Mikel Rodriguez, Saad Ali, Takeo Kanade, "Tracking in Unstructured Crowded Scenes," *The Twelfth IEEE International Conference on Computer Vision*, September 2009.

Tracking in Unstructured Crowded Scenes

Mikel Rodriguez
Computer Vision Lab
University of Central Florida
mikel@ucf.edu

Saad Ali
Robotics Institute
Carnegie Mellon University
saada@cs.cmu.edu

Takeo Kanade
Robotics Institute
Carnegie Mellon University
tk@cs.cmu.edu

Abstract

This paper presents a target tracking framework for unstructured crowded *scenes. Unstructured crowded scenes are defined as those scenes where the motion of a crowd appears to be random with different participants moving in different directions over time. This means each spatial location in such scenes supports more than one, or multi-modal, crowd behavior. The case of tracking in* structured crowded scenes*, where the crowd moves coherently in a common direction, and the direction of motion does not vary over time, was previously handled in [1]. In this work, we propose to model various crowd behavior (or motion) modalities at different locations of the scene by employing Correlated Topic Model (CTM) of [16]. In our construction, words correspond to low level quantized motion features and topics correspond to crowd behaviors. It is then assumed that motion at each location in an unstructured crowd scene is generated by a set of behavior proportions, where behaviors represent distributions over low-level motion features. This way any one location in the scene may support multiple crowd behavior modalities and can be used as prior information for tracking. Our approach enables us to model a diverse set of unstructured crowd domains, which range from cluttered time-lapse microscopy videos of cell populations in vitro, to footage of crowded sporting events.*

1. Introduction

A crowded scene can be divided into two categories: structured and unstructured. In a structured crowded scene, the crowd moves coherently in a common direction, and the direction of motion does not vary over time. That is, each spatial location of the scene supports only one dominant crowd behavior over the video. For instance, a video of a marathon race represents a structured crowded scene because all athletes run along the same path, thus generating a crowd behavior which has a fixed direction of motion/pattern at each location of the path. Other examples of structured crowded scenes include processions, events involving queues of people, and traffic on a road (see Figure1). In an unstructured crowded scene, the motion of the crowd appears to be random, with different participants moving in different directions at different times. That is, in such scenes each spatial location supports more than one, or multi-modal, crowd behavior. For instance, a video of people walking on a zebra-crossing in opposite directions is an example of an unstructured crowded scene because, broadly speaking, at any point on the zebra crossing the probability of observing a person moving from left to right is as likely as observing a person walking from right to left (see Figure 2). Other examples of such scenes include exhibitions, crowds in a sporting event, railway stations, airports, and motion of biological cells (see Figure 1).

Recently Ali *et al.* [1] proposed an algorithm to track objects in structured crowded scenes. Their method is based on the assumption that, in a given scene, all participants of the crowd are behaving in a manner similar to the global crowd behavior. Therefore, at any location in the scene, there is only one direction of motion. This enabled them to learn a higher level constraint or prior on the direction of motion for tracking purposes using the novel construct of 'floor fields.' Given that floor fields could only be learned when there is one dominant direction of motion, the results were reported only on marathon videos. This is a major shortcoming, as floor fields could not be learned for unstructured crowded scenes where each location in the scene supports multiple dominant crowd behaviors. We further explain this point with the aid of an example in which pedestrians are walking on a zebra-crossing (see Figure 2). In the crossing, people walk in both directions and, therefore, each spatial location supports two different dominant types of motion over time which correspond to two different high level crowd behaviors. This means the motion at each spatial location of the zebra-crossing has a multi-modal representation. This is evident from the two types of dominant optical flow vectors for this scene that are shown in Fig. 2(c), and the two corresponding high level crowd behaviors learned by our algorithm that are shown in Fig. 3. In this figure, different slices are shown for different behaviors to emphasize the fact that a single location in the scene can



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

Figure 1. Several instances of structured and unstructured crowded scenes. (a) Structured, (b) Unstructured, (c) Structured, and (d) Unstructured.

support to any number of crowd behaviors with different probabilities.

Floor fields in the form proposed by Ali *et al.* [1] will not be able handle this situation, and therefore, their tracking algorithm is not directly applicable to unstructured crowded scenes. Since such unstructured crowd activities (airports, exhibition halls, and stadiums) are much more common than structured crowd activities (marathons), it is important to develop an algorithm capable of handling multi-modality in crowd behaviors and for using it as a high-level direction prior for tracking.

