Scene-Independent Group Profiling in Crowd

Jing Shao[1] Chen Change Loy[2] Xiaogang Wang[1]
[1]Department of Electronic Engineering, The Chinese University of Hong Kong
[2]Department of Information Engineering, The Chinese University of Hong Kong
jshao@ee.cuhk.edu.hk, ccloy@ie.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk

Abstract

Groups are the primary entities that make up a crowd. Understanding group-level dynamics and properties is thus scientifically important and practically useful in a wide range of applications, especially for crowd understanding. In this study we show that fundamental group-level properties, such as intra-group stability and inter-group conflict, can be systematically quantified by visual descriptors. This is made possible through learning a novel Collective Transition prior, which leads to a robust approach for group segregation in public spaces. From the prior, we further devise a rich set of group property visual descriptors. These descriptors are scene-independent, and can be effectively applied to public-scene with variety of crowd densities and distributions. Extensive experiments on hundreds of public scene video clips demonstrate that such property descriptors are not only useful but also necessary for group state analysis and crowd scene understanding.

1. Introduction

Group dynamics have been extensively studied in socio-psychological [35] and biological [37] research as the primary processes that influence crowd behaviors. In these studies, group dynamics are characterized by both intra- and inter-group properties. *Intra-group properties*, *e.g.* collectiveness, stability, and uniformity, denote internal coordination among members in the same group. Whilst *inter-group properties*, *e.g.* conflict, reflect the external interaction between members in different groups. Such properties widely exist in animal/insect crowd systems (*e.g.* bacterial colonies and bird flocks), and are also frequently researched in socio-psychological studies [25]. For instance, bacterial colonies were found to exhibit collective behavior to achieve a common goal, *i.e.* spreading of diseases [37]. From sociological view-point, conflict occurs for competition of resources or goal incompatibility [35].

In the context of visual surveillance, groups also primarily make up human crowds. Indeed, a rich body of literature [5, 14, 19, 26] suggest that majority of the pedestrians tend to move in groups with their friends and family members. The tendency of forming a coherent group with other pedestrians becomes more prominent in dense crowds, where pedestrians have to align with others to form collective behaviors instead of moving freely [26].

Figure 1. Crowd behavior can be better understood through inherent intra- and inter-group properties. In this study, we show the possibility of quantifying such properties with scene-independent visual descriptors. Best viewed in color.

When pedestrians form groups, they exhibit some interesting properties in their dynamics, which share commonalities with socio-psychological and biological studies (Fig. 1). For instance, collective behavior is observed when pedestrians in a group maneuver towards a common destination. During a crowd disaster, turbulent dynamics in the crowd can be characterized by the stability property. Crowd tends to have non-uniform distribution when its members have different social relationships and walk in less restricted area. Two pedestrian groups with different goals, *e.g.* when crossing roads from different directions, exhibit conflict behavior. Clearly, understanding such properties provides critical mid-representation to crowd motion analysis [3, 24, 17, 15], and could facilitate other high-level semantic analysis such as crowd scene understanding, crowd video classification, and crowd event retrieval.

Our goal is to characterize and quantify these group properties from vision point of view, and study their potentials on crowd behavior analysis and crowd scene understanding. We consider a group beyond just a collection of spatially proximate individuals, but also a dynamic unit that exhibits various fundamental intra- and inter-group proper-

ties, which 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:

1) *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 [11, 39].

2) *Scene-independent group descriptors* - Based on the C-T 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 [9]. Importantly, these descriptors are scene invariant and robust to public scenes with variety of crowdedness.

3) *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 [31, 13, 12]. We also demonstrate their superiority for scene-independent group state analysis and crowd video classification over existing activity descriptors [16].

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[1].

2. Related Work

Most existing imagery-based crowd analysis methods tend to treat a crowd either as a collection of individuals [10, 23, 27] or as an aggregated whole [3, 8, 24, 17]. 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 [17, 22], dynamic texture [8], and grid of particles [3, 24], are useful for learning scene-level spatio-temporal pattern, but not directly applicable for learning group-level properties, which requires finer group segregation.

