A Diffusion and Clustering-Based Approach for Finding Coherent Motions and Understanding Crowd Scenes

Weiyao Lin, Yang Mi, Weiyue Wang, Jianxin Wu, Jingdong Wang, and Tao Mei

Abstract—This paper addresses the problem of detecting coherent motions in crowd scenes and presents its two applications in crowd scene understanding: semantic region detection and recurrent activity mining. It processes input motion fields (e.g., optical flow fields) and produces a coherent motion field named thermal energy field. The thermal energy field is able to capture both motion correlation among particles and the motion trends of individual particles, which are helpful to discover coherency among them. We further introduce a two-step clustering process to construct stable semantic regions from the extracted time-varying coherent motions. These semantic regions can be used to recognize pre-defined activities in crowd scenes. Finally, we introduce a cluster-and-merge process, which automatically discovers recurrent activities in crowd scenes by clustering and merging the extracted coherent motions. Experiments on various videos demonstrate the effectiveness of our approach.

Index Terms—Coherent motion detection, semantic region construction, recurrent activity mining.

I. INTRODUCTION

COHERENT motions, which represent coherent movements of massive individual particles, are pervasive in natural and social scenarios. Examples include traffic flows and parades of people (cf. Figs 1a and 2a). Since coherent motions can effectively decompose scenes into meaningful semantic parts and facilitate the analysis of complex crowd scenes, they are of increasing importance in crowd-scene understanding and activity recognition [2]–[6].

In this paper, we address the problem of detecting coherent motions in crowd scenes, and subsequently using them to understand input scenes. More specifically, we focus

W. Lin and Y. Mi are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wylin@sjtu.edu.cn; deyangmiyang@163.com).

W. Wang is with the Department of Computer Science, University of Southern California, Los Angeles, CA 90089 USA (e-mail: wweiyue@gmail.com).

J. Wu is with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China (e-mail: wujx2001@nju.edu.cn).

J. Wang and T. Mei are with the Microsoft Research Asia, Beijing 100080, China (e-mail: jingdw@microsoft.com; tmei@microsoft.com).
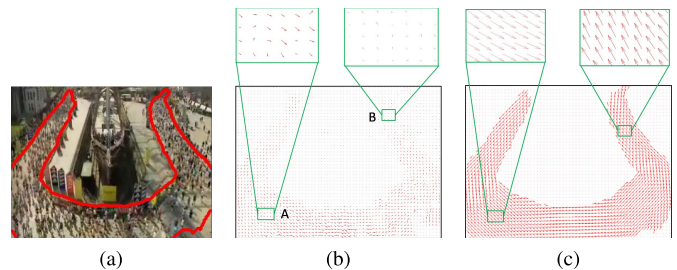
Fig. 1. (a) Example frame of a Marathon video sequence, the red curve is the coherent motion region; (b) Input motion vector field of (a); (c) Coherent motion field from (b) using the proposed approach (Best viewed in color).

on 1) constructing an accurate coherent motion field to find coherent motions, 2) finding stable semantic regions based on the detected coherent motions and using them to recognize pre-defined activities (i.e., activities with labeled training data) in a crowd scene, and 3) automatically mining recurrent activities in a crowd scene based on the detected coherent motions and semantic regions.

First, constructing an accurate coherent motion field is crucial in detecting reliable coherent motions. In Fig. 1, (b) is the input motion field and (c) is the coherent motion field which is constructed from (b) using the proposed approach. In (b), the motion vectors of particles at the beginning of the Marathon queue are far different from those at the end, and there are many inaccurate optical flow vectors. Due to such variations and input errors, it is difficult to achieve satisfying coherent motion detection results directly from (b). However, by transferring (b) into a coherent motion field where the coherent motions among particles are suitably highlighted in (c), coherent motion detection is greatly facilitated. Although many algorithms have been proposed for coherent motion detection [2], [8]–[10], this problem is not yet effectively addressed. *We argue that a good coherent motion field should effectively be able to* 1) *encode motion correlation among particles*, such that particles with high correlations can be grouped into the same coherent region; and, 2) *maintain motion information of individual particles*, such that activities in crowd scenes can be effectively parsed by the extracted coherent motion field. Based on these intuitions, we propose a thermal-diffusion-based approach, which can extract accurate coherent motion fields.

