_i, \tag{1}$$

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i}^{N} (\nabla I_i - \boldsymbol{\mu})(\nabla I_i - \boldsymbol{\mu})^T, \tag{2}$$

where ∇I_i is the 3D gradient vector (containing gradients estimated in the horizontal, vertical, and temporal directions) at pixel location i and N is the number of pixels in the cuboid. We extract this local spatio-temporal motion pattern at each space-time location of the video. Therefore, at spatial location n and temporal location t, the local spatio-temporal motion pattern O_t^n is represented by $\boldsymbol{\mu}_t^n$ and $\boldsymbol{\Sigma}_t^n$.

Other works [9], [14], [18] have used distributions of gradients to describe space-time volumes. Histograms of oriented gradients (HoG) features have been used for human detection [9] and action recognition [18]. The HoG feature is computed on the spatial gradients (the temporal gradient is not used), though they have been extended to the 3D space-time gradient [14]. Rather than modeling a distribution of the gradients' orientations, we use a Gaussian distribution to represent the shape of the 3D gradient points. This retains the capability of estimating the optical flow from the 3D gradients, as we show explicitly in Section 5.1.

The local spatio-temporal motion patterns in a specific location of the frame may change frequently (due to different pedestrians moving in different directions through the same spatial location) or gradually with the time of day (for instance, with the arrival of public transportation). As shown in Fig. 2b, we encode these temporal dynamics by training a HMM at each spatial location of the video. Temporally sequential local motion patterns are strongly related: either they are exhibited by two pedestrians moving in the same direction (i.e., creating a uniform crowd motion) or they represent a change in direction, both of which contribute to the overall crowd motion. Areas of the scene with frequently changing local motion patterns often contain pedestrians changing between a few different directions (e.g., pedestrians moving toward different exits). Such frequent changes between different local spatio-temporal motion patterns are represented by high transition likelihoods in the HMMs. While gradual variations (such as those caused by the arrival of public transportation) are not represented by the HMMs, they do cause variations in pedestrians' movements that impact the transition likelihoods in the HMM. If such events occur regularly, for instance, then the transition between the corresponding local movements of pedestrians would have a significant transition likelihood.

Complex motion representations are often quantized to reduce the computational complexity of HMMs. Quantizing the 3D Gaussians, however, would reduce the rich information that they represent. We retain our local spatio-temporal motion pattern representation (i.e., a 3D Gaussian distribution of spatio-temporal gradients) by modeling a novel HMM that can be trained directly on the 3D Gaussians. Specifically, the hidden states of the HMM are represented by a set of local spatio-temporal motion patterns $\{P_s | s = 1, \ldots, S\}$ that are themselves 3D Gaussian distributions. For a specific hidden state s, the probability of an observed motion pattern O_t^n (defined by $\boldsymbol{\mu}_t^n$ and $\boldsymbol{\Sigma}_t^n$) is

$$p(O_t^n | s) = \frac{1}{\sqrt{\pi\sigma_s^2}} \exp\left[\frac{-\tilde{d}(O_t^n, P_s)^2}{2\sigma_s^2}\right], \tag{3}$$

where P_s is the expected local spatio-temporal motion pattern, σ_s is the standard deviation, and $\tilde{d}(\cdot)$ is the Kullback-Leibler (KL) divergence [17]. The number of hidden states S, as well as an initial classification for each cuboid, is determined by an online clustering algorithm [15]. The expected motion pattern $P_s = \{\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\}$ is

$$\boldsymbol{\mu}_s = \frac{1}{N_s}\sum_{i}^{N_s} \boldsymbol{\mu}_i, \tag{4}$$

$$\boldsymbol{\Sigma}_s = \frac{1}{N_s}\sum_{i}^{N_s} (\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T) - \boldsymbol{\mu}_s \boldsymbol{\mu}_s^T, \tag{5}$$

where N_s is the number of local spatio-temporal motion patterns in the cluster s. The remaining parameters of the HMM are computed using the Baum-Welch algorithm [23].

By training an HMM at each scene location, we directly encode the multiple possible motions that can occur at each location and their temporal relationship with the HMM's transition probabilities. The KL-divergence-based densities emit the mean and covariance matrix of 3D Gaussians, retaining the rich motion information stored in the space-time gradients. This collection of novel HMMs allows us to represent the space-time dynamics of the crowd motion that exists in crowded scenes.

4 Predicting an Individual's Motion Patterns

We train our model of the crowd motion on a video of a crowded scene containing typical crowd behavior. Next, we use it to predict the local spatio-temporal motion patterns at each location of a different video of the same scene. Note that, since we create a scene-centric model based on the changing motion in local regions, the prediction is independent of which individual is being tracked. In fact, we predict the local spatio-temporal motion pattern at all locations of video volume given only the previous frames of the video.

Given a trained HMM at a specific spatial location and a sequence of observed local motion patterns $O_1, \ldots, O_{t-1}$ from the tracking video, we predict the local spatio-temporal motion pattern at the next time instance t. The predictive distribution of O_t is

$$p(O_t|O_1, \ldots, O_{t-1}) = \sum_{s \in S} p(O_t|s) \sum_{s' \in S} p(s|s')\hat{\alpha}_{t-1}(s'), \quad (6)$$

where S is the set of hidden states in the HMM, s and s' are state indices, and $\hat{\alpha}$ is the vector of scaled messages from the forwards-backwards algorithm [23]. We let $\gamma(s)$ represent the second summation in (6) that combines the transition distribution $p(s|s')$ between specific states s and s' and the posterior likelihoods $\hat{\alpha}$. The vector $\gamma(s)$ is defined by

$$\gamma(s) = \sum_{s' \in S} p(s|s')\hat{\alpha}(s'). \quad (7)$$

We predict the next local spatio-temporal motion pattern by taking the expected value of the predictive distribution (6). The predicted local spatio-temporal motion pattern $\widetilde{O}_t$ is

$$\widetilde{O}_t = \mathbb{E}[p(O_t|O_1, \ldots, O_{t-1})] = \sum_{s \in S} \mathbb{E}[p(O_t|s)]\gamma(s). \quad (8)$$

The expected value of the emission probability $p(O_t|s)$ is the local spatio-temporal motion pattern $P_s = \{\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\}$. Therefore, the predicted local spatio-temporal motion pattern (defined by $\widetilde{\boldsymbol{\mu}}_t$ and $\widetilde{\boldsymbol{\Sigma}}_t$) is a weighted sum of the set of 3D Gaussian distributions associated with the HMM's hidden states

$$\widetilde{\boldsymbol{\mu}}_t = \sum_{s \in S} \gamma(s)\boldsymbol{\mu}_s, \quad (9)$$

$$\widetilde{\boldsymbol{\Sigma}}_t = -\widetilde{\boldsymbol{\mu}}_t\widetilde{\boldsymbol{\mu}}_t^T + \sum_{s \in S} \gamma(s)\left(\boldsymbol{\Sigma}_s + \boldsymbol{\mu}_s\boldsymbol{\mu}_s^T\right), \quad (10)$$

where $\boldsymbol{\mu}_s$ and $\boldsymbol{\Sigma}_s$ are the mean and covariance of the hidden state s, respectively.

During tracking, we use the previous frames of the video to predict the local spatio-temporal motion pattern that spans the next M frames (where M is the number of frames in a cuboid). The computational complexity of HMMs is well known [6], but is increased by a constant factor due to use of the KL-divergence in (3) that requires an inversion of the 3×3 covariance matrices. Since the predictive distribution (6) is a function of the HMM's transition probabilities and the hidden states' posteriors, the prediction may be computed online and efficiently during the forward phase of the forwards-backwards algorithm [23]. Specifically, the forward phase computes the posterior likelihoods $\hat{\alpha}$ in (8) efficiently using dynamic programming, and may be computed incrementally. The forward phase used in (6) takes $O(S^2)$ for each prediction where S is the number of hidden states. We train the HMMs using Expectation Maximization [6], which repeats the forwards-backwards algorithm until convergence, and thus requires more time than the prediction.

5 Tracking Individuals

We now use the predicted local motion pattern to track individuals in a Bayesian framework. Specifically, we use the predicted local spatio-temporal motion pattern as a prior on the parameters of a particle filter. Our space-time model of the crowd motion, i.e., the collection of HMMs, enables these priors to vary in space-time and dynamically adapt to the changing motions within the crowd.

