Data-driven Crowd Analysis in Videos

Mikel Rodriguez, Josef Sivic, Ivan Laptev, Jean-Yves Audibert

Data-driven Crowd Analysis in Videos

Mikel Rodriguez[1,4] Josef Sivic[2,4] Ivan Laptev[2,4] Jean-Yves Audibert[3,4]

[1]École Normale Supérieure [2]INRIA [3]Imagine, LIGM, Universite Paris-Est

Abstract

In this work we present a new crowd analysis algorithm powered by behavior priors that are learned on a large database of crowd videos gathered from the Internet. The algorithm works by first learning a set of crowd behavior priors off-line. During testing, crowd patches are matched to the database and behavior priors are transferred. We adhere to the insight that despite the fact that the entire space of possible crowd behaviors is infinite, the space of distinguishable crowd motion patterns may not be all that large. For many individuals in a crowd, we are able to find analogous crowd patches in our database which contain similar patterns of behavior that can effectively act as priors to constrain the difficult task of tracking an individual in a crowd. Our algorithm is data-driven and, unlike some crowd characterization methods, does not require us to have seen the test video beforehand. It performs like state-of-the-art methods for tracking people having common crowd behaviors and outperforms the methods when the tracked individual behaves in an unusual way.

1. Introduction

In recent years, computer vision algorithms have played a growing role in surveillance systems. A common weakness of these systems is their inability to handle densely crowded scenes, such as those depicted in Figure 1. As the density of people in the scene increases, a significant degradation in the performance in terms of object detection, tracking, and event detection, is usually observed. This inability to deal with crowded scenes represents a significant problem given that many public areas are commonly densely populated, and according to a recent report [1], more than half of the world's people live in densely populated areas. One of the most difficult aspects of crowd analysis resides in being able to track a specific individual in a high density crowd. Factors such as the constant interaction among the agents in a crowd, inter-object occlusions, and the complex behavior-driven mechanics of the crowd which depend on specific agent roles (police, protester, marathon runner, etc.), render the tracking of an individual agent within the group a complicated task.

Figure 1. Examples of crowded scenes from our database.

In order to address the difficulty of tracking individuals in high density scenes, tracking algorithms have previously shown improved accuracy when learning a set of collective motion patterns from the data [2, 3]. These collective patterns are typically learned from a specific scene and are then used to constrain the likely locations and motions of individuals in the *same* scene.

While learning motion patterns appears to provide important priors for crowd analysis, learning such priors from the test videos limits applications of the previous methods to the off-line mode. Moreover, strong priors learned from long video sequences may not be useful to model rare events which do not comply with the typical motion of the crowd. The goal of this work is to address these limitations and to pre-learn crowd motion priors from a large dataset of crowd videos comprising both the common and rare crowd behaviors. Building on the recent success of data-driven methods in visual recognition [4–8] as well as image and video restoration/enhancement [9–12], we match test videos with a large database of crowd patterns and transfer corresponding and pre-learned motion priors to improve on-line crowd tracking in test videos.

There are several compelling reasons to search for simi-

[4]WILLOW project, Laboratoire d'Informatique de l'École Normale Supérieure, ENS/INRIA/CNRS UMR 8548.

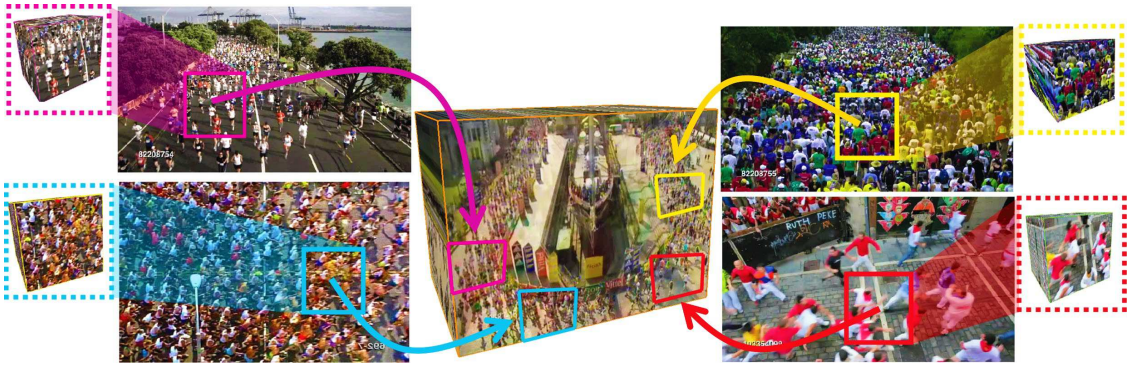

Figure 2. A crowded scene in the middle depicted as a combination of previously observed crowd patches. Each crowd patch contains a particular combination of crowd behavior patterns (people running in a particular direction in this example).

lar behaviors amongst crowd motion patterns. First, the motion of individuals in a high density crowd is often limited both by the physical constraints as well as social conventions of crowd dynamics. Both factors put strong limitations on the possible motions of individuals within a crowd. One can therefore consider the possibility of spanning the space of semantically distinguishable crowd patterns by sampling patches from a large database of crowd motion. An illustration of this idea is provided in Figure 2 where crowd motion patches of one video (middle) are matched to similar motion patches of different crowd scenes.

