can be used to compare group activities across different crowd systems. To our knowledge, this study is the first attempt in computer vision that investigates comprehensively and systematically the universal properties of groups in crowds. We make the following contributions.

- **A robust group detector** - We introduce a novel Collective Transition (CT) prior to capture the underlying dynamics of a group. Based on the prior we formulate a robust group detector that outperforms state-of-the-art methods [12]–[14].

- **Scene-independent group descriptors** - Based on the CT prior, we devise a set of visual descriptors to quantify four fundamental intra- and inter-group properties, namely collectiveness, stability, uniformity, and conflict. These descriptors convey richer group-level information in comparison to the conventional group size and velocity information [15]. Importantly, these descriptors are scene invariant and robust to public scenes with variety of crowdedness.

- **Group-driven crowd scene understanding** - We show that the proposed descriptors are effective in identifying the intrinsic group states (gases, fluids, and solid) following the common analogy employed in crowd modeling literature [16]–[18]. These states are demonstrated to be useful in crowd monitor and control by measuring the transition of group states over time. We also show the superiority of the proposed descriptors in crowd video classification over existing activity descriptors [19]. In addition, the proposed descriptors can be used to retrieve crowd videos across different scenes.

Experiments are conducted on hundreds of video clips collected from over 200 crowded scenes. The dataset and the ground truth are made publicly available to facilitate future research in group-level crowd analysis.

In comparison to our earlier version of this work [20], this paper is substantially improved by providing more technical details and more experimental evaluations and analysis. Specifically: (1) Section III-B provides more details on learning the CT prior. (2) Section VI-B provides ablation study of group detection; (3) Section VII provides more analyses regarding the correlation and scene invariances of the proposed group descriptors; (4) Section VIII explores more potential applications, including crowd dynamics monitoring (i.e. Section VIII-A) and crowd video retrieval (i.e. Section VIII-C).

II. RELATED WORK

Most existing imagery-based crowd analysis methods tend to treat a crowd either as a collection of individuals [21]–[23] or as an aggregated whole [8]–[10], [14]. In contrast to these studies, we analyze crowd at the group-level. The object-centered approaches require explicit detection and segmentation of individuals from crowd. These techniques are infeasible in crowded scenes where inter-object occlusion is severe. The activity representation employed by holistic methods, e.g. optical flow codewords [10], [24], dynamic texture [14], and grid of particles [8], [9], are useful for learning scene-level spatio-temporal pattern, but not directly applicable for learning group-level properties, which requires finer group segregation.

A. Group Detection

Moussaid et al. [7] found that group members tend to walk side-by-side at low crowd density and the formation is bent to a V-shape pattern as the density increases. Some works grouped pedestrians by analyzing their relative distances and moving patterns. Haritaoulou et al. [25] regarded group detection as a graph partition problem, whilst Ge et al. [12] discovered small groups by bottom-up hierarchical clustering (HC) of trajectories based on pairwise objects’ velocity and distance. Another trajectory-based approach was proposed by Zhou et al. [13] and it used a so-called Coherent Filtering (CF) algorithm to segment coherent motion in crowd. Li et al. [26] used trajectory information of multiple objects to learn models for segmenting two group patterns: offensive and defensive patterns.

As shown in our experiments, the above methods are either too sensitive to tracking errors or unscalable to extremely crowded scenes. Importantly, neither of them learn group properties further nor analyze crowd behaviors at the group-level.

B. Group Properties

Unfolding group properties and their descriptors have largely been ignored in computer vision, although they are well-known in other disciplines. In biology, Zhang et al. [2] studied collective motion in bacterial colonies to prevent disease spreading while McPhear [?], [27], [28] treated crowds and collective behavior as synonyms in sociology. Makris et al. [29] conducted quantitative study on the collective spatial and temporal processes formed by ocean shoals. In the field of sociology and psychology, sociologists measured inter-group relationship based on social conflict, discrimination, and prejudice [30], while psychologists explored how intra-group interactions influence the completion of group tasks [31]. Russell et al. [32] stated that a social group should poses “membership stability”. Carley [33] attempted to explain why some groups endure longer and more stable than others. Dahrendorf [34] stated that the social conflict was one of the central themes in social research. Wheelan [1] pointed out that conflict may be caused by competition for resources, goal incompatibility, and contentious influence tactics. Yi [35], [36] proposed multiple attributes to facilitate stationary group analysis. The “stable” attribute, for example, is used for stationary group member invariance measurement, while ours measures the dynamics of the group movements.

C. Crowd Behavior Analysis

There has been a significant amount of work that exploits tracklets or low-level features like optical flow for crowd behavior analysis. From the microscopic level, some approaches [37]–[39] extracted moving pixels as visual features and jointly modeled single-agent behaviors with hierarchical Bayesian models. The well-known agent-based approach
proposed by Helbing [40] has been used in crowd behavior analysis [9], [23], [41] too. For example, Braun et al. [41] extended the social force model by characterizing agents with different personalities and attributes, such as personal ID, dependence level, and desired speed. Beyond particle physics, Kaminka et al. [42] adopted Social Comparison Theory (SCT), a popular social psychology theory, to model crowd behaviors. The theory suggests that pedestrians would evaluate their current states through comparing themselves to others. Using this theory the study [42] generated improved pedestrian movements and accounts for group formation in pedestrians that are inter-related.