Tracking can be formulated in a Bayesian framework [11] by maximizing the posterior distribution of the state $\mathbf{x}_f$ of the target at time f given past and current measurements $\mathbf{z}_{1:f} = \{\mathbf{z}_i | i = 1 \ldots f\}$. Note that the index of each frame f is different from the temporal index t of the local spatio-temporal motion patterns (since the cuboids span many frames). We define state $\mathbf{x}_f$ as a 4D vector $[x, y, w, h]^T$ containing the tracked target's 2D location (in image space), width, and height, respectively. The posterior distribution is

$$p(\mathbf{x}_f|\mathbf{z}_{1:f}) \propto p(\mathbf{z}_f|\mathbf{x}_f) \int p(\mathbf{x}_f|\mathbf{x}_{f-1})p(\mathbf{x}_{f-1}|\mathbf{z}_{1:f-1})d\mathbf{x}_{f-1}, \quad (11)$$

where $\mathbf{z}_f$ is the frame at time f, $p(\mathbf{x}_f|\mathbf{x}_{f-1})$ is the transition distribution, $p(\mathbf{z}_f|\mathbf{x}_f)$ is the likelihood, and $p(\mathbf{x}_{f-1}|\mathbf{z}_{1:f-1})$ is the posterior from the previous tracked frame. The transition distribution $p(\mathbf{x}_f|\mathbf{x}_{f-1})$ models the motion of the target between frames $f-1$ and f, and the likelihood distribution $p(\mathbf{z}_f|\mathbf{x}_f)$ represents how well the observed image $\mathbf{z}_f$ matches the state $\mathbf{x}_f$. Often, the distributions are non-Gaussian, and the posterior distribution is estimated using a Markov chain Monte Carlo method such as a particle filter [11] (please refer to [3] for an introduction to particle filters).

As shown in Fig. 3, we impose priors on the transition $p(\mathbf{x}_f|\mathbf{x}_{f-1})$ and likelihood $p(\mathbf{z}_f|\mathbf{x}_f)$ distributions using the predicted local spatio-temporal motion pattern at the space-time location defined by $\mathbf{x}_{f-1}$. For computational efficiency, we use the cuboid at the center of the tracked target to define the priors, although the target may span several cuboids across the frame.

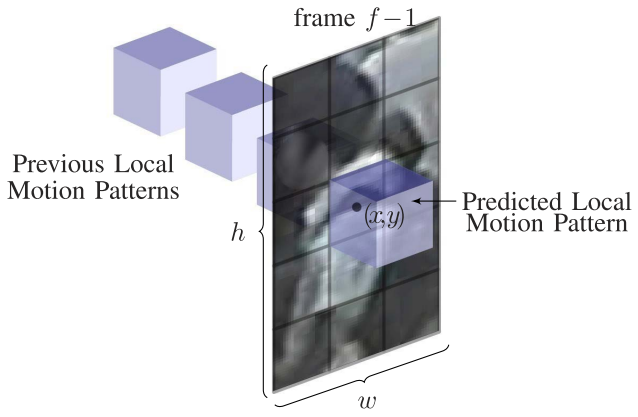

Fig. 3. The state vector $\mathbf{x}_{f-1} = [x, y, w, h]$ defines a window around the tracked target at image location (x, y) with dimensions (w, h). We use the predicted local spatio-temporal motion pattern at the center of the tracked target to estimate priors on the distribution of the next state $\mathbf{x}_f$.

5.1 Transition Distribution

We use the predicted local spatio-temporal motion pattern to hypothesize the motion of the tracked target between frames $f-1$ and f, i.e., the transition distribution $p(\mathbf{x}_f|\mathbf{x}_{f-1})$. Let the state vector $\mathbf{x}_f = [\mathbf{k}_f^T, \mathbf{d}_f^T]^T$ where $\mathbf{k}_f = [x, y]$ is the target's location (in image coordinates) and $\mathbf{d}_f = [w, h]$ the size (width and height) of a bounding box around the target. In this work, we focus on the target's movement between frames and use a second-degree autoregressive model [22] for the transition distribution of the size $\mathbf{d}_f$ of the bounding box.

The transition distribution of the target's location $p(\mathbf{k}_f|\mathbf{k}_{f-1})$ reflects the 2D motion of the target between frames $f-1$ and f. We model this as a normal distribution defined by

$$p(\mathbf{k}_f|\mathbf{k}_{f-1}) = \mathcal{N}(\mathbf{k}_f - \mathbf{k}_{f-1}; \boldsymbol{\omega}, \boldsymbol{\Lambda}), \tag{12}$$

where $\boldsymbol{\omega}$ is the 2D optical flow vector and $\boldsymbol{\Lambda}$ is its covariance matrix.

We estimate the parameters $\boldsymbol{\omega}$ and $\boldsymbol{\Lambda}$ from the predicted local spatio-temporal motion pattern at the space-time location containing $\mathbf{k}_{f-1}$, i.e., we predict the transition distribution using the crowd motion. The predicted motion pattern $\widetilde{O}$ (given by (8)) is defined by a mean gradient vector $\widetilde{\boldsymbol{\mu}}$ and a covariance matrix $\widetilde{\boldsymbol{\Sigma}}$ (we have dropped the n and t for notational convenience). The relationship between space-time gradients and optical flow has been well studied [25], [27], specifically in the form of the structure tensor matrix

$$\widetilde{\mathbf{G}} = \widetilde{\boldsymbol{\Sigma}} + \widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^T. \tag{13}$$

The 3D optical flow can be estimated [25] from the structure tensor by solving

$$\widetilde{\mathbf{G}}\mathbf{w} = \mathbf{0}, \tag{14}$$

where $\mathbf{w} = [u, v, z]^T$ is the 3D optical flow. Assuming that the change in time is 1, the 2D optical flow is $\boldsymbol{\omega} = [u/z, v/z]^T$. The 3D optical flow is the structure tensor's eigenvector with the smallest eigenvalue [27]. Thus, we compute the 2D optical flow $\boldsymbol{\omega}$ from the predicted local motion pattern by solving for the smallest eigenvector of $\widetilde{\mathbf{G}}$.

Next, we estimate the covariance matrix $\boldsymbol{\Lambda}$ of the 2D optical flow from the same predicted structure-tensor matrix $\widetilde{\mathbf{G}}$. As shown in Fig. 4, the 3D optical flow $\mathbf{w}$ is orthogonal to the plane containing the space-time gradients.

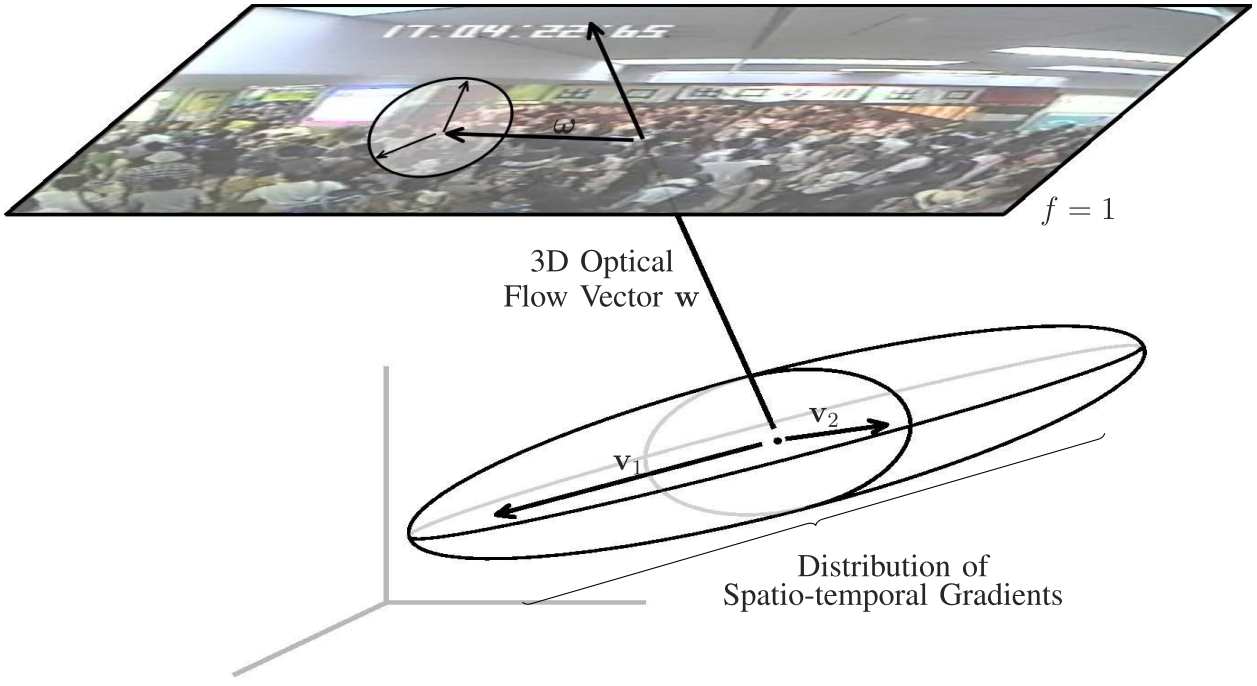

Fig. 4. We hypothesize a full distribution of optical flow for the tracked target using the local spatio-temporal motion pattern predicted by the crowd motion. The 3D optical flow $\mathbf{w}$ is estimated from the predicted local spatio-temporal motion pattern as the structure tensor's eigenvector with the smallest eigenvalue. The 2D optical flow $\boldsymbol{\omega}$ is its projection onto the plane $f = 1$ and represents the predicted movement of the target. We estimate the optical flow's variance by projecting the other two eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ onto the same plane, and by scaling them with respect to the eigenvalues.