State-of-the-art methods [39, 11] achieve group detection through tracklet clustering. Zhou *et al*. [39] present the Coherent Filtering (CF) approach for segmenting coherent motion in crowd, whilst Ge *et al*. [11] discover small groups by hierarchical clustering based on pairwise objects' velocity and distance. As shown in our experiments, the above methods are either too sensitive to tracking noise or unscalable to extremely crowded scenes. Importantly, neither of them learn group properties further nor analyze crowd behaviors at the group-level.

A number of approaches [2] have been proposed for recognizing group activities such as meeting and fighting. These studies tend to analyze small social groups, with specific focus on scenario-specific predicates learning [9], contextual and interaction modeling [4, 10, 19, 21], and social signaling analysis [6]. Moreover, many crowd modeling approaches are scene-specific [17, 34, 41], i.e. activity models learned from a specific-scene cannot be applied to other scenes. Our work differs significantly to the aforementioned studies: (i) we have a different focus on understanding and quantifying the fundamental group properties, which can be well generalized to different crowd systems, and (ii) we focus on crowded scenes where a group may have an arbitrary large number of members (see Fig. 1).

3. Profiling Group Properties

We consider a group as a set of members with a common goal and collective behaviors. Given a short video clip of τ frames, a set of groups $\{\mathcal{G}\}_{i=1}^{m}$ are detected. Each group $\mathcal{G}_i$ encompasses a set of tracklets $\{\mathbf{z}\}$ detected by the KLT feature point tracker. From each detected group, we wish to extract a set of visual descriptors to represent its properties.

3.1. 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. 3.3 that certain group properties can be readily derived from the discovered CT priors.

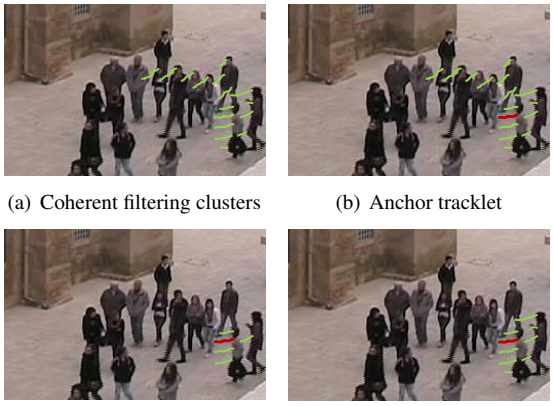
Each pedestrian group has a specific CT prior, which can be discovered from a video clip. More precisely, for n tracklets, $\{\mathbf{z}\}_{k=1}^{n}$, we assume there exist m Markov chains, where $m < n$ and m is inferred automatically. Each Markov chain is a time-series model with the form of

$$\mathbf{z}_k^t = \mathbf{A}\mathbf{z}_k^{t-1} + v^t, \tag{1}$$

where the continuous observation $\mathbf{z}_k^t$ evolves by a transition matrix $\mathbf{A} \in \mathbb{R}^{3\times3}$. Gaussian noise $v^t \sim \mathcal{N}(0, \mathbf{Q})$ is assumed between transition. Let $\mathbf{z}_k^t = [x^t, y^t, 1]^\mathsf{T}$ represent the position of a pedestrian in homogeneous coordinates[2] and the initial observation $\mathbf{z}_k^1$ follows a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. We denote $\Theta = \{\mathbf{A}, \mathbf{Q}, \mu, \Sigma\}$ as the parameters of the chain. $\mathbf{A}$ represents the CT prior, which reveals

[1] http://www.ee.cuhk.edu.hk/~xgwang/CUHKcrowd.html

[2] $\mathbf{A}$ represents affine transforms. Translation, contraction, expansion, dilation, rotation, shear, and their combinations are all affine transforms.



(a) Coherent filtering clusters (b) Anchor tracklet

(c) Seeding tracklets (d) Final group after refinement

Figure 2. (a) Coherent filtering [39] fails to distinguish two subtle groups. We address it through discovering (b) a representative anchor tracklet (in red) and subsequently (c) a set of seeding tracklets to infer a group-specific CT prior. (d) With refinement based on the CT prior, two groups are separated.

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.

3.2. 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 permits fragmented tracklets that fail to sustain over the whole clip. The missing data of $\mathbf{z}_k$ can be inferred with EM. 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 Alg. 1.

Step-1: Generate coherent filtering clusters: We first discover a set of initial tracklet clusters $\{\mathcal{C}\}_{j=1}^{r}$ using CF [39] (Fig. 2(a)). These clusters do not align with our group perception perfectly but can serve as the basis for finding the final tracklet groups $\{\mathcal{G}\}_{i=1}^{m}$.

