Spatio-Temporal Motion Pattern Modeling of Extremely Crowded Scenes

Louis Kratz, Ko Nishino

Spatio-Temporal Motion Pattern Modeling of Extremely Crowded Scenes

Louis Kratz and Ko Nishino

Department of Computer Science, Drexel University
{lak24, kon}@drexel.edu

Abstract. The abundance of video surveillance systems has created a dire need for computational methods that can assist or even replace human operators. Research in this field, however, has yet to tackle an important real-world scenario: extremely crowded scenes. The excessive amount of people and their activities in extremely crowded scenes present unique challenges to motion-based video analysis. In this paper, we present a novel statistical framework for modeling the motion pattern behavior of extremely crowded scenes. We construct a rich yet compact representation of the local spatio-temporal motion patterns and model their temporal behaviors with a novel, distribution-based Hidden Markov Model (HMM), exploiting the underlying statistical characteristics of the scene. We demonstrate that, by capturing the steady-state behavior of a scene, we can naturally detect unusual events as unlikely motion pattern variations. The experiments show promising results in extremely crowded real-world scenes with complex activities that are hard for even human observers to analyze.

1 Introduction

Despite the general interest and active research in motion-based video analysis, an important area of video surveillance has been left widely untackled. Extremely crowded scenes, as shown in Fig. 1, are perhaps in the most need of computational methods for analysis. Crowded, public areas require monitoring of an extreme number of individuals and their activities, a significant challenge for even a human observer. Videos with this level of activity have yet to be analyzed via computational methods. Common video analysis scenes, such as the PETS 2007 database [1], contain less than one hundred individuals in even the most crowded samples. Extremely crowded scenes contain hundreds of people in any given frame, and possibly thousands through the duration of the video.

As illustrated in Fig. 1, the people and objects that compose extremely crowded scenes present an entirely different level of challenges due to their excessive quantity and complex activities. Moving pedestrians can undergo drastic appearance changes between frames, and their sheer number consistently causes excessive local and global occlusions. These characteristics make traditional motion-based approaches such as tracking and segmentation extremely difficult, if not impossible.

In this paper, we derive a novel motion pattern framework for analyzing videos of extremely crowded scenes. Our key insight is to exploit the excessive

Fig. 1. Two frames from an extremely crowded scene that pose significant problems including appearance variations (a), frequent occlusions (b), and concurrent, independent activities (c) and (d). We derive a novel computational framework that models the temporal statistical behavior of local spatio-temporal motion patterns to analyze such videos.

motion patterns within local areas of the entire sequence, modeling the underlying intrinsic structure they form in the video. In other words, we model the motion variation in local space-time volumes and their temporal statistical behaviors that characterize the scene. Our construction is unique in that it captures the large motion pattern variations in the video sequence while maintaining a robust representation of the motion in local video regions. To capture the typical motion behaviors of the scene, we derive a distribution-based HMM that retains the rich motion information within our local spatio-temporal motion pattern representation and describes natural motion transitions within a local video region. We use this to model the stationary structure of motion patterns in the video, i.e. usual events within the scene, and identify atypical events as statistical anomalies.

We capture the excessive motion within the scene by dividing the video into spatio-temporal volumes of a fixed size, which we refer to as cuboids, to represent the local motion behavior. We encode the rich motion patterns in each cuboid with a multivariate Gaussian distribution of its 3D spatio-temporal gradients. Furthermore, we identify prototypical motion patterns in disjoint spatio-temporal locations, providing a faithful yet compact representation of the scene's motion characteristics. Based on these representations, we derive a novel distribution-based HMM that fully utilizes the motion pattern representations, multivariate Gaussian distributions, to characterize their temporal variations. Distribution-based HMMs provide the flexibility to model the motion pattern variations directly from the representation by handling observations that are distributions themselves. The entire video volume is thus modeled as a collection of distribution-based HMMs that encodes the intrinsic statistical structure of local spatio-temporal motion patterns. We model a crowded scene given a training video sequence of usual activity, and detect unusual activities in a query video by identifying local spatio-temporal motion patterns with low likelihoods. To our knowledge, this is the first work to model the motion patterns of truly crowded scenes and successfully detect unusual activities in them.