The ordered eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of $\widetilde{\mathbf{G}}$ encode a confidence of the optical flow estimate [27]. If the video volume contains a single dominant motion vector and sufficient texture, then $\widetilde{\mathbf{G}}$ is close to rank 2 and $\lambda_3 \ll \lambda_1, \lambda_2$. We model the confidence in the optical flow estimate as inversely proportional to its variance, and thus consider the eigenvalues of $\widetilde{\mathbf{G}}$ inversely proportional to the eigenvalues of $\mathbf{\Lambda}$. If $\widetilde{\mathbf{G}}$ is close to full rank, then the gradients are not coplanar and the 3D optical flow may vary in the primary directions of the gradient distributions (i.e., the other eigenvectors of $\widetilde{\mathbf{G}}$). The optical flow vector $\boldsymbol{\omega}$ is a projection of the 3D optical flow $\mathbf{w}$ onto the plane $f = 1$. We let the eigenvectors of $\mathbf{\Lambda}$ be the projection of the other two eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ of $\widetilde{\mathbf{G}}$ onto the same plane. This encodes uncertainty in the primary directions of the spatio-temporal gradients. Therefore, we estimate $\mathbf{\Lambda}$ by

$$\mathbf{\Lambda} = \left[\mathbf{v}_1', \mathbf{v}_2'\right] \begin{bmatrix} \frac{\lambda_3}{\lambda_1} & 0 \\ 0 & \frac{\lambda_3}{\lambda_2} \end{bmatrix} \left[\mathbf{v}_1', \mathbf{v}_2'\right]^{-1}, \tag{15}$$

where $\mathbf{v}_1'$ and $\mathbf{v}_2'$ are the projections of $\mathbf{v}_1$ and $\mathbf{v}_2$ onto the plane $f = 1$. The 2D covariance matrix $\mathbf{\Lambda}$ implies a larger variance when the optical flow estimate is poor, and a smaller variance when the optical flow estimate is reliable.

In summary, we hypothesize the parameters to the transition distribution at the space-time location of the target based on the local spatio-temporal motion pattern predicted from the crowd motion. We use the rich motion information encoded in the space-time gradients of the predicted local motion pattern to hypothesize a full distribution of the 2D optical flow. By doing so, the state-transition distribution is robust to unreliable flow estimates and dynamically adapts to the space-time dynamics of the crowd.

5.2 Likelihood Distribution

Next, we use the predicted local motion pattern as a prior on the likelihood distribution $p(\mathbf{z}_f|\mathbf{x}_f)$. The density of the crowd makes it difficult to discern the target from other pedestrians. In addition, the individual's appearance may change significantly from frame to frame due to their natural body movement or partial occlusion from surrounding pedestrians. The change in a pedestrian's appearance between frames is directly related to the motion at their specific space-time location. Intuitively, as an individual walks through the scene, their body may exhibit movement in different directions [26] or partial occlusions may cause a motion boundary. We use the predicted local motion pattern to hypothesize the amount of variation the pedestrian's appearance will undergo, and use it to model the variance of the likelihood distribution $p(\mathbf{z}_f|\mathbf{x}_f)$ in (11). By doing so, we adjust the likelihood distribution to account for dramatic appearance changes predicted by the crowd motion.

Typical models of the likelihood distribution maintain a template T that represents the target's characteristic appearance in the form of a color histogram [22] or an image [2]. The likelihood distribution $p(\mathbf{x}_f|\mathbf{z}_{1:f})$ is then modeled using the template T and the region R (the bounding box defined by state $\mathbf{x}_f$) of the observed image $\mathbf{z}_f$ by

$$p\left(\mathbf{z}_f|\mathbf{x}_f\right) = \frac{1}{Z}\exp\left[\frac{-d(R,T)^2}{2\sigma^2}\right], \tag{16}$$

where σ is the variance, $d(\cdot)$ is a distance measure, and Z is a normalization constant. Rather than using color histograms [22] or intensity [2] as the defining characteristic of an individual's appearance, we model the template T as an image of the individual's spatio-temporal gradients. This representation is more robust to appearance variations caused by noise or illumination changes. We define the distance measure $d(\cdot)$ as a weighted sum of the angles between the spatio-temporal gradient vectors in the observed region and the template:

$$d(R,T) = \sum_i^M \rho_i^f \arccos(\mathbf{t}_i \cdot \mathbf{r}_i), \tag{17}$$

where M is the number of pixels in the template, $\mathbf{t}_i$ is the normalized spatio-temporal gradient vector in the template, $\mathbf{r}_i$ is the normalized spatio-temporal gradient vector in the region R of the observed image at frame f, and ρ_i^f is the weight of the pixel at location i and frame f,

We model changes in the target's appearance by estimating the weights $\{\rho_i^f | i = 1, \ldots, M\}$ in (17) during tracking. Specifically, pixels that change drastically (due to the pedestrian's body movement or partial occlusions) exhibit a large error between the template and the observed region. We estimate this error E_i^f during tracking to account for a pedestrian's changing appearance. The error at frame f and pixel i is

$$E_i^f = \alpha \arccos(\mathbf{t}_i \cdot \mathbf{r}_i) + (1-\alpha)E_i^{f-1}, \tag{18}$$

where α is the update rate (set to 0.05) and $\mathbf{t}_i$ and $\mathbf{r}_i$ are again the gradients of the template and observed region, respectively. To reduce the contributions of frequently changing pixels to the distance measure, the weight at frame f and pixel i is inversely proportional to the error

$$\rho_i^f = \frac{1}{Z}\left(\pi - E_i^{f-1}\right), \tag{19}$$

where Z is a normalization constant such that $\sum_i \rho_i^f = 1$.

Finally, we hypothesize the variance σ^2 (from (16)) of the likelihood to account for further variations in the pedestrian's appearance (i.e., due to the individual's natural body movement or partial occlusions). This is in contrast to our previous method [16] that assumes a constant variance and uses color information to reduce the sensitivity to appearance changes. Intuitively, tracking a pedestrian whose appearance changes frequently requires a larger variance to account for more variation and avoid drift. Alternatively, tracking a pedestrian whose appearance changes infrequently benefits from a smaller variance for more accurate results. Similarly to the transition distribution, we hypothesize σ^2 using the predicted local spatio-temporal motion pattern at the space-time location that the state $\mathbf{x}_{f-1}$ occurs. The covariance of the predicted local spatio-temporal motion pattern is a 3×3 matrix,

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix}, \tag{20}$$

(a) (b) (c) (d)

Fig. 5. Frames from videos of four different scenes on which we evaluate our method. The first two scenes, (a) and (b), are from a crowded subway station during rush hour and contain pedestrians moving in any number of different directions. The faces in the train station scenes have been blurred to conceal identities. The third scene (c) from the UCF Crowd Data set [1] contains pedestrians avoiding one another. The final scene (d), also from the UCF Crowd Data set, contains pedestrians walking across an intersection in two different directions.

of spatio-temporal gradients. We consider σ_{zz} (i.e., the variance of the temporal gradient) representative of the amount of appearance change that occurs within the cuboid: Temporal gradients from frequent appearance changes have a larger variance. The degree of change, however, depends on the appearance of the specific target. Intuitively, targets with a large amount of texture will naturally exhibit local spatio-temporal motion patterns with larger values of σ_{zz}. Thus, we consider σ_{zz} proportional to the variance σ^2 of the likelihood distribution, and estimate it by

$$\sigma^2 = C\sigma_{zz}^2, \tag{21}$$

where C is a scaling factor selected empirically (typically, results are satisfactory letting $C = 50$).

To summarize, we use the predicted local spatio-temporal motion pattern as a prior on the variance of the likelihood distribution. Intuitively, targets that exhibit larger variations in their motion will have larger variance. Using both our hypothesized state transition distribution and likelihood distribution, we track individuals in videos of extremely crowded scenes based on the crowd motion.