Another compelling reason for searching motion priors amongst a large collection of videos resides in the fact that some motion patterns, such as an individual walking against the flow of traffic, may be observed in only a small number of scenarios. Leveraging motion priors from a large database is therefore expected to provide a better model for analyzing both typical and rare crowd motions.

To operationalize and validate our idea of data-driven crowd analysis, in this paper we address the following challenges. We first consider efficient representation and matching of local crowd patches in video making use of both local motion cues as well as global scene signatures. We next address the transfer of motion priors from the database to the matched patches of the test video. We finally validate our data-driven crowd model on the challenging task of person tracking in high-density crowds. Our experiments show that motion priors learned by the off-line, long-term, unsupervised analysis of a large database of training videos can be used to improve tracking accuracy in previously unobserved testing crowd videos.

2. Related Work

Tracking is one of the most researched areas in computer vision, and a substantial body of work has been devoted to the problem. Most of the proposed algorithms have focused on the general problem of tracking, without specifically addressing the challenges of a crowded scene. Readers interested in a review of the state of the art in tracking are referred to a recent survey by Yilmaz *et al.* [13].

Crowd tracking has been addressed in a variety of contexts, including the study of dense clouds of bats [14] and biological cells in microscopy images [15], as well as medium to high density gatherings of people in monocular video sequences [2, 16–21] and multiple camera configurations [22].

In medium density crowded scenes, research has been done on tracking-by-detection methods [19, 20] in multi-object tracking. Such approaches involve the continuous application of a detection algorithm in individual frames and the association of detections across frames. Several tracking algorithms have centered on learning scene-specific motion patterns, which are then used to constrain the tracking problem. In [2] global motion patterns are learned and participants of the crowd are assumed to behave in a manner similar to the global crowd behavior. Overlapping motion patterns have been studied [3] as a means of coping with multi-modal crowd behaviors. These types of approaches operate in the off-line (or *batch*) mode (i.e. when the entire test sequence is available during training and testing) and are usually tied to a *specific scene*. Furthermore, they are not well suited for tracking rare events that do not conform to the global behavior patterns of the same video.

It is both practical and effective to learn motion priors for a given scene. However, the off-line assumption of previous methods is expected to limit their application in practice. In this work we seek to draw upon long term analysis of other videos for acquiring crowd behavior priors. In particular, we build on the recent progress in large database driven methods, which have demonstrated great promise in providing priors on object locations in complex scenes for object recognition [5–7] or scene completion [11], and recognizing actions of small-scale people [23], as well as predicting and transferring motion from a video to a single image [4,8]. Also related to our work are non-parametric data-driven image/video processing methods that have demonstrated excellent results in denoising [9] and inpainting [10, 12] applications. These works typically find locally similar patches

Figure 3. Representing motion patterns via low-level features. (a) 60 frames of a marathon video. (b) Averaged optical flow across a temporal window of 60 frames. (c) A frame from a crowded march scene. (d) KLT Tracks for 60 frames. (e) KLT tracks of a subregion of the video.

within the same image/video. In contrast, we develop a non-pararametric method driven by a large database of crowd behaviors for transferring different types of motion priors learned from different scenes with the aim to improve tracking.

3. Data-driven Crowd Analysis

We intend to use a large collection of crowd videos to learn crowd motion patterns by performing long-term analysis in an off-line manner. The learned motion patterns can be used in a range of application domains such as crowd event detection or anomalous behavior recognition. In this particular work, we choose to use the motion patterns learned on the database to drive a tracking algorithm. The idea is that any given crowd video can be thought of as being a mixture of previously observed videos. For example, a crowded marathon video, such as the one depicted in the middle of Figure 2, contains regions that are similar to other crowd videos. In it we observe a region of people running in a downwards direction, similar to the video depicted in the top left, as well as a region containing people running towards the right, as in the video depicted in the bottom left. These different videos can provide us with strong cues as to how people behave in a particular region of a crowd. By learning motion patterns from a large collection of crowded scenes, we should be able to better predict the motion of individuals in a crowd. A video narration depicting this scene and the rest of the proposed system can be found in the supplemental material and on the project web page [30].

