Tracking with Local Spatio-Temporal Motion Patterns in Extremely Crowded Scenes

Louis Kratz Ko Nishino
Department of Computer Science
Drexel University
{lak24, kon}@drexel.edu

Abstract

Tracking individuals in extremely crowded scenes is a challenging task, primarily due to the motion and appearance variability produced by the large number of people within the scene. The individual pedestrians, however, collectively form a crowd that exhibits a spatially and temporally structured pattern within the scene. In this paper, we extract this steady-state but dynamically evolving motion of the crowd and leverage it to track individuals in videos of the same scene. We capture the spatial and temporal variations in the crowd's motion by training a collection of hidden Markov models on the motion patterns within the scene. Using these models, we predict the local spatio-temporal motion patterns that describe the pedestrian movement at each space-time location in the video. Based on these predictions, we hypothesize the target's movement between frames as it travels through the local space-time volume. In addition, we robustly model the individual's unique motion and appearance to discern them from surrounding pedestrians. The results show that we may track individuals in scenes that present extreme difficulty to previous techniques.

1. Introduction

Tracking objects or people is a crucial step in video analysis with a wide range of applications including behavior modeling and surveillance. Such automatic video analysis, however, has been limited to relatively sparse scenes primarily due to the limitations of tracking techniques. Extremely crowded scenes present significant challenges to traditional tracking methods due to the large number of pedestrians within the scene. These types of high-density and cluttered scenes, however, are perhaps in the most need of video analysis since they capture a large population of pedestrians in public areas.

Fixed-view surveillance systems continuously capture

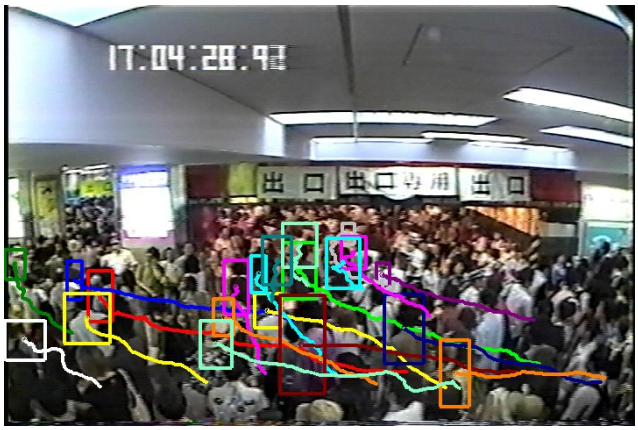

Figure 1. By modeling the spatio-temporal variations in the crowd's motion, we track pedestrians that exhibit significant variations in their motion and their appearance as they move through an extremely crowded scene.

the longitudinal variations of the motions that take place in the scene. These variations result from pedestrians exhibiting different speeds and directions as they traverse the dense crowd. Pedestrian movement in local areas, however, tends to repeat due to the large number of people in the video. As such, the steady-state motion within the scene, *i.e.*, the typical motions exhibited by pedestrians, may be learned from the large amount of video recorded by the camera and used to predict the movement of each individual.

In extremely crowded scenes, the motion of the crowd varies spatially across the frame and temporally throughout the video. The physical environment of the scene constricts the possible movements of pedestrians in specific areas, but the crowd varies naturally over short periods of time. As a result, the crowd's movement in one area of the frame may be completely different from that in another, and the pedestrian motion in the same area of the frame can change dramatically.

In this paper, we exploit the steady-state motion of the crowd, *i.e.*, the underlying structure formed by the spatial and temporal variations in the motion exhibited by pedes-

978-1-4244-6983-3/10/$26.00 ©2010 IEEE

trians, to track individuals in extremely crowded scenes. Our key insight is that the spatio-temporal variations of the crowd's motion can be learned from videos of the same scene to predict an individual's movement. Specifically, we leverage the crowd's motion as a set of priors in a particle filtering-based framework. First, we model the local spatio-temporal motion pattern in each sub-volume of the video defined by a regular grid, similar to [10]. We then train a hidden Markov model (HMM) on the motion patterns at each spatial location of the video to model the steady-state motion of the crowd. Using these HMMs, we predict the motion patterns a subject will exhibit as they move through a video of the same scene, effectively hypothesizing their movement based on the crowd's motion. We then compute a distribution of the subject's motion between frames based on the predicted motion pattern. Finally, we track the individual using their predicted motion pattern and by modeling their unique motion and appearance. We show that our approach faithfully captures the crowd's motion within the scene and provides superior tracking results compared to modern approaches on extremely crowded scenes.

2. Previous Work

Since the literature on tracking is extensive, we only review work that model motion in cluttered or crowded scenes. Previous work [6, 20] track features and associate similar trajectories to detect individual moving entities within crowded scenes. They assume that the subjects move in distinct directions, and thus disregard possible local motion inconsistencies between different body parts. As noted by Okabe *et al*. [20], such inconsistencies cause a single pedestrian to be detected as multiple targets. 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.