Second, constructing meaningful semantic regions to describe activity patterns in a scene is also essential.
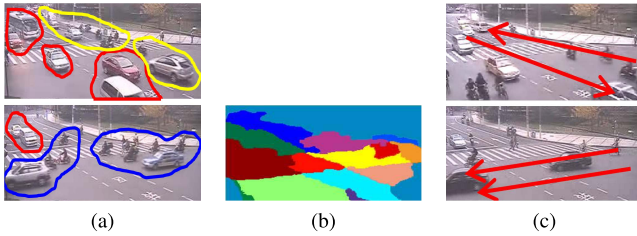
Fig. 2. (a) Example time-varying coherent motions in a scene, where different coherent motions are circled by curves with different color; (b) Constructed semantic regions for the scene in (a); (c) Recurrent activities for the scene in (a), where the arrows represent the major motion flows in each recurrent activity (Best viewed in color).

Coherent motions at different times may vary widely. In Fig. 2a, changing of traffic lights will lead to different coherent motions. Coherent motions alone may not effectively describe the overall semantic patterns in a scene either. Therefore, semantic regions need to be extracted from these time-varying coherent motions to achieve stable and meaningful semantic patterns, as in Fig. 2b. However, most existing works only focus on the detection of coherent motions at some specific time, while the problem of handling time-varying coherent motions is less studied. We proposed a two-step clustering process for this purpose.

Third, mining recurrent activities is another important issue. Many crowd scenes are composed of recurrent activities [11]–[13]. For example, the scene in Fig. 2 is composed of recurrent activities including vertical motion activities and horizontal motion activities, as in Fig. 2c. Automatically mining these recurrent activities is important in understanding scene contents and their dynamics. Although many researches have been done for parsing recurrent activities in low-crowd scenes [14]–[17], this issue is not well addressed in crowd scene scenarios where reliable motion trajectories are unavailable. We proposed a cluster-and-merge process, which can effectively extract recurrent activities in crowd scenes.

Our contributions to crowd scene understanding and activity recognition are summarized as follows.

1) We introduce a coarse-to-fine thermal diffusion process to transfer an input motion field into a thermal energy field (TEF), which is a more accurate coherent motion field. TEF effectively encodes both motion correlation among particles and motion trends of individual particles. To our knowledge, this is the first work that introduces thermal diffusion to detect coherent motions in crowd scenes. We also introduce a triangulation-based scheme to effectively identify coherent motion components from the TEF.

2) We present a two-step clustering scheme to find semantic regions according to the correlations among coherent motions. The found semantic regions can effectively catch activity patterns in a scene. Thus good performance can be achieved when recognizing pre-defined crowd activities based on these semantic regions.

3) We propose a cluster-and-merge process to automatically mine recurrent activities by clustering and merging the coherent motions. The obtained recurrent activities can accurately describe recurrent motion patterns in a crowd scene.

The remainder of this paper is organized as follows. Section II reviews related works. Section III describes the framework of the proposed approach. Sections IV to VI describe the details of our proposed thermal diffusion process, triangulation scheme, two-step clustering scheme, and cluster-and-merge process. Section VII shows the experimental results and Section VIII concludes the paper.

II. RELATED WORKS

Many works [2], [7]–[10], [18]–[22] have been proposed on coherent motion detection. Due to the complex nature of crowd scenes, they are not yet mature for accurate detection of coherent motion fields. Cremers and Soatto [20] and Brox et al. [21] model the intensity variation of optical flow by an objective functional minimization scheme. These methods are only suitable for motions with simple patterns and cannot effectively analyze complex crowd patterns such as the circular flow in Fig. 1a. Other works introduce external spatial-temporal correlation traits to model the motion coherency among particles [8]–[10]. Since these methods model particle correlations in more precise ways, they can achieve more satisfying results. However, most of these methods only consider short-distance particle motion correlation within a local region while neglecting long-distance correlation among distant particles, they have limitations in handling low-density or disconnected coherent motions where the long-distance correlation is essential. Furthermore, without the information from distant particles, these methods are also less effective in identifying coherent motion regions in the case when local coherent motion patterns are close to their neighboring backgrounds. One example of this kind of scenario is showcased in the region B in Fig. 1b.