Differing from the microscopic-level crowd analysis, modeling crowd behaviors from the macroscopic level treats the crowd as a whole. Zhou et al. [43] proposed a Mixture model of Dynamic pedestrian-Agents (MDA) to learning crowd behaviors. Using spatio-temporal features to detect crowd behaviors [10], [19], [44] is another popular macroscopic approach. For instance, Kratz et al. [19] represented the motion pattern in a spatio-temporal volume with a Gaussian distribution. Instead of directly computing spatio-temporal gradients, Zaharescu et al. [44] computed spatio-temporal oriented energy with second derivative of 3D Gaussian filters. Behaviors were modeled by the histogram of relative energy at different orientations and scales. It is worth pointing out that many crowd modeling approaches are scene-specific [10], [37], [45], i.e. activity models learned from a specific-scene cannot be applied to other scenes.

Both microscopic- and macroscopic-level analyses have their own weaknesses. The drawback of microscopic analysis is that it is hard to detect and track individuals precisely in different crowded scenes. Macroscopic methods, on the other hand, have difficulties in capturing detailed group motion patterns. Consequently, a number of approaches [46] have been proposed for recognizing group activities such as meeting and fighting. These studies tend to analyze small social groups, their own weaknesses. The drawback of microscopic analysis of crowded scenes where a group may have an arbitrary large number of members are summarized in Algorithm 1.

III. GROUP DETECTION

We consider a group as a set of members with a common goal and collective behaviors. Given a short video clip of \( \tau \) frames, a set of groups \( \{G_i\}_{i=1}^m \) are detected. Each group \( G_i \) encompasses a set of tracklets \( \{z_{ik}\}_{k=1}^n \) detected by a tracker.\(^2\) We adopt the Kanade-Lucas-Tomasi (KLT) [53] feature point tracker due to its robustness and computational efficiency. From each detected group, we wish to extract a set of visual descriptors to represent its properties.

A. Collective Transition Prior

Precise group detection in crowd is challenging due to complex interaction among pedestrians. We assume that pedestrian movements in a scene are intimately governed by a finite number of Collective Transition (CT) priors. These priors are discovered simultaneously with the group detection process. We show that group detection can be made more robust by considering the temporal smoothness and consistency enforced by the priors. Furthermore, we demonstrate in Sec. IV that certain group properties can be readily derived from the discovered CT priors.

Each pedestrian group has a specific CT prior, which can be discovered from a video clip. More precisely, for \( n \) tracklets, \( \{z_{ik}\}_{k=1}^n \), we assume there exist \( m \) Markov chains, where \( m < n \) and \( m \) is inferred automatically. Each Markov chain is a time-series model with the form

\[
z_k^t = Az_k^{t-1} + v_k^t, \tag{1}
\]

where the continuous observation \( z_k^t \) evolves by a transition matrix \( A \in \mathbb{R}^{3 \times 3} \). Gaussian noise \( v_k^t \sim \mathcal{N}(0, Q) \) is assumed between transition. Let \( z_k^t = [x^t, y^t, 1]^T \) represent the position of a pedestrian in homogeneous coordinates\(^3\) and the initial observation \( z_k^1 \) follows a Gaussian distribution \( \mathcal{N}([\mu, \Sigma]) \). We denote \( \Theta = \{A, Q, \mu, \Sigma\} \) as the parameters of the chain. \( A \) represents the CT prior, which reveals the collective motions of all the members in a group, while \( \{\mu, \Sigma\} \) ensure that group members are spatially proximate at the initial frame. Next, we discuss how to learn this prior and perform robust group detection simultaneously.

B. Group Detection by Collective Transition

The key idea is to search for pedestrian groupings that fit well to the discovered priors within the video clip. The method is robust as it permits fragmented tracklets that fail to sustain over the whole clip. In particular, the missing data of \( z_k \) can be inferred with an Expectation-Maximization (EM) algorithm. It is thus suitable for group detection in dense crowds. In addition, it relies on local spatio-temporal relationships and velocity correlations without assumption on the global shape of the pedestrian group. Therefore it can be applied to scenes with different scales and perspectives.

The key steps of learning the CT priors for group discovery are summarized in Algorithm 1.

Step-1: Generate coherent filtering clusters: We first discover a set of initial tracklet clusters \( \{C_j\}_{j=1}^r \) using Coherent

\(^2\)It should be \( \{z_{ik}\}_{k=1}^n \), but we shorten it by \( \{z_{ik}\}_{k=1}^n \) for brevity.

\(^3\)A represents projective transforms, which include translation, contraction, expansion, dilation, rotation, shear, and their combinations.
Filtering [13], which is a clustering technique for the detection of coherent motion from time-series data and based on a so-called Coherent Neighbor Invariance that characterizes the local spatio-temporal relationships of individuals. These clusters do not align with our group perception perfectly but can serve as the basis for finding the final tracklet groups \( \{G_i\}_{i=1}^m \). Examples are shown in Fig. 2a.

**Step-2: Identify anchor tracklets**: The iterative scheme begins by randomly picking a cluster \( C_i \), and finding its anchor tracklet \( z_i^* \) with long duration and low variance (Fig. 2b).

- **Long duration**: A well-tracked pedestrian owns at least one long-duration trajectory. Thus he has more opportunities to meet with people on his route, and generate a group around him.
- **Low variance**: In most cases, unusual velocity rarely occurs within a group, but over an individual, e.g. running and sudden stop. Therefore, the individual with velocity of high variance cannot be a lasting member in one group.

**Step-3: Discover seeding tracklets**: As shown in Fig. 2c, a set of seeding tracklets, \( S_i \), are selected with the following criteria: (1) they are also from \( C_i \); and (2) have high velocity correlation with \( z_i^* \),

\[
\frac{\langle v_{z_i^*}, v_z \rangle}{\|v_{z_i^*}\| \cdot \|v_z\|} > \eta,
\]

where \( \eta \) is a threshold.

**Step-4: Learning CT prior**: The seeding tracklets discovered in Step-3, together with the anchor tracklet, form a set \( S_i \), which is used to learn a representative CT prior with EM. The CT prior is used to refine the group itself in Step-5.