Data association techniques [4, 7, 12, 22] connect large numbers of short trajectories or detected targets for simultaneous tracking in cluttered scenes. These techniques use the collective tracking information from all targets to improve tracking accuracy, but assume the underlying detection is reliable. Extremely crowded scenes may contain over a hundred pedestrians within 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. In addition, our approach learns the scene's steady-state of motion prior to tracking, rather than combining separate tracking results.

Other work disregard the significant variations of pedestrian movement and appearance that can occur in crowded scenes. Ali and Shah [1] and Rodriguez *et al*. [18] model the motion of individuals across the frame in order to track pedestrians in crowds captured at a distance. Similarly, Pless *et al*. [9] learn a single motion distribution at each frame location from videos of automobile traffic. These approaches impose a fixed number of possible motions at each spatial location in the frame. In extremely crowded scenes, however, pedestrians in the same area of the scene may move in any number of different directions. We encode many possible motions in the HMM, and derive a full distribution of the motion at each spatio-temporal location in the video. In addition, natural body movements appear subtle when captured at a distance but create large appearance changes in near-view scenes, which we explicitly model.

Nestares and Fleet [14] also use neighboring motion patterns to improve tracking. They increase the continuity of the motion boundary tracking from Black and Fleet [5] by including multiple image neighborhoods. In our work, however, we use a dynamic temporal model of sequential motion patterns rather than assuming continuity across spatial locations.

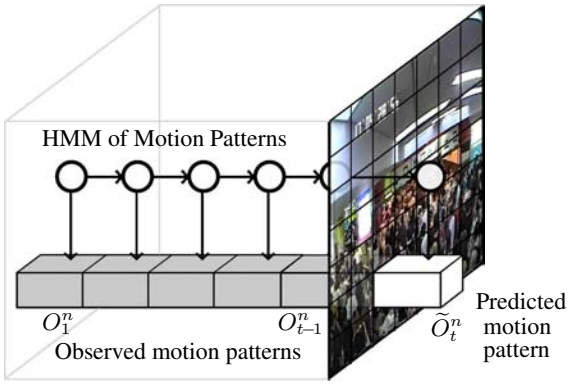

Figure 2. Following [10], the video is subdivided into local spatio-temporal volumes and an HMM is trained on the motion patterns at each spatial location to capture the steady-state motion of the crowd. Using the trained HMM and previously observed motion patterns $O_1^n, \ldots, O_{t-1}^n$, we predict the motion pattern $\widetilde{O}_t^n$ that describes how a target moves through the video.

3. Predicting Motion Patterns

The pedestrians in extremely crowded scenes collectively form a structure in the observed motions at and across different space-time locations. We capture the spatial and temporal variations of the crowd's motion using the local spatio-temporal motion pattern models introduced by Kratz and Nishino [10]. To model the motion in local areas, the video is subdivided into small sub-volumes, or cuboids, defined by a regular grid as shown in Fig. 2. The pedestrian motion through the cuboid at spatial location n and time t (i.e., the local spatio-temporal motion pattern O_t^n) is rep-

resented by a 3D Gaussian $\mathcal{N}\left(\boldsymbol{\mu}_t^n, \boldsymbol{\Sigma}_t^n\right)$ of spatio-temporal gradients

$$\boldsymbol{\mu}_t^n = \frac{1}{M}\sum_i^M \nabla I_i, \;\; \boldsymbol{\Sigma}_t^n = \frac{1}{M}\sum_i^M (\nabla I_i - \boldsymbol{\mu})\,(\nabla I_i - \boldsymbol{\mu})^T \;, \quad (1)$$

where ∇I_i is the 3D gradient at pixel i, and M is the number of pixels in the cuboid.

The spatio-temporal motion patterns at each spatial location vary over time, but tend to exhibit the Markov property. To encode these temporal variations, an HMM is trained on the motion patterns at each spatial location of the video as shown in Fig. 2. The hidden states of the HMM are represented by a set of motion patterns $\{P_s | s = 1, \ldots, S\}$ that are themselves 3D Gaussian distributions. The probability of an observed motion pattern O_t^n (defined by $\boldsymbol{\mu}_t^n$ and $\boldsymbol{\Sigma}_t^n$) given a hidden state s is

$$p\left(O_t^n | s\right) = p\left(\frac{d_{KL}\left(O_t^n, P_s\right)}{\sigma_s}\right) \sim \mathcal{N}\left(0, 1\right) \;, \quad (2)$$

where P_s is the expected motion pattern, σ_s is the standard deviation, and $d_{KL}\left(\cdot\right)$ is the Kullback-Leibler (KL) divergence [11]. Please see [10] for further details.