Step-2: Identify anchor tracklets: The iterative scheme begins by randomly picking a cluster $\mathcal{C}_i$ and finding its *anchor tracklet* $\mathbf{z}_i^*$ with long duration and low variance (Fig. 2(b)).

Step-3: Discover seeding tracklets for learning CT priors: As shown in Fig. 2(c), a set of *seeding tracklets*, $\mathcal{S}_i$, are selected with the following criteria: (1) they are also from $\mathcal{C}_i$; and (2) have high velocity correlation with $\mathbf{z}_i^*$,

$$\frac{\langle v_{\mathbf{z}\in\mathcal{C}_i}, v_{\mathbf{z}_i^*}\rangle}{\|v_{\mathbf{z}\in\mathcal{C}_i}\| \cdot \|v_{\mathbf{z}_i^*}\|} > \eta, \tag{2}$$

Algorithm 1: Group detection by collective transition.

Input: Tracklets $\{\mathbf{z}\}_{k=1}^{n}$ in a video clip.
Output: m tracklet groups, $\{\mathcal{G}\}_{i=1}^{m}$.
Step-1: $i = 1$, generate coherent filtering clusters $\{\mathcal{C}\}_{j=1}^{r}$;
if $\{\mathcal{C}\} \neq \emptyset$ **then**
  Step-2: Identify an anchor tracklet, $\mathbf{z}_i^*$;
  Step-3: Discover seeding tracklets set $\mathcal{S}$ from $\mathcal{C}_i$;
  Learning the collective transition prior $\mathbf{A}_i$ with $\mathcal{S}_i$;
  Step-4: Perform group refinement to discover $\mathcal{G}_i$;
  $\{\mathcal{G}\} = \{\mathcal{G}\} \cup \mathcal{G}_i,\ \{\mathcal{C}\} = \{\mathcal{C}\} \setminus \mathcal{G}_i,\ i = i + 1$;
end

where η is a threshold. $\mathcal{S}_i$ includes reliable tracklets and is used to learn a representative CT prior with EM, which will be used to refine the group itself in Step-4.

Step-4: Group refinement: We fit each tracklet $\mathbf{z}$ in the initial cluster $\mathcal{C}_i$ with $\mathbf{A}_i$ of the i^{th} Markov chain. The fitting error ϵ of a tracklet is defined as

$$\epsilon = \frac{1}{\tau - 1} \sum_{t=1}^{\tau-1} \|\mathbf{A}\mathbf{z}^t - \mathbf{z}^{t+1}\|_2^2. \tag{3}$$

Any tracklet with $\epsilon < \delta$ is retained to construct $\mathcal{G}_i$. Unqualified tracklets will need to repeat the iterative process to be considered for a different group.

3.3. Group Descriptors for Crowd Scenes

We formulate a set of descriptors to quantify group properties (Table 1). The first three quantify the spatio-temporal evolvement of intra-group structure, whilst the fourth characterizes inter-group interaction. Sec. 4 shows that they complement each other to perform well on scene-independent group state analysis and crowd video classification.

Table 1. List of group descriptors.

Property	Descriptor	Equation
Collectiveness	$\phi^{\text{coll}}(\mathcal{G})$	4
Stability	$\Phi^{\text{stab}}(\mathcal{G})$	10
Uniformity	$\Phi^{\text{unif}}(\mathcal{G})$	13
Conflict	$\Phi^{\text{conf}}(\mathcal{G})$	14

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 $\mathbf{W}$, 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 $\mathbf{z}$ as $\mathcal{N}_{\mathbf{z}}^1, \ldots, \mathcal{N}_{\mathbf{z}}^\tau$ at every frame of a given clip. Next we discuss the descriptors in detail.

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 [32, 40]. A collectiveness measurement for the whole video was proposed in [40] using manifold learning. In contrast, we quantify collectiveness at group level with the proposed collective transition prior $\mathbf{A}$, since it captures the coherent motion of all group members. In particular, we compute the collectiveness of group $\mathcal{G}$ as

$$\phi^{\text{coll}}(\mathcal{G}) = \frac{1}{|\mathcal{G}|}\sum\nolimits_{\mathbf{z}\in\mathcal{G}} \epsilon(\mathbf{z}, \mathbf{A}), \tag{4}$$

where $|\cdot|$ denotes the cardinality of the input set, and $\epsilon(\mathbf{z}, \mathbf{A})$ is defined in Eqn. (3).

