Pattern Recognition 75 (2018) 112–127

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Gestalt laws based tracklets analysis for human crowd understanding☆

Weiqi Zhao[a,b], Zhang Zhang[a,b], Kaiqi Huang[a,b,c,*]

[a] CRIPAC & NLPR, CASIA, PR China
[b] University of Chinese Academy of Sciences, PR China
[c] CAS Center for Excellence in Brain Science and Intelligence Technology, PR China

ARTICLE INFO

Article history:
Received 24 November 2016
Revised 15 April 2017
Accepted 11 June 2017
Available online 20 June 2017

Keywords:
Similarity measurement
Group detection
Semantic regions
Hierarchical clustering

ABSTRACT

Crowded scene analysis is a popular research topic due to its great application potentials, such as intelligent video surveillance and crowd density estimation. In this paper, we propose a novel approach to detecting crowd groups and learning semantic regions with a unified hierarchical clustering framework. According to the Gestalt laws of grouping, we propose three priors to define a unified similarity metric to measure the similarities of pairs of original tracklets and pairs of representative tracklets from different crowd groups, so that the short-term crowd groups and the long-term semantic paths commonly composed of several short-term crowd groups can be detected by a bottom-up hierarchical clustering algorithm simultaneously. In order to verify our method at the longer time duration video sequences in the crowded scene, we construct a new crowd database (CASIA crowd database [1]) with various crowd densities in real scenes. Extensive experiments on our CASIA crowd database, Collective Motion Database and CUHK database are performed, and the results demonstrate that our approach is effective and reliable for crowd detection and semantic scene understanding in various crowd densities, especially for the crowd analysis in long temporal video clips.

1. Introduction

On account of the important applications in public security and traffic management, intelligent video surveillance for crowd in public areas with high population density has been widely concerned. Meanwhile, using surveillance tracklets or trajectories data for human behavior analysis [1–3] has drawn a lot of attentions, ranging from activity recognition based on the motion pattern of a single individual or interactions [4] to analysis of the flow of human crowd. For example, Zhang and other researchers [2,5] try to find the source and sink points from semantic region, then cluster tracklets according to the path learnt from the source and sink points, and others try to discover the pathways for abnormal event detection [3]. However, those methods lack of the unified analysis framework of long-term path modeling and short-term crowd group detection. (In our work, the short-term groups consisted of tracklets, and the long-term moving paths are learnt from short-term groups.)

In most of crowded scenes, pedestrian move with certain paths to reach their destinations from one location to another, which can be understood as a number of semantic clusters hierarchically. For example in Fig. 1, the crowd in shopping mall includes several large clusters moving in the different paths, then for each cluster, it can be further divided into some small clusters based on the spatio-temporal distribution along with the path. Therefore, the crowd scene can be modeled by two kinds of clusters in different spatio-temporal scales, one is the spatial locally and short-term crowd groups, the other is the semantic regions over a long-term duration in the whole scene. In crowded scene analysis, detecting moving crowd groups and obtaining their underlying attributes have attracted researchers to devote themselves for many applications, such as crowd tracking [6,7], semantic scene modeling [2,3,5] , crowd activity recognition [4,8], people counting [9–11] and crowd motions detection [1,12,13].

In general, crowd motion can be explained by empirical sociological research [14] where the author find that the crowd in different scenes share some common spatio-temporal characteristics. For example, pedestrian prefer to walk together with their friends in order to communicate with each other conveniently, so they are commonly perceived as one group. While pedestrian tend to keep distance to others they are unfamiliar with Mehdi et al. [15]. Therefore, original small groups appeared according to the scene structure and their intimate relationships. Inspired by this phenomenon, Zhou et. al. [12] apply K-Nearest Neighbors (K-NN) in adjacent temporal frames to find tracklets which are close to each

☆ https://drive.google.com/open?id=0B11LUV32V3OLMW1od19HWUJiYTQ

* Corresponding author.
E-mail addresses: weiqi.zhao@nlpr.ia.ac.cn (W. Zhao), zzhang@nlpr.ia.ac.cn (Z. Zhang), kaiqi.huang@nlpr.ia.ac.cn (K. Huang).

http://dx.doi.org/10.1016/j.patcog.2017.06.020
0031-3203/

Fig. 1. Crowd in different main paths can be considered as some semantic clusters, which can further be divided into small groups based on their spatio-temporal distribution.

other coherently and have approximate velocity correlations. However, this method finds out coherent tracklet groups in short time durations, which miss the structural information of whole trajectories in longer temporal duration. Ge et. al. [1] find small groups between adjacent tracklets by using hierarchical clustering method. But this method only focuses on the tracklets with temporal overlap, and all of these tracklets have same temporal durations instead of different life-spans. Thus, few of works focus on the temporal relevance of tracklets with both overlap part and non-overlap part simultaneously, and cluster tracklets with various life-spans from the tracklet set obtained by long-term tracking.

Motivated by the above analysis, the goal of this work is to learn long-term scene paths and detect short-term crowd groups simultaneously with a unified hierarchical clustering framework. First of all, we propose several priors to guarantee our algorithm more efficient and reasonable based on the temporal overlap and non-overlap part during the whole life-spans of pairs of tracklets. Then, we design a unified similarity measurement according to the spatio-temporal relationships between tracklets. Furthermore, based on this distance measurement, we cluster the tracklets to be different crowd groups and extract representative tracklets from each cluster to be further clustered to the scene path.

This paper is an enhanced version of our previous STS (Spatio-Temporal Similarity) method of the conference paper [16], and there are three main extensions in this work. Firstly, We describe how to determine the cluster numbers in details by applying the Minimum Within-cluster Difference and Maximum Between-cluster Different. We also apply the temporal window to divide the time axis into several parts, and just measure the similarity of the tracklets and cluster them within this temporal window, which is more authentic in a dynamic and random crowded scene. Secondly, in order to demonstrate the effectiveness of our method in long temporal durations, we construct a new CASIA crowd dataset which contains longer crowd video sequences, and the ground truth of group detection and motion paths in the whole temporal duration are manually annotated. Furthermore, more experiments are conducted to verify the proposed method.

The contributions of this paper can be summarized as follows.

- We design three priors according to the Gestalt laws of grouping [17,18], which depend not only spatial information but also temporal information between tracklets. Thus, our method can consider both the temporal overlap part and temporal non-overlap part respect to pairs of tracklets with various life-spans.
- Different from the frame level clustering in [12,19] which may lead to inconsistent clustering labels in time duration, the proposed tracklet-level clustering can obtain more consistent labeling results.
- The proposed unified measurement is more general and intuitive representations of crowded scene in hierarchical structure, where the low level clusters corresponds to some short-term crowd groups and the scene paths can be discovered in the high level clusters in larger spatio-temporal scales, and the number of groups are determined automatically by intra-group tightness and inter-group difference.
- We construct a new crowd database contains longer temporal video sequences, and manually annotated the ground truth of group detection and motion paths in the whole temporal duration, which is more suitable to model the crowd paths in real scenes. At the same time, we label the ground truth of crowd groups and paths in different crowd scenes. For future research on long-term crowded scene analysis, we will publish the new dataset shortly.

This paper is organized as follows. In Section 2, we briefly overview current methods on crowd analysis and tracklet clustering. Section 3 describes the three priors relative to our method and the proposed unified similarity measurement. The experimental results are shown in Section 4, and Section 5 concludes this paper.

2. Related work

Crowd analysis and group detection through the fragments of trajectory (tracklets) or trajectories clustering are an active research topic in computer vision. A number of researchers propose various solutions to certain crowded scenes from different views. For example, Zhou and other researchers [12,19–25] treat the crowd as a collection of individuals. By decomposing a complex behavior pattern according to its temporal characteristics or spatio-temporal visual contexts, they model the decomposed behaviors or detect crowd groups with different priors or models, such as Zhou, et al., [5,12,22]. While Wu and other researchers [26–30] consider the moving crowd as an aggregated whole entity or an aperiodic dynamical system, and apply some physical or fluid dynamics models to analyze the moving crowd properties, such as Chaotic Model [26], Social Force Model(SFM) [27] and Finite Time Lyapunov Exponent (FTLE) field [29]. These methods are useful to abnormal crowd behavior detection or crowd flow segmentation, but not suitable for semantic regions learning.

Lots of methods have been proposed to cluster tracklets or trajectories to learn semantic regions. Previous semantic regions learning approaches can be found in [31,32], but they ignore the attributes along these trajectories. Other state-of-the-art methods use hierarchical clustering to learn the semantic regions or detect crowd groups [1,3,33]. Zhang et. al. [34,35] apply event rule induction to analyze trajectory series and compare six similarity measures for trajectory clustering in outdoor surveillance scenes. Ge et. al. [1] apply hierarchical clustering to find small groups, those groups are measured by a generalized and symmetric Hausdorff distance defined with respect to pairwise proximity and velocity. Zhang et.al. [25] construct a relation graph to discover the relation of trajectories based on slow feature analysis. Shao et. al. [19] use visual descriptors to quantify the group properties and propose the collective transition prior to detect crowd groups. While their work leave out of consideration of temporal non-overlap parts between tracklets.

We classify these related methods into two categories. Spatial distribution based methods [2,36] and spatial distance based methods [3,5,12,19,37,38]. In the first class, Zhang et. al. [2] use co-training algorithm to train two classifiers, one is LDA-based classifier and the other is AdaBoost classifier. Then, an underlying parameter model is used to fit the spatial distribution of trajectories. In the latter class, Zhou et. al. [12] propose the coherent neighbor invariance prior to characterize the local spatio-temporal relationships of tracklets for group detection. However they just focus on instantaneous tracklets clustering frame by frame, and the simple group association method cannot obtain a consistent grouping results over the whole video clip. Zhou et. al. [5] extend the existing Latent Dirichlet Allocation (LDA) topic model with a Markov random field as prior to enforce the spatial and temporal coherence between tracklets during the learning process.

Other researchers cluster tracklets based on the Gestalt laws of grouping [23,39–41]. They define the distances or affinities interrelation between tracklets to obtain a weighted graph, then the spectral clustering will be used to aggregate these tracklets. However, this work does not consider the tracklets with temporal overlap parts and some tracklets are labeled with source and sink as prior information.