There are also other works related to motion modeling. One line of related works is advanced optical flow estimation. These methods try to improve the estimation accuracy of the input motion field by including global constraints over particles [23]–[26]. The focus of our approach is different from these methods. We focus on enhancing the correlation among coherent particles to facilitate coherent motion detection. Thus, the motion vectors of coherent particles are enhanced even if their actual motions are small, such as the region B in Figs 1b and 1c. In contrast, advanced optical flow estimation methods focus on estimating the *actual* motion of particles. They are still less capable of creating precise results when applied to coherent motion detection.

The anisotropic diffusion based methods, used in image segmentation, is also related to our work [27]–[29]. Our approach differs from these methods. First, our approach not only embeds the motion correlation among particles, but also suitably maintains the original motion information from the input motion vector field. Comparatively, the anisotropic-diffusion-based methods are more focused on enhancing the correlation among particles while neglecting the particles' original information. As aforementioned, maintaining particle motion information is important in parsing crowd scenes. More importantly, due to the complex nature of crowd scenes,
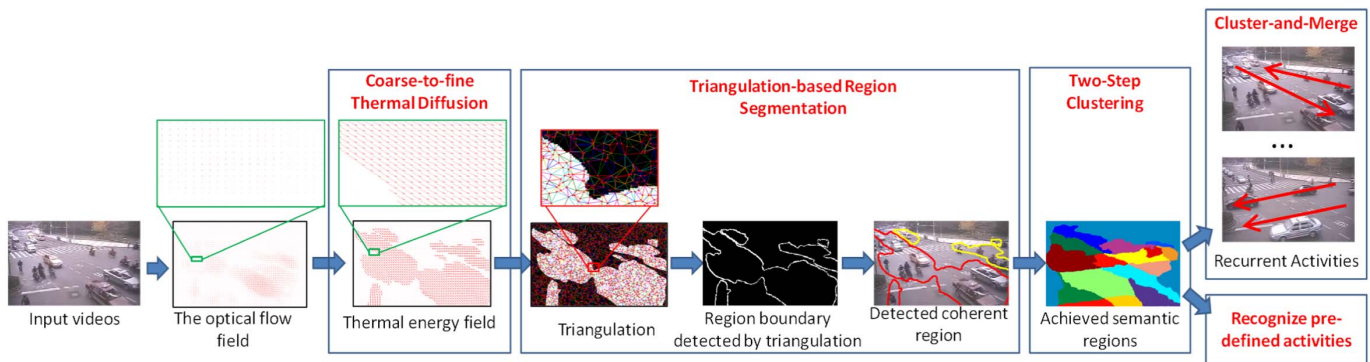
Fig. 3.　The flowchart of the proposed approach (best viewed in color).

many coherent region boundaries are vague, subtle and unrecognizable. Simply applying the anisotropic-diffusion methods cannot identify the ideal boundaries. The proposed thermal diffusion process can achieve more satisfying results by modeling the motion direction, strength, and spatial correlation among particles.

Besides coherent motion detection, it is also important to utilize coherent motions to recognize pre-defined crowd activities. However, most existing coherent motion works only focus on the extraction of coherent motions while the recognition of crowd activities is much less studied. In [7], Ali and Shah detected instability regions in a scene by comparing with its normal coherent motions. However, they assume coherent motions to be stable, while in practice, many coherent motions may vary widely over time, making it difficult to construct stable normal coherent motions. Furthermore, besides the works on coherent motion, there are also other works which directly extract global features from the entire scene to recognize crowd activities [3], [30]. However, since they do not consider the semantic region correlations inside the scene, they have limitations in differentiating subtle differences among activities. Although there are some works [4], [31] which recognize crowd activities by segmenting scenes into semantic regions, our approach differs from them. Our approach finds the semantic regions by first extracting global coherent motion information, while these methods construct semantic regions from the particles' local features. As will be shown later, information from the coherent motions can effectively enhance the correlation among particles, resulting in more meaningful semantic regions to facilitate activity recognition.