After training a collection of HMMs on a video of typical crowd motion, we predict the motion pattern at each space-time location that contains the tracked subject. Given the observed motion patterns $O_1^n, \ldots, O_{t-1}^n$ in the tracking video, the predictive distribution can be expressed by

$$p\left(O_t^n | O_1^n, \ldots, O_{t-1}^n\right) = \sum_{s \in S} p\left(O_t^n | s\right) \omega(s) \;, \quad (3)$$

where S is the set of hidden states, $\omega(s)$ is defined by

$$\omega\left(s\right) = \sum_{s' \in S} p\left(s | s'\right) \hat{\alpha}\left(s'\right) \;, \quad (4)$$

and $\hat{\alpha}$ is the vector of scaled messages from the forwards-backwards algorithm [17]. We compute the predicted local spatio-temporal motion pattern $\widetilde{O}_t^n$ by taking the expected value of the predictive distribution. The expected value becomes

$$\widetilde{O}_t^n = \mathbb{E}\left[p\left(O_t^n | O_1^n, \ldots, O_{t-1}^n\right)\right] = \sum_{s \in S} P_s \; \omega\left(s\right) \;, \quad (5)$$

where P_s is the hidden state's motion pattern from Eq. 2.

Eq. 5 is a weighted sum of the 3D Gaussian distributions associated with the HMM's hidden states. To solve for the predicted motion pattern $\widetilde{O}_t^n$ (defined by $\widetilde{\boldsymbol{\mu}}_t^n$ and $\widetilde{\boldsymbol{\Sigma}}_t^n$), we estimate the weighted expected centroid [13] of the hidden states. The predicted motion pattern becomes

$$\widetilde{\boldsymbol{\mu}}_t^n = \sum_{s \in S} \omega\left(s\right) \boldsymbol{\mu}_s \quad (6)$$

$$\widetilde{\boldsymbol{\Sigma}}_t^n = -\widetilde{\boldsymbol{\mu}}_t^n \left(\widetilde{\boldsymbol{\mu}}_t^n\right)^T + \sum_{s \in S} \omega\left(s\right) \left(\boldsymbol{\Sigma}_s + \boldsymbol{\mu}_s \boldsymbol{\mu}_s^T\right) \;, \quad (7)$$

where $\boldsymbol{\mu}_s$ and $\boldsymbol{\Sigma}_s$ are the mean and covariance of the hidden state s, respectively. Using the collection of HMMs that represents the steady-state motion of the crowd, we have predicted the local spatio-temporal motion patterns at each space-time location of the video.

4. Spatio-Temporal Transition Distribution

After predicting the distribution of spatio-temporal gradients, we use it to hypothesize the motion of the pedestrian from one frame to the next, *i.e.*, track the target. To achieve this, we use the gradient information to estimate the optical flow within each specific sub-volume and track the target in a Bayesian framework. In addition, we estimate the optical flow's variance to compute a full distribution of the subject's movement between frames.

Bayesian tracking [8] can be formulated as maximizing the posterior distribution of the state $\mathbf{x}_t$ of the target at time t given available measurements $\mathbf{z}_{1:t} = \{\mathbf{z}_i, i = 1 \ldots t\}$ by

$$p(\mathbf{x}_t | \mathbf{z}_{1:t}) \propto p(\mathbf{z}_t | \mathbf{x}_t) \int p(\mathbf{x}_t | \mathbf{x}_{t-1})\, p(\mathbf{x}_{t-1} | \mathbf{z}_{1:t-1})\, d\mathbf{x}_{t-1}, \quad (8)$$

where $\mathbf{z}_t$ is the image at time t, $p\left(\mathbf{x}_t | \mathbf{x}_{t-1}\right)$ is the transition distribution, and $p\left(\mathbf{z}_t | \mathbf{x}_t\right)$ is the likelihood. We model the state vector $\mathbf{x}_t$ as the width, height, and 2D location of the target within the image. In this work, we focus on the target's movement between frames and use a 2nd-degree autoregressive model [16] for the transition distribution of the target's width and height.

Ideally, the state transition distribution $p\left(\mathbf{x}_t | \mathbf{x}_{t-1}\right)$ directly reflects the two-dimensional motion of the target between frames $t-1$ and t. Thus we formulate the state transition distribution as

$$p\left(\mathbf{x}_t | \mathbf{x}_{t-1}\right) = \mathcal{N}\left(\mathbf{x}_t - \mathbf{x}_{t-1}; \boldsymbol{\omega}, \boldsymbol{\Lambda}\right) \;, \quad (9)$$

where $\boldsymbol{\omega}$ is the 2D optical flow vector, and $\boldsymbol{\Lambda}$ is the covariance matrix. We estimate these parameters from the predicted local spatio-temporal motion pattern at the space-time location defined by $\mathbf{x}_{t-1}$, *i.e.*, we predict the state transition distribution using the motion of the crowd.

