Anomaly Detection in Extremely Crowded Scenes Using Spatio-Temporal Motion Pattern Models

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

Abstract

Extremely crowded scenes present unique challenges to video analysis that cannot be addressed with conventional approaches. We present a novel statistical framework for modeling the local spatio-temporal motion pattern behavior of extremely crowded scenes. Our key insight is to exploit the dense activity of the crowded scene by modeling the rich motion patterns in local areas, effectively capturing the underlying intrinsic structure they form in the video. In other words, we model the motion variation of local space-time volumes and their spatial-temporal statistical behaviors to characterize the overall behavior of the scene. We demonstrate that by capturing the steady-state motion behavior with these spatio-temporal motion pattern models, we can naturally detect unusual activity as statistical deviations. Our experiments show that local spatio-temporal motion pattern modeling offers promising results in real-world scenes with complex activities that are hard for even human observers to analyze.

1. Introduction

The decreasing costs of video surveillance equipment has resulted in large volumes of video data. This excessive amount of information has not been met with adequate human operators. Extremely crowded scenes, as shown in Fig. 1, require monitoring of an excessive number of individuals and their activities, a significant challenge even for a human observer. Computational approaches that assist human security personnel must be able to handle extremely crowded scenes to be successful in real-world domains.

The excessive number of people and objects that compose extremely crowded scenes presents an entirely different level of challenges. Pedestrians within the scene move in highly irregular motion patterns that result in severe occlusions. The sheer number of subjects within the video makes analyzing each individual's actions a demanding task, even for modern computational systems. In addition, the views recorded by surveillance cameras cover large areas and include hundreds of individuals. As such, a computational approach must be able to isolate activities in different areas of the frame, while retaining structural information regarding the entire scene.

Figure 1. The large number of people moving in irregular directions make extremely crowded scenes, as shown here, significantly difficult to analyze in a computational framework.

Most methods for identifying unusual events in video sequences have been constrained to sequences with only a few subjects. Extremely crowded scenes contain hundreds of individuals in each frame, and thousands throughout the video sequence. Common video analysis scenes, such as the PETS 2007 database [19], contain less than one hundred individuals in even the most crowded samples. Other work have focused on analyzing the entire video frame [3], or extracting subject specific information [2].

In this paper we construct a novel computational framework for modeling videos of extremely crowded scenes, and demonstrate its effectiveness by identifying atypical motion events. Our key insight is to exploit the dense local motion patterns created by the excessive number of subjects and model their spatio-temporal relationships, representing the underlying intrinsic structure they form in the video. In other words, we model the variations of local spatio-temporal motion patterns to describe common behavior within the scene, and then identify the spatial and temporal relationships between motion patterns to charac-

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

terize the behavior of the entire sequence. Using our framework, we are able to identify unusual events as statistical deviations in video sequences of the same scene.

The primary contribution of this paper is the derivation of a novel statistical model that identifies relationships between local spatio-temporal motion patterns while retaining the rich motion information stored in our motion pattern representation. Specifically, we construct motion-pattern distributions that capture the variations of local spatio-temporal motion patterns to compactly represent the video volume. We then derive a distribution-based HMM that describes natural motion transitions within local video regions. Finally, we improve our framework by constructing a coupled HMM that models the spatial relationship of motion patterns surrounding each 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.

2. Previous Work

Approaches to unusual event detection can be categorized into two groups: explicit detection and deviation methods. Deviation approaches model usual activity, detecting unusual events as those that differ from the pretrained model [1, 3, 5, 8, 9, 13, 21, 23, 24, 25]. Event detection approaches model each specific activity for identification in query videos [7, 12, 14]. Detection approaches 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 model extremely crowded scenes would be substantial due to the large variability in unusual activity, making detection significantly costly. Extremely crowded scenes contain a large amount of activity, providing an abundance of data representing the stationary behavior of its constituents, i.e. the usual activities, making them suitable for a deviation approach.