Our data-driven tracking algorithm is composed of three components. We start by learning a set of motion patterns off-line from a large database of crowd videos. Subsequently, given an input video, we proceed to obtain a set of coarsely matching crowd videos retrieved from the large crowd database. Having obtained a subset of videos which roughly match the scale and orientation of our testing sequence, in the second phase of our algorithm, we use this subset of videos to match patches of the input crowded scene. Our goal is to explain input video by the collection of space-time patches of many other videos and to transfer learned patterns of crowd behavior from videos in the database. The final component of our algorithm pertains to how we incorporate local motion priors into a tracking framework. The individual components of the approach are described next.

4. Off-line Analysis of Crowd Video Database

A crowd motion pattern refers to a set of dominant displacements observed in a crowded scene over a given time-scale. These observed motion patterns can be represented either directly, using low-level motion features such as optical-flow, or they can be modeled at a higher level, by a statistical model of flow direction obtained from a long term analysis of a video. In this section we describe each of these representations.

Low-level Representation. Examples of low-level motion features include sparse or dense optical flows (Figure 3-b), spatio-temporal gradients, and feature trajectories obtained using Kanade-Lucas-Tomasi features (Figure 3-d). In this work, a low-level crowd pattern representation is a motion flow field which consists of a set of independent flow vectors representing the instantaneous motion present in a frame of a video. The motion flow field is obtained by first using an existing optical flow method [24] to compute the optical flow vectors in each frame, and then combining the optical flow vectors from a temporal window of frames of the video into a single global motion field (Figure 3-b).

Given an input video, we compute dense optical flow for each frame. The velocity vector of pixel O at time t will be denoted $\mathbf{V}_{O,t}$. The average optical flow $\bar{\mathbf{V}}_{O,t}$ at pixel O and time t is the average velocity across a temporal window of $w = 60$ frames and a (20×20)-cell containing O.

Mid-level Representation. An alternative representation of crowd motion patterns forgoes directly incorporating low-level motion features in favor of a hierarchical Bayesian model of the features. The main thrust behind the use of an unsupervised hierarchical model within this domain is that it allows for long-term analysis of a scene and in contrast to simple averaging can capture both overlapping behaviors at any given location in a scene and spatial dependencies between behaviors. For this purpose, we adopt the representation used in [3] which employs a correlated topic

model (CTM) [25] based on a logistic normal distribution, a distribution that is capable of modeling dependence between its components.

CTM allows for an unsupervised framework for modeling the dynamics of crowded and complex scenes by capturing spatial dependencies between different behaviors in the same scene. One example of this type of spatial dependency is depicted in Figure 6-b. In such a scene 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. 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.

A video is represented by a mixture of behaviors. Each behavior is in turn a distribution over quantized displacements. A video is divided along the temporal domain into non-overlapping short clips. Typically, each resulting clip is close to ten seconds in length. Each scene is sized to 720×480 pixels. Position is quantized by dividing a scene into a grid with 36×24 cells, which are 20×20 pixels in size. For each clip in our dataset, we compute the optical flow as our low-level features. The space of optical flows is partitioned: $\mathbb{R}^2 = \mathcal{V}_0 \sqcup \mathcal{V}_{\text{up}} \sqcup \mathcal{V}_{\text{down}} \sqcup \mathcal{V}_{\text{left}} \sqcup \mathcal{V}_{\text{right}}$, where $\mathcal{V}_0$ corresponds to (close to) zero velocity, $\mathcal{V}_{\text{up}}$ to upwards velocity, $\mathcal{V}_{\text{down}}$ to downwards velocity, and so on.

The input to this representation is a crowd video clip in the form of a collection of motion words, where a motion word consists of both a position and an element in $\{0, \text{up}, \text{down}, \text{left}, \text{right}\}$. A motion word is thus an element in the dictionary $\{1..36\} \times \{1..24\} \times \{0, \text{up}, \text{down}, \text{left}, \text{right}\}$ of size $V = 36 \times 24 \times 5$. The CTM assumes that the motion words of a clip arise from a mixture of K typical behaviors (Figure 4). A (latent) behavior is a probability on the dictionary, in other words, a V-dimensional vector in the $V - 1$ simplex. Thus, given a behavior, one can generate a motion word. The behavior proportions in a clip are modeled by a logistic normal distribution. We refer to [25] for more information about how the parameters of this model are inferred from the video database. At the end, CTM assigns for each image of the video, each position in this image, and each "direction" $\{0, \text{up}, \text{down}, \text{left}, \text{right}\}$, a probability that an individual at that position in this image moves into this direction.