A high value in $\phi^{\text{coll}}(\mathcal{G})$ 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.

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.

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

$$\phi_a^{\text{stab}}(\mathcal{G}) = \frac{1}{|\mathcal{G}|}\sum\nolimits_{\mathbf{z}\in\mathcal{G}} \left(K - |\mathcal{N}_{\mathbf{z}}^1 \backslash \mathcal{N}_{\mathbf{z}}^\tau|\right), \tag{5}$$

where $|\mathcal{N}_{\mathbf{z}}^1 \backslash \mathcal{N}_{\mathbf{z}}^\tau| = |\left\{\mathbf{z} : \mathbf{z} \in \mathcal{N}_{\mathbf{z}}^1 \text{ and } \mathbf{z} \notin \mathcal{N}_{\mathbf{z}}^\tau\right\}|$.

The second stability descriptor is formulated to examine if the members keep consistent topological distance with their nearest neighbors. This is achieved by first ranking the nearest neighbors of a member ($\mathbf{z}$) in accordance to their pairwise affinity, and subsequently applying the Levenshtein string metric distance ($d_{\mathbf{z}}^t$) [20] to compare the rankings at every two consecutive frames. $d_{\mathbf{z}}^t = 0$ if two rankings are the same, and $d_{\mathbf{z}}^t = K$ if the ranking indices of all the members have changed. Through collecting $d_{\mathbf{z}}^t$ over τ frames, we construct its histogram with K bins, $\mathbf{h}(\mathbf{z})$, for each member $\mathbf{z}$. The second stability descriptor is then obtained as an averaged histogram

$$\Phi_b^{\text{stab}}(\mathcal{G}) = \frac{1}{|\mathcal{G}|}\sum\nolimits_{\mathbf{z}\in\mathcal{G}} \mathbf{h}(\mathbf{z}). \tag{6}$$

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

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 $\mathbf{P} \in \mathbb{R}^{n\times n}$ as

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}, \tag{7}$$

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

$$\mathbf{q}_i = \mathbf{e}_i^{\mathsf{T}}\left[(\mathbf{I} - \alpha\mathbf{P})^{-1} - \mathbf{I}\right], \tag{8}$$

where $\mathbf{q} \in \mathbb{R}^{1\times n}$, $\mathbf{I}$ is the identity matrix, and $\mathbf{e}_i = (e_1, \ldots, e_n)^{\mathsf{T}}$ is an indicator vector with $e_i = 1$ and $e_{\backslash i} = 0$. The parameter α has a range of $0 < \alpha < 1/\rho(\mathbf{P})$, where $\rho(\mathbf{P})$ denotes the spectral radius of $\mathbf{P}$. We set $\alpha = 0.9/K$. The stability of i^{th} member is computed by measuring the Kullback-Leibler (KL) divergence [18] of $\mathbf{q}_i$ between the first and final frames. A lower KL-divergence score, s^{kl} suggest higher stability. We compute the the third stability descriptor by averaging the scores across all members

$$\phi_c^{\text{stab}}(\mathcal{G}) = \frac{1}{|\mathcal{G}|}\sum\nolimits_{\mathbf{z}\in\mathcal{G}} s^{\text{kl}}(\mathbf{z}). \tag{9}$$

The final stability descriptor

$$\Phi^{\text{stab}}(\mathcal{G}) = [\phi_a^{\text{stab}}(\mathcal{G}), \Phi_b^{\text{stab}}(\mathcal{G}), \phi_c^{\text{stab}}(\mathcal{G})]. \tag{10}$$

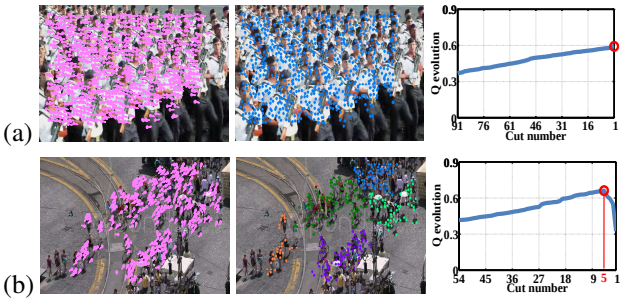


Figure 3. (a) and (b) show a uniform and a non-uniform group, respectively. From left to right, we show the original coherent group detection, the sub-groups obtained through further clustering, and the optimal number of cuts inferred by modularity function.