Typical motion-based video event analysis estimates the optical flow [3, 4] or the motion within spatio-temporal volumes [5, 10, 11, 15]. Flow-based approaches have modeled motion deviations in the form of HMMs [3] or Baye's classifiers [4] 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 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 [3, 24], unfeasible, as the entire frame would be dominated with visual information irrelevant to the specific event of interest. The modeling of motion within spatio-temporal volumes has been limited to volume distance [5, 15] or interest points [11], enforcing an explicit detection model. In addition, most spatio-temporal representations assume that the volume contains a dominant, uniform motion pattern. In extremely crowded scenes it is exactly the non-uniform motion patterns that characterize the crowd behavior.

Most motion based techniques rely on extracting motion information regarding each subject. Trajectory-based approaches [7, 9, 13, 14], for example, track scene objects and describe the motion by their spatial location. 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 motion analysis of trajectory-based approaches focuses on each subject individually, whereas the behavior of extremely crowded scenes depends on the motion of multiple subjects concurrently.

Other work on high density crowded scenes have placed strong restrictions on the behavior of the video subjects. Ali and Shah [2] track subjects in high density crowded scenes that are captured from a distance. They form floor fields that capture expected motion of the video subjects related to the physical nature of the scene. Extremely crowded scenes, though similar in density, have less structure due to the high variability of pedestrian movements. Floor fields for extremely crowded scenes would be highly chaotic, since even the steady state training data does not necessarily impose strong motion directions. In addition, tracking of each individual in pedestrian environments would result in highly inconsistent trajectories, making discriminating between usual and unusual events extremely difficult. Finally, a tracking-based model for extremely crowded scenes would also disregard the important correlation between pedestrians within close proximity of each other.

3. Local Spatio-Temporal Motion Patterns

Extremely crowded scenes contain large amounts of independent activities occurring in different locations within the frame. These activities, however, collectively form the underlying structure of the video sequence. By dividing the video into local spatio-temporal volumes of a fixed size, which we refer to as cuboids, we may isolate the local activities without discriminating between video subjects. We extract a compact motion pattern representation for each spatio-temporal volume of the video, and identify prototypical motion patterns, effectively capturing the motion structure of the entire video.

Extremely crowded scenes provide unique challenges to constructing motion representations. The quantity of pedestrians causes each local activity to be subject to occlusions,

making direct motion extraction challenging, if not impossible. The motion within each local area may be non-uniform and generated by any number of video subjects. A fine-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. Other approaches [15, 22] require the volume to contain a dominant uniform motion pattern. The behavior of extremely crowded scenes, however, is characterized by the non-uniform motion patterns within local spatio-temporal volumes.

We wish to represent the non-uniform local spatio-temporal motion patterns within each cuboid in a compact manner that remains faithful to the rich motion information within the local spatio-temporal region. 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

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

where x, y, and t are the video's horizontal, vertical, and temporal dimensions.

The spatio-temporal gradients of each pixel collectively represent the characteristic motion pattern within the cuboid. Thus we model the distribution of gradients as a 3D Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

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

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 $\boldsymbol{\Sigma}_t^n$. Optical flow estimation techniques have used spatio-temporal gradients [22], however require that the video volume contain a dominant uniform motion. The explicit multivariate Gaussian modeling retains the multiple non-uniform motion observed in the cuboid and allows us to derive sound statistical temporal models for analyzing pattern variations.

To capture the underlying motion structure of the scene, we wish to identify prototypical motion pattern representations and extract the motion variation among the cuboids. To discriminate between local spatio-temporal motion pattern representations we use the symmetric Kullback-Leibler (KL) divergence [16]. Since our motion pattern representation is a three dimensional Gaussian, the divergence has a closed analytical form [17] that requires an inversion of the covariance matrix. We reduce possible numeric instabilities by comparing the condition numbers and the norm of the difference between covariance matrices, and by taking the log of the KL divergence. We retain the positive property of the KL divergence by adding a 1 to the divergence prior to taking the log. Previous approaches have used algebraic metrics which are sensitive to noise [18] or assume that the cuboid volume only consists of a uniform motion [15, 22]. Our approach provides a positive, canonical distance measure that differentiates between the collections of spatio-temporal gradients within each cuboid.