Different from the unsupervised learning based methods [16,19,22,42–45], some methods try to analyze crowd scenes in the supervised way by applying a train stage based on the combination of appearance and motion features or other forms of features [46,47]. Generally, those methods will apply a SVM (Support Vector Machine) based on the designed descriptors to train a classifier. Solera et al. [46] proposes a method for detecting social groups in crowd by correlation clustering procedure on trajectories, and the affinity between crowd members is learnt through an online formulation of the Structural SVM. However, they focus on sparse trajectories instead of dense tracklets, so there is just one trajectory in a group in most instances. This is inappropriate for the crowded scenes which general contain a few hundred tracklets, so this sparse trajectories based method is not suitable to analyze those crowded scenes with dense tracklets.

Our work differs to the above-mentioned studies, we can measure the tracklets with both temporal overlap part and non-overlap part simultaneously, and cluster tracklets with different life-spans from the whole tracklet set. Furthermore, our method can obtain more consistent labels based on the proposed tracklet-level clustering.

3. Proposed methodologies

Our framework is shown in Fig. 2. For a given video sequence, tracklets are extracted by applying the KLT feature point tracker [48]. Before clustering, we remove outliers (i.e., static tracklets and too short tracklets) firstly. Then, we will measure the similarity of pair of tracklets by designing a unified similarity measurement based on the three Gestalt laws priors in certain temporal window. In order to better describe the forming process of crowd, the hierarchical clustering and representative tracklets are applied, and the number of groups is automatically determined by the intra-group tightness and inter-group difference. The representative tracklets possess the similar characteristics of crowd groups, and make the repeated hierarchical clustering process with less computation. Furthermore, the representative tracklets are further used to learn the crowd behaviors. Finally, based on the clustering results, the semantic region (i.e., moving path) will be learnt and further applied to anomaly detection.

Due to the clutter noise and tracking failure in crowded scene, the long trajectories of moving objects are hard to be extracted. Therefore, we use both the fragments of trajectories (tracklets) with short life-span and the long trajectories with long life-span to model the crowd scene in our work, on account of the excellent characteristics that short trajectories carry more dynamic information on individuals and small groups, and long trajectories encode more global information about the motion paths and structures of the scene.

3.1. Crowd clustering priors

Gestalt laws are rules of the organization of perceptual scenes which are first introduced in philosophy and psychology in 1890. According to the theory of Gestalt laws, people usually perceive complex scenes composed of many groups of objects on some backgrounds, with the objects themselves consisting of parts, which may be composed of smaller parts, etc. [17]. It also regarded as a set of principles accounting for the observation that humans naturally perceive objects as organized patterns and objects. Banerjee et. al. [18] consider these principles exist because the mind has an innate disposition to perceive patterns in the stimulus based on certain rules. Inspired by these principles, we define three priors in our method for more reasonable tracklets clustering results. As shown in Fig. 3, there are two rows clustering results with different scales and the corresponding Gestalt laws in the third row. The first rows are the primal clustering results with small groups, and the second rows are the final clustering results with bigger groups. Both of these two rows are the results of hierarchical clustering based on the three priors and the unified similarity metric.

3.1.1. Spatial proximity prior

The law of proximity states that when an individual perceives an assortment of objects, they perceive objects that are close to each other as forming a group. (p1) in Fig. 3 illustrates the Law of proximity, where we perceive the collection of ellipses into four groups [18,49]. Inspired by this principle, we consider that tracklets with Spatial adjacent will prefer to be clustered than those remote. As shown in (a) and (b) of Fig. 3, people walking side-by-side often possess higher similarity than those remote. In addition, besides the calculation of the spatial affinity between them, a higher similarity also should be assigned to the tracklets with temporal co-occurrence, which promises the tracklets with spatio-temporal proximity have higher similarity.

3.1.2. Similarity properties prior

Inspired by the Gestalt law of similarity, which has the tendency to group objects together if there are similar properties such as shape, moving direction, color and shading with each other. For example, (p2) of Fig. 3 illustrates the law of similarity where the ellipses with the same property of color are grouped into one cluster [50]. We assume that tracklets belonging to one crowd group should own approximative life-spans (temporal lengths) and moving directions. Therefore, tracklets with approximative life-spans and moving directions will be given more similarity than others, which is in accordance with the explanation of empirical sociological research interpreted above. In our work, tracklets with similar properties and spatial adjacent will be firstly clustered, then they will be further grouped to find the path in the scene in an upper hierarchy as shown in Fig. 3 (c) and (d).

3.1.3. Spatiotemporal closure prior

The law of closure states that individuals perceive objects such as shapes, pictures, etc., as being a whole when they are not complete. Specifically, when parts of a whole shape are missing, our perception fills in the visual gap. Research shows that the reason the mind completes a regular shape with missing parts is to increase the regularity of surrounding stimuli. For example, the Fig. 3 (p3) that depicts the law of closure where the line fragments are perceived as an ellipse on the left side and a rectangle on the

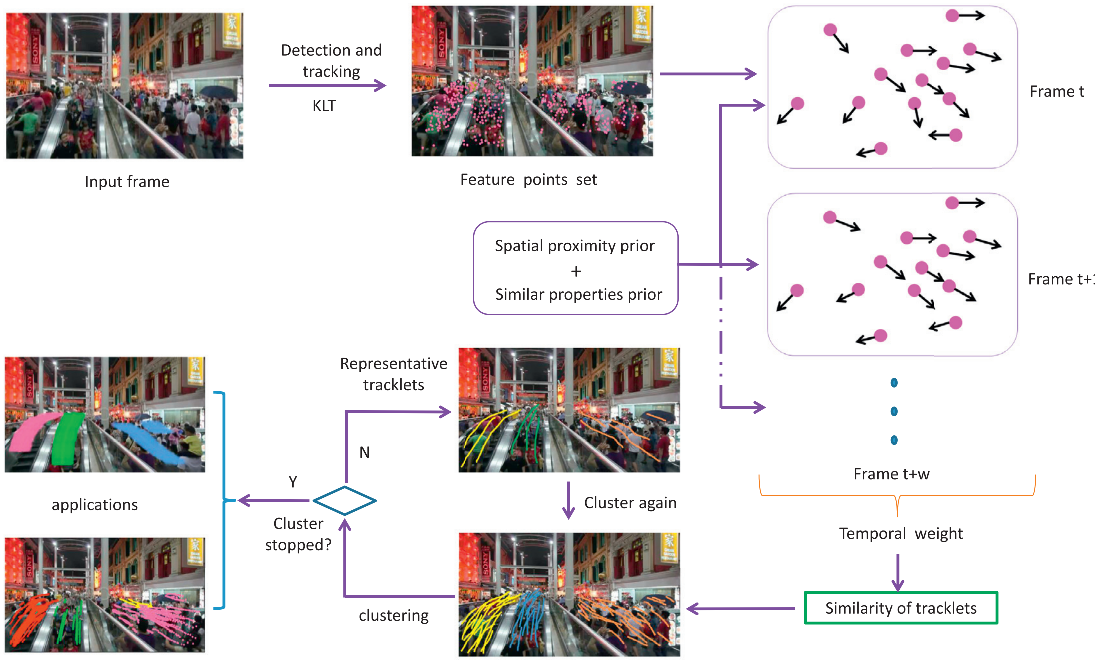


Fig. 2. The framework of the proposed method. Firstly, we apply KLT tracker to extract the tracklets of input video, it illustrated as feature points set in single frame, (i.e., the feature points set means the moment of this frame of tracklets. Secondly, we measure the similarity of feature points according to the first two priors (e.g., related to affinity and velocity correlation), and weighted with the temporal information in certain temporal window w. Then, we cluster the tracklets by applying the hierarchical clustering based on the unified similarity, and we also apply representative tracklets for more convenient and faster in the repeated clustering process. Finally, the learnt crowd groups will be used to further analysis the crowd. The images are selected from CUHK crowd dataset [19].

right side. If the law of closure does not exist, the shape would depict an assortment of different lines with different lengths, rotations, and curvatures. However, with the law of closure, we perceptually combine the lines into whole shapes [50,51].

In our work, we consider that the crowd scene consists of different pathways, and those pathways are further composed of different tracklets groups. Similarly, those tracklets groups also be made up of several long life-span tracklets and smaller groups with short life-span tracklets. However, there are spatial or temporal gaps among these tracklets and groups, where some parts of the whole path are missing. Thus, we consider that tracklets and groups which can be fitted by a certain long path should be clustered as a whole even there are spatio-temporal gaps. In addition, in order to model the paths more reasonable and practical, we prefer that short life-span tracklets (both temporal overlap part and non-overlap part) are firstly clustered, and the long life-span tracklets will serve as a prior guidance of the path.

The whole process could be unified as a hierarchical clustering framework with different spatio-temporal scales in different layers. The learnt low-level groups representing short-term crowd groups will own smaller spatio-temporal scale while the high-level clusters correspond to the global paths in large scale. In addition, the multiscale characteristics can be used to reduce the time cost of pairwise distance computation and filter out those tracklet pairs with far spatio-temporal distances. As shown in Fig. 3, the first two rows interpret this phenomenon with different spatio-temporal scales in different layers.

3.2. Design of similarity metric

The whole life-spans of two tracklets could be divided into two parts according to their temporal relationships, i.e., temporal overlap (co-occurrence) part and temporal non-overlap (temporal gap) part. As shown in Fig. 4, tracklet A and B have both temporal overlap part and temporal non-overlap part, while A and C have temporal non-overlap part only. Therefore, in order to calculate the final similarity of two tracklets, we should analyze their temporal relationships and fuse it to the spatial distance. Thus, based on the above three priors, we combine the spatial distance and temporal correlation to define the unified similarity function of pairs of tracklets as follows:

$$S = (\lambda \cdot F)^{\frac{1}{W}} \tag{1}$$

where F and W are the spatial similarity and temporal weight of two tracklets respectively, and λ is a scale parameter. Then we will illustrate how this similarity function integrates the three priors of Gestalt laws with spatial distance and temporal weight.

3.2.1. Spatial similarity

We treat a tracklet as a series of observations $A = \{\vec{a}_i\}$, where $\vec{a}_i = (x_i^a, y_i^a, t_i^a)$, (x_i^a, y_i^a) is the spatial coordinate and t_i^a is the moment of the ith observation. In addition, we define $T^A = \{t_{start}^a, t_{start+1}^a, \cdots t_{end-1}^a, t_{end}^a\}$ as the temporal indices set of A.