5. Matching

Given a query test video, our goal here is to find similar crowded videos in the database with the purpose of using them as behavior priors. The approach consists of a two-stage matching procedure depicted in Figure 5, which we describe in the remainder of this section.

In order to incorporate previously observed distributions over local motion from other videos, we need to find crowd video regions in our database that not only match the small region of interest around the object that is being tracked, but also roughly match the orientation and scale of the particular testing scene. In other words, we want to avoid using the motion patterns observed in a ground-level video of a crowded marathon scene as priors for an overhead view of a political rally. To help avoid such situations, we first look for videos which are most likely to be visually similar to the video in which we would like to track a particular agent, and then we proceed to match local crowd spatio-temporal regions in these pre-selected videos.

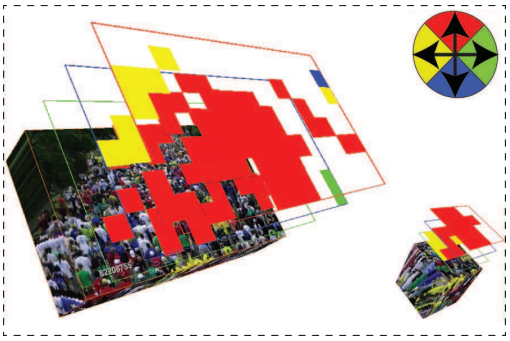

Figure 4. Representing motion patterns via a statistical model of flow direction. Learned clusters of crowd motion are depicted for an entire scene and a subregion of the crowd. Colors denote displacement directions.

Global Crowded Scene Matching. Scene matching is a very active research topic in computer vision, and a range of approaches have been proposed over the years. In this work, we use the Gist scene descriptor [26], given that it has been shown to perform well at grouping similar scenes. Although this scene descriptor does not explicitly encode viewpoint and scale information, our aim in this phase is to select a subset of videos from our dataset that share similar global attributes (Figure 5-b).

Given an input video in which we wish to track an individual, we first compute the Gist descriptor of the first frame. We then select the top 40 nearest neighbors from our database. We found that a Gist descriptor built from 6 oriented edge responses at 5 scales aggregated to a 4×4 spatial resolution to be effective at retrieving crowded scenes that shared similar global visual characteristics. By searching for similar crowded scenes first, instead of directly looking for local matching regions in a crowd video, we avoid searching amongst the several million crowd patches in our database and thus dramatically reduce the memory and computational requirements of our approach.

Local Crowd Patch Matching. Given a set of crowded scenes which roughly match a testing video, we proceed to retrieve local regions that exhibit similar spatio-temporal motion patterns from this subset of videos.

A number of different space-time feature descriptors have been proposed. Most feature descriptors capture lo-

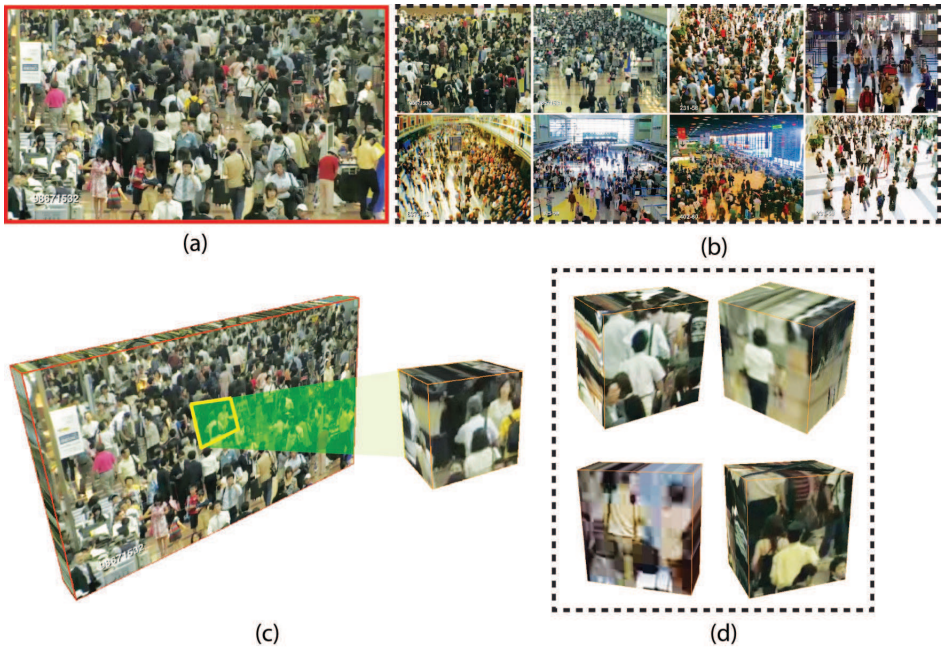

Figure 5. Global and local crowd matching. (a) Testing video. (b) Nearest neighbors retrieved from the database of crowd videos using global matching. (c) A query crowd patch from the testing video. (d) Matching crowd patches from the pool of global nearest neighbor matches.