To overcome the problem of tracking in unstructured crowded scenes, we develop a tracking algorithm that uses Correlated Topic Model (CTM) [16] to capture different overlapping and non-overlapping crowd behaviors in the scene. In our construction, words correspond to low level quantized motion features and topics correspond to crowd behaviors. We used CTM as it provides an elegant way to handle multi-modality of crowd behavior as each location can have a certain probability of belonging to certain crowd behavior (or topic). In addition, it models interactions among topics (or crowd behaviors) which is also desirable, as explained later, for the types of scene that we are handling. Note that, we will use terms 'crowd behavior', 'behavior' and 'topic' interchangeably throughout the paper.

For illustration and understanding purposes we show

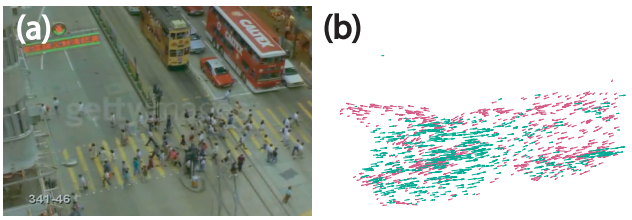


Figure 2. (Frames from a video showing pedestrians crossing the road, and the corresponding optical flow vectors generated by the motion of the crowd. Different colors of optical flow vectors represent two dominant motions of the crowd in this scene. The corresponding two crowd behaviors learned from these optical flow vectors are shown in Fig. 3.

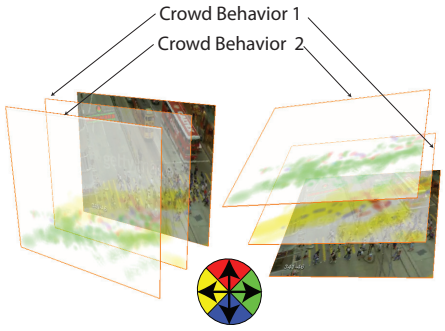


Figure 3. The top two crowd behaviors corresponding to a crosswalk scene capture the multiple behavior modes of pedestrians walking from opposing sides of the street. Behavior 1 captures the dynamics of pedestrians which walk towards the rightmost end of the crosswalk, whereas behavior 2 captures the typical movement of pedestrians which walk towards the leftmost end of the street.

crowd behaviors learned for a typical busy crosswalk scenario in Figure 2. Our model is able to capture different behavior modalities at specific locations in the scene. This can be observed in Figure 3, where we overlay the two most common crowd behaviors learned by our framework. By observing the colors (which represent directions of motion) in each of the crowd behaviors, it can be seen that one of the behaviors corresponds to pedestrians which walk towards the rightmost end of the crosswalk, whereas the other behavior corresponds to pedestrians walking in the opposing direction.

Also note that, in Figure 3 learned crowd behaviors are not spatially mutually exclusive. Therefore, multiple behaviors can occur at different spatial locations in the scene with certain probabilities. Each of these behaviors can then be incorporated as high level information which can aid tracking individuals in this class of scenes. The main contributions of our work are: 1) Extending the idea of using high-level knowledge for tracking in crowds by learning representations of unstructured and multi-modal crowd behavior; 2) Using CTM to solve an existing problem in a crowd tracking framework.

2. Related Work

Tracking is one of the highly researched areas in the field of computer vision. Most tracking algorithms proposed over the years focus on the general problem of tracking, without specifically addressing the challenges of a crowded scene. In this section, we review the tracking methodologies that are specifically designed for crowded situations. The readers interested in a detailed review of the state of the art in tracking are referred to a recent survey by Yilmaz *et al.* [7].