Using the symmetric KL divergence as a distance measure, we identify similar cuboids in the video sequence by associating local spatio-temporal motion patterns that have a small distance between them. Typical clustering approaches such as k-means require the number of prototypical patterns to be known. Extremely crowded scenes, however, have extensive motion pattern variability and may contain any number of prototypical patterns depending on the video length, density of the crowd, and inconsistency of pedestrian motion.

We use an online method that computes the KL distance from each local spatio-temporal motion pattern O_t^n at location n and time t to each prototype P_s 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

$$P_s = \frac{1}{N_s + 1} O_t^n + \left(1 - \frac{1}{N_s + 1}\right) P_s \ , \qquad (3)$$

where N_s is the total number of observations associated with the prototype P_s at time t.

Since the motion patterns O_t^n and P_s are multi-variate Gaussian distributions, the update equation must reflect a weighted sum of the actual data the distributions represent. Thus we update the weighted mean distribution P_s with respect to the KL-divergence using the expected centroid introduced by Myrvoll and Soong [17]. The use of online clustering does not require the number of prototypical patterns to be known ahead of time, only a specified distance threshold d_{KL}. The mean distribution P_s provides a canonical representation of a prototypical motion pattern whose cuboids collectively share similar motion distributions.

The prototypical motion patterns represent a common motion activity within the entire video volume. To fully capture the motion characteristics of the scene, we model the variation of motion pattern representations within each prototypical classification. Each prototype P_s represents a collection of local spatio-temporal motion patterns. Since the motion patterns themselves are modeled as multivariate Gaussian distributions, the motion variation of each prototype can be viewed as a distribution of distributions. We model these collections as 1D normal distributions, specifically by the mean P_s, which is a 3D Gaussian distribution itself, and the standard deviation σ_s, computed using the av-

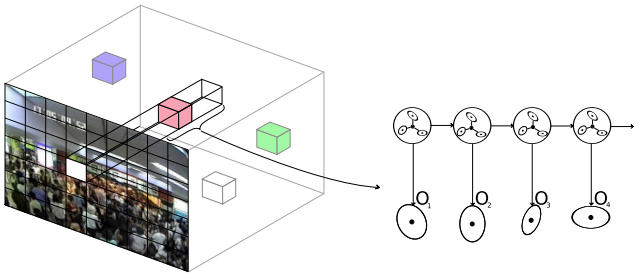

Figure 2. The temporal relationship between local spatio-temporal motion patterns is encoded in a distribution-based HMM at each spatial location.

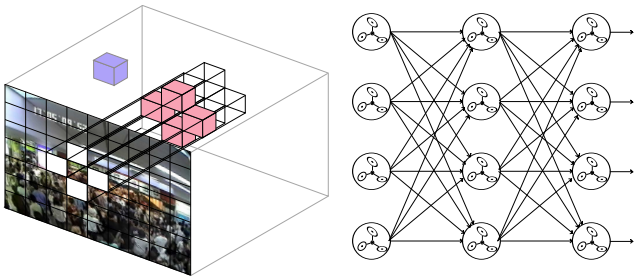

Figure 3. We capture the spatial relationships between local spatio-temporal motion patterns in a coupled HMM that encompasses spatially local tubes.

erage KL distance to the mean. Thus we capture the characteristic motion of a scene by identifying the motion variations within each prototypical pattern, compactly representing the rich information stored in the local spatio-temporal motion patterns.

Now that we have a characteristic representation of the motion within a scene, we may potentially model an extremely 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. Given a local spatio-temporal motion pattern O_t^n, we can evaluate the probability of it belonging to a specific prototypical observation using the KL distance to remove the bias. Thus the probability of an observation O_t^n given prototype s is

$$p\left(O_t^n|s\right) = p\left(\frac{d\left(O_t^n, P_s\right)}{\sigma_s}\right) \sim \mathcal{N}\left(0,1\right) \ , \qquad (4)$$

where d is the KL distance measure.