Uniformity: Uniformity is an important property for characterizing homogeneity of a group in terms of spatial distribution. This 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. 3(a) and 3(b).

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 [38] and the modularity function Q [30] is used to find c^*. Specifically,

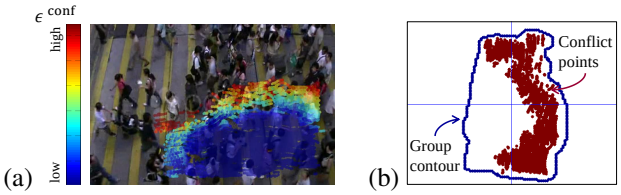


Figure 4. (a) Conflict location in groups. Hot color indicates a high degree of conflict. (b) Conflict distribution of the group. The map is normalized with the group's center, moving direction, and the largest distance from group's contour to its center.

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:

$$c^* = \underset{c \in \{1,\ldots,C\}}{\operatorname{argmax}} \; Q_c \tag{11}$$

$$\text{given} \quad Q_c = \sum_{i=1}^{c} \left[\frac{\mathcal{A}(V_i, V_i)}{\mathcal{A}(V, V)} - \left(\frac{\mathcal{A}(V_i, V)}{\mathcal{A}(V, V)} \right)^2 \right], \tag{12}$$

where $\mathcal{A}(V', V'') = \sum_{i \in V', j \in V''} w(i, j)$. Examples are shown in the last column of Fig. 3. 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 μ_{c^*} and variance σ_{c^*} of the optimal number of cuts over time:

$$\Phi^{\text{unif}}(\mathcal{G}_i) = \{\mu_{c^*}, \sigma_{c^*}\}. \tag{13}$$

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. 4. 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 [7]. 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 [7] 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 [33], the i^{th} conflict point has a distribution $\mathbf{u}_i$ over the vocabulary. We further compute the level of conflict of this conflict point based on the CT prior introduced in Sec. 3.1

$$\epsilon_i^{\text{conf}} = \frac{1}{|\mathcal{N}_i|} \sum\nolimits_{\mathbf{z} \in \mathcal{N}_i} \epsilon(\mathbf{z}, \mathbf{A}), \tag{14}$$

The $\epsilon(\mathbf{z}, \mathbf{A})$ is defined in Eqn. 3, and $\mathbf{A}$ is the CT prior of the group where the conflict point is residing. Intuitively, if the nearest neighbors of a conflict point are mostly external members that do not fit well to $\mathbf{A}$, a high value in ϵ^{conf} is obtained. The final conflict property of a group is computed by max pooling $\{\mathbf{u}_i\}$ weighted by $\{\epsilon_i^{\text{conf}}\}$ as in [33].

4. Applications and Experimental Results

We evaluate group detection, and demonstrate the effectiveness of our descriptors on two applications: group state analysis and crowd video classification. Both are scene-independent.

4.1. 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[3] and Getty Image[4], and 55 clips were captured by us. It is larger than any existing crowd datasets [3, 29, 40] (they are actually covered by our dataset) in terms of scene diversity and clips number. Although the video clips have various length, we only take the first 30 frames from each clip for implementing our approach[5]. The full video clips are available in the dataset. The ground truth of group detection, group state analysis, and crowd video classification are manually annotated and checked by multiple annotators.

4.2. Group Detection

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) [8], hierarchical clustering (HC) [11], and coherent filtering (CF) [39]. Examples of the ground truth and the detection results in comparison are shown in Fig. 5.

[3] http://www.pond5.com/

[4] http://www.gettyimages.com/

[5] We found that considering longer frames does not make significant difference in our evaluation performance. This is because group motions in each segmented clip remain similar across its whole length.