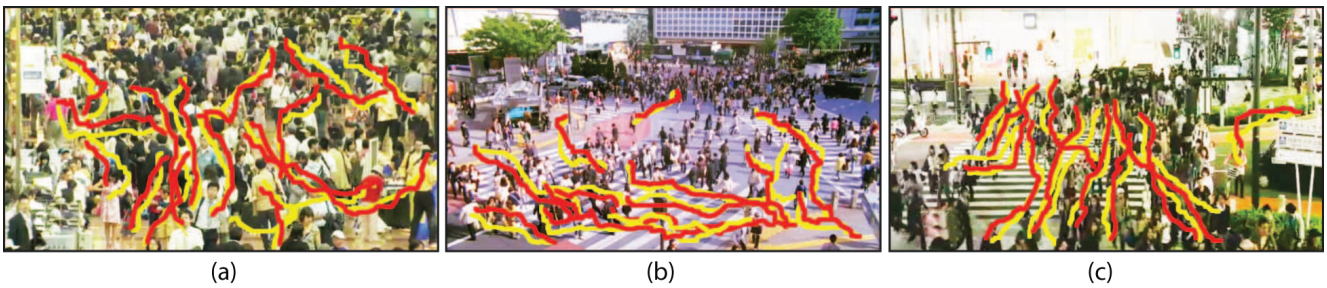

Figure 6. Testing videos along with manually annotated ground truth (yellow) and data-driven tracking results (red). (a) A busy airport scene. (b) The Shibuya crosswalk. (c) A two-way zebra crossing scene.

cal shape and motion in a neighborhood of interest using spatio-temporal image gradients and/or optical flow. In our experiments we employ the HOG3D descriptor [27], which has demonstrated excellent performance in action recognition [28]. In HOG3D, gradients are computed using an integral video representation. Regular polyhedrons are used to quantize the orientation of spatio-temporal gradients. The descriptor, therefore, combines shape and motion information. A given 3D patch is divided into $n_x \times n_y \times n_t$ cells. The corresponding descriptor concatenates gradient histograms of all cells and is then normalized. In our crowd tracking experiments, we use an icosahedron as polyhedron type for quantizing orientations. Descriptors are computed over a range of patch sizes ($80 \times 80 \times 60$, $120 \times 120 \times 60$, $320 \times 320 \times 60$) with a 50% overlap.

Given a region of interest in our testing video (i.e. current tracker position), we compute HOG3D of the corresponding spatio-temporal region of the video. We then proceed to obtain a set of similar crowd patches from the preselected pool of global matching crowd scenes by retrieving the k nearest neighbors, (Figure 5-d) from the crowd patch that belong to the global matching set.

6. Transferring Learned Crowd Behaviors

In this section, we describe how we incorporate the motion patterns obtained by long-term unsupervised analysis of the database videos (Section 4) as motion priors when tracking an individual in a previously unobserved video. We incorporate the off-line learning of motion patterns as a prior over a standard Kalman filter [29].

Tracking Framework. In all of our experiments that evaluate tracking we adopt a linear Kalman filter:

$$\mathbf{x}_k = \Phi_{k-1}\mathbf{x}_{k-1} + \mathbf{w}_{k-1}, \tag{1}$$

where Φ_{k-1} denotes the state transition matrix at time t_{k-1} and $\mathbf{w}_{k-1}$ is the additive system noise (Gaussian in

our experiments). Our state vector $\mathbf{x}$ is given by the 2-D location and displacement of the person being tracked, $\mathbf{x} = (x, y, u, v)$. For an individual located at O at time t, we obtain the next tracker position P as mean of the prediction P_1 of a point in the next time instant from the Kalman filter and the system measurement P_2, which in our experiments is derived from optical flow $\mathbf{V}_{O,t}$ at O.

When there is no behavior prior to be used in tracking, the linear motion model alone drives the tracker and equal weighting is given to the Kalman prediction and measurement. However, if we wish to incorporate information about the learned motion patterns as priors, the Kalman prediction and measurement can be re-weighted to reflect the likelihood of a behavior given the learned motion patterns. We now describe how this is accomplished.