2 Relation to Previous Work

Motion-based video analysis has been an active and popular field of research, using various techniques to represent motion characteristics depending on the

application. Here we review previous work that focuses specifically on motion-based analysis of crowded scenes and unusual event detection.

Approaches to unusual event detection can be categorized into two groups: event detection and deviation methods. Deviation approaches model usual activity, detecting unusual events as those that differ from the model [2–11]. Event detection approaches explicitly model specific activities for identification in query videos [12–14]. These have the capability of differentiating between detected events, however modeling each possible event in extremely crowded scenes is unfeasible. The number of events required to robustly capture the large variability of usual activity in extremely crowded scenes would be substantial. On the other hand, the large amount of activity in extremely crowded scenes provides an abundance of data representing the stationary behavior of the scene's constituents, i.e. the usual activities, making a deviation approach suitable.

Motion-based video analysis approaches model the deviation of motion characteristics throughout the video volume. Trajectory-based approaches [2, 3, 12, 14], for example, track scene objects, describing the motion by a change in spatial location, and spatial deviations are considered unusual. Trajectory-based techniques are suitable for scenes with few moving objects that can easily be tracked, such as infrequent pedestrian or automobile traffic. The tracking challenges posed by extremely crowded scenes, however, makes trajectory-based representations unfeasible. In addition, the motion analysis of trajectory-based approaches focuses on each subject individually, but the behavior of extremely crowded scenes depends on the motion of multiple subjects concurrently.

Other motion pattern approaches [15, 9, 10, 16–20] model the spatio-temporal volume directly. These approaches often estimate the optical flow [15, 9] or the motion within spatio-temporal volumes [10, 16, 17, 19]. Flow-based approaches have modeled motion deviations in the form of HMMs [9] or Baye's classifiers [15] to represent sparse human motion. Unfortunately, the excessive number of subjects present in extremely crowded scenes make the estimation of optical flow unreliable. Furthermore, the variation of activities caused by the large number of individuals makes specific behavior difficult to define and isolate. The nature of extremely crowded scenes, however, requires the ability to capture activity within local scene regions. Extremely crowded scenes may contain any number of concurrent, independent activities taking place in different local areas of the same sequence. This makes global approaches, such as full frame analysis [8, 9], unfeasible, as the entire frame would be dominated with visual information irrelevant to the specific event of interest.

Spatio-temporal approaches [10, 16, 17, 19] directly represent the motion patterns within video volumes. Though these motion pattern representations are well suited to extremely crowded scenes, the modeling of their behavior has been limited to volume distance [16, 10] or interest points [19], requiring the explicit modeling of each event to be detected. In addition, most spatio-temporal representations assume that the volume contains a dominant, uniform motion. In extremely crowded scenes, it is exactly the non-uniform motion patterns that characterize the crowd behavior.

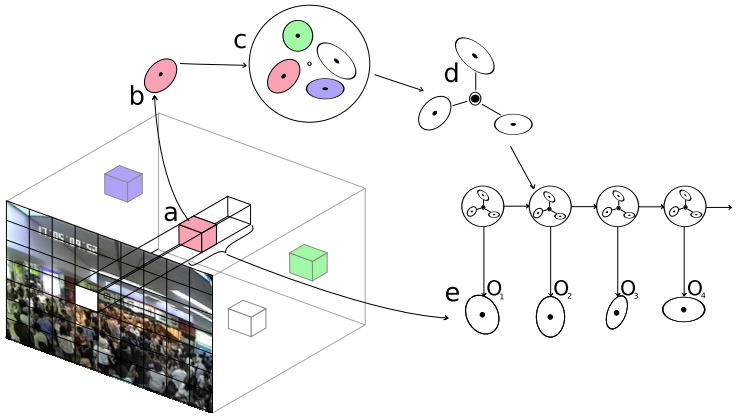

Fig. 2. A schematic overview of the framework. (a) The video sequence is divided into local spatio-temporal volumes (cuboids). (b) The motion pattern is represented by a 3D Gaussian of spatio-temporal gradients. (c) Prototypical motion patterns are identified by grouping similar motion patterns that may lie at disjoint space-time locations. (d) Prototypes model the motion pattern variation in the video with a 1D normal distribution of 3D Gaussians, i.e. a distribution of distributions. (e) A distribution-based HMM is trained for each tube using the prototypical motion patterns for hidden states and 3D Gaussians as observations.