According to the first two priors, two tracklets have higher similarity if those tracklets are spatially close to each other and own similar attributes. Therefore, adjacent tracklets with similar moving direction will prefer to be perceived as a group than those disordered. We apply the modified Hausdorff distance similar to Wang et al. [3] to measure the spatial similarity of pairs of tracklets. We compute the spatial distance by using this distance function instead of the Euclidean distance, because the former is suitable for pairs of the tracklets with different life-spans while the latter not. Consider two tracklets $A = \{\vec{a}_i\}$ and $B = \{\vec{b}_i\}$, for each observation $\{\vec{a}_i\}$ in A, we search the nearest observation $\xi(i)$ in B as follows:

$$\xi(i) = \arg\min_{j \in B} \left\| \left(x_i^a - x_j^b, y_i^a - y_j^b\right) \right\| \tag{2}$$

Then, we compute the Euclidean distance and velocity correlation between these two observations. By summing the distances from the first observation to the last observation and divide the length

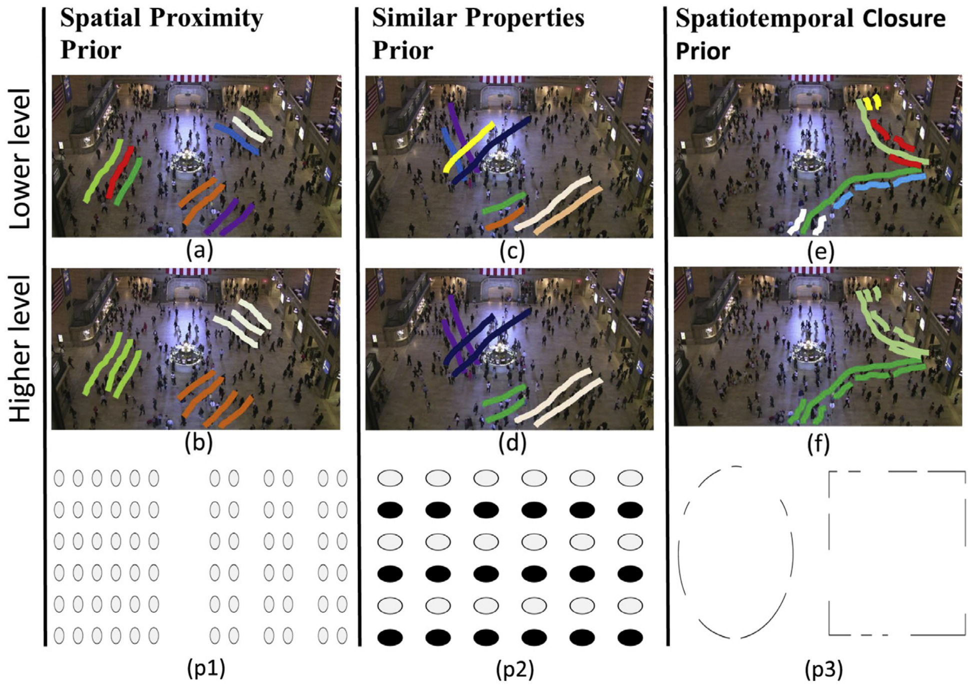


Fig. 3. Illustration of hierarchical clustering in different layers. (a). Most of tracklets are assigned with different groups denoted by different colors in lower clustering level.(b). Tracklets with spatio-temporal proximity will be clustered into a bigger cluster in higher clustering level. (c) and (d). Tracklets with similar properties (i.e., moving directions and lifespan) will be clustered into a bigger cluster in higher clustering level. (e). Short life-span tracklets have higher similarity than that of long temporal tracklets so that short temporal tracklets will be clustered firstly in terms of certain spatio-temporal scale.(f). The tracklet groups will be gathered to bigger clusters to form the moving path in upper layer. (p1)-(p3) are the corresponding Gestalt laws.

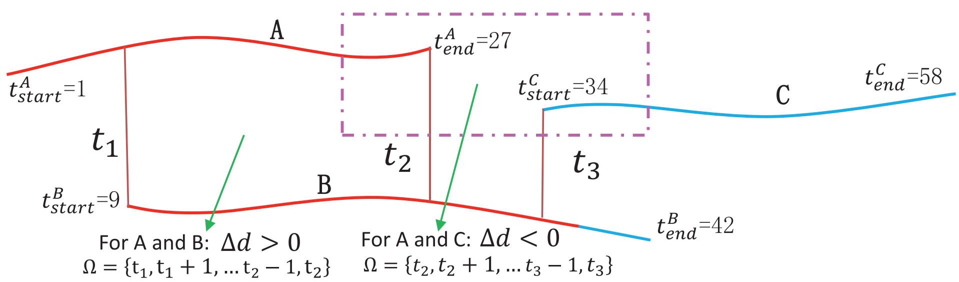


Fig. 4. Temporal relationships of tracklets A, B, C, where A and B have temporal overlap part from t_1 to t_2, and the rest of A and B belong to temporal non-overlap part, and there are left part ($t < t_1$) and right part($t > t_2$) in non-overlap region. Moreover, A and C are temporal non-overlap part only. B and C have both temporal overlap and non-overlap parts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of A, we get the full distance from A to B as follows.

$$d(A,B) = \frac{1}{N_A}\sum_{i\in A}\left(\left\|\vec{a}_i - \vec{b}_{\xi(i)}\right\|_2^2 + \beta\left(1 - \frac{v_i v_{\xi(i)}}{\|v_i\|\left\|v_{\xi(i)}\right\|}\right)\right) \quad (3)$$

where N_A is the number of observations in A, β is the balance parameter to balance the two items of Eq. (3), and v_i and $v_{\xi(i)}$ are the velocities of $\vec{a}_i$ and $\vec{b}_{\xi(i)}$. As the distance of $d(A, B)$, $d(B, A)$ are asymmetric, we also compute the full distance from B to A. The normalized smallest one is the final distance between A and B. Thus the spatial similarity is defined as:

$$F(A,B) = exp(-f(A,B)/\sigma) \quad (4)$$

where

$$f(A,B) = min(d(A,B), d(B,A)) \quad (5)$$

The first part of Eq. (3) enables the proximate tracklets to be assigned with smaller distance, which satisfies the first prior that adjacent tracklets have the precedence to be one group, and also satisfies the third prior that tracklets with longitudinal separation can be clustered together. The second part makes sure that tracklets with similar directions will be assigned with smaller velocity correlation values, (i.e., they are more similar than those with opposite directions), which consistent with the second prior that tracklets with similar directions prefer to be clustered together.

However, on account of the distribution of tracklets and the innate characteristic of the modified Hausdorff distance which prefer to search the nearest point, we find that not all the observations in tracklets are helpful to measure the similarity. For example, if there is a large spatial-temporal gap between two tracklets, only the adjacent parts of these two tracklets (in dotted box of Fig. 4) are useful to measure the similarity. Therefore, we will find an appropriate temporal range for similarity calculation. Given two tracklets, we denote the first tracklet as A if $T^{first}_{start} <= T^{second}_{start}$, and the another is tracklets B. Then, $\Delta d = T^A_{end} - T^B_{start}$, where a positive Δd indicates that there is a temporal overlap part between two tracklets and vice versa. We also define Ω as a temporal moment set related to Δd as shown in Fig. 4, then we have:

$$dt = min\left\{\left\lceil \frac{max\{L_{T^A}, L_{T^B}\}}{f(\Delta d)}\right\rceil, min\{L_{T^A}, L_{T^B}\}\right\} \quad (6)$$

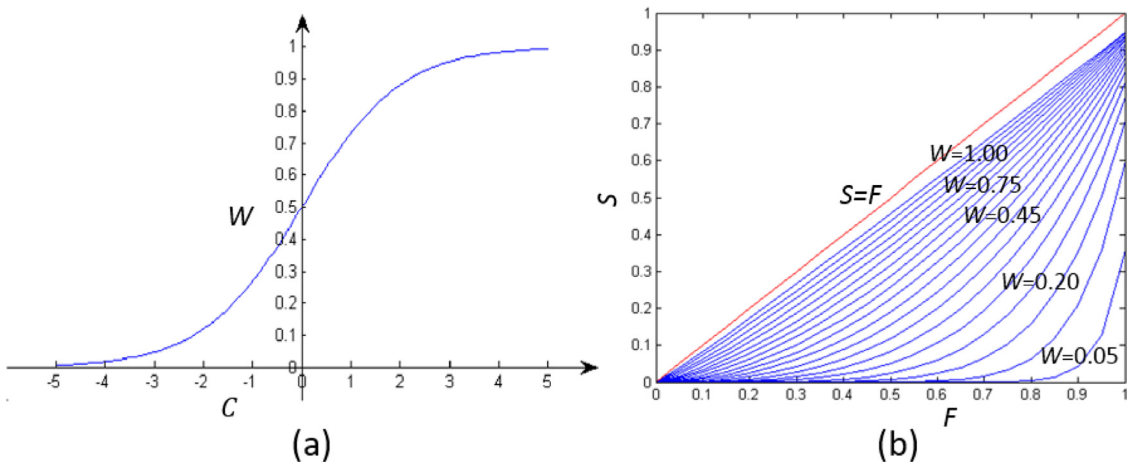


Fig. 5. (a) denotes the weight function of C. (b) is the mapping function from F to S with different W, where $\lambda = 0.95$.

where dt is the number of observations for calculating similarity in each tracklet, $\lceil \cdot \rceil$ means up to the nearest integer, $f(\Delta d) = \gamma^{\Delta d}$ is a monotone decreasing function of Δd, and we set γ=0.8. L_{T^A} and L_{T^B} are the length of T^A and T^B (i.e., the whole temporal length of tracklets A and B) respectively. We define $\Gamma = \{t^{\Omega}_{start} - dt, t^{\Omega}_{end} + dt\}$ as a set of temporal moments, then the temporal range of tracklet A and B are $\Gamma \bigcap T^A$ and $\Gamma \bigcap T^B$ respectively as shown in Fig. 4. For A and B, the red parts are the temporal ranges, and the temporal ranges for A and C are indicated by the dotted box.

3.2.2. Temporal weight

The temporal information is used to weight the spatial similarity to ensure the tracklets can be clustered from small scale to large scale hierarchically, and we also can measure the similarity of pairs of tracklets which are not exist in the same temporal duration simultaneously. Furthermore, this temporal weight enables tracklets with temporal gaps to be clustered as a group, and this is also a vital step to find the crowd paths where the tracklets or groups are not consecutive on.