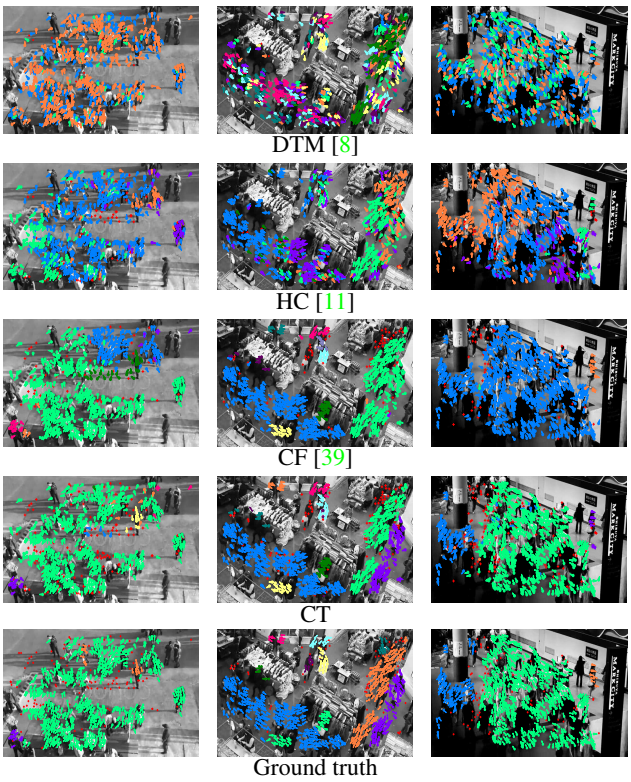


Figure 5. Comparative results of group detection with four methods. Groups are distinguished with colors. Red color indicates outliers. Arrows are moving directions. Best viewed in color.

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 not consider group dynamic prior. It thus leads to more errors than ours. CF detects coherent motions 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 column of Fig. 5, where CF splits a group moving in the same direction into subgroups. Moreover, CF first detects groups with coherent motions between consecutive frames, and then associates the groups through the whole clip. Its errors are therefore accumulated. In the second and third columns of Fig. 5, 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) [36], *Purity* [1], and *Rand Index* (RI) [28]. The comparison is shown in Fig. 6. The bar chart on the right shows the relative improvement of our method compared with DTM, HC, and CF.

Methods	NMI	Purity	RI
DTM [8]	0.30	0.68	0.71
HC [11]	0.27	0.62	0.73
CF [39]	0.42	0.73	0.78
CT	**0.48**	**0.78**	**0.83**

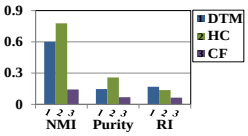


Figure 6. Left: quantitative comparison of group detection methods. Right: relative improvement of our approach (CT) compared with DTM, HC, and CF.

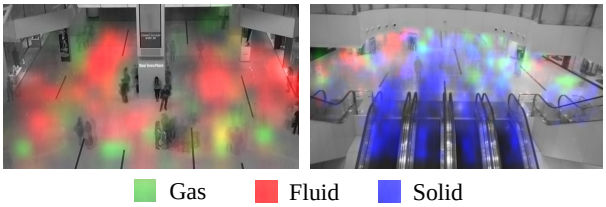


Figure 7. Distributions of different types of groups in a shopping mall and a escalator scene. Colors indicate group states automatically recognized with our descriptors. Best viewed in colour.

4.3. Application I: Group State Analysis

Research on crowd modelling and analysis [31, 13, 12] generally classifies crowd particles into the following states (*impure fluid* is added by us) with an analogy of classifying different phrases of matter in equilibrium statistical mechanics. It is assumed that the underlying physical models are different for different states.

- *Gas*: particles moving in different directions without forming collective behaviors with others.
- *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*: it is similar to pure fluid, but with invasion of particles from other groups.

These states are decided by multiple socio-psychological and physical factors including crowd density, goals, interactions and relationships of group members, and scene structures. As examples shown in Fig. 7, 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 costumers walking around. In the figure on the right, the states of groups transit between solid and fluid at the exits of escalators. In a scene where crowds compete for sources, they behave like fluid. Group states well reflect these factors, which are of interest in various applications. We use the proposed group descriptors to classify crowd groups into states, which is useful in crowd scene understanding.