The nature of extremely crowded scenes, specifically the high concentration and large variability of motion patterns, makes previous motion pattern approaches unsuitable for activity analysis. Other approaches such as statistical background modeling [21, 22] also fail on extremely crowded scenes as the large number of people often completely occlude the background, making the foreground and background indistinguishable. The large amount of motion present in the scene, however, lends itself naturally to a volumetric-based approach, but requires a robust handling of non-uniform motion patterns and a flexible model to capture large motion variations. The construction of a canonical representation of motion patterns within the video sequence is a challenging problem that we explore via the use of novel distribution-based HMMs. The strict motion-pattern requirements presented by extremely crowded scenes makes a direct comparison of our approach to previous work unsuitable, as previous models do not capture large motion pattern variations in local areas of the video scene.

3 Local Spatio-Temporal Motion Patterns

Extremely crowded scenes contain large amounts of independent, local activities. Yet these local activities in the space-time form an underlying intrinsic structure that exhibits characteristic motion patterns and variations in the video. Our goal is to derive a compact representation of the motion patterns that remains faithful to the scene behavior, and provides means to model their temporal variations.

3.1 Spatio-Temporal Motion Representation

We represent the video as a collection of local spatio-temporal volumes as shown in Fig. 2(a) by subdividing the video into volumes of fixed size. The set of

non-uniform motion patterns encoded in these cuboids collectively forms the complex motion in the entire video volume. A finer-grain representation, such as optical flow, would not provide enough motion information. Conversely, the motion of the entire frame would not provide the level of detail necessary for differentiating between independent, concurrent activities. By dividing the video into spatio-temporal volumes, we can extract robust motion information while isolating separate activities to local regions.

We wish to represent the rich, non-uniform, local spatio-temporal motion patterns encapsulated in each cuboid in a compact yet faithful manner. We use the distribution of spatio-temporal gradients as our base representation. For each pixel i in cuboid I, we calculate the spatio-temporal gradient ∇I_i

$$\nabla I_i = [I_{i,x}\ I_{i,y}\ I_{i,t}]^T = \left[\frac{\partial I}{\partial x}\ \frac{\partial I}{\partial y}\ \frac{\partial I}{\partial t}\right]^T \ , \tag{1}$$

where x, y, and t are the video's horizontal, vertical, and temporal dimensions. The set of spatio-temporal gradients from each pixel represents the characteristic motion pattern encoded within the cuboid.

As illustrated in Fig. 2(b), we model the distribution of gradients as a 3D Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

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

Thus for each spatial location n and temporal location t, the local spatio-temporal motion pattern representation O_t^n is defined by $\boldsymbol{\mu}_t^n$ and $\boldsymbol{\Sigma}_t^n$. This representation has been used previously for background modeling [21], and is similar to the Gram Matrix used for analyzing persistent motion patterns [23] and correlating video sequences [24]. The explicit multivariate Gaussian modeling allows us to derive sound statistical temporal models for analyzing pattern variations while faithfully capturing the multiple non-uniform motion encoded in the cuboid.

3.2 Distance Metric

Given a compact representation of the local spatio-temporal motion pattern, we need a canonical distance metric to measure the difference between cuboids in subsequent modeling of the video. Previous approaches have used algebraic metrics which are sensitive to noise [25], or assume that the cuboid volume only consists of a uniform motion [24, 16]. We use an information theoretic distance metric that directly measures the distribution distances: the symmetric Kullback-Leibler (KL) divergence [26]. Since we model the spatio-temporal motion patterns as Gaussian distributions, the divergence has a closed analytical form [27], providing a solid, nonnegative comparison between motion patterns.