The predicted motion pattern $\widetilde{O}_t^n$ 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 motion information encoded in the spatio-temporal gradients can be expressed in the form of the structure tensor matrix

$$\widetilde{\mathbf{G}} = \widetilde{\boldsymbol{\Sigma}} + \widetilde{\boldsymbol{\mu}}\widetilde{\boldsymbol{\mu}}^T. \quad (10)$$

The optical flow can then be estimated from the structure tensor by solving

$$\widetilde{\mathbf{G}}\mathbf{w} = \mathbf{0} \;, \quad (11)$$

where $\mathbf{w} = [u, v, z]^T$ is the 3D optical flow [19]. The optical flow is the structure tensor's eigenvector with the smallest eigenvalue [21]. Assuming that the change in time is 1, the 2D optical flow is $\boldsymbol{\omega} = [u/z, v/z]^T$.

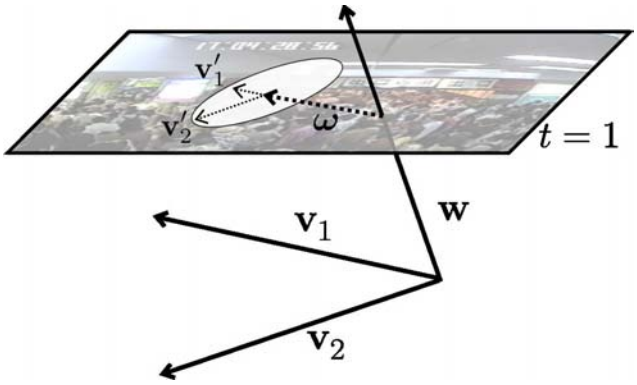

Figure 3. The 3D optical flow $\mathbf{w}$ is estimated from the predicted structure tensor matrix as the eigenvector with the smallest eigenvalue. The 2D optical flow $\boldsymbol{\omega}$ is its projection onto the plane $t = 1$, representing the predicted movement of the subject through the cuboid. 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 scaling them with respect to the estimation confidence.

The estimated optical flow from the structure tensor provides the expected motion vector for the state transition distribution. It does not, however, provide the covariance matrix $\mathbf{\Lambda}$ that encodes the variance of the flow vector. We exploit the rich motion information encoded in the structure tensor to estimate the covariance based on the shape of the distribution of gradients. Specifically, the magnitude of the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of $\widetilde{\mathbf{G}}$ provide a confidence of how much gradient information is present in the directions of the eigenvectors. If the video volume contains a single dominant motion vector and sufficient texture, then $\widetilde{\mathbf{G}}$ is rank 2 and $\lambda_3 \ll \lambda_1, \lambda_2$. In addition, the eigenvectors of the structure tensor indicate the primary directions of the spatio-temporal gradients. As shown in Fig. 3, the optical flow vector $\boldsymbol{\omega}$ is a projection of the 3D optical flow $\mathbf{w}$ onto the plane $t = 1$. Thus we consider the eigenvectors of $\mathbf{\Lambda}$ to be the projection of the eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ of $\widetilde{\mathbf{G}}$ onto the same plane, providing uncertainty in the primary directions of the spatio-temporal gradients. Since the eigenvalues of $\widetilde{\mathbf{G}}$ indicate a confidence in the flow estimation, we consider them inversely proportional to the eigenvalues of $\mathbf{\Lambda}$, as a higher confidence value indicates less variance. Thus $\mathbf{\Lambda}$ can be estimated by

$$\mathbf{\Lambda} = [\mathbf{v}'_1, \mathbf{v}'_2] \begin{bmatrix} \frac{\lambda_3}{\lambda_1} & 0 \\ 0 & \frac{\lambda_3}{\lambda_2} \end{bmatrix} [\mathbf{v}'_1, \mathbf{v}'_2]^{-1} \ , \qquad (12)$$

where $\mathbf{v}'_1$ and $\mathbf{v}'_2$ are the projections of $\mathbf{v}_1$ and $\mathbf{v}_2$ onto the plane $t = 1$.

By estimating the state transition distribution using the predicted motion pattern, we have predicted the pedestrian's motion as they pass through the cuboid at each space-time location. Intuitively, we have imposed a spatio-temporal prior on the movement of pedestrians based on the motion of the crowd.

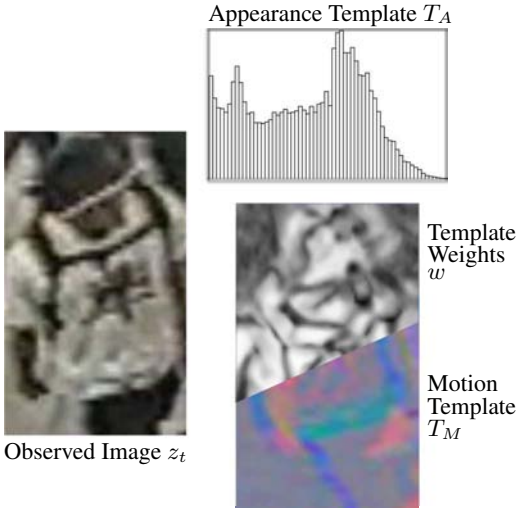

Figure 4. The pedestrian's appearance template T_A and motion template T_M are represented by a color histogram and an image of spatio-temporal gradients, respectively. The observed frame z_t is used to estimate the motion model's error at each pixel, illustrated by the weight image w. The pedestrian's natural movement causes lower weights in moving areas such as the head and shoulders. Pixels behind the pedestrian also change significantly as indicated by the dark background area.