6 Results

We evaluate our method on videos[1] of four scenes: two real world extremely crowded scenes[2] and two videos from the UCF Crowd Data set [1], frames of which are shown in Fig. 5. The first, which is from a train station's concourse, contains a large number of pedestrians moving in different directions. The second, from the station's ticket gate, contains pedestrians moving through turnstiles and then changing directions after exiting the gate. The faces in the train station scenes have been blurred to conceal identities. The third scene contains many pedestrians captured in a relatively close view on a sidewalk moving in a number of different directions. The fourth contains pedestrians moving in different directions as they cross an intersection. We use a sampling importance resampling particle filter as in [11] with 100 particles to estimate the posterior in (11). The computation of the transition distribution in (14) and (15) require a singular value decomposition on a 3×3 matrix which may be done efficiently using any standard numerical technique. Our implementation of the tracker runs at roughly 10 frames per second on an Intel Xeon X5355 2.66 GHz processor, though we have made no attempt to optimize the software.

We train a collection of HMMs on a video[3] of each scene, and use it to track pedestrians in videos of the same scene recorded at a different time. The training videos for each scene have 300, 350, 300, and 120 frames, respectively. The training videos for the concourse, ticket gate, and sidewalk scenes have a large number of pedestrians moving in a variety of directions. The video for the intersection scene has fewer frames due to the limited length of video available. In addition, the training video of the intersection scene contains only a few motion samples in specific locations as many of the pedestrians have moved to other areas of the scene in that point in time. Such sparse samples, however, still result in a useful model since most of the pedestrians are only moving in one of two directions (either from the lower left to the upper right, or from the upper right to the lower left).

Due to the perspective projection of many of the scenes, which is a common occurrence in surveillance, the sizes of pedestrians varies immensely. As such, the initial location and size of the targets are selected manually. Many methods exist for automatically detecting pedestrians and their sizes [9] even in crowded scenes [19]. We consider pedestrian detection beyond the scope of this paper, and focus solely on the problem of tracking from an initial size and location. We are confident, however, that any pedestrian detector designed for crowded scenes may be used to automatically initialize our method.

The motion represented by the local spatio-temporal motion pattern depends directly on the size of the cuboid. Ideally, we would like to use a cuboid size that best represents the characteristic movements of a single pedestrian. Cuboids the size of a single pedestrian would faithfully represent the pedestrian's local motion and therefore enable the most accurate prediction and tracking. The selection of the cuboid size, however, is entirely scene dependent since the relative size of pedestrians within the frame depends on the camera and physical construction of the scene. In addition, a particular view may capture pedestrians of different sizes due to perspective projection. We use a cuboid of size $10 \times 10 \times 10$ on all scenes so that a majority of the cuboids are smaller than the space-time

1. Please see the supplementary material for video results, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2011.173.

2. The original videos courtesy of Nippon Telegraph and Telephone Corporation.

3. Available in the online supplementary material for the training videos.

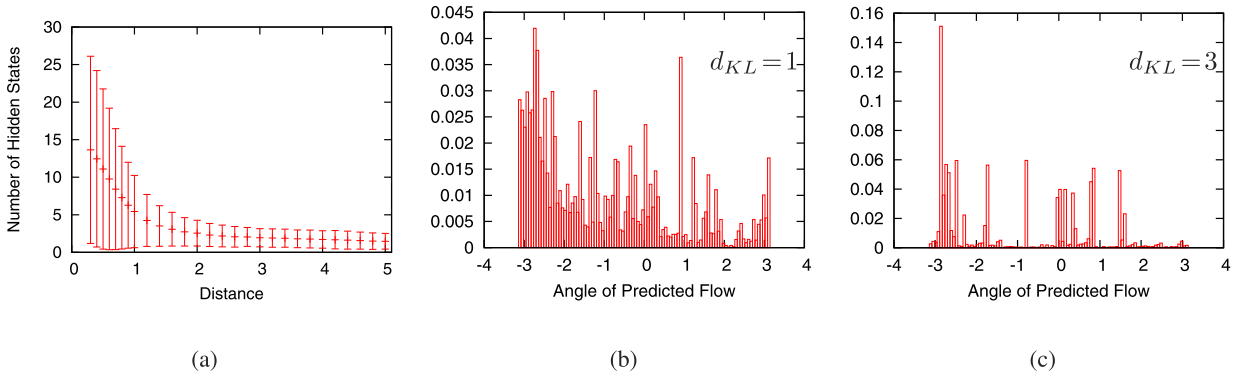

Fig. 6. Effects of varying the distance threshold d_{KL} on the training video of the concourse scene. (a) The average number of hidden states for all tubes using different values of d_{KL}. The bars show the standard deviation. The distribution of the direction of the predicted optical flow vector for (b) $d_{KL}=1$ and (c) $d_{KL}=3$. The number of states decreases with larger distance values and results in a more rigid model, as indicated by the sparse distribution of predicted flow directions.

region occupied by a moving pedestrian. By doing so, the cuboids represent the motion of a single pedestrian but still contain enough pixels to accurately estimate a distribution of spatio-temporal gradients. We will explore the automatic detection of cuboid sizes in future work, but note that, since the cameras recording the scenes are static, the sizes must be determined only once for each scene prior to training. Therefore, the cuboid sizes may be determined by a semi-supervised approach that approximates the perspective projection of the scene.

The accuracy of our predictions depends heavily on the number of hidden states in each HMM. The online clustering algorithm [15] uses a distance threshold d_{KL} to vary the number of hidden states depending on the local motion patterns within the video. Large variations in the flow may result in an excessive number of hidden states. For example, the same pedestrian moving at two different speeds may result in two different motion patterns, and such behavior from many pedestrians may vastly increase the number of hidden states (and therefore the complexity) of our model.

Fig. 6a shows the number of hidden states, averaged over all of the tubes in the video, for different values of d_{KL} on the training video from the concourse scene. Small values of d_{KL} result in a larger number of hidden states, but require a longer training video to achieve stable HMMs. As d_{KL} approaches 2, the number of hidden states stabilizes, suggesting a small number of distinct motion patterns in the video. Figs. 6b and 6c show the histograms of the predicted optical flow directions for the query video of the concourse scene using HMMs trained with $d_{KL}=1$ and $d_{KL}=3$, respectively. Larger values of d_{KL} impose a more limited number of directions on the transition distribution, as shown by the sparser distribution in Fig. 6c. We empirically select the value of d_{KL} near 1 for each training video to represent a large number of different motion patterns within each tube while ensuring sufficient flexibility and stability of the model.

We measure the accuracy of the predicted local spatio-temporal motion patterns using the angular error [4] between the optical flow estimated from the predicted structure tensor $\tilde{\mathbf{G}}$ and the optical flow estimated from the actual structure tensor matrix $\mathbf{G}$. Fig. 7 shows the angular error averaged over the entire video for each spatial location in all four scenes. Noisy areas with little motion, such as the concourse's ceiling, result in higher error due to the lack of reliable gradient information. High motion areas, however, have a lower error that indicates a successful prediction of the local spatio-temporal motion patterns. The sidewalk scene contains errors in scattered locations due to the occasional visible background in the videos and close view of pedestrians. There is a larger amount of error in high-motion areas of the intersection scene since a relatively short video was used for training.

Fig. 8 shows the predicted optical flow, colored by key in the lower left, for four frames from the sidewalk scene. Pedestrians moving from left to right are colored red, those moving right to left are colored green, and those moving

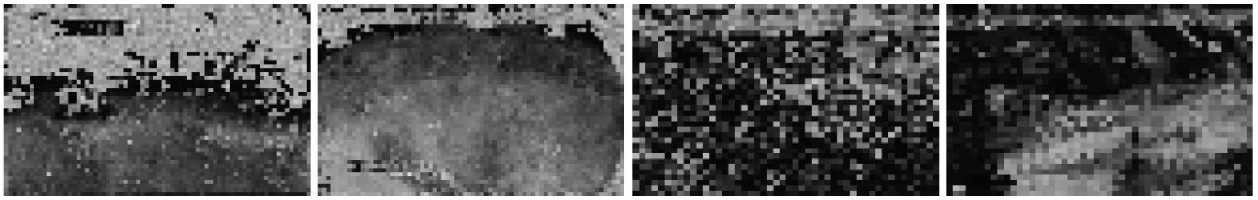

Fig. 7. The angular error between the predicted optical flow vector and the observed optical flow vector, averaged over all frames in the video, for the concourse, ticket gate, sidewalk, and intersection scenes, respectively. White indicates a high error for predicted motion patterns in areas with little texture, such as the ceiling of the train station. Areas in the concourse and ticket gate scenes where pedestrians are present, however, contain enough spatio-temporal gradient information for a successful prediction. The sidewalk scene contains errors due to the background occasionally being visible or pedestrians with little texture. The intersection scene contains larger errors in high motion areas of the frame due to the small amount of training data.

Fig. 8. Four frames from the sidewalk scene, colored to indicate the direction of the predicted optical flow using the key on the lower left. Our space-time model dynamically adapts to changes in the motion of pedestrians, as indicated by the changing directions of flow. Noise corresponds to poor predictions and occur in areas of little texture, such as the visible sidewalk.