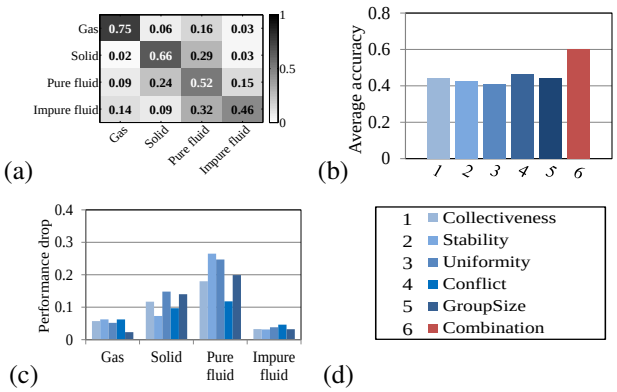


Figure 8. (a) Confusion matrix of classifying group states by combining all the group descriptors. (b) Average accuracy of using each descriptor and combining them. (c) Performance drop by using only one descriptor on classifying each of the group states. A lower bar indicates that the descriptor is more effective on classifying a particular group state. (d) Legend for (b) and (c). $1 \sim 5$ are single descriptors, and 6 is combination of the five descriptors.

There are 927 groups manually labeled as ground truth: 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"[6] are combined as features input to a SVM classifier. The confusion matrix averaged over 10 trials is shown in Figure 8(a). The average accuracy[7] 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 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. Figures 8(b) and 8(c) show the effectiveness of each group descriptor on classifying different group states. It is observed that stability and conflict are the most effective on classifying solid groups. Collectiveness and conflict are the most effective on classifying pure fluid groups. Group size is effective for gas groups.

4.4. Application II: 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 [16, 34] on using holistic descriptors to classify crowd video clips. For instance, Kratz *et al*. [16] 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 2. 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 ($4 \sim 8$) which are of particular interest in crowd management and wish to distinguish them from the remaining crowd videos. Therefore, classes $1 \sim 3$ have excluded videos from classes $4 \sim 8$. 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 it tests the cross-scene generalization capability. If a video has multiple groups, we take the average of a descriptor over groups as the video descriptor. SVM is used for classification. The confusion matrices are shown in Figure 9. The average accuracy of our approach is shown 70%, much higher than that of random guess (12.5%) and the result of using

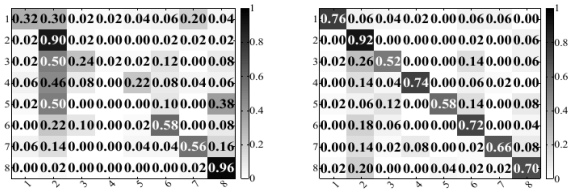


Figure 9. Confusion matrices of crowd video classification. Left: using holistic features in [16]. The average accuracy is 44%. Right: using our descriptors. The average accuracy is 70%.

Table 2. List of crowd video classes.

	Class name
1	Highly mixed pedestrian walking
2	Crowd walking following a mainstream and well organized
3	Crowd walking following a mainstream but poorly organized
4	Crowd merge
5	Crowd split
6	Crowd crossing in opposite directions
7	Intervened escalator traffic
8	Smooth escalator traffic

[6]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.

[7]We first calculate the accuracy within each class and then average them. So the biggest class will not dominate the average accuracy.

the holistic crowd scene descriptor proposed in [16] (44%).

5. Conclusions

In this paper, we systematically study the fundamental and universal group properties, which exist in various crowd systems, from the vision point of view. They are motivated by the socio-psychological studies and importance in crowd scene understanding. A robust group detection algorithm and a rich set of group-property visual descriptors are proposed through learning the collective transition prior. They are well applied to scene-independent group states analysis and crowd video classification. This research will also inspire new applications in the future work, such as cross-scene crowd event detection and modeling pedestrian dynamics with group context.

Acknowledgement: This work is supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project No. CUHK 417110, CUHK 417011, CUHK 429412).