An interesting body of work tries to track sparse crowds of ants [4] and people [8], hockey players [6], crowds of densely packed people [5, 11, 12], a dense flock of bats [2], and biological cells [3]. In their work, Brostow *et al.* [8] tracked and clustered feature points over time and used them to generate a separate trajectory for each individual. In [4], Khan *et al.* employed a Markov chain Monte Carlo based particle filter to deal with interactions among targets in a crowded scenario. They used the intuitive notion that in a crowded situation the behaviors of targets are influenced by the proximity and/or behavior of other targets. Cai *et al.* [6] proposed a mutli-target tracking algorithm for tracking hockey players in a video. In [11, 12], Lin *et al.* advocated a different paradigm for tracking groups of people by treating them as a near-regular texture (NRT). Recently, Betke *et al.* [2] proposed an algorithm to track a dense crowd of bats in thermal imagery. They combined multiple techniques such as multi-target track initiation, recursive Bayesian tracking, clutter modeling, event analysis, and multiple hypotheses filtering for this purpose. Tracking of multiple interacting and crowded objects has been attempted in the area of biological cell tracking as well. For instance, Li *et al.* [3] have recently developed an algorithm for tracking thousands of cells in phase contrast time-lapse microscopy images. Another approach for tracking in crowded scenes using selective visual attention is proposed by Yang *et al.* [13]. Most tracking algorithms described so far only use low-level image information for tracking purposes. Surprisingly, little has been done in exploiting high-level cues for human detection and tracking in complex crowded situations. One of the few works on this topic is that of Antonini *et al.* [14], which used discrete choice models (DCM) [15] as motion priors to predict human motion patterns and fused this model in a visual tracker for improved performance. The other work that used high-level motion priors for tracking is by Ali *et al.* [1] which we already described in the previous section.

In contrast to above mentioned body of work, our method addresses the problem of tracking in high density crowds by learning high level direction priors for *unstructured crowded scenes*. To the best of our knowledge, this has not been attempted before.

3. Unstructured Crowded Scene Model

In this work, our goal is to develop a framework for modeling the dynamics of crowded and complex scenes. In general, an effective crowded scene model will need to be capable of both capturing the correlation amongst different patterns of behavior as well as allowing for the multi-modal nature of crowded scenes over time. The importance of correlation among themes can be explained as follows. At an intersection of roads, the presence of pedestrians walking from one end of the crosswalk to the other will likely coincide with a crowd behavior which corresponds to pedestrians crossing from the opposite side of the crosswalk. That is these two behaviors are correlated. On the other hand, in the presence of a behavior which corresponds to vehicle traffic, it is not likely that we will observe pedestrians walking across the scene.

In this work, we model the given crowd scene using Correlated Topic Model (CTM) which is adopted from the text processing literature. CTM offers an elegant framework within which multi-modality of crowd behaviors and correlations among them can be handled. Another benefit of using topic Models like CTM is that they enables us to bypass the need for object detection within this class of crowded scenes in favor of direct processing on low level flow vectors and at the same time allow us to connect these low level features with high level crowd behaviors. It is pertinent to mention here that we are interested in capturing the interrelationships between different behaviors in the scene as well, therefore it may not be appropriate to rely on topic models based on Latent Dirichlet Allocation (LDA) [17] to model scene dynamics [20]. This is due to the fact that these models assume a near independence of topics (or behaviors). On the other hand, CTM addresses this limitation by introducing a logistic normal prior of topics instead of the Dirichlet prior and by using the covariance matrix of the variables in the logistic normal model to capture correlations among topics (or crowd behaviors).

The elements of CTM and their conditional dependencies are depicted in the graphical model shown in Figure 4. In this figure, shaded variables represent the observed variables (motion words), while the unshaded variables represent the latent variables. Edges encode the conditional dependencies of the generative process. In the following section we detail the terminology of this graphical model and describe how it enables us to capture the multi-modal nature of a crowded scene.

3.1. Notation and Terminology Overview

In this section, we explain various terminologies and demonstrate the mapping between original CTM model and our scenario. The only observable random variable that we consider is the low-level motion feature x, which cor-

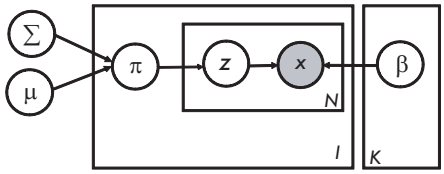


Figure 4. Graphical model used for modeling various crowd behaviors and correlation between them.

responds to a quantized optical flow vector and location. A video is represented by N such low-level visual features (motion words) $X = \{x_1, x_2, \ldots, x_N\}$ where x_n is the nth motion word in the sequence, defined by a displacement and location in the scene (u, v, x, y). The number of words in the ith video is represented by N_I. The crowd behavior β is a distribution over the vocabulary of motion words of size V. It represents a point on a $V-1$ simplex. A model with K crowd behaviors is represented by a matrix β of size $K \times V$, where the ith row represents a distribution of the ith crowd behavior over the vocabulary. Each motion word x is associated with a crowd behavior z drawn from one of the K crowd behaviors. The behavior assignment $z_{i,n}$ is associated with the nth motion word and ith sequence.