The symmetric KL distance measure contains the inverse of both covariance matrices. Cuboids with little or no motion and texture have small covariance

matrices, introducing large errors in the KL distance. We define a less sensitive distance measure between two distribution A and B

$$\tilde{d}(A,B) = \begin{cases} 0 & \mathcal{K}(\boldsymbol{\Sigma}_a) > d_{\mathcal{K}}, \mathcal{K}(\boldsymbol{\Sigma}_b) > d_{\mathcal{K}}, \parallel \boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_b \parallel_F < d_\Sigma, \parallel \boldsymbol{\mu}_a - \boldsymbol{\mu}_b \parallel < d_\mu \\ d(A,B) & \text{otherwise} , \end{cases} \tag{3}$$

where $d(A,B)$ is the KL Divergence and $\mathcal{K}()$ is the condition number of the matrix, confirming that the inverse is unstable. The norms of the covariance and mean differences are calculated to ensure that the distributions are within acceptable distances (d_Σ, d_μ). We refer to this measure as the KL distance for the remainder of this paper.

3.3 Prototypical Local Spatio-Temporal Motion Patterns

By directly modeling the video as a collection of normal distributions, we reduce the size of the video representation from a set of raw pixels to a collection of Gaussian parameters. As illustrated in Fig. 2(c), we further reduce the size of this representation by exploiting common motion patterns occurring in disjoint space-time locations. We model the recurrence of similar motion patterns by extracting prototypical local spatio-temporal motion patterns (prototypes) that characterize the activities of the scene. Note that prototypical motion patterns can occur at disjoint spatio-temporal locations, and such recurrences actually form the usual activities of the scene.

Many methods exist that can extract a video's prototypes. Off-line clustering techniques such as K-means, however, require that all of the observations be available. We use an online method that computes the KL distance from each local spatio-temporal motion pattern O_t^n to the prototypes as we parse the video. If the KL distance is greater than a specified threshold, d_{KL}, for all prototypes $\{P_s | s = 1, \ldots, S\}$, then the cuboid is considered a new prototype. Otherwise, the prototype distribution P_s is updated with the new observation O_t^n by

$$P_s = \frac{1}{N_s + 1} O_t^n + \left(1 - \frac{1}{N_s + 1}\right) P_s \ , \tag{4}$$

where N_s is the total number of observations associated with the prototype P_s at time t. To solve Eq. 4 with respect to the KL-divergence, we use the expected centroid presented by Myrvoll et al. [27] to compute the mean distribution P_s. The use of the expected centroid provides an efficient way to compute prototypical representations that reflect the distance measure between motion pattern representations, ensuring a canonical calculation of motion patterns.

By extracting prototypical local spatio-temporal motion patterns, we effectively construct a canonical representation of the video motion patterns. The local spatio-temporal motion patterns are classified by the prototypes, which represent the expected observed motion patterns within a specific KL distance. The variation in KL space between these observations describes the distribution of motion patterns associated with a prototype.

4 Statistical Characterization of Crowded Scenes

While the set of prototypes provides a picture of similar activities in the scene, it does not capture the relationship between their occurrences. By modeling the temporal behavior of local spatio-temporal motion patterns, we characterize a given video by its motion pattern variation. We assume that cuboids in the same spatial location exhibit Markov property in the temporal domain since the scene is comprised of physically moving objects. In order to achieve a localized model, we observe each spatial location separately creating a single HMM for each tube of observations in the video. However, since each cuboid is a distribution of gradients (3D Gaussian representing local motion patterns), and the video a collection of prototypes representing distributions of distributions, we must derive an HMM that can handle observations that are distributions themselves.