Since motion patterns that occur regularly in one spatial location of the video may be unusual in another, observations are only evaluated for prototypical distributions that occurred in the same spatial location n, or tube, in the training video. We compute the confidence measure for each cuboid as the maximum likelihood given the possible prototypical distributions within a tube. We then identify unusual motion patterns by thresholding low confidence values. Since extremely crowded scenes may contain larger variations in one location than another, we normalize the measure by the minimum confidence value of the training set in each spatial location n.

4. Capturing Temporal Statistics in Distribution Based Hidden Markov Models

While the set of prototypes provides a picture of similar activities in the scene, it does not capture the relationship between their occurrences. As a result, we cannot assume that the approach in the previous section will lead to robust detection of unusual activities. We will now consider modeling and leveraging the temporal dynamics of the motion patterns. Since the scene is comprised of physically moving objects we assume that cuboids in the same spatial location exhibit the Markov property in the temporal domain. In order to achieve a localized model, we observe each spatial location separately, creating a single HMM for each tube of observations as shown in Fig. 2.

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 $n = 1, \ldots, N$. The set of possible observations $\mathbf{O}^n$ is a continuous range of 3D Gaussian distributions. Complex observations for HMMs are often quantized, however this would significantly reduce the rich motion information in each cuboid. We associate the hidden states H^n with the prototypes S^n in the tube n, and use Eq. 4 to evaluate the emission probabilities. Note that, while a single HMM is created for each tube, the emission probability density functions are created using samples from the entire video volume. This construction permits the observations to remain continuous 3D Gaussian distributions, thus capturing the temporal relationships between motion patterns while maintaining their dense motion pattern representation. We do not re-train the emission densities for the HMM, as the prototypes already provide a good approximation and re-estimation is computationally costly. The parameters $\mathbf{A}^n$ and $\boldsymbol{\pi}^n$ are estimated by expectation maximization.

The likelihood of an observation sequence given an HMM is traditionally evaluated by the forwards-backwards algorithm [20]. However, we would like to evaluate each individual cuboid. Primarily, we are interested in using temporal statistics to indicate unlikely transitions between cuboids. Thus we evaluate a specific cuboid using the predictive likelihood and a single motion pattern following it. Our temporal confidence measure ρ_t^n for observation O_t^n is

$$\begin{aligned} \rho_t^n \quad = \quad & \log\left(\sum_{s_t \in S^n} p\left(s_{t+1}|s_t\right) p\left(s_t|O_t\right)\right) + \\ & \log\left(p\left(O_t^n|O_1^n, \ldots, O_{t-1}^n\right)\right) \ , \end{aligned} \qquad (5)$$

Figure 4. The two data sets courtesy of Nippon Telegraph and Telephone Corp. from an extremely crowded subway station during rush hour. The ticket gate data set (top) and concourse data set (bottom) both contain thousands of pedestrians, extreme occlusions, and irregular motion patterns.

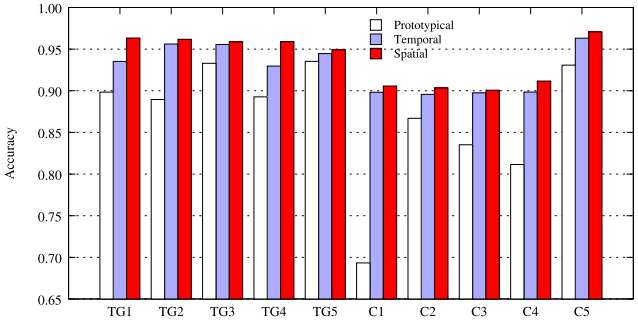

Figure 5. Our local spatio-temporal motion pattern modeling successfully detects irregular motion patterns in the ticket gate (TG) and concourse (C) data sets. The inclusion of temporal and spatial statistics significantly improvements detection on our test sets.

where s_{t+1} is the most likely prototypical class of observation O^n_{t+1}. This approach uses the state transition information captured in the HMM to include temporal information in the confidence measure for each cuboid.

By training each distribution-based HMM, we capture the steady-state temporal motion relationships within the extremely crowded scene. Incorporating such temporal information into our confidence measure enables the detection of motion patterns with unusual temporal transitions, specifically those that do not typically occur adjacent to one another. The increase in performance accuracy over the static approach described in the previous section is illustrated in Fig. 5. The results show that by modeling the relationships between surrounding cuboids, we can further identify anomalies in extremely crowded scenes as unusual temporal sequences. Temporal statistics, however, do not capture the relationship between cuboids surrounding specific motion patterns, only those before and after it.