In our model, the notion of document is generated by dividing each video sequence is into short clips (documents), and then each clip is associated with a set of crowd behavior proportions θ_i, which represents a point on $K-1$ simplex. θ_i represents the probability with which motion words are drawn from each behavior. It is obtained by mapping the behavior probability vector π to a simplex as $\theta = \frac{exp(\pi)}{\sum_i exp(\pi_i)}$, and thus obtaining a multinomial parameter.

3.2. Generative Process

Our model assumes that the N motion words from a video sequence (short clip) i arise from the following generative process:

- Randomly draw a $k-$dimensional vector $\pi \backsim p(\pi \mid \Sigma, \mu)$ that determines the distribution of intermediate spatio-temporal crowd behaviors. Here, π is a parametrization of the multinomial distribution ($\pi = log(\frac{\theta_i}{\theta_K})$) that captures the covariance among these behaviors, while μ and Σ are the mean and covariance of the normal distribution.
- For each motion word x_n in the sequence
  - Choose a behavior $z_n \backsim \text{Mult}(\frac{exp(\pi)}{\sum_i exp(\pi_i)})$. z_n is a K-dimensional unit vector where $z_n^k = 1$ indicates that the kth behavior is selected.
  - Choose a low-level motion feature $x_n \backsim p(x_n \mid z_n, \beta)$, where β is distribution over the vocabulary of motion words.

In order to perform parameter estimation for our model, we use a collection of training video sequences and adopted the variational expectation maximization (EM) algorithm proposed in [16]. We refer reader to this reference for further details on the parameter estimation algorithm.

4. Tracking Framework

In this section we describe implementation details of various steps involved in our tracking framework.

4.1. Scene Codebook

Given a video of a specific scene, we uniformly divide it along the temporal domain into non-overlapping short clips. In our framework, each of these video clips is treated as a document. For each clip in our dataset, we compute optical flow as our low-level features. All moving pixels in each video sequence are quantized according to a codebook in the following manner: Each moving pixel has two features which correspond to its position and its direction of motion respectively. Position is quantized by dividing a scene into a grid with cells which are 10×10 pixels in size. The motion of a moving pixel is quantized into four directions of motion. Therefore, for a scene which is digitized to a size of 320×240 the size of the codebook is $32 \times 24 \times 4$, and thus each detected moving pixel is assigned to a word from the codebook based on rough position and motion direction.

The size of the codebook depends on the granularity of the spatial and motion direction quantization, a choice which represents a balance between the descriptive capability and complexity of the model. We found that increasing the size of the codebook resulted in diminishing returns. Therefore, for all of our experiments we maintain the quantization described above.

In the next section we describe how we formulate tracking in complex scenes by incorporating the high-level information about a scene, which is captured by estimating parameters of CTM described previously.

4.2. Formulation

Let the observed measurements of m objects in the scene at time instance i be given by $\Omega = \{\omega_i^1, \omega_i^2, ..., \omega_i^m\}$, and let the predicted states of the previously observed s objects be given by $\Theta = \{\theta_i^1, \theta_i^2, ..., \theta_i^m\}$. In the proposed work, the analysis is performed on a feature space which consists of a pair of 2-D locations of the centroid of object before and after transition, and time taken to execute the transition.

For each object observed at time instance $i-1$, we obtain the next tracker position as a weighted mean of the next observation and the tracker prediction by incorporating the learned high-level scene dynamics as weights. Specifically, the state of the tracker at time instance i given all previous tracks is given by:

$$\sum^{j=i} p(x_{\omega_j^k}|\Sigma,\mu,\beta)\omega_i^k + p(x_{\theta_l^k}|\Sigma,\mu,\beta)\theta_l^k \qquad (1)$$

where $x_{\omega_j^k}$ corresponds to motion word of the displacement which commences at the location given by the tracker at $i-1$ and the current observation (ω_j^k). Similarly, $x_{\theta_l}^k$ is the codebook entry (motion word) that corresponds to the displacement vector from the previous tracker position to the current tracker prediction (θ_l^k).