4.1 Distribution-Based Hidden Markov Models

Ordinary HMMs are defined by five parameters $M = \{H, \mathbf{o}, \mathbf{b}, \mathbf{A}, \boldsymbol{\pi}\}$, where H is the number of hidden states, $\mathbf{o}$ the possible values of observations, $\mathbf{b}$ a set of H emission probability density functions, $\boldsymbol{\pi}$ an initial probability vector, and $\mathbf{A}$ a transition probability matrix. We model a single HMM $M^n = \{H^n, \mathbf{O}^n, \mathbf{b}^n, \mathbf{A}^n, \boldsymbol{\pi}^n\}$ for each spatial location $n = 1, \ldots, N$ and associate the hidden states H^n with the number of prototypes S^n in the tube. The set of possible observations $\mathbf{O}^n$ is a continuous range of 3D Gaussian distributions. Complex observations for HMMs are often quantized for use in a discrete HMM. However, such quantization of observations would significantly reduce the rich motion information present in each cuboid. Our construction of distribution-based HMMs permits the observations to remain continuous 3D Gaussian distributions. Therefore, the emission probability density function for each prototype must be a distribution of distributions.

We model the local spatio-temporal motion pattern variation with a 1D normal distribution using the KL distance to remove the bias. Thus the probability of an observation O_t^n is

$$\mathrm{p}(O_n^t|b_s) = \mathrm{p}(\frac{\tilde{d}(O_t^n, P_s)}{\sigma_s}) \sim \mathcal{N}(0,1) \ , \tag{5}$$

where P_s is the average of the observed motion patterns assigned to prototype s, given in Eq. 4. This retains the rich motion information represented by each observation within the emission probability distribution, and provides a probability calculation consistent with our distance measure. The estimation of P_s is given in Eq. 4 and the standard deviation by the maximum likelihood estimator

$$\sigma_s^2 = \frac{1}{N_s} \sum_j^{N_s} \tilde{d}(O_j, \tilde{P}_s)^2 \ , \tag{6}$$

representing the average KL distance between the mean distribution and observed motion patterns. In practice, however, there may be too few cuboids in a

given classification to capture specific behavior with reliable variation. In such an occasion, we use a 99.7 percent window around d_{KL}, letting $\sigma_s = 3d_{\mathrm{KL}}$.

This distribution-based HMM models the temporal behavior of local spatio-temporal motion patterns in a sound statistical framework. The emission probability distributions are created using Eq. 4 and 6 for each prototype in the scene. Note that, while a single HMM is computed for each tube, the emission probability density functions are created using samples from the entire video volume. The parameters $\mathbf{A}^n$ and $\boldsymbol{\pi}^n$ are estimated by expectation maximization.

4.2 Atypical Events as Statistical Anomalies

Modeling the temporal behavior of the local spatio-temporal motion patterns in a statistical framework provides a natural way to define usual activities as statistical stationary behavior. This in turn enables the detection of unusual events in the scene as statistical deviations from a learnt model of a video sequence that captures the mostly usual activities. Given a query video, the likelihood of a temporal sequence of local spatial-temporal motion patterns can be calculated via the forwards-backwards algorithm [28]. Let ζ_t^n be the likelihood of the t^{th} temporal sequence of observations within a given video tube n. Thus

$$\zeta_t^n = \mathrm{p}(O_t^n, \ldots, O_{t+w}^n | M^n) \ , \tag{7}$$

where M^n is the HMM at tube n, w is the sequence length, and $O_t^n, \ldots, O_{t+w}^n$ is the sequence of observed local spatio-temporal motion patterns. Ideally, we would like to focus this measure to each individual cuboid. Since ζ_t^n is calculated for every sequence within the tube of length w, each observation O_t^n is associated with w likelihood measures. We define an ensemble function that selects a measure from the set of likelihoods the observation contributed to. We use a window size of 2 and let the ensemble function maximize over the likelihoods. This correctly classifies the cuboids with the exception of one case, when a usual cuboid is temporally sandwiched between two unusual cuboids, which is rare and errs on the side of caution (a false positive). Observations with a likelihood less than a threshold ξ^n are considered unusual. We define ξ^n using the training data $\tilde{O}_t^n$ in the tube n:

$$\xi^n = \rho \times \min_t \mathrm{p}(\tilde{O}_t^n | M^n) \ , \tag{8}$$

where ρ is a constant term accounting for noise in the training data.