We detail the EM optimisation as follows. Given the Markov chain defined in Equation (1), we have the transition probability as

\[
p(z_k^i | z_k^{i-1}) = \mathcal{N}(z_k^i | Az_k^{i-1}, Q),
\]

Assuming that the noise introduced by the observation is Gaussian distributed, we have the observation probability represented as \( p(x_k^i | z_k^i) = \mathcal{N}(x_k^i | z_k^i, R) \), where \( R \) is a given diagonal matrix. Let the initial observation follows a Gaussian distribution \( z_k^i \sim \mathcal{N}(z_k^i | \mu, \Sigma) \), we can formulate the joint likelihood with respect to the parameters \( \Theta = (A, Q, \mu, \Sigma) \) for all the observations \( X^T_k = \{x_1^k, \ldots, x_T^k\} \) and the hidden tracklet \( Z^T_k = \{z_1^k, \ldots, z_T^k\} \) as

\[
p(X^T_k, Z^T_k; \Theta) = \prod_{t=2}^{T} p(z_t^k | z_t^{k-1}) \prod_{t=1}^{T} p(x_t^i | z_t^i).
\]

The parameters \( \Theta \) can be effectively estimated by EM algorithm aiming at maximizing the marginal likelihood

\[ L(\Theta; X^T_k) = \sum_{z_k^i} p(X^T_k; \Theta) \]

**E-Step**: \( Q(\Theta | \Theta^{(n)}) = \mathbb{E}_{Z^T_k | X^T_k, \Theta^{(n)}} [\log L(\Theta; X^T_k, Z^T_k)] \)

**M-Step**: \( \Theta^{(n+1)} = \arg \max_\Theta Q(\Theta | \Theta^{(n)}) \).

The optimisation typically converges in a few iterations.

**Step-5: Group refinement**: We fit each tracklet \( z \) in the initial cluster \( C_i \) with \( A_i \) of the \( i \)-th Markov chain. The fitting error \( \epsilon \) of a tracklet is defined as

\[
\epsilon = \frac{1}{T-1} \sum_{t=1}^{T-1} \|Ax^t - z^{t+1}\|^2.
\]

Any tracklet with \( \epsilon < \delta \) is retained to construct \( G_i \). Unqualified tracklets will need to repeat the iterative process to be considered for a different group. A refined group is shown in Fig. 2d.

**IV. GROUP DESCRIPTORS**

We formulate a set of descriptors to quantify group properties (Table I). The first three quantify the spatio-temporal evolution of intra-group structure, whilst the fourth characterizes inter-group interaction. Sections VII and VIII show that they complement each other to perform well on scene-independent group state analysis and crowd video classification.
TABLE I

<table>
<thead>
<tr>
<th>Property</th>
<th>Descriptor</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collectiveness</td>
<td>$\phi^{\text{coll}}(G)$</td>
<td>8</td>
</tr>
<tr>
<td>Stability</td>
<td>$\phi^{\text{stab}}(G)$</td>
<td>14</td>
</tr>
<tr>
<td>Uniformity</td>
<td>$\phi^{\text{unif}}(G)$</td>
<td>17</td>
</tr>
<tr>
<td>Conflict</td>
<td>$\phi^{\text{conf}}(G)$</td>
<td>18</td>
</tr>
</tbody>
</table>

To facilitate explanation, we make an analogy between a point and a member. A detected group has $n$ members in a frame, which form a K-NN graph, $G(V, E)$, whose vertices $V$ represent the members, and member pairs are connected by edges, $E$. The edges are weighted by an affinity matrix $W$, with elements $w_{ij} = \exp(-d_{ij}^2/\sigma^2)$, where $d_{ij}$ is the spatial distance between two members. We denote the set of nearest neighbors of a member $z$ as $N^1_z$, $\ldots$, $N^M_z$ at every frame of a given clip. Next we discuss the descriptors in details.

A. Collectiveness

The collectiveness property indicates the degree of individuals acting as a union in collective motion. It is a fundamental and universal measurement for various crowd systems [54], [55]. A collectiveness measurement for the whole video was proposed in [55] using manifold learning. In contrast, we quantify collectiveness at the group level with the proposed collective transition prior $A$, since it captures the coherent motion of all group members. In particular, we compute the collectiveness of group $G$ as

$$\phi^{\text{coll}}(G) = \frac{1}{|G|} \sum_{z \in G} \epsilon(z, A),$$  

where $|\cdot|$ denotes the cardinality of the input set, and $\epsilon(z, A)$ is defined in Eqn. (7).

A low value in $\phi^{\text{coll}}(G)$ suggests that the members of a group move coherently towards a common destination. The descriptor is useful for distinguishing low-collectiveness groups, e.g., in a train station or wet market, from high-collectiveness groups, e.g., observed during a marathon or on an escalator track.

B. Stability

The stability property characterizes whether a group can keep internal topological structure over time. It is analogous to molecules stability in a chemical system. In particular, stable members tend to (1) maintain a similar set of nearest neighbors; (2) keep a consistent topological distance with its neighbors throughout a clip; and (3) a member is less likely to leave its current nearest neighbor set. Following this idea, we formulate three stability descriptors.

(1) We compute the first stability descriptor by counting and averaging the number of the invariant neighbors of each member in the $K$-NN graph over time

$$\phi^{\text{stab}}_a(G) = \frac{1}{|G|} \sum_{z \in G} (K - |N^1_z \setminus N^\tau_z|),$$  

where $|N^1_z \setminus N^\tau_z| = |\{ z : z \in N^1_z \text{ and } z \notin N^\tau_z \}|$.

