Extracting Descriptive Motion Information from Crowd Scenes

Muhammad Saqib
School of Software,
University of Technology Sydney,
NSW 2007, Australia
muhammad.saqib@student.uts.edu.au

Sultan Daud Khan
Makkah Technology Valley,
Saudi Arabia
sdkhan@gistic.org

Nabin Sharma, Michael Blumenstein
School of Software,
University of Technology Sydney,
NSW 2007, Australia
nabin.sharma@uts.edu.au
michael.blumenstein@uts.edu.au

Abstract—An important contribution that automated analysis tools can generate for management of pedestrians and crowd safety is the detection of conflicting large pedestrian flows: this kind of movement pattern, in fact, may lead to dangerous situations and potential threats to pedestrian’s safety. For this reason, detecting dominant motion patterns and summarizing motion information from the scene are inevitable for crowd management. In this paper, we develop a framework that extracts motion information from the scene by generating point trajectories using particle advection approach. The trajectories obtained are then clustered by using unsupervised hierarchical clustering algorithm, where the similarity is measured by the Longest Common Sub-sequence (LCS) metric. The achieved motions patterns in the scene are summarized and represented by using color-coded arrows, where speeds of the different flows are encoded with colors, the width of an arrow represents the density (number of people belonging to a particular motion pattern) while the arrowhead represents the direction. This novel representation of crowded scene provides a clutter free visualization which helps the crowd managers in understanding the scene. Experimental results show that our method outperforms state-of-the-art methods.

Moreover, tracking is also not feasible in the high-density crowd. The tracker is lost easily and requires re-initialization at regular interval. The reason for suboptimal performance is that there is less number of pixels per person and inter-object occlusions, which make the detection and tracking unreliable solutions. Therefore in such scenarios, a holistic understanding of the scene is required. The study of crowd behavior includes the study of flows, dominant flows/common pathways, and identification of inflows/source and sinks/outflows. This study of behavior can be beneficial as a preprocessing step in the understanding of crowd behavior in extraordinary situations such as congestion.

It is very critical for a computer vision system to capture accurate motion information for understanding the behavior of crowd. Thus the performance of the system is dependent upon how the motion is represented. The desired motion representation should generate long and most importantly reliable trajectories. These trajectories can then be used to describe the flow for the whole video. Traditionally, optical flow is used for computing the pixel-wise flow between the two frames [4], [12]. However, there are some inherent problems in the direct usage of optical flow for motion representation. Firstly, optical flow produces ambiguous results on the boundaries of conflicting flows and perform poorly in the case of very slow moving objects. Secondly, optical flow does not represent the long-range spatio-temporal motion representation required in many applications. Typically obtaining complete trajectories is a difficult task from crowd motion in the high-density crowd. Therefore to get complete trajectories a notion of tracklet is used to capture short-term motion. A tracklet is a fragment of trajectory obtained by the tracker for the object. As previously mentioned, detection of the object and their tracking does not perform well in a high-density crowd. In most of the previous studies, tracklets are extracted from feature points and subsequently tracked for the very short duration to generate tracklets. After tracklets generation, various approaches such as Markov Random Field (MRF) [22] are used to ensure spatio-temporal dependencies between the tracklets [22], [23]. The tracklets give a robust high-level representation of the crowd and are less like likely to drift as compared to complete trajectories [11].

In this paper, we propose a framework for the study of
crowd behavior by analyzing the flows in the crowd and describing the flows quantitatively and qualitatively. Our proposed framework used particle flow computed by moving the grid of particles with the optical flow, initialized by the dynamical system \[17\]. Numerical integration of the dynamical system over the segment of the video is used to generate tracklets. The first step in the crowd analysis is to generate reliable long trajectories. Firstly, we make use of tracklets to generate short term reliable trajectories for a small segment of video and then use clustering both spatially and temporally to produce flows for the whole video. By using the resultant flows, we can identify source/inflow and outflow/sink of the scene. Once the reliable flows are achieved, we compute basic parameters like speed, density and direction of the dominant flows. These parameters can fully characterize pedestrian movement and measuring these will help the crowd managers in understanding over all motion in the scene. Once dominant flow are describe by these parameters, we can visually analyze speed-density relationship that serve basis for understanding the behavior of complex system. The remainder of the paper is organized as follows. In section II, discusses the most related work in the literature. In section III we discuss the proposed methodology. In section IV, we present and analyze the experimental results. Finally, the paper is concluded in section V.