According to the first and third prior, the temporal weight should make sure that tracklets with spatio-temporal proximity will be assigned larger similarity and short life-span tracklets should be firstly clustered. Then the weight is defined as:

$$W = 1/(1 + exp(-C)) \tag{7}$$

where

$$C = \Delta d \cdot (\eta - \eta_{\_threshold})/(max\{L_{T^A}, L_{T^B}\})^k \tag{8}$$

where k is a scale parameter, and η is the tracklet life-span ratio which is defined as the temporal length of the shorter tracklet divides the longer one: $\eta = L_{T_{shorter\ one}}/L_{T_{longer\ one}}$, L_{T^A} and L_{T^B} are the whole temporal length of Tracklets A and B respectively. We set $k = 2$ and $\eta_{\_threshold} = 0.4$ to the preferable results. According to the second prior, η can make sure that tracklets with equal temporal length will be assigned larger temporal weight. As interpreted in the middle column of Fig. 3, tracklets with approximately equal temporal lengths and same moving directions will have the precedence to be grouped together. In addition, according to the third prior (the last column in Fig. 3), Δd can make sure tracklets with temporal adjacent own larger temporal weight, so the tracklets and groups which can be fitted by a certain semantic region will be clustered as a whole even there are spatio-temporal gaps, in this way we can find some semantic regions of the crowded scene (i.e., the moving paths).

Eq. (1) maps F and W into a new similarity space as shown in Fig. 5 (b). A small temporal weight will lead to a smaller similarity,

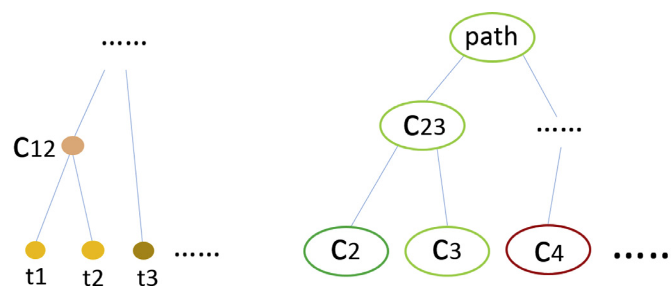


Fig. 6. Left: in low-level clustering layer, single tracklet will be clustered according to the three priors. Right: small groups are merged to be bigger one in the higher clustering layer, and these bigger groups are used to further learning the crowd path.

no matter how large F is, which is reasonable since tracklets with large temporal gap should be assigned smaller similarity based on the first prior. Therefore, a larger S means that the pair of tracklets are spatially and temporally close, as well as similar moving directions so that they have precedence to be clustered together.

3.3. Clustering and path modeling

Each pair of tracklets will be assigned a finial spatio-temporal similarity value by applying the designed similarity measurement. Therefore, given a set of existing long pedestrian trajectories and short tracklets, we can form an $N \times N$ affinity matrix where each element contains a similarity value to indicate the affinity of pairs of tracklets, and the affinity measure is a combination of spatial and temporal terms. Therefore, this is a symmetrical affinity matrix where the diagonal elements are zero. We identify groups based on a bottom-up hierarchical clustering approach that starts with individuals as separate clusters and gradually builds larger groups by merging two clusters with the strongest intergroup closeness (i.e., the smallest spatio-temporal similarity value in the affinity matrix), a brief process as Fig. 6 shows. Alternatively, one could take a top-down approach, starting with the entire crowd as a whole group and iteratively splitting into subgroups based on the same distance measurement. We choose the bottom-up approach because it is more efficient in crowds composed of small groups [1].

Compared with other clustering approaches (e.g., spectral clustering or K-means), our approach does not require a predefined cluster number. To automatically decide when the clustering process stopped and discover the optimal number of groups in different crowded scenes, we apply the Minimum Within-cluster Difference and Maximum Between-cluster Difference based on the affinity matrix. For a given cluster number n, the Within-cluster Difference $S_W(c_i)$ is:

$$S_W(c_i) = \sum_{p,q\in c_i, p\neq q} (S_{pq} - \mu_{c_i}) \tag{9}$$

where

$$\mu_{c_i} = \frac{1}{N_i} \sum_{p,q\in c_i, p\neq q} (S_{pq}) \tag{10}$$

Then we have S_W as follows:

$$S_W = \frac{1}{n} \sum_{i=1}^{i=n} (S_W(c_i)) \tag{11}$$

where c_i is the ith cluster, $i \in \{1, 2, \cdots, n-1, n\}$, N_i is the total tracklets numbers of cluster c_i, and S_{pq} is the final spatio-temporal similarity in affinity matrix. In the same way, the Between-cluster Difference is:

$$S_B = \frac{1}{n} \sum_{i,j=1, i\neq j}^{n} \left(\mu_{c_i} - \mu_{c_j}\right) \tag{12}$$

Then we can obtain the cluster number measurement Φ from Eqs. (11) and (12) as:

$$\Phi = \frac{S_B}{S_W} \tag{13}$$

Therefore, we can draw a conclusion about when to stop the clustering process and how to find the optimal number of groups automatically by finding the maximum Φ. It is reasonable that the maximum Φ means the Maximum Between-cluster Difference and Minimum Within-cluster Difference, which illustrates the preferable clustering results of specific crowd scene based on their intra-group tightness and inter-group difference, and also provides a more principled way to determine when to stop the clustering process instead of manually setting a threshold.

According to the hierarchical clustering process, we gather some low-level clusters firstly, then we extract the representative tracklets from each cluster, and these representative tracklets will be our input data in a higher level procedure. Algorithm 1 gives

Algorithm 1 Extract representative tracklets.

```
N is the number of tracklets which is similar to tracklet i in
  certain cluster based on the similarity threshold S_threshold, T(·)
  is the lifespan set of certain cluster, and L is the lifespan set
  with descending order.
for cluster k in cluster set do
    L ← T(k)    ▷ sort the lifespan of cluster (k) and reserve for L(k)
    for i in cluster (k) do
        calculate N of tracklets i in L
        if N of L(i) > σ_threshold then
            repre_trks ← L(i)
        end if
    end for
end for
return repre_trks
```

the process to extract representative tracklets, where $S_{threshold} = 3 \times S_{max}/4$, S_{max} is the maximum similarity value in current clustering, and $\sigma_{threshold} = 5$ empirically. In order to make sure the representative tracklets have longer life-span instead of short fragments, we sort the life-spans of tracklets and give preference to a longer one of given cluster. Then we apply the unified similarity measurement to measure the similarity among these tracklets or representative tracklets. As a result, we will obtain some high-level clusters with similarity attributes, and these clusters represent the high-level information of the crowd (i.e., scene structure and semantic regions). Therefore, we can find the motion path by modeling these high-level clusters in a long temporal duration.

The whole process is shown from the left to right in Fig. 7. In the beginning, we cluster tracklets based on the three priors and the unified similarity measurement elaborated before. Thus, the co-occurring tracklets and short term tracklets will be clustered firstly in small spatio-temporal scale. While in higher layers of the hierarchical clustering, some small groups will be merged into a bigger one, then the representative tracklets are extracted, and these representative tracklets are used to further learning of the semantic regions (i.e., the moving path).

3.4. Temporal window partition

Given a long temporal crowd scene, it is inapplicable to cluster all the tracklets simultaneously when it comes to the dynamic change and random motion. Furthermore, there are two aspects to illustrate the drawbacks of clustering all the tracklets simultaneously for a long temporal video sequence: Firstly, there is no need to calculate the similarity of pairs of tracklets which are far from each other in temporal space. According Eqs. (7) and (8), tracklets with large temporal gaps will be assigned small C, which leads to small temporal weight. As a result, there is a smaller final similarity according to Fig. 5. Secondly, the crowd is a dynamic collection, and the velocity and motion trails among individuals are different. Thus, it means the tracklets of the same pedestrian may belong to different groups in different temporal intervals, so it is unreasonable to assign a tracklets into a constant cluster in the whole period of time.

In order to cluster tracklets more reasonable and closer to the actual truth, we apply a temporal window ω to divide the time axis into several temporal intervals. Therefore, we will calculate the final similarity and cluster these tracklets based on the affinity matrix illustrated in the previous sections within each temporal partition. In this way, we just measure the similarity of the tracklets and cluster them within certain temporal window, and left out the tracklets with large temporal gaps. On the other hand, the temporal window partition operation satisfies the actual situation that the same tracklets may belong to different groups in different periods of temporal, which is more authentic in a dynamic and random crowd scene.

4. Experimental results

We evaluate our method on the large-scale CUHK crowd dataset [19], the Collective Motion Database [52] as well as our dataset, where some tasks, i.e., crowd group detection, collectiveness measuring, path modeling and anomaly detection, are performed.

4.1. Crowd database

CUHK crowd dataset: This dataset is constructed by CUHK in [19]. 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. Although the video clips have various length, the quantitative evaluation proceed on 300 manually annotated videos clips (only the first 30 frames from 300 videos clips are manually annotated with ground truth) [19].

CASIA crowd dataset: Different from all of the existing crowd datasets which contain too short crowd video sequences and incomplete labeled tracklets or groups, our new CASIA crowd dataset

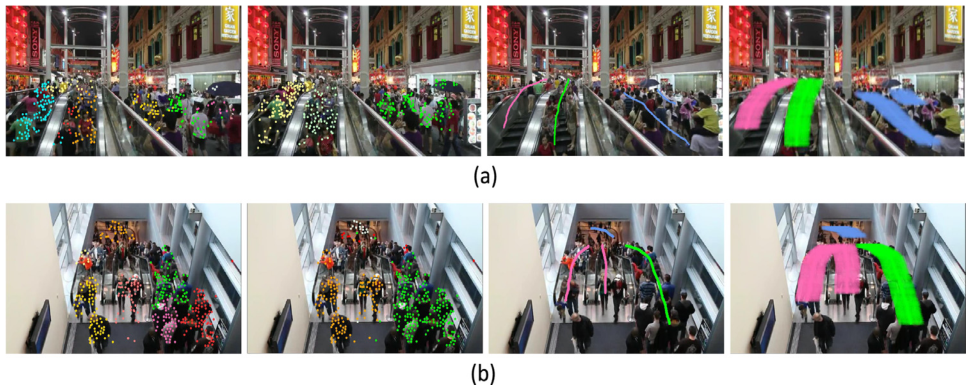


Fig. 7. The first two columns are the representative frames of clustering results from bottom-top cluster process. The third column is the extracted representative tracklets based on the clustering results in higher layer. The last column is the learnt moving path based on the clustering results and representative tracklets.