(2) The second stability descriptor is formulated to examine if the members keep consistent topological distance with their nearest neighbors, as shown in Fig. 3. This is achieved by first ranking the nearest neighbors of a member ($z$) in accordance to their pairwise affinity, and subsequently applying the Levenshtein string metric distance ($d^2_k$) [56] to compare the rankings at every two consecutive frames. $d^2_k = 0$ if two rankings are the same, and $d^2_k = K$ if the ranking indices of all the members have changed. Through collecting $d^2_k$ over $\tau$ frames, we construct its histogram with $K$ bins, $h(z)$, for each member $z$. The second stability descriptor is then obtained as an averaged histogram

$$\phi^{\text{stab}}_b(G) = \frac{1}{|G|} \sum_{z \in G} h(z).$$  

It reveals information about the change of topological distances between members in a group.

(3) The third stability descriptor measures how likely a member would depart from its existing nearest neighbor set. We assume a random walk behavior on all the group members, i.e. we allow the members to transit freely within the group and join other members to form new neighborhood. We then measure the stability of a member as the difference between its initial and final transition probabilities. We initialize the transition probability matrix $P \in \mathbb{R}^{n \times n}$ as

$$P = D^{-1}W,$$

where $D$ is a diagonal matrix whose elements are $D_{ii} = \sum_j w_{ij}$. The probability distribution of the $i^{\text{th}}$ member ‘walks’ to and ‘joins’ other members is defined by

$$q_i = e_i^T \left[(I - \alpha P)^{-1} - I \right],$$

where $q_i \in \mathbb{R}^{1 \times n}$, $I$ is the identity matrix, and $e_i = (e_{i1}, \ldots, e_{in})^T$ is an indicator vector with $e_{i1} = 1$ and $e_{ij} = 0$. The parameter $\alpha$ has a range of $0 < \alpha < 1/\rho(P)$, where $\rho(P)$ denotes the spectral radius of $P$. We set $\alpha = 0.9/K$. The stability of $i^{\text{th}}$ member is computed by measuring the Kullback-Leibler (KL) divergence [57] of $q_i$ between the first and final frames. A lower KL-divergence score, $\delta^{kl}$ suggests
higher stability. We compute the third stability descriptor by averaging the scores across all members
\[
\phi_c^{stab}(\mathcal{G}) = \frac{1}{|\mathcal{G}|} \sum_{z \in \mathcal{G}} s^k(z). \tag{13}
\]
The final stability descriptor is
\[
\Phi^{stab}(\mathcal{G}) = [\phi_a^{stab}(\mathcal{G}), \phi_b^{stab}(\mathcal{G}), \phi_c^{stab}(\mathcal{G})]. \tag{14}
\]

The shortcoming can be complemented by three stability descriptors. For example, as shown in Fig. 3, only with the second descriptor, it cannot measure stability well from both of them have the same Levenshtein distance as 6. And this can be complemented by another two stability descriptors.

C. Uniformity

Uniformity is an important property for characterizing homogeneity of a group in terms of spatial distribution. This property is in contrast to the two previous properties that measure temporal aspects. A group is uniform if their members stay close with each other and are evenly distributed in space. A non-uniform group has a tendency to be further divided into subgroups. A comparative example of uniform and non-uniform groups is shown in Fig. 4a and 4b.

We quantify uniformity by inferring the optimal number \( c^* \) of graph cuts on the K-NN graph. A higher \( c^* \) suggests a higher degree of non-uniformity. A hierarchy of clusters (\( \mathcal{H} \)) is generated with agglomerative clustering [58] and the modularity function \( Q \) [59] is used to find \( c^* \). Specifically, given a cluster number \( c \), a graph partition \( \{V_1, \ldots, V_c\} \) is obtained from \( \mathcal{H} \). Computing \( Q_c \) for \( c \in \{1, \ldots, C\} \) and its maximum value suggests the optimal number of cuts:
\[
c^* = \arg \max_{c \in \{1, \ldots, C\}} Q_c \tag{15}
\]

given
\[
Q_c = \sum_{i=1}^{c} \left[ A(V_i, V_i) - \frac{A(V_i, V)}{A(V, V)} \right] ^2, \tag{16}
\]
where \( A(V', V'') = \sum_{w(i,j) \in V'} w(i,j) \). Examples are shown in the last column of Fig. 4. They show that a non-uniform group has a relatively higher number of cuts.

Since the uniformity of a group may change as group evolves, we measure the its uniformity by the mean \( \mu_{c^*} \) and variance \( \sigma_{c^*} \) of the optimal number of cuts over time:
\[
\Phi^{unif}(\mathcal{G}_t) = \{\mu_{c^*}, \sigma_{c^*}\}. \tag{17}
\]

D. Conflict

The conflict property characterizes interaction/friction between groups when they approach each other. The spatial distribution and level of conflict experienced by a group can be visualized on a 2D normalized map as shown in Fig. 5. Such a map is informative for crowd understanding as it contains rich information about different natures of inter-group interactions observed in different scenes. On this map, the group contour is obtained as the outer boundary of the internal members, whereas a conflict point is defined as a member with external group members in its K-NN set, \( \mathcal{N} \). Note that the K-NN sets defined here differ from those we employed earlier, as the current sets are allowed to include members from external groups.