II. LITERATURE REVIEW

There has been a body of work on the usage of tracklets for understanding the collective movement of pedestrians in the crowd. Zhou et al. \[22\] used Random Field Topic (RFT) to obtain semantic regions by non-parametric clustering of the tracklets. The RFT model ensures spatio-temporal coherence between the tracklets. The information about source and sink in the scene are given to the models as a highly semantic-level priors during the inference. However, our framework does not assume sources and sinks to be known a priori. In a variation of this approach, Zhou et al. formulated the whole crowd with the agent-based model called Mixture model of Dynamic pedestrian Agents (MDA) \[23\]. In this model, each pedestrian is modeled as an agent with the three factors. The factors included the belief of starting and the destination position of the pedestrian, movement dynamics, and the timing of pedestrian entering the scene. The collective behavior is analyzed using agent after learning from real data. However, this agent-based approach is only suitable for the low-density crowd, where agents can easily model each pedestrian movement. Blei and Frazier \[3\] used Distance Dependent Chinese Restaurant Process (DD-CRP) to cluster the coherent tracklets in the first stage. In the second stage, the clusters are aggregated temporally to obtain pathways, or dominant motion flows in the scene. Next, a geometric analysis is applied on the pathway shape to determine the sources and sinks in the scene. Hassanein et al. \[7\] considered tracklet as a directed line segment and used line geometry to find similarity between the tracklets. Upon clustering, semantic regions are extracted which include source, sink, and common pathways. Wang et al. analysed motion patterns by clustering the hybrid generative-discriminative feature maps \[20\]. The extracted feature map associate motion pattern with low-level feature. Finally, hierarchical clustering is used to cluster feature maps into semantic regions. However, all of the above mentioned approaches only describe the crowd flows qualitatively without further details about the motion in the scene. Meta-tracking is another technique employed to discover the scene structure in the crowd \[9\]. The Orientation Distribution Functions (ODFs) model is used to find out the direction of the flow at each pixel. A particle meta-tracking is used to obtain particle trajectories called meta-tracks. Finally, meta-tracks are clustered to get dominant motion flows in the crowd. This kind of motion representation is complex for extracting dense trajectories.

The raw data about path taken by pedestrian is considered noisy. For instance, Nedrich et al. used weak tracking system to identify source and sink through the mean shift clustering algorithm \[15\]. Manifold is a denoising algorithm which is also used for clustering of the data. Xu et al. proposed a shrinkage-based framework for unsupervised trajectory clustering using adaptive kernel-based estimation \[21\]. The framework can be viewed as the combination of mean-shift clustering and manifold base model. The center of the cluster represent manifold, and the structure of cluster shows motion patterns exhibited by the pedestrians in the crowd. Trajectories in the cluster are considered as noisy traversal of the manifold. Thus suppressing the noise and reconstructing manifold is seen as clustering of the trajectories.

Some approaches have been proposed to extract feature point for extracting trajectories called feature trajectories. The features are extracted using SIFT, HOG, and Histogram of Oriented Flow (HOF) or combination of various feature descriptors \[19\]. SIFT features are used to find correspondence between two frames for trajectories. Similarly, Sun et al. used SIFT descriptor along with KLT tracker to track feature of the sampled points from dense SIFT descriptor for dense trajectories \[18\]. Nawaz et al. proposed an end-to-end approach for trajectory clustering in aerial videos \[14\]. Firstly, camera motion is compensated before clustering of trajectories obtained from the co-efficient of Discrete Wavelet Transform (DWT). In another approach motion is analyzed and performed by the clustering of the trajectories in the frequency domain \[8\]. However, these approaches are not able to capture whole motion information because only few feature points are detected which are not sufficient for the extraction of dense trajectories. The most related work to our proposed work used KLT tracker to extract motion trajectories and used hierarchical clustering to obtain dominant flows \[5\], \[6\], \[13\]. However, the approaches are not able to capture the long-term motion required for understanding the behavior of the crowd. Furthermore, these approaches do not identify source and sink locations of the dominant flows. Similar approach is followed by Khan et al. \[10\] for extracting dominant motion flows and identification of sources and sinks. However, the motion flows do not describe scene in terms of speed and density. Together, these studies provide useful insight.
into the study of crowd flow regarding motion representation, similarity measures and crowd behavior qualitatively such as the identification of sources, sinks and dominant motion pattern or common pathways. However, previous studies did not quantify the speed, density, and direction of the flows in the crowd. Our proposed framework extract dense motion trajectories using tracklets as a basic motion unit. Furthermore, the flows are described and visualized by the speed and direction of dominant motion patterns.