Fig. 9. The direction of the predicted optical flow for the same spatial location in different frames of the sidewalk scene. Our temporally varying model accurately predicts the changing direction of pedestrians as the video progresses.

from the bottom of the screen to the top are colored blue. As time progresses, our space-time model dynamically adapts to the changing motions of pedestrians within the scene as shown by the changing cuboid colors over the frames. Poor predictions appear as noise, and occur in areas of little texture such as the visible areas of the sidewalks or pedestrians with little texture.

Fig. 9 shows a specific example of the changing predicted optical flow on six frames from the sidewalk scene. In the first two frames the predicted flow is from the left to the right, correctly corresponding to the motion of the pedestrian. In later frames the flow adjusts to the motion of the pedestrian at that point in time. Only by exploiting the temporal structure within the crowd motion are such dynamic predictions possible.

Fig. 10 shows frames of a target during tracking and the variance of the likelihood distribution that we hypothesize from the predicted local spatio-temporal motion patterns. The frames shown in the top row are indicated on the plot by blue boxes. When the target's appearance changes, for instance due to the partial occlusions on frames 56, 112, and 201, the variance is larger. The target's appearance has less variation in frames 29, 71, 154, and 171, and the hypothesized variance is also lower, allowing for more accurate tracking due to a narrower likelihood distribution.

Fig. 11 shows a visualization[4] of our tracking results on videos from each of the different scenes. Each row shows four frames of our method tracking different targets whose trajectories are shown up to the current frame by the colored curves. The different trajectories in the same spatial locations of the frame demonstrate the ability of our approach to capture the temporal motion variations of the crowd. For example, the green target in row 1 is moving in a completely different direction than the red and pink targets, although they share the spatial location where their trajectories intersect. Similarly, the pink, blue, red, and green targets in row 2 all move in different directions in the center part of the frame, yet our method is able to track each of these individuals. Such dynamic variations that we model using an HMM cannot be captured by a single motion model such as a "floor fields" [2]. Spatial variations are also handled by our approach, as illustrated by the targets concurrently moving in completely different directions in rows 5 and 6. In addition, our method is robust to partial occlusions as illustrated by the pink target in row 1, and the red targets in rows 3, 5, and 6.

Fig. 12 shows a failure case due to a severe occlusion. In these instances, our method begins tracking the individual that caused the occlusion. This behavior, though not desired, shows the ability of our model to capture multiple motion patterns since the occluding individual is moving in a different direction. Other tracking failures occur due to poor texture. In the sidewalk scene, for example, the occasional viewable background and lack of texture on the pedestrians cause poorly predicted local spatio-temporal motion patterns. On such occasions, a local motion pattern that describes a relatively static structure, such as black clothing or the street, is predicted for a moving target. This produces nonsmooth trajectories, such as the pink and red targets in row 5 or the red target in row 6 of Fig. 11.

Occasionally, an individual may move in a direction not captured by the training data. For instance, the pedestrian shown on the left of Fig. 13 is moving from left to right, a motion not present in the training data. Such cases are difficult to track since the space-time model cannot predict the pedestrian's motion. On such occasions, the posteriors (given in (7)) are near identical (since the emission

4. Available in the online supplementary material for video results.

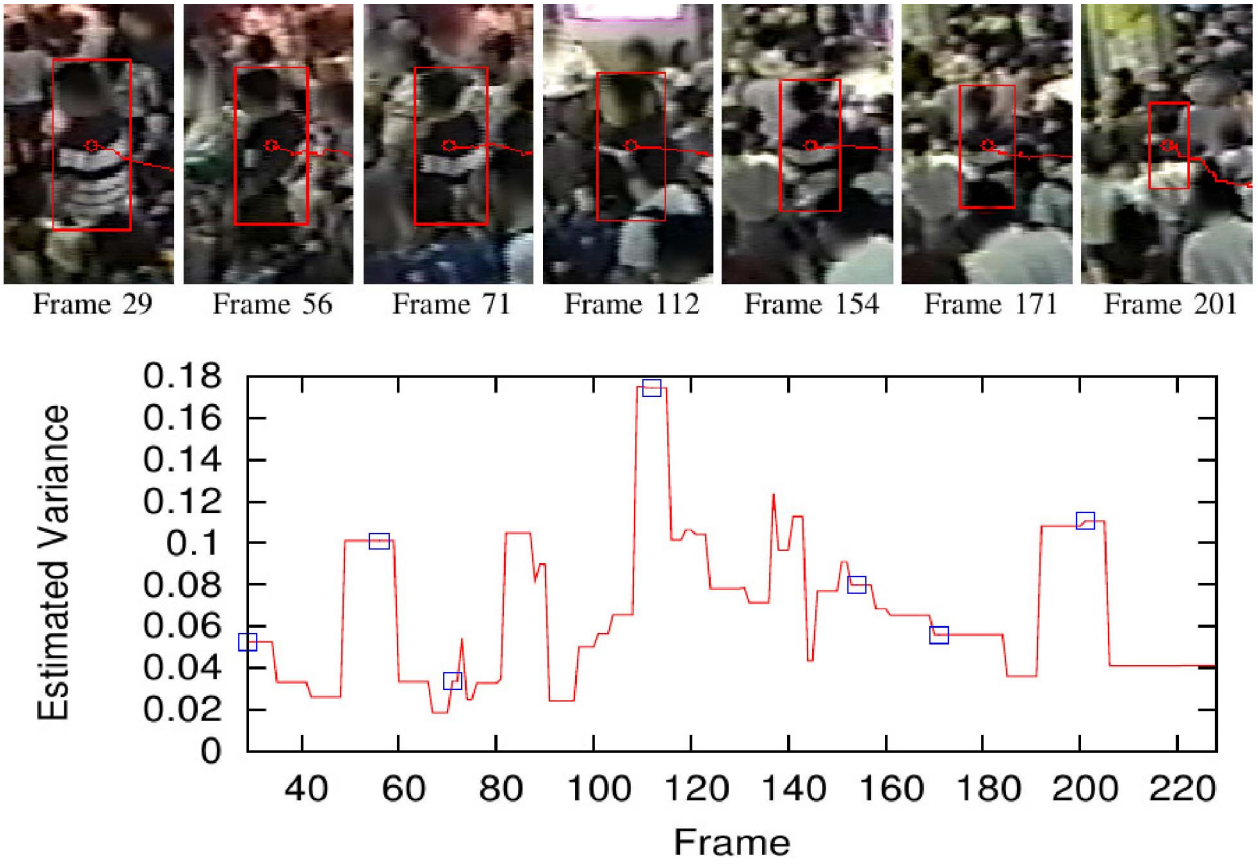

Fig. 10. The predicted variance of our likelihood distribution (bottom) and some sampled frames of the tracked target (top). When the target's appearance changes due to partial occlusions, the variance increases, creating a smoother likelihood function of the target's appearance.

probabilities are all close to 0), and thus the predicted optical flow is unreliable. This does not mean the targets cannot be tracked, as shown by the correct trajectories in Fig. 13, but the accuracy of the tracking depends heavily on the target's unique appearance.

We hand labeled ground truth tracking results for 40 targets, 10 from each scene, to quantitatively evaluate our approach. Each target is tracked for at least 120 frames. The ground truth includes the target's position and the width and height of a bounding box. The concourse and ticket gate scenes contain many pedestrians whose lower bodies are not visible at all over the duration of the video. On such occasions, the ground truth boxes are set around the visible torso and head of the pedestrian. Given the ground truth state vector $\mathbf{y}_t$, we measure the error of the tracking result $\hat{\mathbf{x}}_t$ as $\|\mathbf{y}_t - \hat{\mathbf{x}}_t\|_2$.

Fig. 14 shows the error of our method for each labeled target, averaged over all of the frames in the video, compared to a particle filter using a color-histogram likelihood and second-degree autoregressive model [22] (labeled as Pérez). In addition, we show the results using our predicted state-transition distribution with a color-histogram likelihood (labeled as Transition Only). On many of the targets our state transition distribution is superior to the second-degree autoregressive model, though nine targets have a higher error. Our full approach improves the tracking results dramatically and consistently achieves a lower error than that of Pérez et al. [22].

Fig. 15 compares our approach with the "floor fields" method by Ali and Shah [2] and the topical model from Rodriguez et al. [24]. Since the other methods do not change the target's template size, we only measured the error in the x, y location of the target. Our approach more accurately tracks the pedestrian's locations in all but a few of the targets. The single motion model by Ali and Shah completely loses many targets that move in directions not represented by their single motion model. The method of Rodriguez et al. [24] models multiple possible movements, and thus achieves better results than that of Ali and Shah, but is still limited since it does not include temporal information. Our temporally varying model allows us to track pedestrians in scenes that exhibit dramatic variations in the crowd motion.