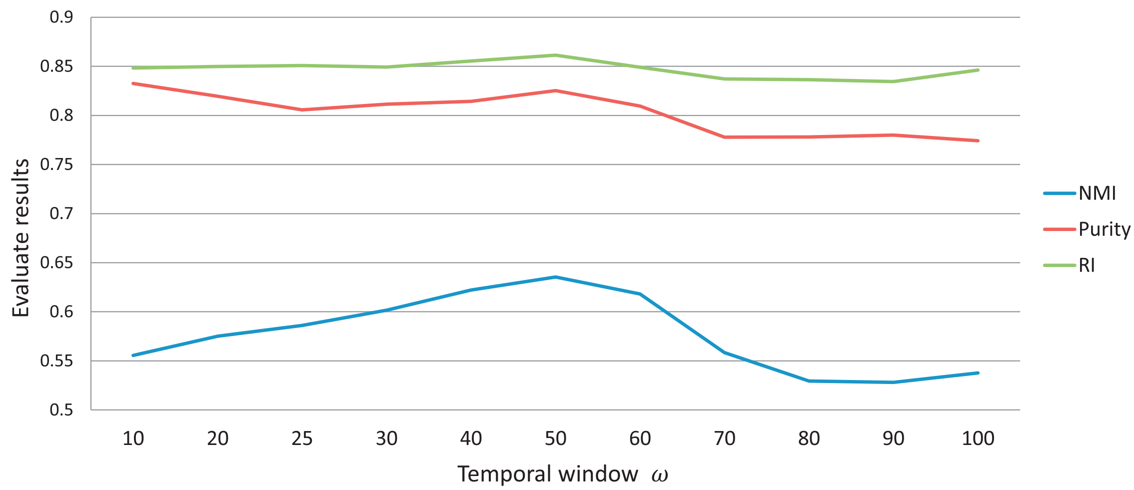


Fig. 8. The comparison results within different temporal windows, where $\omega = 50$ achieves the best performance.

contains 9 long crowd video sequences with various crowd densities from three real scenes. The length of crowd video sequences various from 16s to 70s and totally includes 6980 frames and 15,421 tracklets. The ground truth of group detection and motion paths in the whole temporal duration are manually annotated and checked by multiple annotators. In order to maintain the crowd dynamic that one tracklet may belong to different groups in different temporal durations, we update the annotated ground truth in each 5 frames, and the tracklets not belong to any groups are annotated as outliers.

Collective Motion Database: The Collective Motion Database consists of 413 video clips from 62 crowded scenes and each video clip consists of 100 frames. To get the ground truth, 10 subjects are invited to rate the level of collective motions in a video from three options: low, medium, and high, and score the video as 0, 1, 2 respectively [52].

4.2. Group detection

Before conducting other experiments, we will decide the optimal temporal window firstly. In addition, group detection issue is also considered as clustering process, so we apply three useful evaluation criterions in clustering field. They are Purity [53], Rand Index (RI) [54], and Normalized Mutual Information (NMI) [39].

We present the clustering results under different temporal windows on our CASIA crowd dataset as shown in Fig. 8, where the results illustrate that the temporal window $\omega = 50$ can obtain the best performance. Therefore, we measure the unified similarity of pairs of tracklets and cluster groups within each 50 frames duration in our following experiments. We conduct the temporal window parameter choosing on the CASIA crowd dataset instead of CUHK crowd dataset because our method can learn both the short temporal and long temporal crowd behaviors, while the CUHK crowd dataset only contains short temporal sequences and labels certain frames which is inappropriate to learn long temporal crowd information.

In order to evaluate our method qualitatively, we compare our method with other two outstanding algorithms: coherent filtering (CF) [12], and collective transition priors (CT) [19] on the CUHK crowd dataset and the CASIA crowd dataset. Figs. 9 and 10 show the results of crowd group detection of representative frames from different crowd scenes on these two crowd datasets. The first column is the result of CF method which can detect most of tracklets near by the front, and it prefers to cluster tracklets into many small groups because of the strict restriction of velocity correlation. The second column is the result of CT method and it can detect groups fitly, but it also loses the remote tracklets and many groups. However, from the visual comparison of our method and the ground truth, we can see that our method detects the tracklets and groups well, and possesses more comprehensive detection and clustering results. Furthermore, we lose less tracklets, and discover more small groups and details, which demonstrates that our method is more effective and practical.

In quantitative evaluations, we compare our results with five outstanding algorithms: mixture of dynamic texture (DTM) [30], hierarchical clustering (HC) [1], coherent filtering (CF) [12], collec-

Fig. 9. The results of crowd group detection of representative frames on CUHK crowd dataset, the qualitative comparisons with other outstanding methods illustrates that our method possesses more comprehensive detection and clustering results. The red color indicates outliers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tive transition priors (CT) [19] and our previous work (STS) [16] on the CUHK crowd dataset and the CASIA crowd dataset. For the CUHK crowd dataset, the result is shown as Fig. 11. The NMI of STS is lower than CT because it clusters tracklets on the whole tracklet-level in different hierarchy and measures the similarity of tracklets in the whole temporal length directly, so it is more sensitive by fickle tracklets. However, the temporal window ω can make our method more robust to outliers and ensure the results more effective and practical. Therefore, our method can find more connotative groups and deep-seated relationships between tracklets.

For the CASIA crowd dataset, the NMI value and RI value have a great improvement than CF and CT methods as shown in Fig. 12. This is reasonable that our method is good at the long temporal duration situation, especially there is spatio-temporal gap of pairs of tracklets, while the CF and CT methods cluster tracklets frame by frame and can only measure the tracklets co-occurring. When it comes to compute purity, each cluster is assigned to the class which is most frequent in the cluster, so there are two situations to achieve high purity value. One is that we assign the samples of each cluster to the real class correctly, the other one is when the number of clusters is large - in particular, purity is 1 if each document gets its own cluster. Thus, the purity value of CF method in Fig. 12 is the largest because the CF method tends to obtain too many small clusters as shown in Figs. 9 and 10. However, our method has a high purity with larger crowd groups, which means we cluster the samples to their real class correctly. Therefore, from all views, our method achieves better comprehensive results qualitatively and quantitatively.

From Figs. 13 and 14 we can find that the cluster number obtained from our method are closer to the ground truth. The cluster number of CF method and the CT method are decided by setting a predefined threshold. Furthermore, CT method apply the Collective Transition prior to refine the groups learnt by CF, so the final cluster number less than CF and more precise. However, our method does not require a predefined number of clusters, which can find the optimal cluster number according to the tightness of groups in crowd scenes.

4.3. Evaluation of collectiveness

Collectiveness, which indicates the degree of individual acting as a union in collective motion, is a fundamental and universal measurement for various crowd systems [44]. Similar to the Measuring Crowd Collectiveness (MCC) [55] and Measuring Collectiveness via Refined Topological Similarity (RTS) [52], where they evaluate the collectiveness by measuring the similarity of pairs of tracked feature points (i.e., the tracklets at a specific moment) based on their velocity correlation and spatial information, our method also measures the similarity of pairs of tracklets based on their interrelation. Therefore, we evaluate the collectiveness in the same way of Li et al. [52] based on the similarity we designed to demonstrate the effectiveness of our method. As shown in Fig. 15, we compare the averaged precision, recall and F-measure and precision-recall curves with MCC, CT and RTS. We can see that our method has better discriminative capability than MCC, CT and RTS.

MCC and CT are the baselines in this task. However, they all neglect the temporal information. RTS incorporate the temporal infor-

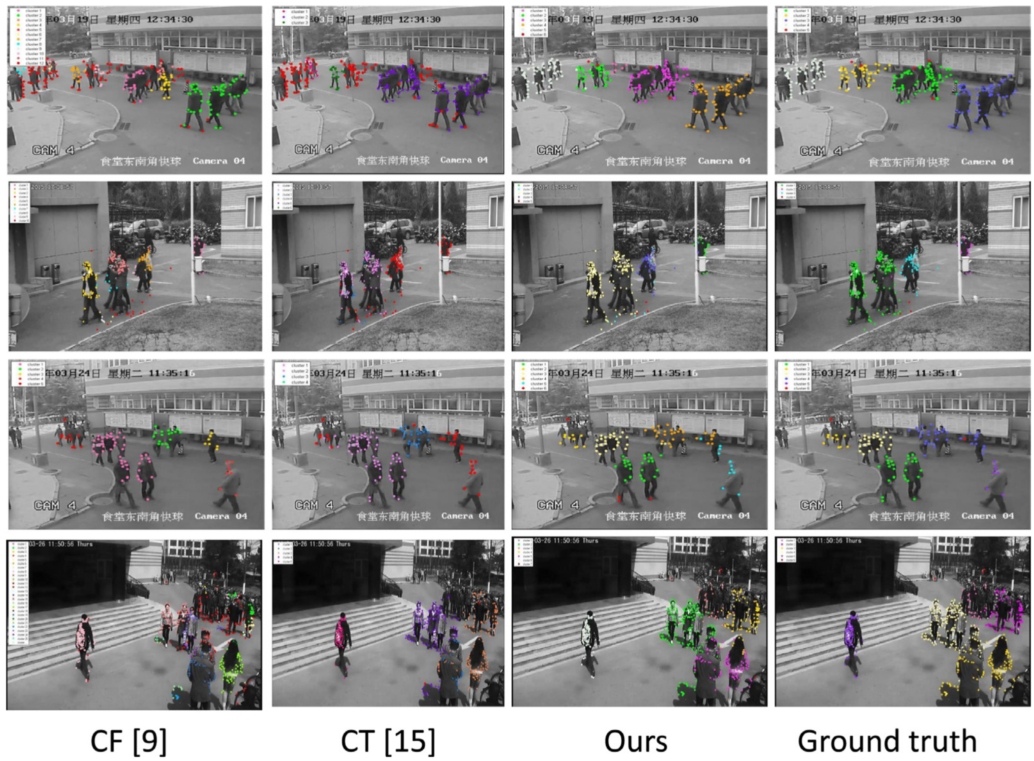
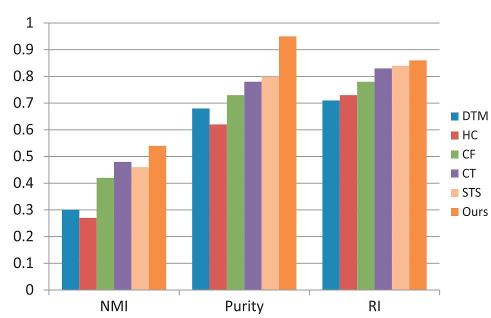


Fig. 10. The results of crowd group detection of representative frames on CASIA crowd dataset, the qualitative comparisons with other outstanding methods illustrates that our method can find more details and achieve the best result. The red dots indicate outliers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Methods	NMI	Purity	RI
DTM [28]	0.30	0.68	0.71
HC [1]	0.27	0.62	0.73
CF [11]	0.42	0.73	0.78
CT [18]	0.48	0.78	0.83
STS [15]	0.46	0.80	0.84
Ours	**0.54**	**0.95**	**0.86**

(a)

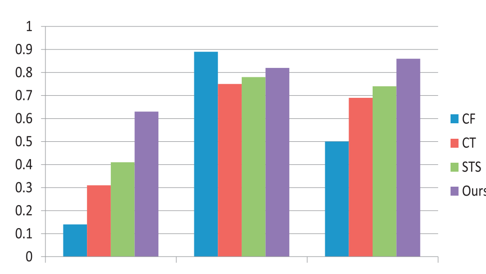


(b)

Fig. 11. The quantitative comparison on CUHK crowd dataset. (a) The result of our method with other three state-of-the-art algorithms. (b) The visual results comparison.