5. Coupling of Spatial Relationships

The connected structure of extremely crowded scenes results in a strong correlation between spatially neighboring motion patterns. The motion of pedestrians across the frame exhibits relationships among patterns in close proximity, and the physical construction of the scene causes similar motion patterns to occur in spatially neighboring locations. For example, if traffic flow is interrupted by a physical obstruction, pedestrians must pause or navigate around the affected area. The density of extremely crowded scenes causes pedestrians to react to activities directly surrounding them. We model this strong relationship between spatially neighboring cuboids to identify motion patterns with spatial interaction.

In order to capture the correlation between spatially neighboring cuboids, we construct a coupled Hidden Markov Model among surrounding tubes as illustrated in Fig. 3. We create a coupled HMM using motion patterns from the tubes above, below, and to either side of each spatial location n. As with the temporal statistics, the emission probabilities are computed from the prototypical motion pattern distributions, retaining the rich motion information within the local spatio-temporal motion pattern representation. Our goal is to model the activity surrounding the local motion pattern, thus we exclude location n from the coupling. Inclusion of the motion patterns within the coupled model, however, would not provide additional information, merely mimic the behavior of larger cuboids that encompass all of the pixels in surrounding tubes. Alternative models, such as a Markov Random Field, would model the relationship between surrounding cuboids in the same temporal frame. Since the scene is comprised of physically moving objects, we assume that the primary correlation between motion patterns occurs across temporal frames.

In order to efficiently compute a spatial confidence measure, we use the N-heads dynamic programming inference algorithm introduced by Brand [6]. We are specifically interested in measuring the likelihood of neighboring motion patterns and their association strengths. The spatial confidence measure ρ^n_s is computed by

$$\rho^n_s = \log\left(\sum_{s_t \in S^{n'}} p\left(O^{n'}_t|s_t\right) \sum_{s_{t-1} \in S^{n'}} p\left(s_t|s_{t-1}\right) q_{t-1,k_{s_t}}\right), \tag{6}$$

where n' is the neighborhood of location n, q is the partial posterior, and k the associated sidekick as discussed by Brand [6]. Thus our spatial confidence measure captures the likelihood of surrounding motion patterns and the temporal transitions across tubes.

Coupled HMMs for spatially neighboring tubes capture the spatial-temporal correlation between neighboring local spatio-temporal motion patterns. Thus our confidence measure indicates anomalies in surrounding areas of the video frames. In order to directly measure the effects of surrounding tubes with the motion pattern at location n, we use the

Figure 6. Detection of unusual motion patterns in the concourse video (top) and the ticket gate (bottom). Correctly classified unusual cuboids are highlighted in blue, usual cuboids in green, false negatives in red, and false positives in magenta. The intensity of magenta blocks indicates the severity of false positives. Individuals reversing direction and moving in irregular patterns are correctly classified as unusual.

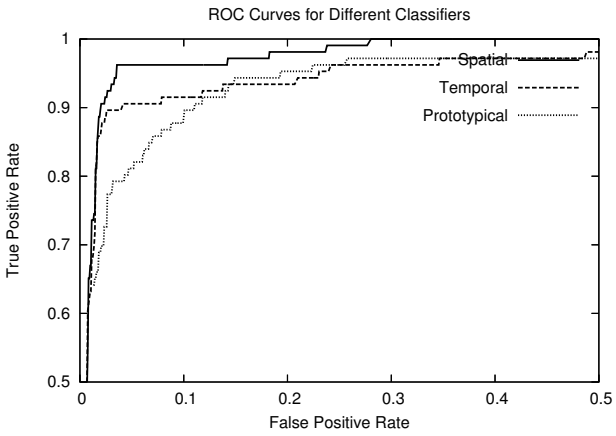

Figure 7. ROC curves on the first ticket gate data set query video. The temporal and spatial approaches (Section 4 and 5, respectively) provide superior results compared to merely using prototypical distributions (Section 3).

spatial confidence measure in conjunction with the temporal measure given in Eq. 5 for classification. Specifically, we take a linear combination $\alpha\rho_t^n + (1-\alpha)\,\rho_s^n$ of the spatial and temporal confidence measures, combining the spatial and temporal statistical information to identify unusual relationships between motion patterns.