Fig. 16 compares our results to two variants of our method. First, we compute the transition distribution from all of the cuboids spanned by the pedestrian (All Cuboids), rather than simply the center cuboid. All of the predicted motion patterns spanned by the pedestrian's bounding box are averaged using Gaussian weights. The results using all cuboids are negligibly different for the increased computational cost. This is primarily due to the fact that the cuboid at the center of the pedestrian is less likely to be corrupted by their articulated motion or a partial occlusion. Second, we compare to a model that only uses optical flow vectors. We train a collection of HMMs on the dominant optical flow vector estimated

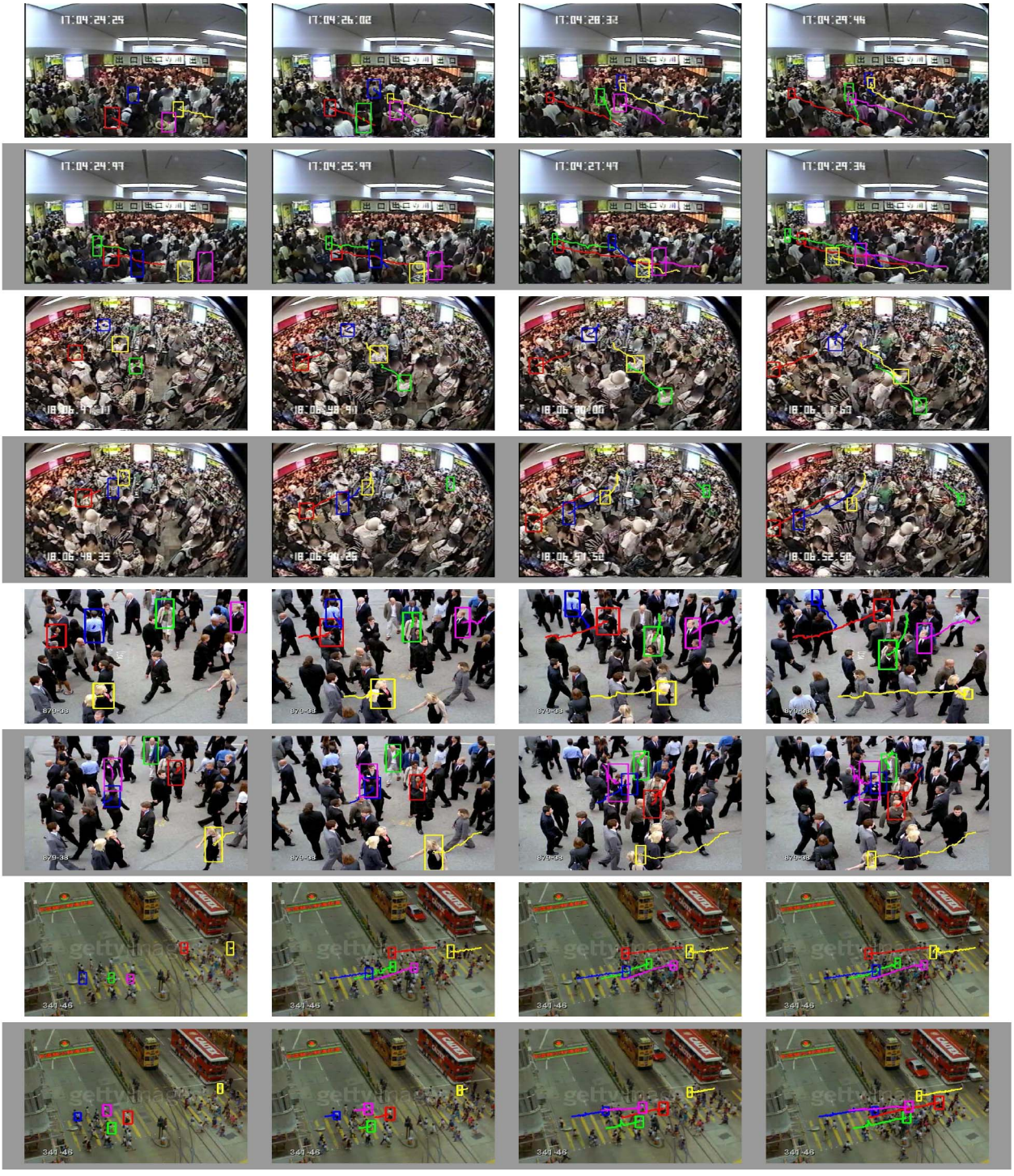

Fig. 11. Frames showing our method of tracking pedestrians in videos of the concourse, ticket gate, sidewalk, and intersection scenes, respectively. The faces in the train station scenes have been blurred to conceal identities. The video sequence in each row progresses from left to right, and each curve shows the target's trajectory up to the current frame. Please see the online supplementary material for video results.

Fig. 12. Four frames showing a tracking failure due to a severe occlusion. Our method begins tracking the occluding pedestrian. The faces have been blurred to conceal identities.

Fig. 13. Occasionally, pedestrians move in directions not in the training data, such as the one moving from left to right in the concourse scene (left) or reversing direction in the ticket gate scene (right). On such occasions, the predicted optical flow is unreliable and tracking entirely depends on the likelihood distribution. Our method is able to track such pedestrians, as shown above, but is less robust to partial occlusions and appearance changes.

from each cuboid. The number of hidden states is set to the same number of hidden states as our motion pattern HMMs, and the emission distributions are multivariate Gaussians. Such a model does not retain enough information to hypothesize a covariance matrix of the optical flow or the variance of the likelihood, and thus has higher tracking errors. The optical flow model performed well on many targets in the intersection scene since most of the pedestrians moved in one of two directions.

Fig. 17 shows the tracking error over time, averaged over all of the targets, using our approach, that of Ali and Shah [2], and that of Rodriguez et al. [24]. The consistently lower error achieved by our approach indicates that we may track subjects more reliably over a larger number of frames. Our temporally varying model accounts for a larger amount of directional variation exhibited by the targets, and enables accurate tracking over a longer period of time.

7 Conclusion

In this paper, we use a novel, space-time model of the crowd motion as a set of priors for tracking individuals in videos of extremely crowded scenes. We have shown that by modeling the crowd motion as a space-time dynamical system, we may represent scenes containing pedestrians whose speed and direction vary spatially across the frame and temporally over the video. Furthermore, our novel statistical models of local spatio-temporal motion patterns provide the flexibility to hypothesize full distributions of optical flow, as well as the amount of variation in the pedestrians appearance, both of which increase robustness while tracking in videos of crowded scenes. Our results show that our collection of HMMs is able to accurately predict the changing motion of pedestrians by encoding the scene's underlying temporal dynamics, and that our method is more accurate than those that do not model the space-time variations in the crowd motion. In addition, we have shown that the crowd motion may be learned from a