Methods	NMI	Purity	RI
CF [11]	0.14	**0.89**	0.50
CT [18]	0.31	0.75	0.69
STS [15]	0.41	0.78	0.74
Ours	**0.63**	0.82	**0.86**

(a)



(b)

Fig. 12. The quantitative comparison on CASIA crowd dataset. (a) The result of our method with other three state-of-the-art algorithms. (b) The visual results comparison.

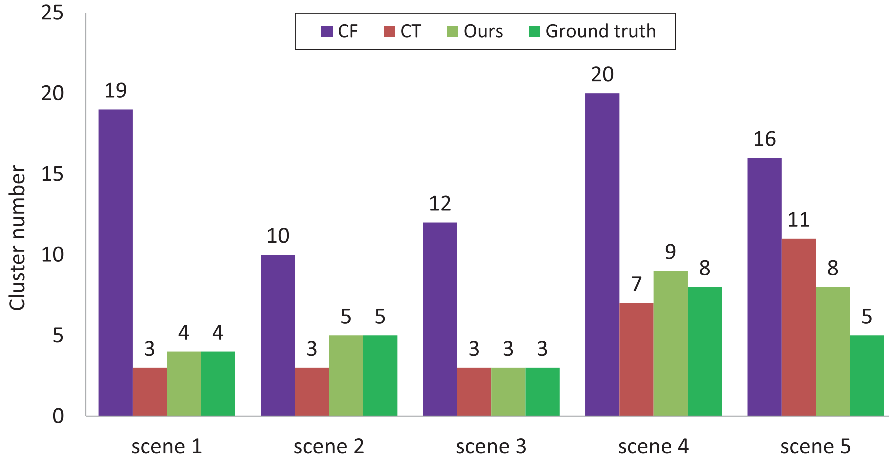


Fig. 13. Cluster number comparison with other methods on CUHK crowd dataset. The scenes are showed in Fig. 9 from the first raw to the last.

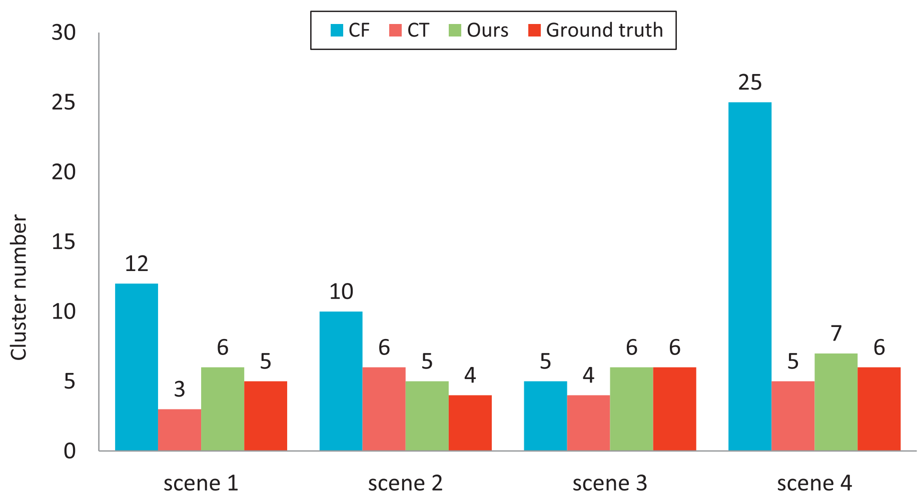


Fig. 14. Cluster number comparison with other methods on CASIA crowd dataset. The scenes are showed in Fig. 10 from the first raw to the last.

mation into similarity definition, and detect the unsteady points. Nevertheless, this method just considers two adjacent frames, which is not enough to find deeper temporal connections for a dynamic system. All of the three methods share the same shortcoming in that they are between-frame association based, so they cannot dig the long-term temporal relationships of tracked points. However, the similarity we designed can find their long-term temporal relationships within specific temporal window. As a result, the proposed similarity based on spatio-temporal interrelation outperforms other methods.

4.4. Path modeling and anomaly detection

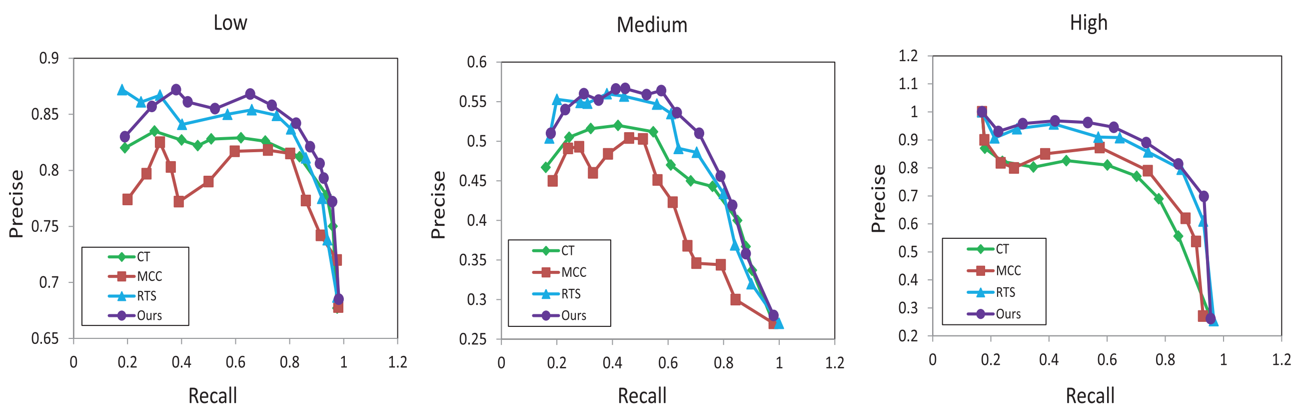
As shown in Fig. 16 (a) from left to right, which is the merging process of the bottom-up hierarchical clustering, and the different colored areas denote different crowd groups. Then we will find the final groups when the clustering process is stopped according to the intra-group tightness and inter-group difference. After the groups are detected based on the unified similarity measurement, the representative tracklets will be extracted from these crowd groups and used to further learning of the long-term scene paths.

It may be a moving path if most of the crowd groups move on it during a long temporal duration. Therefore, we will count the moving crowd groups with certain main directions as shown in Fig. 16 (b). Then, we find that there are four different main directions and the number of crowd groups in each direction in the whole temporal duration. Therefore, we can obtain the main moving paths by setting a number threshold. For example, there are two paths in Fig. 16 under the number threshold of 4, and there are other learnt paths from six different scenes in the CUHK crowd dataset and CASIA crowd dataset as shown in Fig. 17. Note that there are overlaps within paths because the crowd groups may crossover each other or move in the same spatial place in different temporal durations.

Anomaly detection is also a significant issue on intelligent surveillance. Our method can learn the crowd paths in different crowd scenes based on the unified similarity measurement presented in previous section. After the scene structure is learnt, we can use this model to detect abnormal tracklets, and the tracklets which are obviously differ from the learnt crowd paths and own a long temporal duration are considered as abnormal tracklets. In the first row of Fig. 18, three men riding bicycle going in the direction forbidden by traffic regulations are also detected as abnormal actions by our model. In the second row, most of pedestrians walk on the direction of main crowd path from the bottom to up, and meanwhile a man go in an opposite direction to the crowd flow. Our model can automatically cluster and label his tracklets and treat this is an anomalous activity. The right part of Fig. 18 is the detected abnormal tracklets number and their ground truth, and this result illustrates that our method can detect the abnormal tracklets effectively.

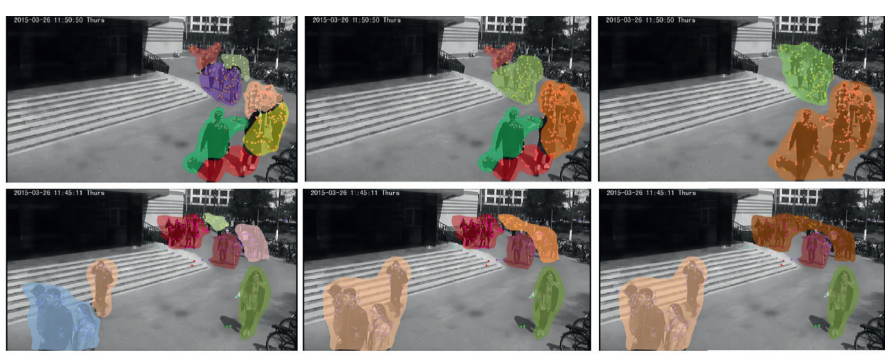
	Low			Medium			High		
Methods	Precise	Recall	F-measure	Precise	Recall	F-measure	Precise	Recall	F-measure
CT	0.80	0.59	0.67	0.45	0.57	0.49	0.69	0.46	0.56
MCC	0.75	0.61	0.63	0.38	0.56	0.57	0.73	0.51	0.59
RTS	**0.83**	0.66	0.72	0.47	0.61	0.51	0.81	**0.58**	0.67
Ours	**0.83**	**0.67**	**0.74**	**0.50**	**0.62**	**0.55**	**0.84**	**0.58**	**0.68**

(a)

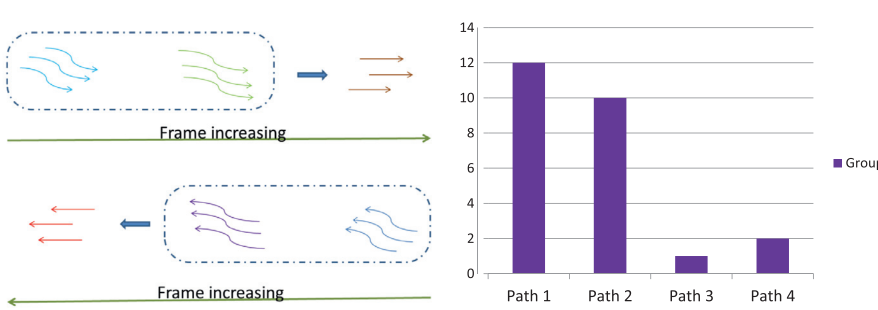


(b)

Fig. 15. Comparison of the averaged precise and recall value, and the final F-measure on the Collective Motion Database [55].(a) Results in the low, medium, and high level collective scenes. (b) The corresponding precision-recall curves.