III. PROPOSED METHODOLOGY

In this paper, we propose a framework for describing flows exhibited by pedestrians in the crowd. In the first step, a stable, dense trajectories are generated in the crowded scene. Secondly, the generated trajectories are clustered together spatially and temporally to get Long Dense Trajectories (LDT). These flows help at a later stage in the identification of sources and sinks for the behavior analysis. According to the proposed framework, input video is divided into multiple equal length segments as shown in the Figure 1. To generate reliable tracklets, the first frame of each segment is overlaid with the grid of particles. After initialization of the dynamical system, the particles are moved through the segment by optical flow [17]. The integration of the dynamical system gives reliable particle trajectories for particular video segment. The trajectories are then clustered into local tracks using the Longest Common Sub-sequence (LCS) metric. Global tracks are combined into the global tracks using the same hierarchical clustering algorithm. We call these global tracks as the identification of sources, sinks and dominant motion flows. The source and sink of the dominant flows are identified which help in describing the speed, density, and mean direction of the flows.

A. Motion flow-field computation

Consider a feature point \( i \) in the frame of the video segment at time \( t \) is given by the following equation 1

\[
Z_{i(t)} = (X_{i(t)}, V_{i(t)})
\]  

Where \( Z_{i(t)} \), \( X_{i(t)} \), \( V_{i(t)} \) are the flow vector, location and velocity of the \( i^{th} \) feature point at time \( t \) respectively. Furthermore, \( V_{i(t)} = (v_{x_{i(t)}}, v_{y_{i(t)}}) \) incorporate both change in x and y directions. Similarly, \( X_{i(t)} = (x_{i(t)}, y_{i(t)}) \) is the initial location of the particle in the grid and \( X_{i(t+1)} = (x_{i(t+1)}, y_{i(t+1)}) \) is next location at time \( t + 1 \) approximated for all particles in the grid by solving the dynamical system given by the following equation 3.

\[
X_{i(t+1)} = F(X_{i(t)}) + X_{i(t)}
\]  

B. Particle advection

In this step, a grid of particle is overlaid on the first frame and advect it on the optical flow field for each segment of the video. The grid resolution is kept at a lower resolution to make the process less computationally expensive. Let \( X_{i(t)} = (x_{i(t)}, y_{i(t)}) \) is the initial location of the particle in the grid. The next location \( X_{i(t+1)} = (x_{i(t+1)}, y_{i(t+1)}) \) can be approximated for all the particle in the grid by solving the following equation 3.

\[
X_{i(t+1)} = F(X_{i(t)}) + X_{i(t)}
\]  

Let \( \Omega_i \) is the trajectory of the particle, which is essentially the movement of a particle from one frame to another frame according to the flow map. The pair of the flow maps are computed for each particle in the video segment as in [17]. In the unstructured crowd, the particle may drift from the flow of pedestrian and background noise can also induce its effect in the particle flow. Therefore a threshold term \( B_i \) is included in equation 4 to remove erroneous particles out of the computation. As a result of particle advection short trajectories called tracklets are obtained.

\[
X_{i(t+1)} = F(X_{i(t)}) + X_{i(t)} * B_i
\]  

C. Clustering of tracklets

Tracklets only provide information about the local movements. Because the tracklets are short-term, therefore to obtain dense trajectories the tracklets are clustered together to generate dominant motion flows. The source and sink of the dominant flows are identified which help in describing the speed, density, and mean direction of the flows.