Furthermore, pre-defining or labeling crowd activities requires lots of human labors, making it desirable to automatically discover activity patterns in a crowd video without human intervention. In [16], Morris and Trivedi clustered trajectories into groups and modeled the spatio-temporal dynamic patterns of each trajectory group by Hidden Markov Models. Wang et al. [14] and Hu et al. [15] further introduced Dirichlet processes to model the activity patterns of different trajectory groups. However, since these methods extract recurrent activities from motion trajectories, they are not suitable for crowd scene scenarios where reliable trajectories are difficult to achieve. Besides using motion trajectories, other researches tried to find recurrent activities by extracting low-level or short-term motion features. For example, Zhou et al. [13] extracted fragments of trajectories (called tracklets) and utilized a Latent Dirichlet Allocation topic model to infer recurrent activities. Emonet et al. [11] and Jagannadan et al. [12] extracted low-level motion flows as motion descriptors and introduced a Probabilistic Latent Sequential Motif (PLSM) model to achieve recurrent activities. Although these methods can be applied in crowd scenes, they still have limitations in obtaining precise recurrent activity patterns under scenes with complex motions. Our approach differs from the previous methods in that 1) Our approach utilizes coherent motions to discover recurrent activities. Since coherent motions can effectively catch the local activity pattern in each frame, more precise recurrent activities can be achieved by our approach, 2) Our approach also extracts flow curves to describe and visualize recurrent activities. Compared with the previous methods which described recurrent activities by trajectory clusters or probability densities, the flow curves derived by our approach can visualize recurrent activity patterns in a clearer and more straightforward way.

III. Overview of the Approach

The framework of the proposed approach is shown in Fig. 3. The optical flow fields [7], [32] are first extracted from input videos. Secondly, the coarse-to-fine thermal diffusion process is applied to transfer the input motion fields into coherent motion fields, i.e., thermal energy fields (TEFs). Thirdly, the triangulation-based scheme is applied to identify coherent motions. Fourthly, with the obtained coherent motions, the two-step clustering scheme is performed to cluster coherent motions from multiple TEFs and construct semantic regions for the target scene. Finally, based on these semantic regions, we can extract effective features to describe crowd activities in the scene and recognize pre-defined crowd activities accordingly. At the same time, the cluster-and-merge process is also applied based on the extracted coherent motions and semantic regions to discover recurrent activities in the target scene. These proposed techniques are described in the following sections in detail.

IV. Finding Coherent Motions

In order to find accurate coherent motions, it is important to construct a coherent motion field to highlight the motion

correlation among particles while still maintaining the original motion information. To achieve this requirement, we introduce a *thermal diffusion process* to model particle correlations. Given an input optical flow field, we view each particle (i.e., each pixel in a frame) as a "heat source" and it can diffuse energies to influence other particles. By suitably modeling this thermal diffusion process, precise correlation among particles can be achieved. The formulation is motivated by the following intuitions:

1) Particles farther from heat source should achieve fewer thermal energies;
2) Particles residing in the motion direction of the heat source particle should receive more thermal energies;
3) Heat source particles with larger motions should carry more thermal energies.

A. Thermal Diffusion Process

Based on the above discussions, we borrow the idea from physical thermal propagation [33] and model the thermal diffusion process by Eq. 1:

$$\frac{\partial \mathbf{E}_{\mathbf{P},l}}{\partial l} = k_p^2 \left(\frac{\partial^2 \mathbf{E}_{\mathbf{P},l}}{\partial x^2} + \frac{\partial^2 \mathbf{E}_{\mathbf{P},l}}{\partial y^2} \right) + \mathbf{F}_{\mathbf{P}} \qquad (1)$$

where $\mathbf{E}_{\mathbf{P},l} = [E_{\mathbf{P},l}^x, E_{\mathbf{P},l}^y]$ is the thermal energy for the particle at location $\mathbf{P} = (p^x, p^y)$ after performing thermal diffusion for l seconds, $\mathbf{F}_{\mathbf{P}} = [f_{\mathbf{P}}^x, f_{\mathbf{P}}^y]$ is the input motion vector for particle $\mathbf{P}$, k_p is the propagation coefficient.