This approach results in the assignment of larger weights for tracks which are made up of transitions which are more likely given the learned crowd behaviors. The first term in equation 1 weights the displacements of the observations based on the learned crowd behaviors, whereas the second term weights the displacements predicted by the tracker. These weights help establish correspondences such that the probability of an object's track is maximized based on the typical behavior modalities observed in the scene.

4.3. Experiments and Results

4.3.1 Datasets

In this work our data consists of crowded and complex video sequences which contain many interactions amongst agents. We explore different crowd domains, which range from cluttered time-lapse microscopy videos of cell populations in vitro to footage of crowded sporting events. In each of these domains objects move in complex patterns, such that any one location in the scene may host multiple modalities of motion direction at different times throughout the scene.

4.3.2 Tracking Human Crowds

A first round of experiments was geared towards assessing the performance of the proposed crowd model in improving tracking in the presence of large crowds of humans. The first scene we considered can be seen in Figure 8-a. It consists of a crowded baseball bleachers scene in which fans move in complex patterns across the frame. By inspecting the top behaviors which are learned from the model, we observe that fans typically move in bidirectional aisles which either move up and down the bleachers or laterally across the aisles. Therefore, most locations in the scene host different behavior modes throughout different times in the scene. Unlike models which assume that participants of the crowd behave in a consistent global manner, here we learn the complex patterns of motion and incorporate this high-level information directly into our tracking framework.

A set of trajectories generated by our tracking algorithm is shown in Figure 8-b. Quantitative analysis of the tracking was performed by generating ground-truth trajectories for 60 fans, which were selected randomly from the set of

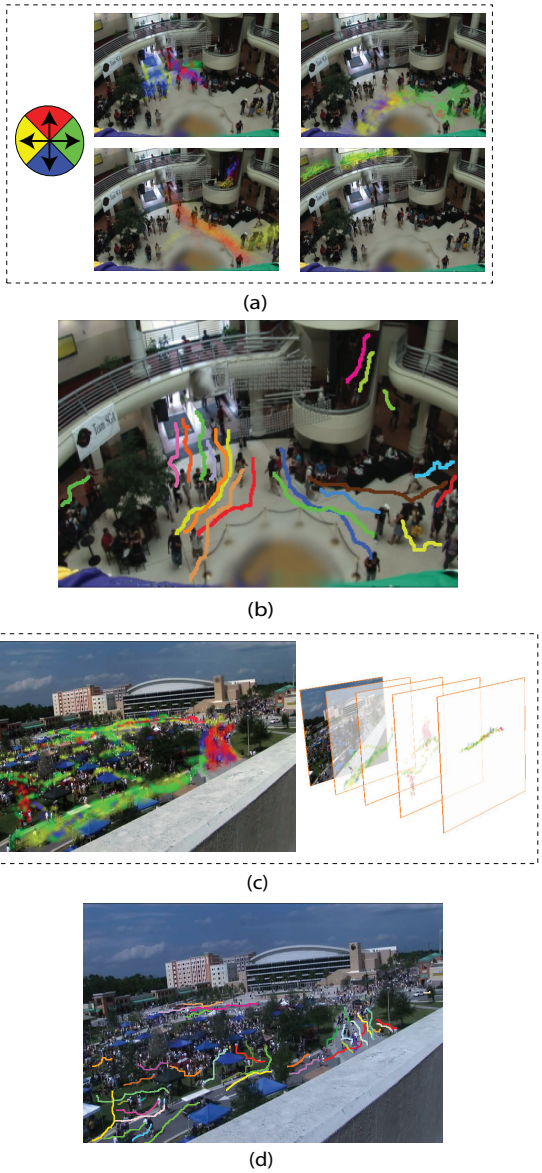


Figure 5. Tracking humans in unstructured crowded scenes. (a) The learned behavior modes for a crowded student union scene. (b) A subset of the tracks generated by our framework. (c) The learned behavior modes for a tailgating scene. (d) Tracking results.

all moving fans. The ground-truth was generated by manually tracking the centroid of each selected fan. The average tracking error obtained using the proposed model was 35 pixels, whereas an average error of 57 pixels was observed when tracking using a Kalman tracker (Figure 6-a).