6. Results

We evaluate our approach on two extremely crowded real-world scenes from a subway station during rush hour[1]. The first set, shown on the top of Fig. 4, is from a ticket gate. The second, on the bottom of Fig. 4, is from the stations concourse. Both data sets contain large numbers of pedestrians moving in irregular motion patterns with frequent occlusions. We use a total of 10 query videos with hand-labeled ground truth to thoroughly and quantitatively evaluate the effectiveness of local spatio-temporal motion pattern models. The distance threshold d_{KL} is selected empirically by evaluating the performance of the prototypical distributions. The temporal and spatial analysis is then evaluated using the best performing distance threshold, and the mixing coefficient α is selected empirically. The cuboid size is set to $30 \times 30 \times 20$ and $40 \times 40 \times 20$ for the ticket gate and concourse data set, respectively. The length of training data varied for each example between 27 and 150 observations, depending on the real-world data available.

Abnormal events such as pedestrians moving in irregular directions, individuals obstructing traffic, or a lack of pedestrians in an otherwise highly-crowded area are detected. The results are shown in Fig. 5. For all three approaches, we remove variations in the training data by normalizing the confidence measure with the minimum training data likelihoods in each tube. We measure accuracy by the maximum average of the sensitivity and specificity for varied confidence thresholds. Our prototypical motion pattern approach performs well for most of the query videos, achieving over eighty percent accuracy for all but one example. The receiver operator characteristic curves for the TG1 data set are shown in Fig. 7. The temporal approach vastly improves this accuracy, and the incorporation of spatial statistics proves superior.

Two frames from query videos are shown in Fig. 6. In the top frame, from the ticket gate data set, a pedestrian who has reversed direction is successfully detected as unusual. The physical environment in the ticket gate data set forces pedestrians to move forward through the gate during usual activity. Other areas of the frame, such as the upper and lower regions, contain less structure, however are still modeled correctly by the local spatio-temporal motion patterns. In the lower frame, station employees moving against the general flow of pedestrians are successfully detected. Pedestrians in the concourse data set have fewer physical limitations imposed by the environment, resulting in much larger motion pattern variations than the ticket gate data set.

Lack of motion in areas typically containing high motion are also detected as unusual, as shown in Fig. 8. In the top frame, pedestrians are not using specific ticket gates, an area of usual high traffic. In the lower frame, the far right area typically contains large amounts of pedestrians. Fig. 10 shows successful detection of pedestrians loitering in high traffic areas.

False positives occur in both experiments for slightly irregular motion patterns that may not have been captured in the training data due to unusual textures, as shown in Fig. 8. Some motion pattern variations have similar severity to

[1]The original video sequence courtesy of Nippon Telegraph and Telephone Corporation. Please see supplementary video available at `http://www.cs.drexel.edu/~lak24/cvpr09`.

Figure 8. Areas lacking in pedestrian motion that usually contain it are successfully detected in both data sets. False positives indicate a sensitivity to unusual textures.

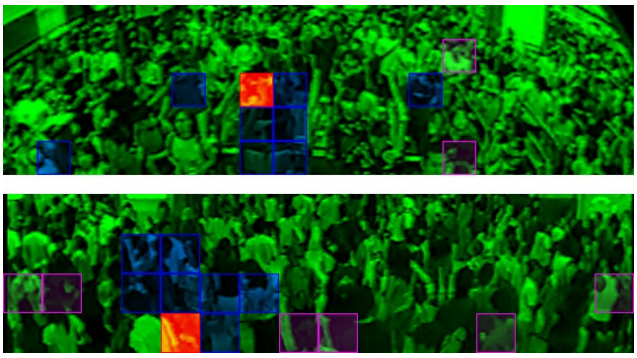

Figure 10. Pedestrians loitering or not moving in areas of high traffic are successfully detected. A few false negatives are adjacent to true positives, thus we consider them harmless in practical scenarios.