5 Experimental Results [1]

We evaluate the effectiveness of our approach on three data sets: one simulated crowded scene and two real-world extremely crowded scenes. We train a set of the distribution-based HMMs on a sequence of the usual state of the scene and use them to detect unusual activities in query videos of the same scene. The ground truth was hand-labeled for all data sets to quantitatively evaluate

[1] **A movie of results is located at** `http://www.cs.drexel.edu/~lak24/mlvma08`**.**

Fig. 3. Two extremely crowded real-world scenes include a railway station concourse (left) and ticket gate (right). The top half of the original frames are excluded from modeling, as they only contain views of the station ceiling.

Fig. 4. Detection of unusual events in the concourse video (left) and the ticket gate (right). Correctly classified unusual cuboids are highlighted in blue, usual cuboids in green, and false positives in magenta. The intensity of magenta blocks indicates the severity of false positives. Individuals reversing direction, loitering, and moving in irregular patterns are correctly classified as unusual. False positives indicate a sensitivity to low motion irregular patterns. A few false negatives are adjacent to true positives, thus we consider them harmless in practical scenarios.

the performance, and all threshold values were selected empirically. The size of cuboids is set to $40 \times 40 \times 20$ for all experiments.

The synthetic crowded scene was generated by translating a texture of a crowd across the frame, resulting in large motion variations and nonuniform motion along border areas. The image sequence consists of 216 tubes and $9,072$ total cuboid observations. Several smaller images were inserted moving in arbitrary directions to simulate unusual motion activities. The thresholds used in this experiment were $d_{\mathrm{KL}} = 0.02$, $d_{\Sigma} = 5$, $d_{\mu} = 1$, and $d_{\mathcal{K}} = 400$. The Receiver Operating Characteristic (ROC) curve for this example is shown in Fig. 5, and is produced by varying values of the weighting scalar ρ in Eq. 8. Our results achieve a false positive rate of 0.009 and a true positive rate of 1.0 using a weighted scalar of 0.005. The false detections occurred in cuboids of no motion and texture. Though Eq. 3 reduces these sensitivities, the synthetic nature of the data set results in small, numerically stable distributions, increasing the KL distance enough to cause a false positive.

Videos of two extremely crowded real-world scenes were also used to evaluate the performance. For both data sets, thresholds in the KL distance were set to $d_{\Sigma} = 1$, $d_{\mu} = 1$, and $d_{\mathcal{K}} = 1000$. The first, illustrated on the left in Fig. 3, is a video sequence from an extremely crowded concourse of a railway station. The scene contains a large number of moving and loitering pedestrians and employees directing traffic flow. The query video contains station employees walking against the flow of traffic and irregular pedestrian motions. The training video contains $9,664$ cuboids and the query video $1,216$. The threshold d_{KL} is .06. The second real-world data set, shown on the right in Fig. 3, is a wide-angle view of pedestrian traffic at a railway terminal. The motion of the crowd occurs in a more constant direction than the first real-world data set, however still contains excessive occlusions and motion variations. The query video contains

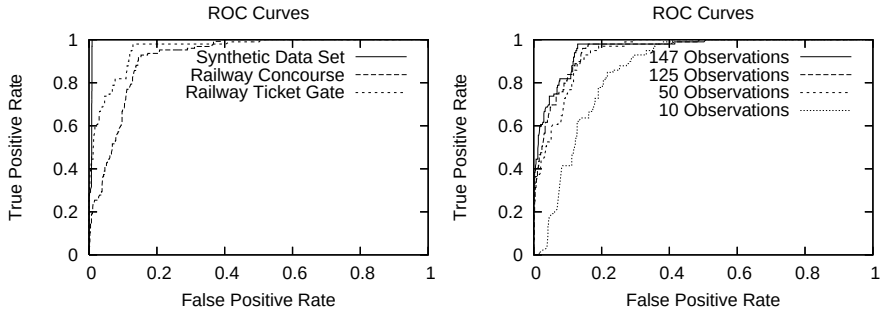

Fig. 5. ROC curves for the synthetic and real-world extremely crowded scenes (a). Effects of increased training data on the Ticket Gate dataset (b).

instances of people reversing direction and pausing in traffic. The training video contains $8,288$ cuboids and the query video 840. The threshold d_{KL} is 0.05.