The first step in the clustering in getting the dense tracks is by extending shorter tracklets through the segments of the video. The assumption here is that these shorter tracklets are located spatially close together and can have plausible continuation if they can be connected to each other in a head-to-tail manner for a long trajectory construction. For the construction of long trajectory, the first tracklet called a query tracklet (that need to be extended) is compared with the potential candidates tracklets from the neighborhood of query tracklet’s position on the grid. The best among the candidate is selected for the continuation of query tracklet. The selected candidate tracklet is now considered to be a query tracklet, and the same process is repeated until there are no more tracklets to be extended. Once the set of these tracklets are achieved, which are then connected by mean of \( K^{th} \) order least square polynomial regression. As shown in the Figure 1 after long trajectory construction, tracks are fed to the hierarchical clustering algorithm which clusters them into local tracks using Longest Common Sub-sequence as a similarity metric (LCS). The LCS is an efficient technique for matching tracks of unequal length to determine the longest common portion between the tracks. Finally, the local tracks are combined into the global tracks using the same hierarchical clustering algorithm. We call these global tracks as Longest Dense Trajectories (LDT).
Once the LDTs or global tracks (dominant flows) are achieved, the next step is to compute basic parameters that describe the flows. The first parameter that describes the flow is average speed of the flow. Speed together with the density can serve as a base for understanding the behavior of complex systems. Let $G_1$ and $G_2$ are the two dominant flows obtained after clustering. Let $G_i = \{t_1, t_2, ..., t_n\}$, where $t_j$ are the tracklets that belong to the dominant flow $G_i$. In order to find the average speed of $G_i$, we first compute the speed of each tracklet $t_j$ belongs to $G_i$. The instantaneous speed $v_k$ of each point $p_i \in t_j$ can be computed as $v_k = \frac{\Delta d}{\Delta t}$, where $\Delta d$ is the displacement among two consecutive points $p_i$ and $p_j$ that belong to $t_j$ and $\Delta t$ is the frame rate of video. The overall speed $v$ of tracklet $t_j$ can be computed by $v = \frac{1}{n} \sum_{k=1}^{n} v_k$ where $n$ is the total number of points in $t_j$. In this way we compute the speed of every tracklet that belongs to $G_i$ and finally average speed of $V_i = \frac{1}{N} \sum_{m=1}^{N} v_m$, where $N$ is the number of tracklets belong to $G_i$.

In the next step, we compute the mean direction of the flow in the same way as average speed is computed. Here, we compute instantaneous flow $f_k$ of each point $p_i \in t_j$ since $f_k$ is a vector showing displacement $u$ in x-dimension and displacement $v$ in y-dimension, therefore we compute the angle $\theta_k = tan^{-1} \frac{u}{v}$. The mean direction of $t_j$ is $\frac{1}{n} \sum_{k=1}^{n} \theta_k$ and finally average direction of $G_i$ becomes $\frac{1}{N} \sum_{m=1}^{N} \theta_m$. We compute the density $D_i$ of $G_i$ by $\frac{N}{M}$, where $N$ is total number of tracklets belong to $G_i$ and $M$ is the total number of tracklets belonging to all dominant flows. The summarized representation of dominant flows is shown in Fig 2(a). It is obvious in the figure that multiple dominant flows with their respective descriptions are detected by our algorithm. From the figure, we can easily visualize the inverse relationship of speed-density. The red-thick and long arrow show the main dominant flow which implies that density on this path is high relative to other paths in the scene.

IV. EXPERIMENTAL RESULTS

We have conducted several experiments to evaluate the performance of the proposed framework on publicly available datasets. All the algorithms are implemented in MATLAB and the experiments are carried out on PC with the configuration of 2.7GHz (core i7) processor and 8 GB of RAM. The datasets include both indoor/outdoor scenes comprising of structured/unstructured [16] crowd of varying density level. The complete details of the datasets are shown in Table I. The input to our proposed framework is the segments of the video. We have divided every video into the segments of 50 frames. The ground truth for the all the videos is established manually by the human observer so that obtained results can be compared to the ground truth. As we are employing particle advection for extracting dense, long trajectories. The resolution of the grid overlaid on the segment of the videos is kept at 1/10th of the actual resolution determined experimentally. The increase in resolution increases the computation, and there is no significant improvement seen in the extraction of trajectories.