To represent the conflict map compactly with invariance to scales, we formulate a Conflict Shape Context (CSC) descriptor inspired by shape context [60]. The first step is to capture the spatial distribution for each conflict point by computing a histogram of the relative coordinates of group contour points. This is achieved by introducing a polar coordinate system [60] centered on each conflict point, and computing the frequency of contour points in the bins. 8 equally spaced angle bins and 5 equally spaced radius bins are used. The second step is to perform K-means clustering over training clips to build a vocabulary on the histograms, and produce Bag of Words (BoW) representation. Using locally constrained linear coding [61], the \( i^{th} \) conflict point has a distribution \( u_i \) over the vocabulary. We further compute the level of conflict of this conflict point based on the CT prior introduced in Sec. III-A
\[
\epsilon^{conf}_i = \frac{1}{|\mathcal{N}_i|} \sum_{z \in \mathcal{N}_i} \epsilon(z, A), \tag{18}
\]
The \( \epsilon(z, A) \) is defined in Eqn. 7, and \( A \) is the CT prior of the group where the conflict point is residing. Intuitively, if
Fig. 6. Comparative results of group detection with four methods. Groups are distinguished with colors. Red color indicates outliers. Arrows are moving directions. Best viewed in color. (from [20])

<table>
<thead>
<tr>
<th>Methods</th>
<th>NMI</th>
<th>Purity</th>
<th>RI</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTM [14]</td>
<td>0.30</td>
<td>0.71</td>
<td>0.68</td>
</tr>
<tr>
<td>HC [12]</td>
<td>0.27</td>
<td>0.73</td>
<td>0.62</td>
</tr>
<tr>
<td>CF [13]</td>
<td>0.42</td>
<td>0.78</td>
<td>0.73</td>
</tr>
<tr>
<td>CT</td>
<td><strong>0.48</strong></td>
<td><strong>0.83</strong></td>
<td><strong>0.78</strong></td>
</tr>
</tbody>
</table>

Fig. 7. Left: quantitative comparison of group detection methods. Right: relative improvement of our approach (CT) compared with DTM, HC, and CF. (from [20])

the nearest neighbors of a conflict point are mostly external members that do not fit well to $A$, a high value in $\epsilon^{conf}$ is obtained. The final conflict property of a group is computed by max pooling $\{u_i\}$ weighted by $\{\epsilon_i^{conf}\}$ as in [61].

V. CROWD DATABASE

Evaluations are conducted on a new CUHK Crowd Dataset. It includes crowd videos with various densities and perspective scales, collected from many different environments, e.g., streets, shopping malls, airports, and parks. It consists of 474 video clips from 215 scenes, among which 419 clips were collected from Pond5\(^4\) and Getty Image\(^5\), and 55 clips were captured by us. It is larger than any existing crowd datasets [8], [55], [62] (they are actually covered by our dataset) in terms of scene diversity and clips number. Although the video clips have various lengths, we only take the first 30 frames from each clip for implementing our approach\(^6\). In all experiments, we guarantee that there is no overlap on scenes in training and test sets to demonstrate the scene-independence of the proposed descriptors. The full video clips are available in the dataset. The resolutions of videos are different, varying from $480 \times 360$ to $1920 \times 1080$. We keep the original resolutions of all videos. The frame rates are also different, varying from 20 to 30 frames per second. We conduct all experiments on the extracted frames with respect to their corresponding frame rates.

The ground truth of group detection, group state analysis, crowd video classification, and crowd video retrieval are manually annotated and checked by multiple annotators. We implemented the proposed descriptor and all the experiments on a MATLAB platform. The results reported in the following sections were measured on a 3.3 GHz Intel Core i7 processor with 16 GB RAM.

VI. EXPERIMENTAL RESULTS ON GROUP DETECTION

A. Comparison Results

Tracklets from 300 video clips are manually annotated into groups for evaluation based on the criterion that members in the same group have a common goal and form collective movement. Tracklets not belonging to any group are annotated as outliers. We compare our group detection method of using Collective Transition priors (CT) with three state-of-the-art approaches: mixture of dynamic texture (DTM) [14], hierarchical clustering (HC) [12], and coherent filtering (CF) [13]. Examples of the ground truth and the detection results in comparison are shown in Fig. 6.

DTM well separates background and simple group motions, but it performs poorly on complex and mixed group motions. Besides, it requires manual specification of group number (we provides ground truth as input) and for each clip it takes hundred-fold longer time than our method. HC hierarchically clusters tracklets with velocity and spatial constraints and does

\(^4\)http://www.pond5.com/
\(^5\)http://www.gettyimages.com/
\(^6\)We also compare the video classification results in Section VIII-B by 30 frames with that by 75 frames as used in [63].
not consider group dynamic prior. It thus leads to more errors than ours. CF detects coherent motion with a neighborhood measurement without modeling dynamics shared by the whole group. It is thus sensitive to tracking failures. This can be observed in the first row of Fig. 6, where CF splits a group moving in the same direction into subgroups. Moreover, CF first detects groups with coherent motion between consecutive frames, and then associates the groups through the whole clip. Its errors are therefore accumulated. In the second and third rows of Fig. 6, CF associates two groups moving in different directions into one due to errors made in single frames.

For quantitative evaluation, we consider group detection as a clustering problem, and adopt three widely used measurements in clustering evaluation, i.e., Normalized Mutual Information (NMI) [64], Purity [65], and Rand Index (RI) [66]. The comparison is shown in Fig. 7. The bar chart on the right shows the relative improvement of our method compared with DTM, HC, and CF.

B. Ablation Study

We further measure the influence of parameters to group detection, and analyze the significance of different steps. We use the same evaluation criteria, namely NMI, Purity, and RI.

Parameter analysis.