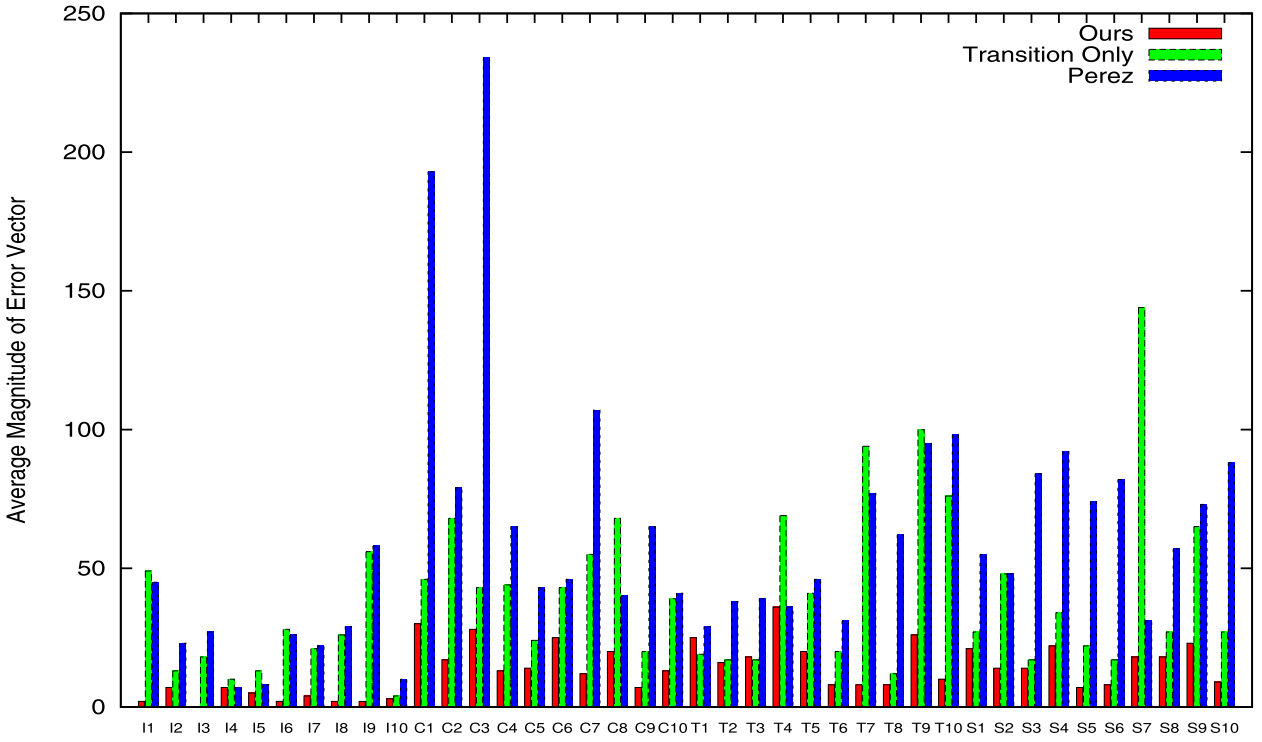

Fig. 14. The average magnitude of the error vector using our method, a color-based particle filter with a second-degree autoregressive model [22] (Pérez) and using only our transition distribution with a color-based likelihood (TransOnly) for targets in the concourse (C), ticket gate (T), sidewalk (S), and intersection (I) scenes. Our method achieves consistently lower errors.

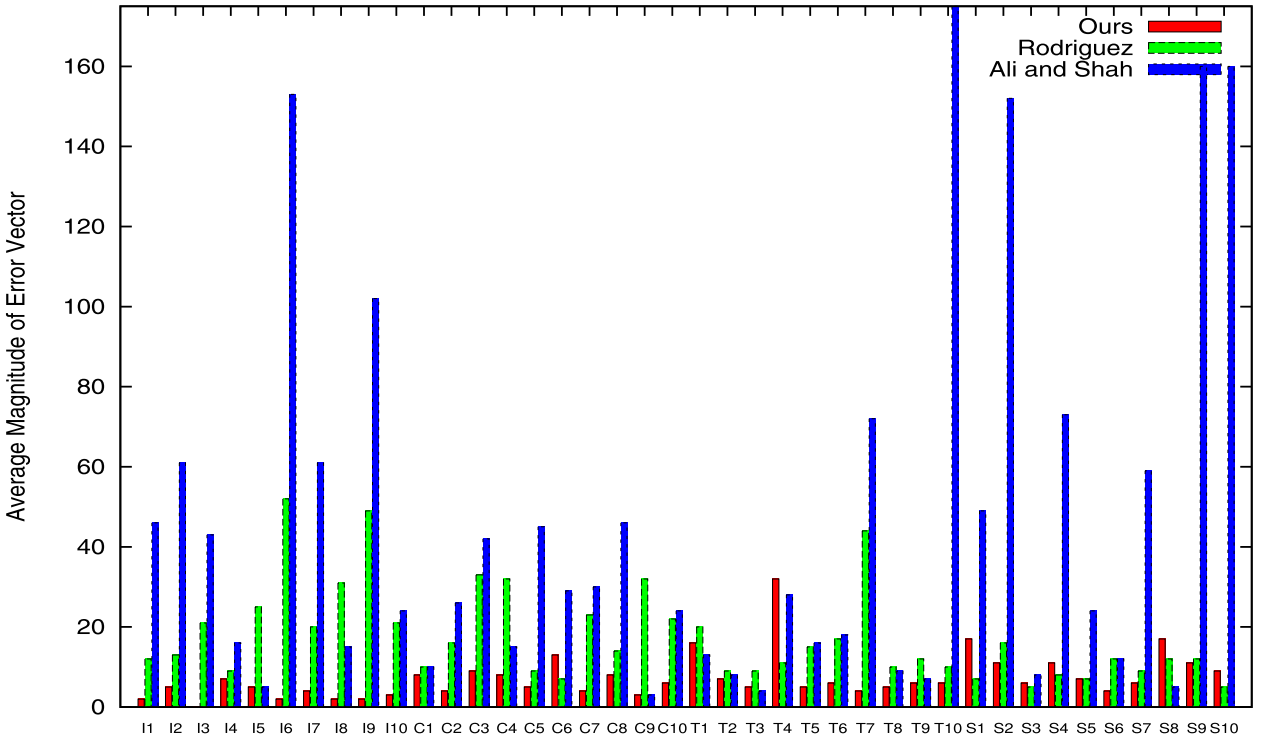

Fig. 15. The magnitude of the error vector for 40 targets, 10 from each scene, using our approach, that of Ali and Shah [2], and that of Rodriguez et al. [24]. Our approach achieves a lower error on almost all targets.

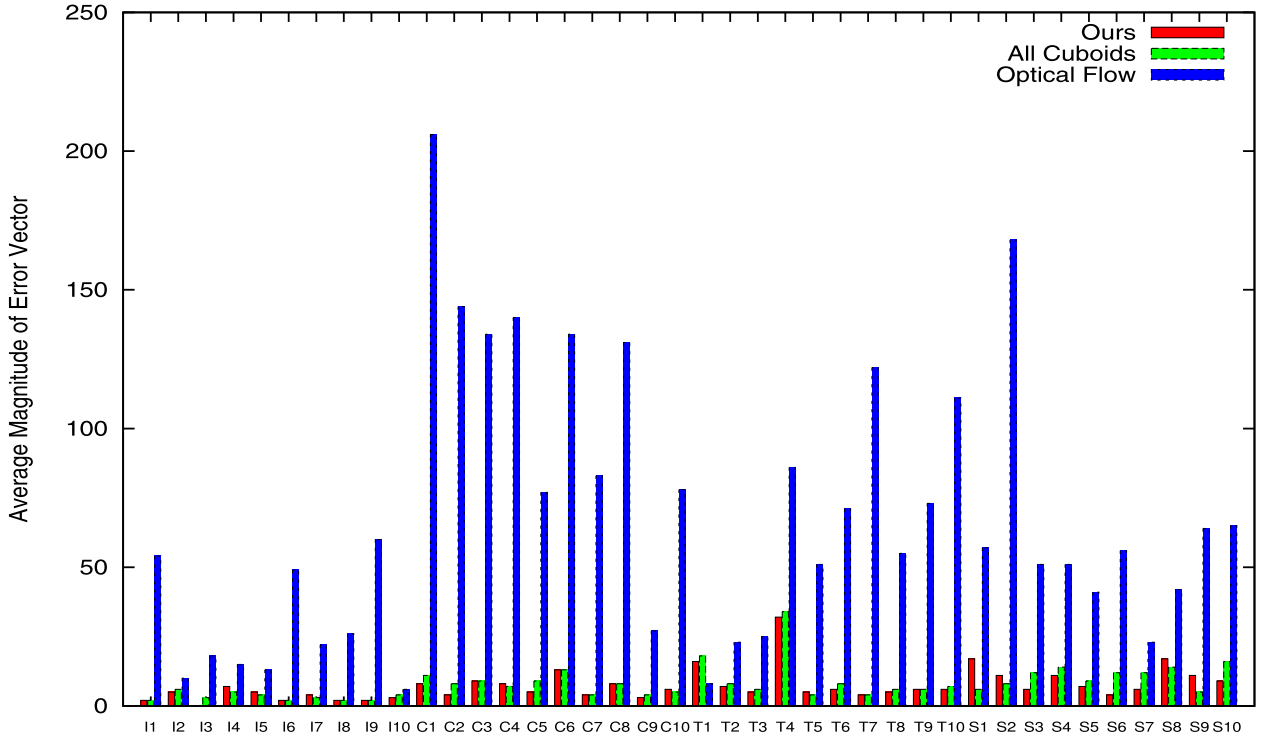

Fig. 16. The average magnitude of the error vector using our approach, our approach using all cuboids the target spans (All Cuboids), and using only an optical-flow based prediction (Optical Flow). No discernible benefit is achieved by using all of the cuboids a pedestrian spans. HMMs trained on the optical flow vectors do not retain the rich distribution of flow within the local spatio-temporal motion patterns, and result in high errors.

typical video of the scene prior to tracking, and presented a method that may operate efficiently and online. We believe that our space-time model may be further leveraged to provide robustness to failure cases, such as severe occlusions, and are exploring such possibilities for future work. Finally, we view tracking as only a single step in understanding and automatically analyzing videos of crowded scenes, a challenge that we believe our space-time model of local spatio-temporal motion patterns may further address.