The effectiveness of our proposed framework is compared with the relevant current approaches in the extraction of
TABLE I

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Resolution</th>
<th>Location</th>
<th>Crowd Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airport [1]</td>
<td>480X360</td>
<td>Indoor</td>
<td>Medium-density</td>
</tr>
<tr>
<td>Gallery [10]</td>
<td>1812X1461</td>
<td>Indoor</td>
<td>High-density</td>
</tr>
<tr>
<td>Unimib [2]</td>
<td>480X360</td>
<td>Indoor</td>
<td>Low-density</td>
</tr>
<tr>
<td>Escalator [6]</td>
<td>480X480</td>
<td>Indoor</td>
<td>Low-density</td>
</tr>
<tr>
<td>Hajj [10]</td>
<td>2946X1656</td>
<td>Outdoor</td>
<td>High-density</td>
</tr>
<tr>
<td>Station [6]</td>
<td>957X719</td>
<td>Indoor</td>
<td>Low-density</td>
</tr>
</tbody>
</table>

dominant flows/global tracks and also in the identification of sources and sinks of the dominant flows. The extraction of accurate flows is vital to the understanding of crowd scene. Our framework uses the extracted information to describe the speed, density, and mean direction of the dominant flows. Firstly, we compare the global tracks/dominant flows generated by our framework with ground-truth dominant flows. Moreover, a similar comparison is drawn between the current approaches and ground-truth flows using a flow similarity metric. The flow similarity metric is used to measure the similarity between dominant flows generated and the ground-truth flows. The flow similarity (Sim) is given by the equation

$$Sim = \frac{\sum_{i=1}^{N_{gf}} \arg \max_{j \in [1,M_{df}]} \frac{LCS(G_j,G_{th,i})}{\text{Length}(G_{th,i})}}{N_{gf}}$$ (5)

Whereas $N_{gf}$ and $M_{df}$ represent the total number of ground truth flows $G_{th}$ and detected number of dominant flows $G$ respectively. The Fig 4 shows the performance of our method encoded with the colors. The darker the color which corresponds to lower value on color bar, the more similar is the extracted global tracks with the ground truth flows. Similarly, a lighter color indicates lower resemblance of the detected global tracks with the detected one.

Secondly, the second metric identifies how closely the locations of sources/sinks of extracted trajectories lie together to the ground truth sources/sinks. For this comparison, merely using distance metric such as euclidean will not work as it is hard to relate changes in the scene to the changes in the real-world.

Therefore, we have used an association matrix to relate the locations of sources and sinks by their joint probability distribution. We have constructed an association matrix for sources and sinks identified by the ground-truth, our method, and other current methods. The measure of Kullback-Leibler (KL) divergence also referred to as relative entropy is adapted to compute the difference between the association matrices of compared methods with the ground truth matrices. The Kullback-Leibler formula is given below in equation 6

$$D_{KL}(AM_{Gth}||AM_{Gtrack}) = \sum_i AM_{Gth}(i) \ln \frac{AM_{Gth}(i)}{AM_{Gtrack}(i)}$$ (6)

Whereas $AM_{Gth}$ and $AM_{Gtrack}$ are the association matrices of ground truth flows and dominant flow produced by the compared methods respectively. The comparative analysis is shown in the Fig 4. The figure explores the difference between the relative entropies of our method with other methods encoded with colors. The smaller value of the entropies for our method indicates that the sources and sinks locations are closed to the locations established by the ground-truth entropies as compared to the other state-of-the-art methods.

We have computed the average speed and density for three video sequences after successfully extracting the dominant flows with sources and sinks. The arrow overlaid on the sample frames represent the dominant flows as shown in the Fig 2(a) and 3(a) 3(b). The overall speed calculated for the dominant flows is by finding the average speed of the tracklets comprising it. The associated color bar with the frames indicates the speed. The density of the crowd can be visualized by the width of the arrows representing dominant
flows, and the arrowhead gives the mean direction of the flow. Our proposed framework can extract useful descriptive information which is essential for understanding of the crowd behavior.

V. CONCLUSION

In this paper, we have proposed a framework for extracting descriptive motion information from the crowd scenes. The framework makes sure that reliable, dense and trajectories are extracted. The descriptive information includes speed, density, and mean direction of dominant flows.

REFERENCES