5. Discerning Pedestrians

The large number of individuals in extremely crowded scenes also presents unique challenges to modeling the likelihood distribution $p(\mathbf{z}_t|\mathbf{x}_t)$. The density of the crowd makes it difficult to discern a specific individual from pedestrians not being tracked. In addition, the individual's appearance may change significantly due to their natural body movements. To robustly represent an individual, we model the likelihood function using their motion in conjunction with their appearance.

Typical models of the likelihood distribution $p(\mathbf{z}_t|\mathbf{x}_t)$ maintain a template T that represents the target's characteristic appearance using either a color histogram or an image. The distribution is modeled using the difference between a region R (defined by state $\mathbf{x}_t$) of the observed image $\mathbf{z}_t$ and the template. The likelihood is computed by

$$p(\mathbf{z}_t|\mathbf{x}_t) = \frac{1}{Z} \exp \frac{-d(R, T)}{\sigma} \ , \qquad (13)$$

where σ is the variance, $d(\cdot)$ is a distance measure, and Z is a normalization term. We assume pedestrians exhibit consistency in their appearance and their motion, and model them in a joint likelihood by

$$p(\mathbf{z}_t|\mathbf{x}_t) = p_{\mathrm{A}}(\mathbf{z}_t|\mathbf{x}_t)\, p_{\mathrm{M}}(\mathbf{z}_t|\mathbf{x}_t) \ , \qquad (14)$$

where p_{A} and p_{M} are the appearance and motion likelihoods, respectively. We model each of these distribu-

tions in the form of Eq. 13, with parameters $\{\sigma_A, T_A\}$ and $\{\sigma_M, T_M\}$ for the appearance and motion distributions, respectively. For our appearance distribution, we use a color histogram model [16].

As illustrated in Fig. 4, we model the motion template T_M with an image of spatio-temporal gradients to represent the target's movement between frames. Other motion representations use feature trajectories [20], assuming that the motion across the entire target is uniform. Human movement, however, is typically not uniform when captured at a close distance and causes gradual changes in the spatio-temporal gradients. We capture these variations by updating the motion template T_M during tracking. After tracking in frame t, we update each pixel i in the motion template by

$$T^t_{M,i} = \alpha \nabla R_{E[\mathbf{x}_t|\mathbf{z}_{1:t}],i} + (1-\alpha)\, T^{t-1}_{M,i} \;, \qquad (15)$$

where T^t_M is the motion template at time t, $\nabla R_{E[\mathbf{x}_t|\mathbf{z}_{1:t}]}$ is the region of spatio-temporal gradient defined by the tracking result (*i.e.*, the expected value of the posterior), and α is the learning rate. We also update the appearance histogram T_A using the same learning rate to model the subject's appearance changes.

To further discern the target from the surrounding crowd, we identify frequently changing areas of the motion template and reduce their contribution to the computation of the motion likelihood. We measure the change in motion at time t and pixel i by the angle between the spatio-temporal gradient from the tracking result $R_{E[\mathbf{x}_t|\mathbf{z}_{1:t}]}$ and the motion template T^{t-1}_M. Pixels that change drastically due to the pedestrian's body movement or partial occlusions produce a large angle between vectors. Since the pedestrian's motion changes gradually, we update this error measurement during tracking. The error at pixel i and time t becomes

$$E^t_i = \alpha \arccos\left(\mathbf{t}_i \cdot \mathbf{r}_i\right) + (1-\alpha)\, E^{t-1}_i \;, \qquad (16)$$

where $\mathbf{t}_i$ and $\mathbf{r}_i$ are the normalized gradient vectors of the motion template and the tracking result at time t, respectively, and α is the learning rate from Eq. 15. To reduce the contributions of frequently changing pixels to the computation of the motion likelihood, we weigh each pixel in the likelihood's distance measure. The weight of each pixel i is inversely proportional to the error, and computed by

$$w^t_i = \frac{1}{Z}\left(\pi - E^{t-1}_i\right) \;, \qquad (17)$$

where Z is a normalization term such that $\sum_i w^t_i = 1$. Fig. 4 shows an example of a subject, their motion template, and the weights. The distance measure of the motion likelihood distribution becomes

$$d\left(R, T_M\right) = \sum_i w^t_i \arccos\left(\mathbf{t}_i \cdot \mathbf{r}_i\right) \;, \qquad (18)$$

where again $\mathbf{t}_i$ and $\mathbf{r}_i$ are the normalized gradients.

Figure 5. We use videos of two real-world extremely crowded scenes containing a large amount of individuals to evaluate our method. The concourse scene (left) has few physical restrictions on pedestrian movement, resulting in irregular trajectories. The ticket gate scene (right) has more structure, but still contains a large number of individuals moving in different directions.