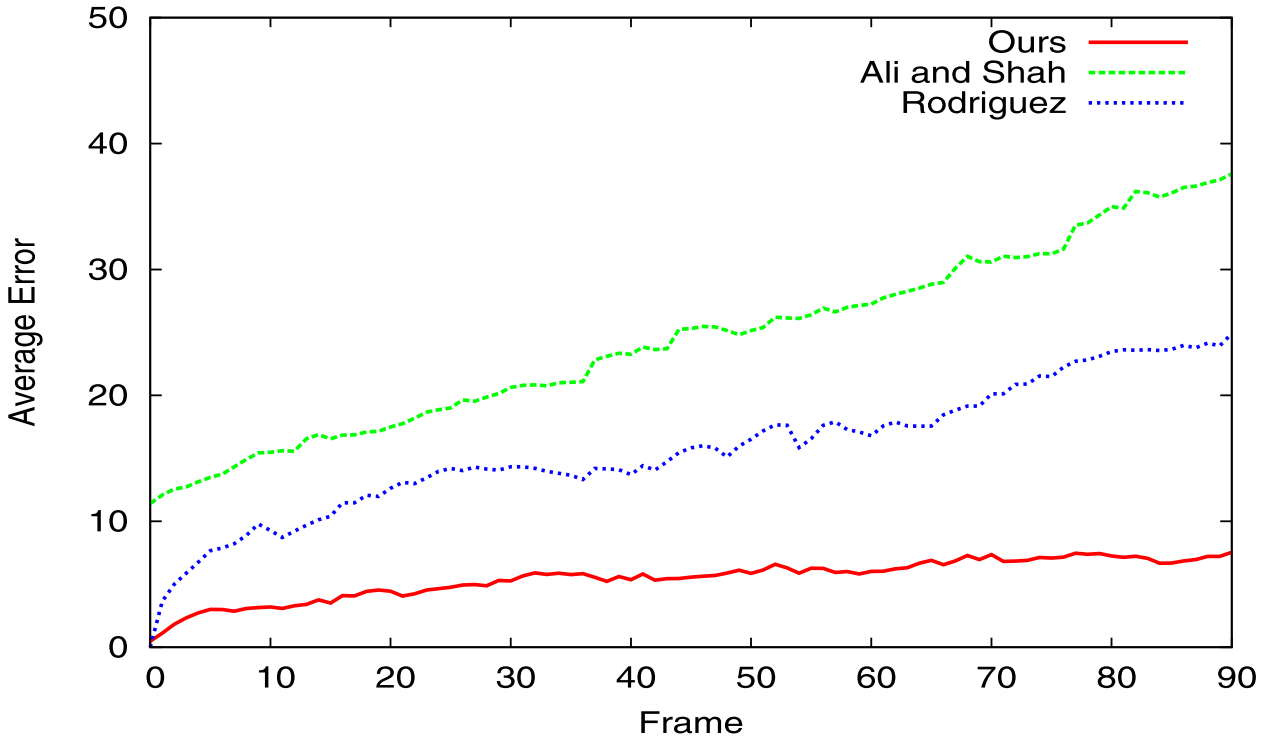

Fig. 17. The average error of all 40 targets over time using our approach, that of Ali and Shah [2], and that of Rodriguez et al. [24]. Our approach maintains a consistently lower error.

ACKNOWLEDGMENTS

This work was supported in part by US National Science Foundation grants IIS-0746717 and IIS-0803670, and Nippon Telegraph and Telephone Corporation. The authors thank Nippon Telegraph and Telephone Corporation for providing the train station videos.

REFERENCES

[1] S. Ali and M. Shah, "A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* pp. 1-6, 2007.

[2] S. Ali and M. Shah, "Floor Fields for Tracking in High Density Crowd Scenes," *Proc. European Conf. Computer Vision,* 2008.

[3] S.M. Arulampalam, S. Maskell, and N. Gordon, "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking," *IEEE Trans. Signal Processing,* vol. 50, no. 2, pp. 174-188, Feb. 2002.

[4] J. Barron, D. Fleet, and S. Beauchemin, "Performance of Optical Flow Techniques," *Int'l J. Computer Vision,* vol. 12, no. 1, pp. 43-77, 1994.

[5] M. Betke, D.E. Hirsh, A. Bagchi, N.I. Hristov, N.C. Makris, and T.H. Kunz, "Tracking Large Variable Numbers of Objects in Clutter," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* pp. 1-8, 2007.

[6] C.M. Bishop, *Pattern Recognition and Machine Learning.* Springer, Oct. 2007.

[7] M.J. Black and D.J. Fleet, "Probabilistic Detection and Tracking of Motion Boundaries," *Int'l J. Computer Vision,* vol. 38, no. 3, pp. 231-245, July/Aug. 2000.

[8] G.J. Brostow and R. Cipolla, "Unsupervised Bayesian Detection of Independent Motion in Crowds," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 594-601, June 2006.

[9] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2005.

[10] C. Hue, J.-P. Le Cadre, and P. Pérez, "Posterior Cramer-Rao Bounds for Multi-Target Tracking," *IEEE Trans. Aerospace and Electronic Systems,* vol. 42, no. 1, pp. 37-49, Jan. 2006.

[11] M. Isard and A. Blake, "CONDENSATION-Conditional Density Propagation for Visual Tracking," *Int'l J. Computer Vision,* vol. 29, no. 1, pp. 5-28, Aug. 1998.

[12] Z. Khan, T. Balch, and F. Dellaert, "MCMC-Based Particle Filtering for Tracking a Variable Number of Interacting Targets," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 11, pp. 1805-1819, Nov. 2005.

[13] Z. Khan, T. Balch, and F. Dellaert, "MCMC Data Association and Sparse Factorization Updating for Real Time Multitarget Tracking with Merged and Multiple Measurements," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 12, pp. 1960-1972, Oct. 2006.

[14] A. Kläser, M. Marszałek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," *Proc. British Machine Vision Conf.,* pp. 995-1004, 2008.

[15] L. Kratz and K. Nishino, "Anomaly Detection in Extremely Crowded Scenes Using Spatio-Temporal Motion Pattern Models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1446-1453, 2009.

[16] L. Kratz and K. Nishino, "Tracking with Local Spatio-Temporal Motion Patterns in Extremely Crowded Scenes," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition,* 2010.

[17] S. Kullback and R.A. Leibler, "On Information and Sufficiency," *The Annals of Math. Statistics,* vol. 22, no. 1, pp. 79-86, 1951.

[18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.

[19] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian Detection in Crowded Scenes," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* June 2005.

[20] Y. Li, C. Huang, and R. Nevatia, "Learning to Associate: HybridBoosted Multi-Target Tracker for Crowded Scene," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2009.

[21] O. Nestares and D.J. Fleet, "Probabilistic Tracking of Motion Boundaries with Spatiotemporal Predictions," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 358-365, 2001.

[22] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-Based Probabilistic Tracking," *Proc. European Conf. Computer Vision,* pp. 661-675, 2002.

[23] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE,* vol. 77, no. 2, pp. 257-286, Feb. 1989.

[24] M. Rodriguez, S. Ali, and T. Kanade, "Tracking in Unstructured Crowded Scenes," *Proc. 12th IEEE Int'l Conf. Computer Vision,* 2009.
[25] E. Shechtman and M. Irani, "Space-Time Behavior Based Correlation," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 405-412, 2005.
[26] D. Sugimura, K. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Using Individuality to Track Individuals: Clustering Individual Trajectories in Crowds Using Local Appearance and Frequency Trait," *Proc. 12th IEEE Int'l Conf. Computer Vision,* 2009.
[27] J. Wright and R. Pless, "Analysis of Persistent Motion Patterns Using the 3D Structure Tensor," *Proc. IEEE Workshop Motion and Video Computing,* pp. 14-19, 2005.
[28] B. Wu and R. Nevatia, "Tracking of Multiple, Partially Occluded Humans Based on Static Body Part Detection," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 951-958, 2006.
[29] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and Tracking of Multiple Humans in Crowded Environments," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 7, pp. 1198-1212, July 2008.

Louis Kratz received the BS and MS degrees in computer science from Drexel University in 2006. He is currently working toward the PhD degree at Drexel University. His research interests include machine learning and computer vision, with specific interests in video analysis and surveillance. He is a student member of the IEEE.

Ko Nishino received the BE and ME degrees in information and communication engineering (ECE) and the PhD degree in information science (CS) from the University of Tokyo in 1997, 1999, and 2002, respectively. He joined the Department of Computer Science at Drexel University as an assistant professor in 2005 and was promoted to associate professor in 2011. Prior to joining Drexel University, he was a postdoctoral research scientist in the Department of Computer Science at Columbia University. His research interests include computer vision and its intersections with computer graphics, machine learning, and digital archaeology. The main focus of his research is on photometric and geometric modeling of real-world objects and scenes. He has published a number of papers on related topics, including physics-based vision, image-based modeling and rendering, geometry processing, and video analysis. He received the US National Science Foundation CAREER Award in 2008. He is a member of the IEEE and the ACM.