There are two parameters $\eta$ and $\delta$ for group detection: $\eta$ defines the threshold of velocity correlation and $\delta$ determines the range of fitting errors of each tracklet during group refinement. As shown in Fig. 9a, we keep the value of $\delta$ unchanged and let $\eta$ increase, the three evaluation criteria show that the parameter $\eta$ does not have a remarkable impact on the detection results when $\eta$ is smaller than 0.9. This is because the first condition in seeding tracklets selection has already discarded some tracklets seldom correlating with the anchor tracklet. Even though a small $\eta$ might make a large set of seeding tracklets, the refinement step can help further filter out non-coherent tracklets. As $\eta$ increases beyond 0.9, the curves significantly decrease. $\eta = 1$ leads to an extreme case that almost only one tracklet (i.e. anchor tracklet) is selected as the seeding tracklet. In this case, the criteria achieve the worst since too few seeding tracklets cannot infer accurate priors. Meanwhile, the running time decreases when $\eta$ increases. This is because the set of seeding tracklets becomes smaller as $\eta$ increases, and with a smaller and more accurate set of seeding tracklets, it costs less time to get EM algorithm converged. For instance, the average running time with $\eta = -1$ is 12.36
seconds, while with $\eta = 0.99$ is 7.65 seconds. In the same way, by fixing $\eta$, the evaluation results with respect to different $\delta$ are shown in Fig. 9b. As shown in the first (NMI) and third (RI) figures in Fig. 9b, a small $\delta$ might result in poor detection results due to the missing of many coherent tracklets. Noted that the second figure (Purity) in the Fig. 9b varies little as $\delta$ changes, and it presents a decreasing trend when $\delta < 5$. High purity might be achieved when the number of clusters is large (e.g. Purity is 1 if each document gets its own cluster). And compared to the Purity, NMI can trade off the quality of the clustering against the number of clusters. Furthermore, all the three curves nearly stop increasing when $\delta > 5$. When we continue increasing $\delta$, the curves slightly decrease, and the gap between the poorest score and the best score is within 0.2. It shows that the proposed group detection algorithm is not too sensitive on $\delta$. The qualitative results are shown in Fig. 8. Results in the first row are the worst because small $\delta$ makes more clusters, and when $\delta$ is larger than 5, it can achieve a stable detection results.

**Step analysis.**

As stated in Sec III-B, there are four steps in the group detection algorithm. The key step is to select seeding tracklets to learn CT priors. We prove its significance by comparing the detection results when omitting this step from the proposed method (named as CT w/o seeding), with the complete one (named as CT). The quantitative results are shown in Table II and their running times are shown in Fig. 10a. Both detection accuracy and running time of CT w/o seeding are worse than those of CT. Without seeding tracklets, CT priors learned from the initial tracklets clusters are not accurate enough since these tracklets are too disordered to perform well in CT prior learning. Therefore, the detection results will be messed up if the prior cannot correctly measure the collective behavior of a group. On the contrary, the set of seeding tracklets can ensure the tracklets used in prior learning are collective and coherent which makes the detection results better. Moreover, the set of seeding tracklets is smaller than the initial tracklets cluster and the seeding tracklets are much less disordered, both of which encourage the effectiveness and efficiency in learning CT prior with EM algorithm. As a by-product, this step renders a shorter running time than that of CT w/o seeding. In addition, Fig. 10b shows a statistical distribution of running time with different tracklet numbers and group numbers, which reveals that the running times are generally shorter with the smaller number of tracklets and groups.

### VII. Analysis on Group Descriptors

We are interested to know more about the group descriptors, with questions like: (1) Are different group descriptors correlated and complement each other? (2) Can these descriptors identify intrinsic group states (gases, fluids, and solid)? (3) Are these descriptors scene-independent? The following three subsections provide the quantitative and qualitative results to answer these questions.

#### A. Group Descriptors Correlation

In this section, we examine the correlation of different group descriptors and show that they each play an important role and complement each other to describe a scene. We take collectiveness as an example and compare it separately with stability and conflict descriptors.

We first rank 927 labeled groups in an ascending order based on their respective “collectiveness” (i.e. $\phi^{col}$) value, and take the ranked group ID to form the $x$-axis in Fig. 12a and Fig. 12b. Since the descriptors of “stability” and “conflict” are high-dimensional, we cannot directly compare their correlations with “collectiveness”. Consequently, we reduce their dimensions to 1 by projecting them onto their first principal subspace.

The correlation between stability and collectiveness is shown in Fig. 12a and Fig. 12b, respectively. Fig. 12a suggests that the increase of $\phi^{col}$ usually corresponds to the decrease of $\phi^{stab}$. As stated in Section IV-A, higher $\phi^{col}$ value represents lower collectiveness. It shows that collectiveness, as a whole, presents a weak positive correlation with stability – the higher...
the collectiveness is, the higher the stability of the group will be. However, if we examine in more detail, groups with a similar level of collectiveness may not own similar stability. Two examples are shown at the bottom of Fig. 12a. The first example is a video of marathon in which pedestrians run coherently and stably, while the group (marked in pink color) in the second example also coherently moves towards the same direction, but intervened by the other surrounding groups, its structure is not stable. It shows that the collectiveness and stability both play an important role and complement each other to describe a scene.

Similarly, we reduce the dimension of conflict and show its correlation with collectiveness in Fig. 12b. Different from the stability, the correlation between conflict and collectiveness is much lower. Interestingly, there are many data around −0.3, since some groups do not have “conflict” interactions with other groups. Nevertheless, their collectiveness are not guaranteed to be similar. The examples show that groups with conflict may also have high collectiveness.

In the next section, we show that the proposed group descriptors, thanks to their complementary nature, are effective in identifying different intrinsic group states.

B. Capturing Intrinsic States with Group Descriptors

Research on crowd modelling and analysis [16]–[18] generally classifies crowd particles into the following states with an analogy of classifying different phrases of matter in equilibrium statistical mechanics, as shown in Fig. 13. It is assumed that the underlying physical models are different for different states.

- **Solid**: particles moving in the same direction collectively. Their relative positions remain unchanged, bounded by internal forces.
- **Pure fluid**: particles moving towards the same direction; however, their relative positions change constantly due to the lack of inter-particle forces.
- **Impure fluid**: this state is added by us to better capture the crowd characteristics. It is similar to pure fluid, but with invasion of particles from other groups.
- **Gas**: particles moving in different directions without forming collective behaviors with others.