The second scene we consider is a crowded student union. As can be seen in Figure 5-a, there are multiple patterns of behavior observed over time. We observe a collection of patterns which correspond to students entering and exiting the building through the same entrance. We also see multiple modes of behavior in the center of the scene, where students avoid stepping over the university logo (which is

blurred in the image). It can also be seen that the framework learns the different patterns of behavior observed at the winding stairwell, in which people ascend and descend at any given moment throughout the video sequence. A subset of the tracks generated by our framework can also be seen in Figure 5-b. We manually annotated 50 tracks and used them to compute mean tracking errors (in pixels) of the automatic tracking, which was found to be 17 pixels on average (Figure 6-b).

Another set of experiments in tracking human crowds was centered around a crowded football tailgating event which is depicted in Figure 5-c. In this scene we see how the learned behavior modes correspond to the unmarked lanes of pedestrian traffic across the lawn as well as the structured lanes of vehicle traffic. The ability to learn multiple modes of behavior is particularly important in regions of the scene that typically contain multiple patterns of motion as opposed to a single global behavior. This is the case within the lawn area of this scene, where there are no marked paths, and therefore fans tend to walk in more than one pattern. The quantitative analysis of the tailgating scene was based on 80 trajectories which were manually ground-truthed. At each frame we compare ground-truth centroids with the current tracks generated by our framework. Over the 1040 frames we report a mean tracking error of 39 pixels (Figure 6-c).

4.3.3 Tracking Cell Populations

A second round of experiments focused on automated tracking of crowded cell populations in vitro recorded using phase-contrast time-lapse microscopy. In these experiments we utilized three video sequences of human MG-63 osteosarcoma cells recorded by an 8-bit CCD camera on a Zeiss IM35 microscope. The sequences last for 43.5 hours with a frame interval of 15 minutes, corresponding to 180 frames/sequence. The frame dimensions are 400×400 pixels with a resolution of 3.9 μm/pixel at 5:1 magnification.

As can be seen in Figure 9-b, cells in these videos move in complex patterns throughout the scene, such that at any given location in the scene different behaviors are observed over time. Further complicating the tracking process is the large number of cells per video. The cell population in each of the sequences is in the range of 350-750 cells per frame. Finally, the appearance of individual cells within video sequences contains very little intra-class variation, proving it difficult to rely on appearance as a means of tracking.

Given that each cell population video sequence contains different motion dynamics, each video is considered to be a separate scene. Therefore, we learn the probabilistic crowd model described in Section 3 for each video and perform tracking in batch mode on the same sequence.

In order to assess the effectiveness of the proposed approach, we compare tracking results obtained using our tracking framework with tracks obtained using a Kalman tracker. We also assess the effectiveness of explicitly capturing the correlation amongst behaviors by replacing the logistic normal distribution in our probabilistic crowd model with a Dirichlet distribution, therefore resembling the popular Latent Dirichlet Allocation (LDA)[17].

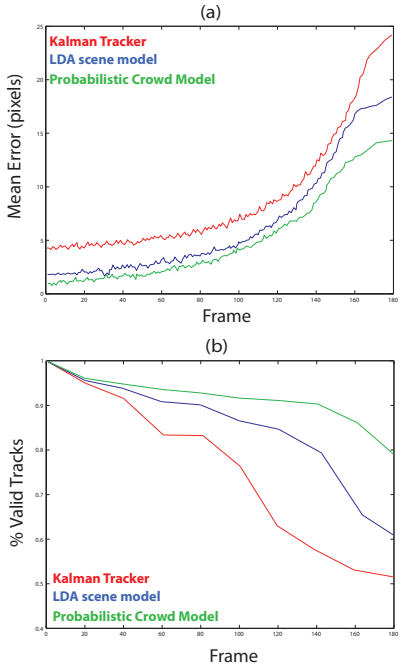


Figure 7. (a) The mean distances between the manually and the automatically tracked trajectories. (b) The percentages of cells successfully tracked.

In a first set of experiments on this dataset our tracking process was compared to that achieved by a human operator. This comparison was based on 120 cell trajectories for which manual tracks are available as groundtruth. For each cell trajectory we computed the average distance (in pixels) at each time step between the manually annotated cell locations and those computed by the algorithm. Figure 7-a shows the distribution of these distances (means and standard errors computed on the trajectories analyzed) according to time. It can be observed that a slight and progressive increase in distance occurs as the experiment progresses in time and probably results from error accumulation. The final mean distance obtained using the proposed method was 14.67, considerably less than the one obtained using the Kalman tracker and the LDA-based probabilistic crowd model. In fact, the mean error is near to the average radius of cells during mitosis, indicating that the location errors are small compared to the size of the cells. Figure 7-b displays the percentages of cells successfully tracked by the algorithm according to time. Although we observed that sometimes the algorithm loses and then recovers a cell, to sim-