As illustrated in Fig. 5, our approach achieves a false positive rate of 0.15 and a true positive rate of 0.92 with a weighting scalar of 10^{-91} for the first data set, and a false positive rate of 0.13 and a true positive rate of 0.97 at a weighting scalar of 10^{-42} for the second. A visualization of the detection results for this data set is shown in Fig. 4. Employees crossing the scene against the pedestrian traffic flow activity in the concourse, and pedestrians loitering and reversing direction in the ticket gates are successfully detected as unusual. The weight scalar values represent how much variability occurs in the motion patterns. The larger value in the second experiment indicates the usual motion patterns are more defined, and directly relates to the motion sensitivity a human operator would likely observe. Intuitively, unusual behavior beyond what is detected here, such as that warranting personnel intervention, would be satisfied with an even larger weighting scalar.

False positives occur in both experiments for slightly irregular motion patterns, such as after pedestrians exit the gate, and areas of little motion, such as where the floor of the station is visible. These motion pattern variations have similar severity to those detected as unusual. The few false negatives in both real-world examples occur adjacent to true positives, which suggests they are harmless in practical scenarios. We expect these can be avoided by accounting for the statistical information in surrounding local areas, which we plan to pursue as future work.

The effects of increasing the training data size for the Ticket Gate dataset are shown in Fig. 5. As expected, the performance increases with the training data size, and approaches a steady state with more than 100 observations per tube. The performance with only 10 observations per tube achieves a false positive rate of 0.22 and true positive rate of 0.84. This is largely due to the window around d_{KL} for σ_s in Eq. 6 when there are insufficient observations for specific prototypes, and the inclusion of disjoint motion patterns in the prototype construction. The performance with such small training data reflects the robust motion pattern representation captured by the distribution-based HMMs.

6 Conclusion

In this paper we introduced a novel statistical framework for analyzing local spatio-temporal motion patterns in extremely crowded scenes. Motion patterns are represented with multivariate Gaussian distributions, and prototypical occurrences of these motion patterns are identified, canonically representing the motion variations in the video sequence. The temporal statistical variation of the motion patterns is modeled in a robust, statistical framework using a novel distribution-based Hidden Markov Model.

Our results indicate that local spatio-temporal motion patterns are a suitable representation for analyzing extremely crowded scenes. Their use is demonstrated on real-world data of extremely crowded scenes, and successfully detects unusual motion patterns in pedestrian behavior including movement against the normal flow of traffic, loitering, and traffic congestion. We believe the proposed framework can play an important role in video analysis of extremely crowded scenes. We are currently investigating methods for incorporating local spatial statistics and believe this will lead to a more vivid description of the local activity and improved detection accuracy.

Acknowledgments. This work was in part supported by Nippon Telegraph and Telephone Corporation and National Science Foundation under ITR award IIS-0426674.

References

1. PETS: Tenth IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance. `http://www.pets2007.net/` (2007)
2. Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., Maybank, S.: A System for Learning Statistical Motion Patterns. IEEE Trans. on Pattern Analysis and Machine Intelligence **28**(9) (Sep. 2006) 1450–1464
3. Dee, H., Hogg, D.: Detecting Inexplicable Behaviour. In: Proc. of British Macine Vision Conf. (2004) 477–486
4. Cui, P., Sun, L.F., Liu, Z.Q., Yang, S.: A Sequential Monte Carlo Approach to Anomaly Detection in Tracking Visual Events. In: IEEE Workshop on Visual Surveillance. (2007) 1–8
5. Zhou, H., Kimber, D.: Unusual Event Detection via Multi-camera Video Mining. In: Proc. of International Conf on Pattern Recognition. (2006) 1161–1166
6. Xiang, T., Gong, S.: Online Video Behaviour Abnormality Detection Using Reliability Measure. In: Proc. of British Macine Vision Conf. (2005)
7. Salas, J., Jimenez Hernandez, H., Gonzalez Barbosa, J., Hurtado Ramos, J., Canchola, S.: A Double Layer Background Model to Detect Unusual Events. In: Proc. of Advanced Concepts for Intelligent Vision Systems. (2007) 406–416
8. Zhong, H., Shi, J., Visontai, M.: Detecting Unusual Activity in Video. In: Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition. (2004) 819–826
9. Andrade, E., Blunsden, S., Fisher, R.: Modelling Crowd Scenes for Event Detection. In: Proc. of International Conf on Pattern Recognition. (2006) 175–178
10. Boiman, O., Irani, M.: Detecting Irregularities in Images and in Video. In: Proc. of IEEE Int'l Conf on Computer Vision. (2005) 462–469

11. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors. IEEE Trans. on Pattern Analysis and Machine Intelligence **30**(3) (Mar. 2008) 555–560
12. Chan, M.T., Hoogs, A., Schmiederer, J., Petersen, M.: Detecting Rare Events in Video Using Semantic Primitives with HMM. In: Proc. of International Conf on Pattern Recognition. (2004) 150–154
13. Gryn, J., Wildes, R., Tsotsos, J.: Detecting Motion Patterns via Direction Maps with Application to Surveillance. In: IEEE Workshop on Motion and Video Computing. (2005) 202–209
14. Johnson, N., Hogg, D.: Learning the Distribution of Object Trajectories for Event Recognition. In: Proc. of British Macine Vision Conf. Volume 2. (1995) 583–592
15. Black, M.: Explaining Optical Flow Events with Parameterized Spatio-Temporal Models. In: Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition. (1999) 326–332
16. Ke, Y., Sukthankar, R., Hebert, M.: Event Detection in Crowded Videos. In: Proc. of IEEE Int'l Conf on Computer Vision. (2007) 1–8
17. DeMenthon, D., Doermann, D.: Video Retrieval using Spatio-Temporal Descriptors. In: Proc. of the Eleventh ACM Int'l Conf on Multimedia. (2003) 508–517
18. Chomat, O., Crowley, J.: Probabilistic Recognition of Activity using Local Appearance. In: Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition. (1999) 104–109
19. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior Recognition via Sparse Spatio-Temporal Features. In: Int'l Workshop on Performance Evaluation of Tracking and Surveillance. (2005) 65–72
20. Laptev, I., Lindeberg, T.: Velocity Adaptation of Spatio-Temporal Receptive Fields for Direct Recognition of Activities: An Experimental Study. Image and Vision Computing **22**(2) (Feb. 2004) 105–116
21. Pless, R.: Spatio-temporal Background Models for Outdoor Surveillance. EURASIP Journal on Applied Signal Processing **2005**(14) (2005) 2281–2291
22. Zhong, J., Sclaroff, S.: Segmenting Foreground Objects from a Dynamic Textured Background via a Robust Kalman Filter. In: Proc. of IEEE Int'l Conf on Computer Vision. (2003) 44–51
23. Wright, J., Pless, R.: Analysis of Persistent Motion Patterns Using the 3D Structure Tensor. In: IEEE Workshop on Motion and Video Computing. (2005) 14–19
24. Shechtman, E., Irani, M.: Space-Time Behavior Based Correlation. In: Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition. (2005) 405–412
25. Nishino, K., Nayar, S.K., Jebara, T.: Clustered Blockwise PCA for Representing Visual Data. IEEE Trans. on Pattern Analysis and Machine Intelligence **27**(10) (Oct. 2005) 1675–1679
26. Kullback, S., Leibler, R.A.: On Information and Sufficiency. The Annals of Mathematical Statistics **22**(1) (1951) 79–86
27. Myrvoll, T., Soong, F.: On Divergence Based Clustering of Normal Distributions and its Application to HMM Adaptation. In: Proc. of European Conf Speech Communication and Technology. (2003) 1517–1520
28. Rabiner, L.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. of the IEEE **77**(2) (Feb. 1989) 257–286