These states are decided by multiple socio-psychological and physical factors including crowd density, goals, interactions and relationships of group members, and scene structures. There are 927 groups manually labeled as ground truth:

![Confusion matrix of classifying group states by combining all the group descriptors (darker color represents higher accuracy).](image)

![Average accuracy of using each descriptor and combining them.](image)

![Performance drops by using only one descriptor on classifying each of the group states. A lower bar indicates that the descriptor is more effective on classifying a particular group state.](image)

![Legend for (b) and (c).](image)

![Illustration of different group states.](image)

![Representative frame with group detection results. Colors indicate different groups.](image)

128 gas groups, 291 solid groups, 349 pure fluid groups, and 159 impure fluid groups. Half of the data is randomly selected for training and the remaining for test (the training and test sets do not contain the same scenes). All of our proposed descriptors together with “group size”\(^7\) are combined as features inputed to a non-linear SVM classifier. The confusion matrix averaged over 10 trials is shown in Fig. 11a. The average accuracy\(^8\) is 60% while the chance of random guess is 25%. The result shows the effectiveness of our group descriptors and their generalization power across scenes. It is understandable that pure and impure fluid groups are the

---

\(^7\)Group size is the number of tracklets in a group normalized by the total number of tracklets in a scene. It is useful for classifying gas groups. Some groups have one pedestrian with multiple feature points.

\(^8\)We first calculate the accuracy within each class and then average them. So the biggest class will not dominate the average accuracy.
most confusing classes, since some pure fluid groups have interactions with other groups on their boundaries. The major but subtle difference between the two classes is the spatial distributions of conflict points. Tracking errors also increase difficulty in separating these two classes. Fig. 11b and Fig. 11c show the effectiveness of each group descriptor on classifying different group states. It is observed that stability and conflict are most effective on classifying solid groups. Collectiveness and conflict are most effective on classifying pure fluid groups. Group size is effective for gas groups.

As examples shown in Fig. 14, in a large open area, pedestrians behave more like gas and fluid, while move as flying solid on an escalator track or in a queue. In the figure on the left, fluid groups appear frequently on the paths connecting entrances and exits regions, while gas groups locate randomly and they are isolated customers walking around. In the figure on the right, the states of groups transit between solid and fluid at the exits of escalators. These examples show that the proposed group descriptors capture the group states well to reflect various activities in a scene.

C. Scene-Independence of Group Descriptors

In order to evaluate the scene-independence of the proposed descriptors, we cluster arbitrary regions across scenes and examine if each cluster could successfully captures coherent regional flows observed in different scenes.

More precisely, for each crowd video, we divide its image space into N cells, and extract group descriptors within each cell. Multiple groups might pass through the same cell. Thus, we gather the descriptors conveyed by all the groups that traverse this cell, and the average descriptors depict the dynamic of crowds occurs in this cell.

We use K-means algorithm to cluster the cells of different videos. Examples of three clusters are shown in Fig. 15. For each cluster (displayed in row), there are three representative cells marked with red boxes. Similar to [67], these cell clusters have semantic meanings related to the scene layout like (a) bottleneck (b) blocking/crossing (c) lane, indicating that the particular locations inside the marked cells always present some unique crowd dynamics constrained by the facilities or layout in the scenes. Importantly, the fact that each cluster captures coherent regional flow verifies the scene-independence of the proposed descriptors.

VIII. APPLICATIONS

In this section, we explore different applications to demonstrate the potentials of the proposed group descriptors.

A. Crowd Dynamics Monitoring

The group states defined in Section VII-B can help to monitor crowd dynamics over time. Typically, the transition of group states within one entire group or between groups exert valuable impact on the detection and prediction of crowd events. We extract group descriptors for each group inside a crowded scene, and use the scene-independent model learned in Sec. VII-B to recognize the group states over time.

Intra-group state transition. As shown in the first scene in Fig. 16a, the state of pink-marked group turns from pure fluid into impure fluid when the group collide with another group from the opposite direction. There is only one group in the second scene in Fig. 16a with its state transferred from pure fluid to solid. The runners accelerate at the beginning of a race with different velocities, and then keep a stable running pattern. Both examples suggest that critical points or events in crowd can be inferred through group state transition.

Inter-group state transition. The pedestrians inside the red box in the first example of Fig. 16b are moving out of escalator along the time line. They belong to the pink-marked “solid” group on the escalator. When they move out of the escalator
TABLE III
List of Crowd Video Classes.

<table>
<thead>
<tr>
<th>Class name</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly mixed pedestrian walking</td>
<td>0.3200</td>
</tr>
<tr>
<td>Crowd walking following a mainstream and well organized</td>
<td>0.0200</td>
</tr>
<tr>
<td>Crowd walking following a mainstream but poorly organized</td>
<td>0.0200</td>
</tr>
<tr>
<td>Crowd merge</td>
<td>0.0200</td>
</tr>
<tr>
<td>Crowd split</td>
<td>0.0200</td>
</tr>
<tr>
<td>Crowd crossing in opposite directions</td>
<td>0.0200</td>
</tr>
<tr>
<td>Intervened escalator traffic</td>
<td>0.0200</td>
</tr>
<tr>
<td>Smooth escalator traffic</td>
<td>0.0200</td>
</tr>
</tbody>
</table>

Fig. 17. Confusion matrices of crowd video classification (darker color represents higher accuracy). Left: using holistic features in [19]. The average accuracy is 44%. Right: using our descriptors. The average accuracy is 70%. (from [20])

and walk away, they split from the pink-marked group and turn to be a distinct violet-mark “gas” group. In the second example shown in Fig. 16b, pedestrians inside the red box originally belong to the orange-marked group and move towards the escalator. When they approach the escalator, they deviate from the orange-marked group and join into the group entering escalator and marked by pink. Their state then turns from the impure fluid (orange-marked group) to the solid (pink-marked group) state. These examples further demonstrate that either profiling the entire group or its members is useful and significant in crowd understanding and control.