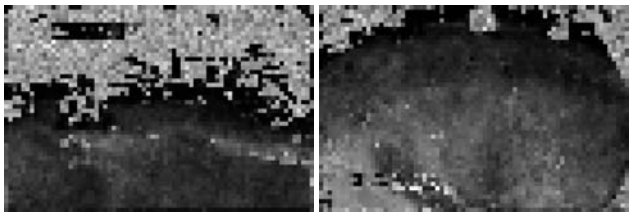

Figure 6. The angular error between the predicted optical flow vector and the observed optical flow vector, averaged over all frames in the video, for the concourse (left) and ticket gate (right) scenes. White indicates a high error for predicted motion patterns in areas with little texture such as the ceiling. Areas where pedestrian targets are present, however, contain enough spatio-temporal gradient information for a successful prediction.

Table 1. Tracking Error on Extremely Crowded Scenes

	PF	**OF**	**OA**	**Ours**
TG	82	42	26	**23**
C	126	71	80	**21**

6. Experimental Results

We evaluated our approach on a number of videos of two real-world extremely crowded scenes of a train station [10], frames of which are shown in Fig. 5. The first scene, from the station's concourse, contains a large number of pedestrians moving in different directions. The second, from the station's ticket gate, contains pedestrians moving in similar directions through the turnstiles and changing directions after exiting the gate. For each scene, we train the HMMs on a video of consistent crowd activity, and then track individuals in a different video from the same scene. The training video for the concourse scene contains 300 frames (about 10 seconds of video), and the video for ticket gate scene contains 350 frames. We set the cuboid size to $10 \times 10 \times 10$ for both scenes. The learning rate α, appearance variance σ_A, and motion variance σ_M are 0.05.

We evaluated the angular error [3] between the flow estimated from the predicted structure tensor $\widetilde{\mathbf{G}}$ and the actual gram matrix $\mathbf{G}$ to measure the accuracy of the predicted motion patterns. The median angular error (in radians) is

Figure 7. Scenes that lack temporal variations in motion, such as the one shown here from the data set used in [1], exhibit a single motion vector at each spatial location. Such scenes are special cases of reduced complexity that can easily be handled with our approach. The superimposed trajectories show successful tracking of individuals in the scene.

0.47 for the ticket gate scene, and 0.24 for the concourse scene. Fig. 6 shows the angular error averaged over the entire video for each spatial location in both scenes. Areas with little motion, such as the concourse's ceiling, result in higher error due to the lack of gradient information. High motion areas, however, have a lower error that indicates a successful prediction of the local spatio-temporal motion patterns.

A visualization of our tracking results[1] is shown in Fig. 8. Each row shows 4 frames of our method tracking different subjects in crowded areas of the scene. The different trajectories in similar areas of the frame demonstrate the ability of our approach to capture the temporal motion variations of the crowd. For example, the yellow target in row 2 is moving in a completely different direction than the red and green targets, though they share the spatial location where their trajectories intersect. Such dynamic variations in the movement of targets cannot be captured by a single motion model such as a floor fields [1]. For instance, Fig. 7 shows a marathon video from the data set used by Ali and Shah [1]. The marathon runners at each frame location tend to move in the same direction, as shown by the similar green and yellow trajectories. Our approach successfully tracks pedestrians in such scenes since they are special cases requiring only a single possible motion at each frame location.

Our motion and appearance likelihood successfully represents subjects with varying body movement. For example, the face of the yellow target in row 1 of Fig. 8 is visible in the second and third column, but not in the first and fourth. Our method is also robust to partial occlusions, as demonstrated by the blue and red targets in row 1 of Fig. 8.

Errors occur in the presence of severe occlusions, as shown by the red targets in rows 2, 4, and 5 of Fig. 8. The third frame of these rows shows the target occluded, and 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.

[1]Please see the supplementary video for tracking results.

Figure 9. Pedestrians moving against the crowd in the concourse (top) and ticket gate (bottom) scenes are successfully tracked, even though their motion patterns deviate from the learned model.

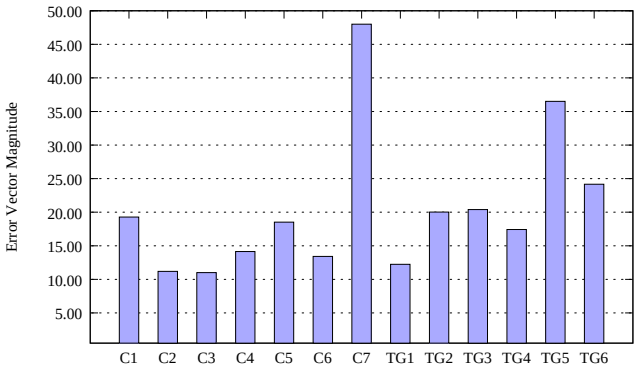

Figure 10. The magnitude of the error vector (difference between the ground truth state vector and tracking result) for subjects in the concourse (C) and ticket gate (TG) scenes. Our tracker achieves a consistently low error, except for on two very challenging cases (C7 and TG5) that undergo almost complete occlusion.