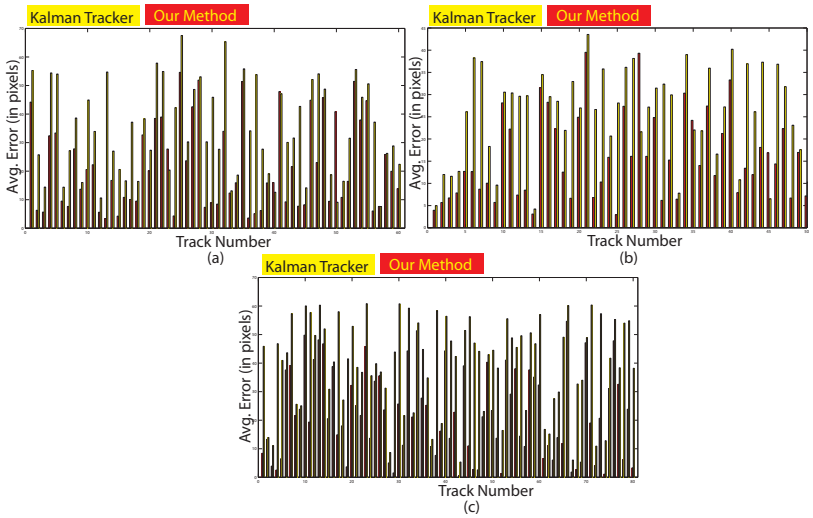


Figure 6. Comparison of the tracking error of our method against the Kalman tracker for the baseball sequence (a), crowded student union sequence (b), and the tailgating sequence (c).



Figure 8. Distribution of low-level motion features of the top four behaviors of a scene (a). The model captures the typical scene dynamics, i.e. aisles of traffic and walkways (b). We incorporate high-level scene information to improve tracking.

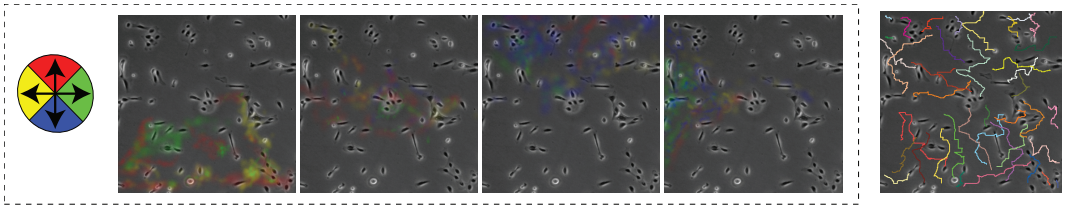
Table 1. Cell Tracking Accuracy Comparison

Sequence	Li et al[18]	Our method
A	86.4%	84.2%
B	91.2%	89.0%
C	88.2%	79.1%

plify this evaluation a cell was considered definitively "lost" by the tracking algorithm the first time that the distance between the groundtruth and automatic centroid locations exceeded a given threshold value (in our experiments we used 30 pixels). As expected, the percentages of lost cells increased with time. However, there is a significant difference in the final percentage of valid tracks between the proposed tracking framework and the comparison methods. This significant difference in performance can be attributed to the way in which the proposed model is capturing both the correlation amongst local motion as well as the multi-modality of displacements at different locations in the scene. As can be seen in Figure 7, the probabilistic crowd model is able to capture the multi-modality of local displacements in the scene. The top behaviors in the scene cover typical behavior of cells. Any given location in the scene may be included in different behaviors, each of which may capture a displacement mode of cells, a task which cannot be accomplished with crowd models which assume that all participants behave in a manner similar to a global crowd behavior.

A possible explanation for these "crowd behavior" patterns that have been learned for the MG-63 osteosarcoma cells can be found in Wang et al's work [19], in which they find that the character of cell motility is different in Sarcoma and chondrosarcoma cells. In the former, cells move over each other, and the direction of motility is not linear as has been learned in our model. Instead, according to [19] it appears that this class of cell motility is guided by collagen fibers in association with vessels.