B. Crowd Video Classification

We also demonstrate the robustness and effectiveness of the proposed group descriptors in the application of classifying crowd videos instead of individual groups. There exist research studies [19], [37] on using holistic descriptors to classify crowd video clips. For instance, Kratz et al. [19] divided a video clip into spatio-temporal cubiod and extracted motion features from each cube. We show that our descriptors specially designed for quantifying group properties are much more effective than generic features.

All the 474 video clips in our dataset are manually assigned into 8 classes as shown in Table III. The 8 classes are commonly seen in crowd videos and some are of special interest in crowd management and traffic control. For example, crowd merge and crowd crossing may cause traffic congestion and crowd disasters such as stampede. It is also important to keep escalator traffic smooth at the entrance and exit regions to avoid blocking, collisions, and potential dangers. In class 1, pedestrians in a scene walk in multiple directions with highly mixed behaviors. In classes 2 and 3, most pedestrians follow the main stream. In class 2, the relative positions of pedestrians are stable and there are rarely overtake events, while pedestrians in class 3 are not well organized. Most crowd videos can be generally classified into the above three categories. However, we identify a few classes (4 ~ 8) which are of particular interest in crowd management and wish to distinguish them from the remaining crowd videos. Therefore, classes 1 ~ 3 have excluded videos from classes 4 ~ 8. All the 8 categories are classified together.

Leave-one-out evaluation is used. Each time one scene (which may include multiple video clips) is selected for test, and the remaining scenes for training. Thus the cross-scene generalization capability is evaluated. If a video has multiple groups, we take the average of a descriptor over groups as the video descriptor. The non-linear SVM with RBF-kernel is used for classification. The confusion matrices are shown in Fig. 17. The average accuracy of our approach is 70%, much higher than that of random guess (12.5%) and the result of using the holistic crowd scene descriptor proposed in [19] (44%).

We perform some further analyses. First, we compare non-linear and linear SVMs. In comparison to non-linear SVM (70%), classification by linear SVM only yields an accuracy rate of 65%, which shows the effectiveness of adopting non-linear SVM. Second, we evaluate the effect of using different numbers of frames in video classification. Specifically, we compare results obtained by using the default 30 frames with 75 frames as suggested in [63]. We found that using more frames does not improve video classification accuracy significantly. Although longer frames contain more temporal information that can help motion modeling, the result by 75 frames is 72% and has only subtle improvement over that by 30 frames. We observe that the group motions in each segmented clip remain similar inside one crowd video. A satisfactory video classification result can therefore be obtained based on fewer frames.

To further evaluate our method qualitatively, Fig. 18 shows several examples and the probability of being classified into each class. The video will be assigned to the class with
the largest prediction probability. From the results, holistic features [19] often mistakenly classify the crowd video into the second class and cannot recognize other classes well. This observation is consistent to the corresponding confusion matrix depicted in Fig. 17, which has a high score in recognizing class 2 and class 8 but poor in other classes.

C. Crowd Video Retrieval

We also report the performance of the crowd video retrieval by our proposed descriptors. Each time a query video is picked from the CUHK Crowd Dataset, while the remained 473 videos form the set for retrieval. Similar to Sec. VIII-B, we exploit an average descriptor over all the groups inside a video as the video descriptor. We apply the Euclidean distance to measure the similarity between the corresponding scalar descriptor components, while $\chi^2$-distance for histogram-styled descriptor components. The performance is evaluated by the average precision in the top $k$ retrieved samples (AP@$k$). The retrieved annotations adopt the ground truth labels introduced in Sec. VIII-B.

Fig. 19a shows that our proposed descriptors outperform the holistic features [19] by enhancing the average top 100 precision (AP@100) over all video classes from 22% to 32% (a relative improvement of 45%). We also show the performance of the average precision of each video class defined in Sec. VIII-B. As shown in Fig. 19b, compared with the holistic features, our proposed descriptors have superior performance over [19] on the first 6 classes, and have a comparable AP on the 7th class, while poorer on the last class. This is consistent with the classification results shown in Fig 17. The holistic features [19] can only well recognize the second and the last class, and especially had bad performance on the recognition of the 4th and 5th classes.

IX. CONCLUSION

In this paper, we systematically study the fundamental and universal group properties, which exist in various crowd systems, from the vision point of view. These properties are motivated by the socio-psychological studies and important in crowded scene understanding. A robust group detection algorithm is proposed through learning the collective transition prior. From a graph-driven view, we design a rich set of group-property visual descriptors, including geometric structure, topological structure, and collective degree. They are well applied to scene-independent group states analysis, crowd dynamics monitoring, crowd video classification, and crowd video retrieval.

As the first few attempts on profiling group properties and descriptors from computer vision, this research will inspire new applications and extensions in the future work. The proposed group detection and group descriptors are both based on tracklets. Thus some crowded scenes with bad tracking results (e.g. time-lapse crowd videos and crowd videos shot from horizontal view) or without tracklets (e.g. crowds sitting still watching performance) are difficult to characterize group properties precisely. The future work can substitute tracklets by other medium conveying motion information. The analysis of group states transition can be extended to detect cross-scene crowd events and monitor pedestrian dynamics. The proposed descriptors not only exist in the group, but can be also extended to entire scene. The most important characteristic of these descriptors is scene-independent. Therefore, the enhanced descriptors can be applied to cross-scene crowd video retrieval.

References