As noted by Kratz and Nishino [10], pedestrians moving against the crowd create anomalous motion patterns that deviate from the model. Nevertheless, our approach is able to track these pedestrians, as shown in Fig. 9, since the predicted motion pattern reflects the deviation from the model.

In order to quantitatively evaluate our approach, we hand-labeled ground truth tracking results for 7 targets in the concourse scene and 6 targets in the ticket gate scene. Given the ground truth state vector $\mathbf{y}_t$ (defined by the width, height, and 2D frame location of the target), the error of the tracking result $\hat{\mathbf{x}}_t$ is $||\mathbf{y}_t - \hat{\mathbf{x}}_t||_2$. Fig. 10 shows the tracking error for each subject averaged over all of the frames in the video. Our approach achieves an error less than 25 for

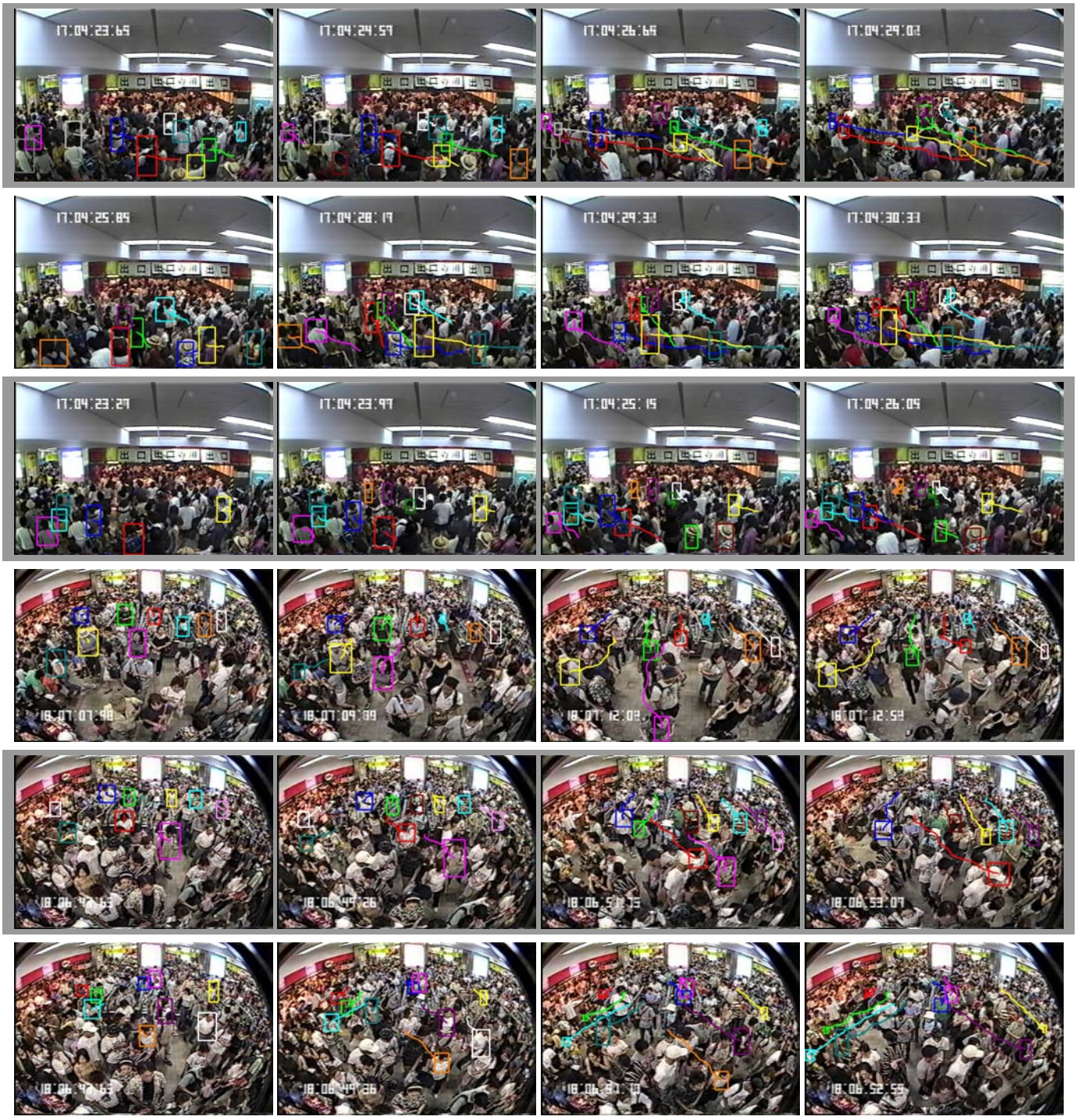

Figure 8. Example frames from tracking results for the concourse (rows 1-3) and ticket gate (rows 4-6) videos. The video sequence in each row progresses from left to right, and the colored curves show the trajectories of the targets up to the current frame. Individuals are successfully tracked in the extremely crowded areas of both scenes. The diversity of trajectories show the ability of our approach to capture spatial and temporal variations in the movement of pedestrians. Local motion variations are successfully captured, as shown by the head motion of the yellow target in row 1. The red targets in rows 2, 4, and 5 are lost due to almost complete occlusions. Please see the supplementary video for tracking results.