(a)



(b)

Fig. 16. (a). Both two rows are the bottom-up hierarchical clustering process, and the final groups in the third column are decided automatically and used to further learning the moving path. (b) illustrates the forming of path by analyzing the moving groups during different main directions in the whole temporal duration.

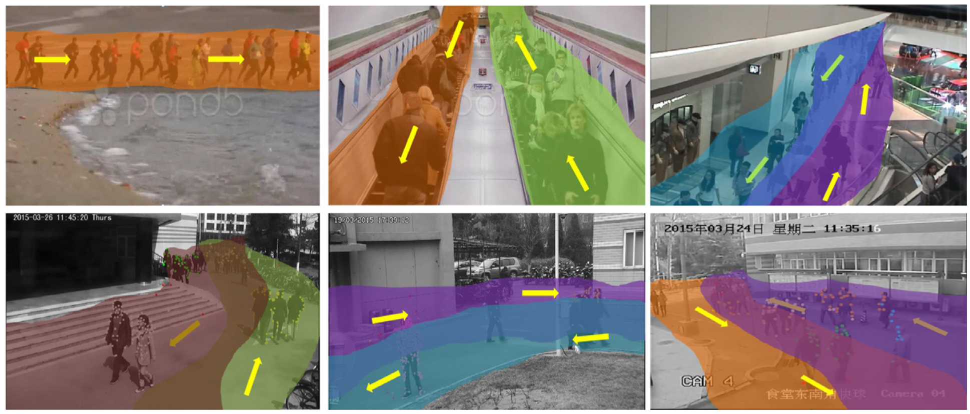


Fig. 17. The learnt moving paths based on our method. The first row is from the CUHK crowd dataset and the second from the CASIA crowd dataset.

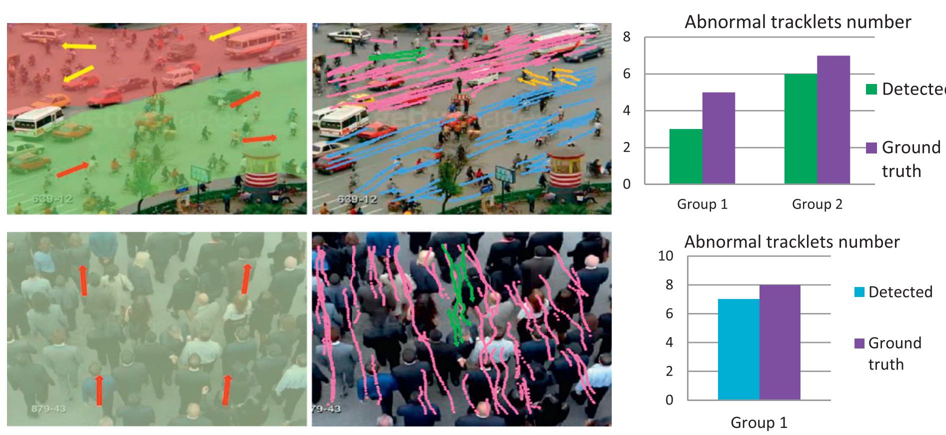
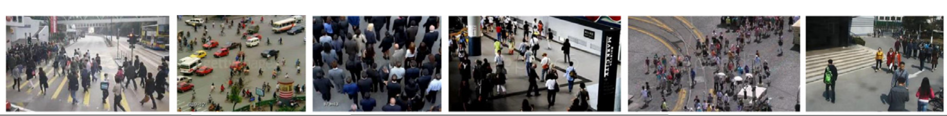


Fig. 18. Some abnormal crowd actions are detected by the learnt crowd path, and the whole anomalous tracklets and their number are represented in dynamic scenes. The arrows indicate the direction of the main crowd path.



	S1	S2	S3	FS4	FS5	S6	Mean
Dt-trks	11vs13	11vs9	7vs6	8vs5	16vs12	14vs13	- -
TP-trks	8vs8	9vs6	7vs5	5vs4	12vs10	14vs12	- -
GT	9	12	8	6	14	15	- -
Precise	**0.73**vs0.62	**0.82**vs0.67	**1.00**vs0.83	0.63vs**0.80**	0.75vs**0.83**	**1.00**vs0.92	**0.82**vs0.78
Recall	**0.89**vs**0.89**	**0.75**vs0.50	**0.88**vs0.63	**0.83**vs0.67	**0.86**vs0.71	**0.93**vs0.80	**0.86**vs0.70

Fig. 19. The quantitative evaluation (Our method vs CF method) for anomaly detection in six crowd scenes selected from bath CUHK dataset and CASIA dataset. Dt-trks means detected tracklets, TP-trks (True Positive trackles) denotes the real anomaly tracklets detected in Dt-trks, and GT is the ground truth.

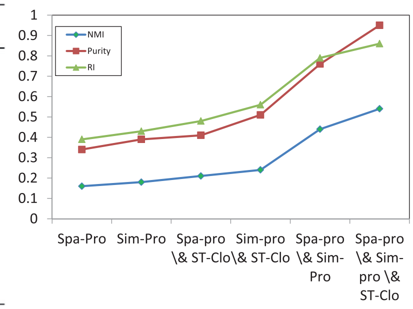
In order to give a quantitative evaluation for anomaly detection, we manually annotate the anomaly tracklets (totally, 64 anomaly tracklets are labeled) in six crowd scenes selected from both CUHK dataset and CASIA dataset. In addition, we evaluate the validity of anomaly detection by precise and recall, and make a comparison with CF method. As shown in Fig. 19, the results demonstrate that our method outperform CF in this task.

4.5. Further analysis of proposed priors

Our similarity measurement is based on the Spatial Proximity Prior (Spa-Pro), Similar Properties Prior (Sim-Pro) and Spatio-temporal Closure Prior (ST-Clo). Therefore, we further analyze how the proposed three priors compensate with each other. Figs. 20 and 21 are the group detection results of each prior and their interactive compensation on the CUHK crowd dataset and the CASIA crowd dataset.

From Figs. 20 and 21 we can find that only applying one of the Spatial Proximity Prior or Similar Properties Prior to measure the similarity leads to poor results on both datasets. The performance can be improved after weighted the similarity with temporal information derived from the Spatio-temporal Closure prior. However, the contribution of the temporal weight is limited, this is reasonable because the Spatial Proximity Prior decides the spatial neighbor relations of tracklets, which clusters tracklets close to each other to be a cluster. While the Similar Properties Prior can separate tracklets with different moving directions, even those tracklets are clustered together based on their spatial neighbor relations. Thus, only one of Spatial Proximity Prior or Similar Properties Prior

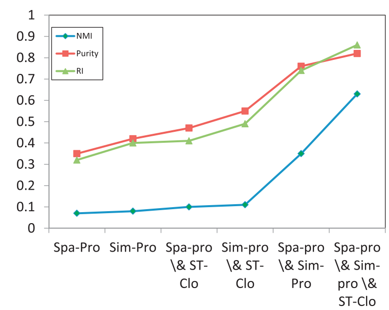
Methods	NMI	Purity	RI
Spa-Pro	0.16	0.34	0.39
Sim-Pro	0.18	0.39	0.43
Spa-Pro & ST-Clo	0.21	0.41	0.48
Sim-Pro & ST-Clo	0.24	0.51	0.56
Spa-Pro & Sim-Pro	0.44	0.76	0.79
Spa-Pro & Sim-Pro & ST-Clo	**0.54**	**0.95**	**0.86**



(a) (b)

Fig. 20. (a).The results of the proposed priors compensate with each other on the CUHK crowd dataset. (b). The perspicuous comparison of all situations. Spa-Pro means Spatial Proximity Prior, Sim-Pro means Similar Properties Prior and ST-Clo means Spatio-temporal Clpsure Prior.

Methods	NMI	Purity	RI
Spa-Pro	0.07	0.35	0.32
Sim-Pro	0.08	0.42	0.40
Spa-Pro & ST-Clo	0.10	0.47	0.41
Sim-Pro & ST-Clo	0.11	0.55	0.49
Spa-Pro & Sim-Pro	0.35	0.77	0.74
Spa-Pro & Sim-Pro & ST-Clo	**0.63**	**0.82**	**0.86**



(a) (b)

Fig. 21. (a).The results of the proposed priors compensate with each other on the CASIA crowd dataset. (b). The perspicuous comparison of all situations. Spa-Pro means Spatial Proximity Prior, Sim-Pro means Similar Properties Prior and ST-Clo means Spatio-temporal Clpsure Prior.

weighted with temporal information will lose either spatial affinity or velocity correlation information. Therefore, a greater improvement can be achieved when we combine the Spatial Proximity Prior (i.e., spatial affinity) and Similar Properties Prior (i.e., velocity correlation) to measure the similarity, and the performance has been comparable with CF and CT methods as shown in Figs. 11 and 12. Finally, our method achieves the best result as the combination of all three priors which illustrates that the temporal weight is very helpful to measure the crowd groups.

5. Conclusions

In this paper, we propose three priors based on the Gestalt law to measure the similarity of tracklets with spatial affinity and temporal relevance, which is very useful and practical to detect crowd groups and learn long temporal representative tracklets. Then, we design a unified similarity measurement to cluster tracklets in multiple spatio-temporal scales hierarchically within each temporal window, and they are well applied to group detection, collectiveness measuring, path modeling and abnormal tracklets detection. The proposed algorithm with temporal window achieves the state-of-the-art results in the CUHK crowd dataset, Collective Motion Database and our new CASIA crowd dataset.