Incorporating Motion Priors. The learned motion patterns described in Section 4 can be incorporated into the above mentioned tracking framework by considering a weighted mean of P_1 and P_2. The weighted mean incorporates the crowd motion pattern, which is a 5-dimensional vector $(\alpha_0, \alpha_{\text{up}}, \alpha_{\text{down}}, \alpha_{\text{left}}, \alpha_{\text{right}})$. The next tracker position P at time $t+1$ is obtained by:

$$\overrightarrow{OP} = \frac{w(P_1)\overrightarrow{OP_1} + w(P_2)\overrightarrow{OP_2}}{w(P_1) + w(P_2)}, \qquad (2)$$

where $w(X) = \sum_{i \in \{0,\text{up},\text{down},\text{left},\text{right}\}} \alpha_i \mathbf{1}_{\overrightarrow{OX} \in \mathcal{V}_i}$, where $\mathbf{1}$ stands for the indicator function. The crowd motion pattern is obtained by first uniformly averaging up to k CTM representations, and then use the 5-dimensional vector normalized to one corresponding to the pixel O. The proposed tracking algorithm thus combines (i) the linear Kalman filter on the test video, (ii) the two-step matching process described in Section 5 where a set of videos is selected using gist, and a local matching of these videos are obtained by using the HOG3D descriptor, and (iii) the CTM of the local parts of the selected videos through (2).

7. Experiments and Results

This section evaluates our approach on a challenging video dataset collected from the web and spanning a wide range of crowded scenes. In order to quantitatively assess the effectiveness of drawing upon other videos in order to acquire crowd behavior priors, we focus on two testing scenarios. The first testing scenario involves tracking individuals performing typical crowd behaviors, such as walking with the flow of traffic in a political rally, or crossing a busy crosswalk scene. A second testing scenario focuses on rare events, such as walking against the flow of traffic, sudden crowd dispersion, and mass panic scenes.

Crowd Video Database. To track individuals in a wide range of crowd scenes, we aim to sample the set of crowd videos as broadly as possible. To this end, we construct our crowd video collection by crawling and downloading videos from search engines and stock footage websites (such as gettyimages and youtube) using text queries such as "crosswalk," "political rally," "festival," and "marathon." We discard duplicate videos, as well as time-lapse videos and videos taken with tilt-shift lenses. Our database contains 520 unique videos varying from two to five minutes (10 hours 24 minutes in total) and resized to 720x480 resolution. This data is available on the project website [30].

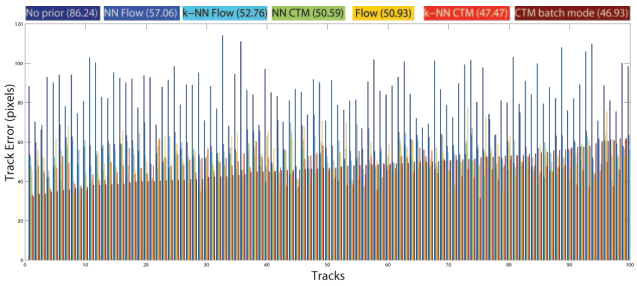

Figure 7. Comparison of average tracking errors for 100 tracks. Tracks are sorted in increasing order based on batch mode tracking baseline errors (best viewed in PDF).

Tracking Typical Crowd Behaviors. The first testing scenario was geared towards assessing the performance of the proposed data-driven model in the presence of large crowds of people which exhibit typical crowd behaviors. In these experiments we track several individuals in previously unobserved videos. The first scene we considered is depicted in Figure 5-a. The scene consists of a crowded airport terminal in which passengers move in complex patterns across the frame. The top nearest neighbors resulting from the global scene matching are depicted in Figure 5-b.

A set of trajectories generated by our tracking algorithm is shown in Figure 6. Quantitative analysis of the tracking was performed by generating ground-truth trajectories for 100 individuals, which were selected randomly from the set of all moving people. The ground-truth was generated by manually tracking the centroid of each selected person (Figure 6). Tracking error is measured (in pixels) by comparing the tracker position at each frame with respect to the position indicated by the ground truth (Figure 7).