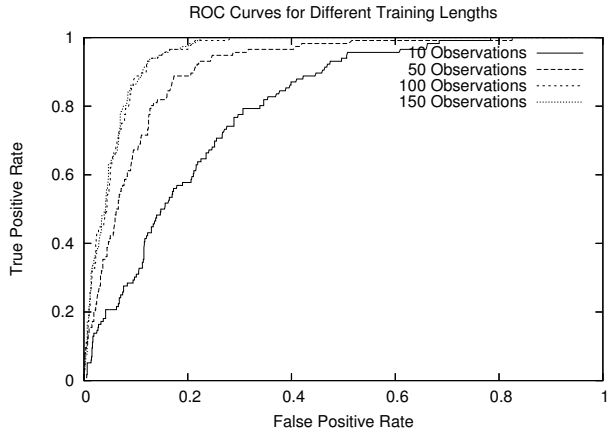

Figure 9. Effects of increased training data on the query video C1. Our approach performs well with as little as 50 observations.

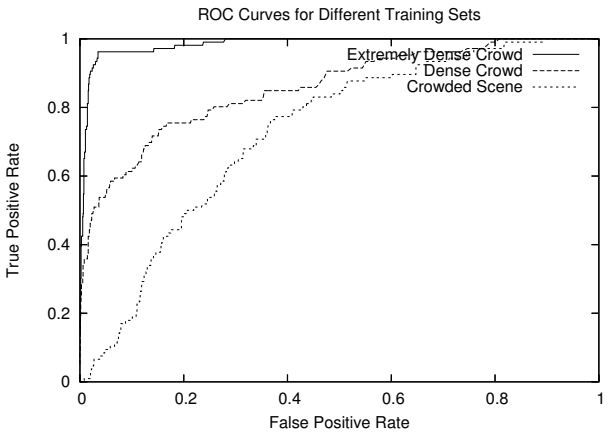

Figure 11. Results for query video TG1 using three different training sets .

those detected as unusual. The few false negatives in both real-world examples always occur adjacent to true positives, as shown in Fig. 10, which suggests they are harmless in practical scenarios. Many of the detected events are subtle, and as such result in higher false positive rates than unusual events requiring personnel intervention. Their subtle nature, however, shows the ability of our approach to capture the steady state motion of the scene.

The effects of increasing the training data size for the concourse dataset are shown in Fig. 9. As expected, the performance increases with longer training data, and approaches a steady state with 100 observations per tube. The performance with only 50 observations per tube achieves a false positive rate of 0.17 and true positive rate of 0.89. The strong performance with such small training data reflects the robust motion pattern representation captured by the distribution-based models.

Due to the diversity of our query data, different training sequences are selected to ensure correct modeling of the steady-state motion of the scenes. The concourse data set, for example, has frequent changes of crowd density throughout the duration of the video. The effects of using different training sets on sequence TG1 are illustrated in Fig. 11. We used three different training data sets: a crowded scene, a densely crowded scene, and an extremely dense crowded scene. Even the least crowded scene in this example contains hundreds of individuals within the frame. In a real-world implementation, all training sets would be accounted for to handle non-subtle unusual events. As expected, the extremely crowded scene performs best, however even the densely crowded scene performs with significant accuracy as a result of the rich descriptive motion information captured in our distribution-based models.

7. Conclusion

Extremely crowded scenes provide unique challenges to video event analysis, primarily due to the high density of the crowds and frequent occlusions. In this paper, we introduced a novel framework for modeling the motion patterns of extremely crowded scenes and detect-

ing unusual events. We represent the rich, non-uniform, localized motions patterns with 3D Gaussian distributions of spatio-temporal gradients. The temporal relationship between local spatio-temporal motion patterns is captured via a distribution-based HMM, and the spatial relationship by a coupled HMM. 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 videos of extremely crowded scenes in which they successfully detect unusual motion patterns in pedestrian behavior including movement against the normal flow of traffic, loitering, and traffic congestion. We believe the proposed framework plays an important role in video analysis of extremely crowded scenes. Currently, we are investigating the use of varying size cuboids for scenes with strong perspective projections.