In the future work, we will find out more crowd attributes, such as the crowd coherence, to analyze crowd motion in fine grained scale. Meanwhile, we will also explore more reliable priors to make it more robust and practical to noise such as drifting tracklets.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (Grant No. 61473290, Grant No. 61673375 and Grant No. 61403383), the National Key Research and Development Program of China (2016YFB1001004, 2016YFB1001005), 863 Program (Grant No. 2015AA042307), and the Projects of Chinese Academy of Science (Grant No. QYZDB-SSW-JSC006, Grant No. 173211KYSB20160008).

References

[1] W. Ge, T. Collins, R. Ruback, Vision-based analysis of small groups in pedestrian crowds, Pattern Anal. Mach. Intell. (2012) 1003–1016.
[2] T. Zhang, H. Lu, S. Li, Learing semantic scene models by object classification and trajectory clustering, in: CVPR, 2009, pp. 1940–1947.
[3] X. Wang, T. Kinh, G. Eric, Learning semantic scene models by trajectory analysis, in: ECCV, 2006, pp. 110–123.
[4] S. Wu, O. Oreifej, M. Shah, Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories, in: CVPR, 2011, pp. 1419–1426.
[5] B. Zhou, X. Wang, X. Tang, Random field topic model for semantic region analysis in crowded scenes from tracklets, in: CVPR, 2011, pp. 3441–3448.
[6] M. Rodriguez, S. AliAli, T. Kanade, Tracking in unstructured crowded scenes, in: ICCV, 2009, pp. 1389–1396.
[7] A. Dehghan, H. Idrees, A. Zamir, M. Shah, Automatic detection and tracking of pedestrians in videos with various crowd densities, in: Pedestrian and Evacuation Dynamics, 2014, pp. 3–19.
[8] Z. Wu, R.J. Radke, Improving counterflow detection in dense crowds with scene features, Pattern Recognit. Lett. 44 (2014) 152–160.
[9] L. Cao, X. Zhang, W. Ren, K. Huang, Large scale crowd analysis based on convolutional neural network, Pattern Recognit. 48 (10) (2015) 3016–3024.
[10] D. Ryan, S. Denman, C. Fookes, S. Sridharan, Scene invariant multi camera crowd counting, Pattern Recognit. Lett. 44 (2014) 98–112.
[11] K. Chen, C.C. Loy, S. Gong, T. Xiang, Feature mining for localised crowd counting., in: BMVC, 1, 2012, p. 3.
[12] B. Zhou, X. Tang, X. Wang, Coherent filtering: detecting coherent motions from crowd clutters, in: ECCV, 2012, pp. 857–871.
[13] A. Fagette, N. Courty, D. Racoceanu, J.-Y. Dufour, Unsupervised dense crowd detection by multiscale texture analysis, Pattern Recognit. Lett. 44 (2014) 126–133.
[14] H. Chaiklin, The myth of the madding crowd, Nerv. Mental Dis. (1963) 342–343.

[15] M. Mehdi, N. Perozo, S. Garnier, D. Helbing, G. Theraulaz, The walking behaviour of pedestrian social groups and its impact on crowd dynamics, (2010) 5, (4) e10047.
[16] W. Zhao, Z. Zhang, K. Huang, Joint crowd detection and semantic scene modeling using a gestalt laws-based similarity, in: ICIP, 2016, pp. 2381–8549.
[17] D. Todorovic, Gestalt principles, Scholarpedia 3 (12) (2008) 5345 http://www.scholarpedia.org/article/Gestalt_principles.
[18] J. Banerjee, Gestalt theory of perception, in: Encyclopaedic Dictionary of Psychological Terms, 1994, pp. 107–109. ISBN 978-81-85880-28-0
[19] J. Shao, C. Chen, X. Wang, Scene-independent group profiling in crowd, in: CVPR, 2014, pp. 2227–2234.
[20] C. Loy, T. Xiang, S. Gong, Detecting and discriminating behavioural anomalies, Pattern Recognit. (2011) 117–132.
[21] S. Pellegrini, A. Ess, K. Schindler, L. Gool, Youl never walk alone: modeling social behavior for multi-target tracking, in: ICCV, 2009, pp. 261–268.
[22] W. Choi, S. Savarese, A unified framework for multi-target tracking and collective activity recognition, in: ECCV, 2012, pp. 215–230.
[23] T. Brox, J. Malik, Object segmentation by long term analysis of point trajectories, in: ECCV, 2010, pp. 282–295.
[24] R. Chaker, Z. Al Aghbari, I.N. Junejo, Social network model for crowd anomaly detection and localization, Pattern Recognit. 61 (2017) 266–281.
[25] Z. Zhang, K. Huang, T. Tan, P. Yang, J. Li, Red-sfa: relation discovery based slow feature analysis for trajectory clustering, in: CVPR, 2016, pp. 752–760.
[26] S. Wu, E. Moore, M. Shah, Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes, in: CVPR, 2010, pp. 2054–2060.
[27] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: CVPR, 2009, pp. 935–942.
[28] D. Kuettel, M. Breitenstein, L. Gool, V. Ferrari, What going on? discovering spatio-temporal dependencies in dynamic scenes, in: CVPR, 2010, pp. 1951–1958.
[29] S. Ali, M. Shah, A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis, in: CVPR, 2009, pp. 935–942.
[30] A. Chan, N. Vasconcelos, Modeling, clustering, and segmenting video with mixtures of dynamic textures, Pattern Anal. Mach. Intell. (2008) 909–926.
[31] J. Fernyhough, G. Anthony, C. David, Generation of semantic regions from image sequences, in: ECCV, 1996, pp. 475–484.
[32] D. Makris, T. Ellis, Path detection in video surveillance, Image Vision Comput. 20 (2002) 859–903.
[33] X. Wang, X. Ma, W. Grimson, Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models, in: PAMI, 2009, pp. 39–555.
[34] Z. Zhang, K. Huang, T. Tan, L. Wang, Trajectory series analysis based event rule induction for visual surveillance, in: CVPR, 2007, pp. 1–8.
[35] Z. Zhang, K. Huang, T. Tan, Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes, in: ICPR, 2006, pp. 1135–1138.
[36] X. Wang, K. Ma, G. Ng, E. Grimson, Trajectory analysis and semantic region modeling using a nonparamet-ric bayesian model, in: CVPR, 2008, pp. 1–8.
[37] I. Junejo, H. Foroosh, Trajectory rectification and path modeling for video surveillance, in: ICCV, 2007, pp. 1–7.
[38] Z. Fu, W. Hu, T. Tan, Similarity based vehicle trajectory clustering and anomaly detection, in: ICIP, 2, 2005, p. 602.
[39] K. Komodakis, in: Principles of Gestalt Psychology, Hartcourt Brace Jovanovich, 1935, pp. 39–555.
[40] P. Ochs, J. Malik, T. Brox, Segmentation of moving objects by long term video analysis, in: PAMI, 36, 2014, pp. 1187–1200.
[41] K. Fragkiadaki, J. Zhang, G. Shi, Video segmentation by tracing discontinuities in a trajectory embedding, in: CVPR, 2012, pp. 1846–1853.
[42] Y. Yuan, J. Wan, Q. Wang, Congested scene classification via efficient unsupervised feature learning and density estimation, Pattern Recognit. 56 (2016) 159–169.
[43] W. Liu, R.W. Lau, D. Manocha, Robust individual and holistic features for crowd scene classification, Pattern Recognit. 58 (2016) 110–120.
[44] B. Zhou, X. Tang, X. Wang, Measuring crowd collectiveness, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3049–3056.
[45] B. Cancela, M. Ortega, M.G. Penedo, J. Novo, N. Barreira, On the use of a minimal path approach for target trajectory analysis, Pattern Recognit. 46 (7) (2013) 2015–2027.
[46] F. Solera, S. Calderara, R. Cucchiara, Socially constrained structural learning for groups detection in crowd, IEEE Trans. Pattern Anal. Mach. Intell. 38 (5) (2016) 995–1008.
[47] M. Manfredi, R. Vezzani, S. Calderara, R. Cucchiara, Detection of static groups and crowds gathered in open spaces by texture classification, Pattern Recognit. Lett. 44 (2014) 39–48.
[48] B. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proceedings of Imaging Understanding Workshop, 81, 1981, pp. 674–679.
[49] H. Stevenson, Emergence: The Gestalt Approach to Change, Unleashing Executive and Orzanizational Potential, 2012.
[50] M. Soegaard, Gestalt Principles of Form Perception, Interaction Design, 2012.
[51] P. Dirac, The lorentz transformation and absolute time, Physica 19 (1953) 888–896, doi:10.1016/S0031-8914(53)80099-6. 1-2.
[52] X. Li, M. Chen, Q. Wang, Measuring collectiveness via refined topological similarity, ACM Trans. Multimedia Comput.Commun. Appl. (TOMM) 12 (2) (2016) 34.
[53] C. Aggarwal, A human-computer interactive method for projected clustering, in: PAMI, 2004, pp. 448–460.
[54] D. Todorovic, Objective criteria for the evaluation of clustering methods, Am. Stat.Assoc. 66 (336) (1971) 846–850.
[55] B. Zhou, X. Tang, H. Zhang, X. Wang, Measuring crowd collectiveness, Pattern Anal. Mach. Intell. (2014) 1586–1599.

Weiqi Zhao, born in 1990, received the B.S. degree in electrical engineering and automation from Beijing Jiaotong University, Beijing, China, in 2013, He is currently pursuing the M.S. degree in pattern recognition and intelligent systems with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. His research interests include Compute Vision, Pattern Recognition and Machine Learning.

Zhang Zhang, born in 1980, received the B.S. degree in computer science and technology from Hebei University of Technology, Tianjin, China, in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China in 2008. Currently, he is an associate professor at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include activity recognition, video surveillance, and time series analysis. He has published a number of papers at top venues including the IEEE Transactions on Pattern Analysis and Machine Intelligence, CVPR, and ECCV. He is a member of the IEEE.

Kaiqi Huang, born in 1977, received the M.S. degree in electrical engineering from Nanjing University of Science and Technology, Nanjing, China, and the Ph.D. degree in signal and information processing from Southeast University, Nanjing. He is currently a Professor in Institute of Automation, Chinese Academy of Science (CASIA). He has published more than 150 papers on TPAMI, TIP, TCSVT, TSMCB, CVIU, Pattern Recognition and ICCV, CVPR, and ECCV. His interests include visual surveillance, image and video analysis, human vision and cognition, computer vision, etc. Dr. Huang is IEEE Senior Member and Program Committee Member of more than 50 international conferences and workshops, he also severed several AEs of journals such as IEEE Systems, Man, and Cybernetics: Systems.