We evaluated several baseline algorithms in addition to the proposed system. The first baseline is a linear Kalman tracker (described in Section 6), for which $w(P_1) = w(P_2) = 1/2$ in (2). This baseline tracking configuration achieves a mean tracking error of 86.24 pixels (Table 1). A second baseline followed the batch mode paradigm which is used in many existing crowd tracking algorithms: motion patterns are learned on the entire testing video itself and are used as motion priors. The first of these batch mode baseline algorithms couples the Kalman tracker with a motion prior which is drawn from the average optical flow across 60 frames of the testing video itself. Precisely, in (2), we take $w(P_1) = w(P_2) = 1/2$ but with P_2 obtained by us-

Table 1. Tracking error comparison for typical crowd behaviors

		mean ($\pm$ SD)
No prior		86.24 ($\pm$ 3.65)
Learned on test video	Optical flow	50.93 ($\pm$ 1.92)
	CTM [3]	46.93 ($\pm$ 1.22)
Transfer (Learned on database)	1^{st}-nn optical flow	57.06 ($\pm$ 2.05)
	k-nn optical flow	52.76 ($\pm$ 2.05)
	1^{st}-nn CTM	50.59 ($\pm$ 1.83)
	k-nn CTM	47.47 ($\pm$ 1.27)

ing the average optical flow $\bar{\mathbf{V}}_{O,t}$ instead of the optical flow $\mathbf{V}_{O,t}$ (described in Section 4). This configuration results in a mean tracking error of 50.93 pixels. A second batch mode baseline followed the experimental setup described in [3], in which motion priors are learned via a correlated topic model (CTM) from the same testing sequence: the weights $w(P_1)$ and $w(P_2)$ in (2) are then the probability respectively assigned to the directions $\overrightarrow{OP_1}$ and $\overrightarrow{OP_2}$ at pixel O. The resulting mean tracking error of this configuration was 46.93 pixels across the 100 tracks.

Finally, we evaluated the proposed data-driven approach. In these experiments, all of the motion priors are transferred from database videos. We do not rely on long-term observations in the test scene. Instead, we rely on off-line, long-term analysis of the database videos. The first of these setups uses the average optical flow of the nearest neighboring crowd patch from the pool of globally matching crowd scenes, resulting in a mean error of 57.06 pixels. Another setup uses the mean of the $k = 3$ average optical flows of the k NN video patches resulting in an average error of 52.76 pixels. The third configuration weights $\overrightarrow{OP_1}$ and $\overrightarrow{OP_2}$ in (2) according to the CTM crowd behavior model of the nearest neighboring crowd patch from the pool of globally matching crowd scenes. In this configuration we observe a mean tracking error of 50.59 pixels. The fourth data-driven configuration averages the CTM weights of the $k = 3$ nearest neighbors as described in Section 6, resulting in a mean error of 47.47 pixels (Table 1).

In this first round of experiments, we observe that learning motion priors from the testing video itself leads to low tracking errors, as would be expected in most typical crowd behavior scenarios. However, we also observed that drawing crowd behavior priors from other videos in our database led to high tracking accuracy that approached that of batch mode configurations without using test videos for training.

Tracking Rare and Abrupt Events. A second testing scenario focused on tracking rare and abrupt behaviors of individuals in a crowd. This class of behaviors refers to motions of an individual within a crowd that do not conform to the global behavior patterns of the same video, such as an individual walking against the flow of traffic. These events are not common in most videos. Therefore, there may only exist a few examples throughout the course of a video sequence. In these scenarios, the effect of the data-driven tracking approach is expected to be even more prominent. This is due to the fact that the test videos alone are not likely to contain sufficient repetitions of rare events in order to effectively learn motion priors for this class of events.

Figure 8. Example tracks of rare and abrupt behaviors. (a) A woman (red) walks perpendicular to the main flow of traffic, consisting of a crowd of people walking towards the left (blue). (b) A cameraman walks against the follow of traffic.

Figure 8 depicts examples of relatively rare crowd events. In order to assess the performance of the proposed data-driven model in tracking this class of event, we selected a set of 21 videos containing instances of relatively rare events. Quantitative analysis of tracking in these scenarios closely followed the experimental configuration described in Section 7. A first baseline tracking algorithm consisted of the linear Kalman tracker with no additional behavior prior. The second baseline learned motion priors on the testing video itself (batch mode) using the CTM motion representation. Finally, the proposed data-driven approach transferred motion priors from the top k matching database videos, for which motion patterns had been learned off-line using the CTM motion representation.

The tracking errors for this round of experiments are depicted in Figure 9. It can be seen that batch mode tracking is unable to effectively capture strong motion priors for temporally-short events that only occur once throughout a video (with a mean tracking error of 58.82 pixels), whereas data-driven tracking (with a mean tracking error of 46.88 pixels) is able to draw motion priors from crowd patches that both roughly match the appearance of the tracked agent, and exhibit a strongly defined motion pattern. This is evident in Figure 10, which depicts a successfully tracked individual moving perpendicular to the dominant flow of traffic in a political rally scene. The corresponding nearest neighbors (Figure 10-b) are crowd patches that, for the most part, contain upwards-moving behaviors from the crowd database. Besides, it can be noted that the retrieved crowd patches belong to behaviors which are commonly repeated throughout the course of a clip, such as crossing a busy intersection in the upwards direction. By matching a rare event in a testing video with a similar (yet more commonly observed) behavior in our database, we are able to incorporate these strong motion cues as a means of improving tracking performance.