References

[1] C. C. Aggarwal. A human-computer interactive method for projected clustering. *TPAMI*, 16(4):448–460, 2004. 6

[2] J. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3):16, 2011. 2

[3] S. Ali and M. Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*, 2007. 1, 2, 5

[4] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *ECCV*, 2012. 2

[5] A. F. Aveni. The not-so-lonely crowd: Friendship groups in collective behavior. *Sociometry*, 40:96–99, 1977. 1

[6] L. Bazzani, M. Cristani, G. Paggetti, D. Tosato, G. Menegaz, and V. Murino. Analyzing groups: a social signaling perspective. In *Video Analytics for Business Intelligence*, pages 271–305. 2012. 2

[7] S. Belongie, G. Mori, and J. Malik. Matching with shape contexts. In *Statistics and Analysis of Shapes*, pages 81–105. 2006. 5

[8] A. B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *TPAMI*, 30(5):909–926, 2008. 2, 5, 6

[9] M.-C. Chang, N. Krahnstoever, and W. Ge. Probabilistic group-level motion analysis and scenario recognition. In *ICCV*, 2011. 2

[10] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *ECCV*, 2012. 2

[11] W. Ge, R. T. Collins, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *TPAMI*, 34(5):1003–1016, 2012. 2, 5, 6

[12] D. Helbing and A. Johansson. Pedestrian, crowd and evacuation dynamics. In *Extreme Environmental Events*, pages 697–716. 2011. 2, 6

[13] L. F. Henderson. The statistics of crowd fluids. *Nature*, 229:381–383, 1971. 2, 6

[14] I. Karamouzas and M. Overmars. Simulating and evaluating the local behavior of small pedestrian groups. *TVCG*, 18(3):394–406, 2012. 1

[15] S. Kim, S. J. Guy, and D. Manocha. Velocity-based modeling of physical interactions in multi-agent simulations. In *SIGGRAPH*, pages 125–133. ACM, 2013. 1

[16] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, 2009. 2, 7, 8

[17] D. Kuettel, M. D. Breitenstein, L. Van Gool, and V. Ferrari. What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *CVPR*, 2010. 1, 2

[18] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. 4

[19] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *TPAMI*, 34(8):1549–1562, 2012. 1, 2

[20] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady*, 10:707–710, 1966. 4

[21] R. Li, R. Chellappa, and S. K. Zhou. Recognizing interactive group activities using temporal interaction matrices and their riemannian statistics. *IJCV*, 101(2):305–328, 2013. 2

[22] C. C. Loy, T. Xiang, and S. Gong. Modelling multi-object activity by gaussian processes. In *BMVC*, pages 1–11, 2009. 2

[23] C. C. Loy, T. Xiang, and S. Gong. Detecting and discriminating behavioural anomalies. *Pattern Recognition*, 44(1):117–132, 2011. 2

[24] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009. 1, 2

[25] M. Moussaid, S. Garnier, G. Theraulaz, and D. Helbing. Collective information processing and pattern formation in swarms, flocks, and crowds. *Topics in Cognitive Science*, 1:469–497, 2009. 1

[26] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PloS*, 5(4), 2010. 1

[27] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009. 2

[28] W. M. Rand. Objective criteria for the evaluation of clustering methods. *JASSA*, 66(336):846–850, 1971. 6

[29] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert. Data-driven crowd analysis in videos. In *ICCV*, 2011. 5

[30] S. Smyth and S. White. A spectral clustering approach to finding communities in graphs. In *SDM*, 2005. 4

[31] J. Toner, Y. Tu, and S. Ramaswamy. Hydrodynamics abd phases of flocks. *Annals of Physics*, 318:170–244, 2005. 2, 6

[32] T. Vicsek and A. Zafeiris. Collective motion. *arXiv*, 1010.5017, 2012. 4

[33] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 5

[34] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *TPAMI*, 31(3):539–555, 2009. 2, 7

[35] S. A. Wheelan. *The handbook of group research and practice*. Sage, 2005. 1

[36] M. Wu and B. Schölkopf. A local learning approach for clustering. In *NIPS*, 2006. 6

[37] H.-P. Zhang, A. Beér, E.-L. Florin, and H. L. Swinney. Collective motion and density fluctuations in bacterial colonies. *PNAS*, 107(31):13626–13630, 2010. 1

[38] W. Zhang, X. Wang, D. Zhao, and X. Tang. Graph degree linkage: agglomerative clustering on a directed graph. In *ECCV*. 2012. 4

[39] B. Zhou, X. Tang, and X. Wang. Coherent filtering: detecting coherent motions from crowd clutters. In *ECCV*. 2012. 2, 3, 5, 6

[40] B. Zhou, X. Tang, and X. Wang. Measuring crowd collectiveness. In *CVPR*, 2013. 4, 5

[41] B. Zhou, X. Wang, and X. Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *CVPR*, 2012. 2