all but 2 of the targets. Target C7 and TG5 are lost during tracking due to severe occlusion. Table 1 shows the error (averaged over all videos) of our approach compared to that of a color particle filter (PF) [16], an optical flow tracker

(OF) [2], and an occlusion-aware tracker (OA) [15]. The occlusion-aware tracker also has a low error on the ticket gate videos, but fails on the concourse videos due to significant appearance variations.

7. Conclusion

In this paper, we derived a novel probabilistic method that exploits the inherent spatially and temporally varying structured pattern of a crowd's motion to track individuals in extremely crowded scenes. Using a collection of HMMs that encode the spatial and temporal variations of local spatio-temporal motion patterns, the method successfully predicts the motion patterns within the video. The results show that leveraging the steady-state motion of the crowd provides superior tracking results in extremely crowded areas. We believe the crowd's motion can be further leveraged to track across temporary full occlusions, which we plan to investigate next.

Acknowledgement

This work was supported in part by National Science Foundation grants IIS-0746717 and IIS-0803670. We acknowledge Nippon Telegraph and Telephone Corporation for providing the train station videos.

References

[1] S. Ali and M. Shah. Floor Fields for Tracking in High Density Crowd Scenes. In *ECCV*, 2008.

[2] S. Baker and I. Matthews. Lucas-Kanade 20 Years On: A Unifying Framework. *IJCV*, 56(1):221–255, March 2004.

[3] J. Barron, D. Fleet, and S. Beauchemin. Performance of Optical Flow Techniques. *IJCV*, 12(1):43–77, 1994.

[4] M. Betke, D. Hirsh, a. Bagchi, N. Hristov, N. Makris, and T. Kunz. Tracking Large Variable Numbers of Objects in Clutter. In *CVPR*, pages 1–8, 2007.

[5] M. J. Black and D. J. Fleet. Probabilistic Detection and Tracking of Motion Boundaries. *IJCV*, 38(3):231–245, July 2000.

[6] G. Brostow and R. Cipolla. Unsupervised Bayesian Detection of Independent Motion in Crowds. In *CVPR*, volume 1, pages 594–601, June 2006.

[7] A. Gilbert and R. Bowden. Multi Person Tracking Within Crowded Scenes. In *IEEE Workshop on Human Motion*, pages 166–179, 2007.

[8] M. Isard and A. Blake. CONDENSATION-Conditional Density Propagation for Visual Tracking. *IJCV*, 29(1):5–28, August 1998.

[9] N. Jacobs, M. Dixon, and R. Pless. Location-specific Transition Distributions for Tracking. In *IEEE Workshop on Motion and Video Computing*, 2008.

[10] L. Kratz and K. Nishino. Anomaly Detection in Extremely Crowded Scenes Using Spatio-Temporal Motion Pattern Models. In *CVPR*, pages 1446–1453, 2009.

[11] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[12] Y. Li, C. Huang, and R. Nevatia. Learning to Associate: HybridBoosted Multi-Target Tracker for Crowded Scene. In *CVPR*, 2009.

[13] T. Myrvoll and F. Soong. On Divergence Based Clustering of Normal Distributions and Its Application to HMM Adaptation. In *Proc. of European Conf Speech Communication and Technology*, pages 1517–1520, 2003.

[14] O. Nestares and D. J. Fleet. Probabilistic Tracking of Motion Boundaries With Spatiotemporal Predictions. In *CVPR*, pages 358–365, 2001.

[15] H. T. Nguyen and A. W. M. Smeulders. Fast Occluded Object Tracking By A Robust Appearance Filter. *TPAMI*, 26(8):1099–1104, 2004.

[16] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-Based Probabilistic Tracking. In *ECCV*, pages 661–675, 2002.

[17] L. Rabiner. A Tutorial On Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE*, 77(2):257–286, Feb. 1989.

[18] M. Rodriguez, S. Ali, and T. Kanade. Tracking in Unstructured Crowded Scenes. In *ICCV*, 2009.

[19] E. Shechtman and M. Irani. Space-Time Behavior Based Correlation. In *CVPR*, pages 405–412, 2005.

[20] 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. In *ICCV*, 2009.

[21] J. Wright and R. Pless. Analysis of Persistent Motion Patterns Using the 3D Structure Tensor. In *IEEE Workshop on Motion and Video Computing*, pages 14–19, 2005.

[22] B. Wu and R. Nevatia. Tracking of Multiple, Partially Occluded Humans Based On Static Body Part Detection. In *CVPR*, pages 951–958, 2006.