The results above provide a compelling reason for searching a large collection of videos for motion priors when tracking events that do not follow the global crowd

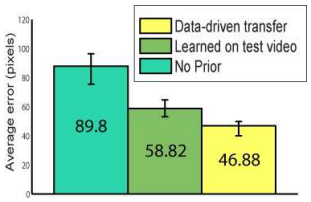

Figure 9. Comparison of average tracking errors when tracking people in rare crowd events based on 21 tracks and $k = 3$.

Figure 10. (a) Data-driven track of a person walking across a crowded demonstration (green), ground-truth (red), batch mode tracking (yellow). (b) Matching crowd patches from the database.

behavior pattern. Searching for similar motion patterns in our large database has proven to provide better motion priors which act as strong cues that improve accuracy when tracking rare events.

8. Conclusion

In this work we have approached crowd analysis from a new direction. Instead of learning a set of collective motion patterns which are geared towards constraining the likely motions of individuals from a specific testing scene, we have demonstrated that there are several advantages to searching for similar behaviors amongst crowd motion patterns in other videos. Our experiments have shown that by leveraging a large database of previously observed crowd behavior patterns we are able to accurately track individuals in a crowd exhibiting both typical and rare behaviors.

Acknowledgements. This work was partly supported by the Quaero, OSEO, MSR-INRIA, ANR DETECT (ANR-09-JCJC-0027-01) and the CROWDCHECKER project. We thank Pierre Bernas, Philippe Drabczuk, and Guillaume Ne from Evitech for the helpful discussions and the testing videos.

References

[1] M. Montgomery. The urban transformation of the developing world. *science*, 319(5864):761, 2008.

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

[3] M. Rodriguez, S. Ali, and T. Kanade. Tracking in unstructured crowded scenes. In *ICCV*, 2009.

[4] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. Sift flow. In *Proc. ECCV*, 2008.

[5] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing. In *CVPR*, 2009.

[6] B. Russell, A. Torralba, C. Liu, R. Fergus, and W. T. Freeman. Object recognition by scene alignment. In *NIPS*, 2007.

[7] B. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Segmenting scenes by matching image composites. In *NIPS*, 2009.

[8] J. Yuen and A. Torralba. A data-driven approach for event prediction. In *Proc. ECCV*, 2010.

[9] B. Baudes, A. Coll and J. M. Morel. A non-local algorithm for image denoising. In *CVPR*, 2005.

[10] A. Efros and T. Leung. Texture synthesis by non-parametric sampling. In *ICCV*, 1999.

[11] J. Hays and A. Efros. Scene completion using millions of photographs. *SIGGRAPH*, 2007.

[12] Y. Wexler, E. Schechtman, and M. Irani. Space-time video completion. In *CVPR*, 2004.

[13] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4), 2006.

[14] M. Betke, D. Hirsh, A. Bagchi, N. Hristov, N. Makris, and T. Kunz. Tracking Large Variable Numbers of Objects in Clutter. In *CVPR*, 2007.

[15] K. Li and T. Kanade. Cell population tracking and lineage construction using multiple-model dynamics filters and spatiotemporal optimization. In *IWMIAAB*, 2007.

[16] G. Gennari and G. Hager. Probabilistic data association methods in visual tracking of groups. In *CVPR*, 2007.

[17] W. Lin and Y. Liu. Tracking Dynamic Near-Regular Texture Under Occlusion and Rapid Movements. *ECCV*, 2006.

[18] G. Brostow and R. Cipolla. Unsupervised Bayesian detection of independent motion in crowds. In *CVPR*, 2006.

[19] B. Leibe, K. Schindler, and L. Van Gool. Coupled detection and trajectory estimation for multi-object tracking. In *ICCV*, 2007.

[20] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *ICCV*, 2010.

[21] T. Zhao, R. Nevatia, and B. Wu. Segmentation and tracking of multiple humans in crowded environments. *TPAMI*, 30(7):1198–1211, 2008.

[22] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *TPAMI*, 30(2):267–282, 2007.

[23] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, 2003.

[24] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 3, pages 674–679, 1981.

[25] D. Blei and J. Lafferty. A correlated topic model of science. *AAS*, 1(1):17–35, 2007.

[26] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[27] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *BMVC*, 2008.

[28] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[29] R. Kalman et al. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[30] http://www.di.ens.fr/willow/research/datadriven/.