Finally, for this dataset we compare our cell tracking re-



(a) (b)

Figure 9. (a) The top behaviors learned by the model. Each distribution over behaviors may capture different cell motion modalities for a given location in the scene. (b) Tracking results for the first thirty frames.

sults with the current state-of-the-art [18], a domain-specific tracking method which incorporates cell detection as well as cell division recognition to perform tracking in this specific domain. As can be seen in Table 1 (where we depict the percentage of valid tracks for each video in the dataset), despite the fact that we do not exploit domain-specific information (i.e cell and mitosis detection), our general approach is comparable to the best results on this dataset.

5. Conclusion

We have presented a framework for tracking individual targets in high density unstructured crowded scenes, a class of crowded scenes where the motion of the crowd at any given location is multi-modal over time. To this end we adopted the Correlated Topic Model (CTM) in which each scene is associated with a set of behavior proportions, where behaviors represent distributions over low-level motion features. Unlike some existing formulations, our model is capable of capturing both the correlation amongst different patterns of behavior as well as allowing for the multi-modal nature of unstructured crowded scenes. In order to test our approach we performed experiments on a range of unstructured crowd domains, from cluttered time-lapse microscopy videos of cell populations in vitro to videos of sporting events. In each of these domains we found that explicitly modeling the interrelationships between different behaviors in the scene allowed us to improve tracking predictions.

References

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

[2] M. Betke et al., *Tracking Large Variable Numbers of Objects in Clutter*, IEEE CVPR, 2007.

[3] K. Li and T. Kanade, *Cell Population Tracking and Lineage Construction Using Multiple-Model Dynamics Filters and Spatiotemporal Optimization*, International Workshop on Microscopic Image Analysis with Applications in Biology, 2007.

[4] Z.Khan et al., *An MCMC-based Particle Filter for Tracking Multiple Interacting Targets*, ECCV, 2004.

[5] G. Gennari and G. D. Hager, *Probablistic Data Association Methods in Visual Tracking of Groups*, IEEE CVPR, 2004.

[6] Y.Cai et al., *Robust Visual Tracking of Multiple Targets*, ECCV, 2006.

[7] A. Yilmaz et al., *Object Tracking: A Survey*, ACM Journal of Computing Surveys, Vol. 38, No. 4, 2006.

[8] G. Brostow and R. Cipolla, *Unsupervised Bayesian Detection of Independent Motion in Crowds*, IEEE CVPR, 2006.

[9] T. Zhao and R. Nevatia, *Bayesian Human Segmentation in Crowded Situations*, IEEE CVPR, 2003.

[10] T. Zhao and R. Nevatia, *Tracking Multiple Humans in Crowded Environment*, IEEE CVPR, 2004.

[11] W. Lin et al., *Tracking Dynamic Near-regular Textures under Occlusion and Rapid Movements*, European Conference on Computer Vision (ECCV), 2006.

[12] W. Lin et al., *A Lattice-based MRF Model for Dynamic Near-regular Texture Tracking*, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Vol. 29, No. 5, 2007.

[13] M. Yang, J. Yuan, and Y. Wu, *Spatial Selection for Attentional Visual Tracking*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[14] G. Antonini, S. V. Martinez, M. Bierlaire, and J. P. Thiran, *Behavioral Priors for Detection and Tracking of Pedestrians in Video Sequences*, International Journal of Computer Vision (IJCV), Vol. 69, No. 2, 2006.

[15] M. Ben-Akiva and M. Bierlaire, *Discrete Choice Methods and Their Applications to Short-term Travel Decisions*, In Handbook of Transportation Science, pp.534, R.Hall(ed.), Kluwer, 1999.

[16] Blei, D.M. and Lafferty, J.D. *A Correlated Topic Model of Science*, Annals of Applied Statistics, Vol. 1, No.1, 2007.

[17] D. Blei et al., *Latent dirichlet allocation*, The Journal of Machine Learning Research, Vol. 3, 2003.

[18] Li, K. and Chen, M. and Kanade, T. and Miller, E.D. and Weiss, L.E. and Campbell, P.G. *Cell population tracking and lineage construction with spatiotemporal context*, Medical Image Analysis, 2008.

[19] W. Wang et al., Single cell behavior in metastatic pri- mary mammary tumors correlated with gene expression patterns revealed by molecular profiling. Cancer Research, 62(21), 2002.

[20] X. Wang, K. Ma, G. Ng, and W. Grimson. Trajectory analysis and semantic region modeling using a nonparametric Bayesian model. In CVPR, 2008.