Acknowledgement

This work was supported in part by National Science Foundation grants IIS-0746717 and IIS-0803670, and Nippon Telegraph and Telephone Corporation. We thank Kensaku Fujii and Hiroyuki Arai of Nippon Telegraph and Telephone Corporation for providing the train station videos.

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(3):555–560, Mar. 2008.

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

[3] E. Andrade, S. Blunsden, and R. Fisher. Modelling Crowd Scenes for Event Detection. In *Proc. of International Conf on Pattern Recognition*, pages 175–178, 2006.

[4] M. Black. Explaining Optical Flow Events with Parameterized Spatio-Temporal Models. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, pages 326–332, 1999.

[5] O. Boiman and M. Irani. Detecting Irregularities in Images and in Video. In *Proc. of IEEE Int'l Conf on Computer Vision*, pages 462–469, 2005.

[6] M. Brand. Coupled Hidden Markov Models for Modeling Interacting Processes. Technical report, MIT, 1996.

[7] M. T. Chan, A. Hoogs, J. Schmiederer, and M. Petersen. Detecting Rare Events in Video Using Semantic Primitives with HMM. In *Proc. of International Conf on Pattern Recognition*, pages 150–154, 2004.

[8] P. Cui, L. F. Sun, Z. Q. Liu, and S. Yang. A Sequential Monte Carlo Approach to Anomaly Detection in Tracking Visual Events. In *IEEE Workshop on Visual Surveillance*, pages 1–8, 2007.

[9] H. Dee and D. Hogg. Detecting Inexplicable Behaviour. In *Proc. of British Macine Vision Conf*, pages 477–486, 2004.

[10] D. DeMenthon and D. Doermann. Video Retrieval using Spatio-Temporal Descriptors. In *Proc. of the Eleventh ACM Int'l Conf on Multimedia*, pages 508–517, 2003.

[11] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior Recognition via Sparse Spatio-Temporal Features. In *Int'l Workshop on Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005.

[12] J. Gryn, R. Wildes, and J. Tsotsos. Detecting Motion Patterns via Direction Maps with Application to Surveillance. In *IEEE Workshop on Motion and Video Computing*, pages 202–209, 2005.

[13] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A System for Learning Statistical Motion Patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(9):1450–1464, Sep. 2006.

[14] N. Johnson and D. Hogg. Learning the Distribution of Object Trajectories for Event Recognition. In *Proc. of British Macine Vision Conf*, pages 583–592, 1995.

[15] Y. Ke, R. Sukthankar, and M. Hebert. Event Detection in Crowded Videos. In *Proc. of IEEE Int'l Conf on Computer Vision*, pages 1–8, 2007.

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

[17] 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.

[18] K. Nishino, S. K. Nayar, and T. Jebara. Clustered Blockwise PCA for Representing Visual Data. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10):1675–1679, Oct. 2005.

[19] PETS. Tenth IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance. `http://www.pets2007.net/`, 2007.

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

[21] J. Salas, H. Jimenez, J. Gonzalez-Barbosa, J. Hurtado-Ramos, and S. Canchola. A Double Layer Background Model to Detect Unusual Events. In *Proc. of Adv. Concepts for Intelligent Vision Systems*, pages 406–416, 2007.

[22] E. Shechtman and M. Irani. Space-Time Behavior Based Correlation. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, pages 405–412, 2005.

[23] T. Xiang and S. Gong. Online Video Behaviour Abnormality Detection Using Reliability Measure. In *Proc. of British Macine Vision Conf*, 2005.

[24] H. Zhong, J. Shi, and M. Visontai. Detecting Unusual Activity in Video. In *Proc. of IEEE Int'l Conf on Computer Vision and Pattern Recognition*, pages 819–826, 2004.

[25] H. Zhou and D. Kimber. Unusual Event Detection via Multi-camera Video Mining. In *Proc. of International Conf on Pattern Recognition*, pages 1161–1166, 2006.