The first term in Eq. 1 models the propagation of thermal energies over free space such that the spatial correlation among particles can be properly enhanced during thermal diffusion. The second term $\mathbf{F}_{\mathbf{P}}$ can be viewed as the external force added on the particle to affect its diffusion behavior, which preserves the original motion patterns. The inclusion of this term is one of the major differences between the proposed approach and the anisotropic-diffusion methods [29]. Without the $\mathbf{F}_{\mathbf{P}}$ term, Eq. 1 can be solved by:

$$\mathbf{E}_{\mathbf{P},l} = \frac{1}{wh} \sum_{\mathbf{Q} \in \mathbf{I}, \mathbf{Q} \neq \mathbf{P}} \mathbf{e}_{\mathbf{P},l}(\mathbf{Q}) \qquad (2)$$

where $\mathbf{E}_{\mathbf{P},l}$ is the final diffused thermal energy for particle $\mathbf{P}$ after l seconds, $\mathbf{I}$ is the set of all particles in the frame, w and h are width and height of the frame. The individual thermal energy $\mathbf{e}_{\mathbf{P},l}(\mathbf{Q}) = [e_{\mathbf{P},l}^x(\mathbf{Q}), e_{\mathbf{P},l}^y(\mathbf{Q})]$ is diffused from the heat source particle $\mathbf{Q} = (q^x, q^y)$ to particle $\mathbf{P}$ after l seconds, defined as:

$$e_{\mathbf{P},l}^{\gamma}(\mathbf{Q}) = u_{\mathbf{Q}}^{\gamma} \cdot e^{\frac{-k_p}{l} ||\mathbf{P} - \mathbf{Q}||^2} \qquad (3)$$

where $\gamma \in \{x, y\}$, $\mathbf{U}_{\mathbf{Q}} = (u_{\mathbf{Q}}^x, u_{\mathbf{Q}}^y)$ is the current motion pattern for the heat source particle $\mathbf{Q}$ and it is initialized by $\mathbf{U}_{\mathbf{Q}} = \mathbf{F}_{\mathbf{Q}}$, $||\mathbf{P} - \mathbf{Q}||$ is the distance between particles $\mathbf{P}$ and $\mathbf{Q}$. In this paper, we fix l to be 1 to eliminate its effect.

However, when $\mathbf{F}$ in Eq. 1 is non-zero, it is difficult to get the exact solution for Eq. 1. So we introduce an additional term $e^{-k_f |\mathbf{F}_{\mathbf{Q}} \cdot (\mathbf{P} - \mathbf{Q})|}$ to approximate the influence of $\mathbf{F}_{\mathbf{Q}}$ where k_f is a force propagation factor. Moreover, in order to prevent unrelated particles from accepting too much heat from $\mathbf{Q}$,
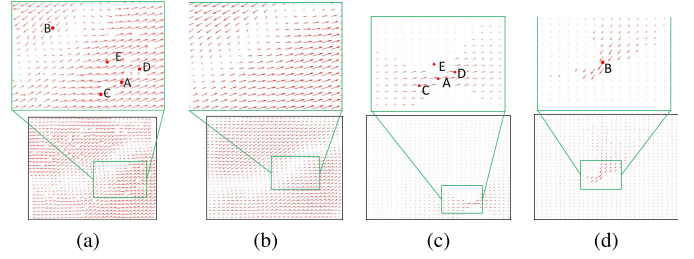

Fig. 4. (a),(b): One input optical flow field and its thermal energy field; (c), (d): Individual thermal diffusion result by diffusing from a single heat source particle A and B to the entire field.

we restrict that only highly correlated particles will propagate energies to each other. The final individual thermal energy from $\mathbf{Q}$ to $\mathbf{P}$ is:

$$e_{\mathbf{P},l}^{\gamma}(\mathbf{Q}) = u_{\mathbf{Q}}^{\gamma} \times e^{-k_p ||\mathbf{P} - \mathbf{Q}||^2} \times e^{-k_f |\mathbf{F}_{\mathbf{Q}} \cdot (\mathbf{P} - \mathbf{Q})|} \qquad (4)$$

if $\cos(\mathbf{F}_{\mathbf{P}}, \mathbf{F}_{\mathbf{Q}}) \geq \theta_c$ and is 0 if otherwise, where $\mathbf{F}_{\mathbf{P}}$ and $\mathbf{F}_{\mathbf{Q}}$ are the input motion vectors of the current particle $\mathbf{P}$ and the heat source particle $\mathbf{Q}$, and $\cos(\mathbf{F}_{\mathbf{P}}, \mathbf{F}_{\mathbf{Q}})$ is the cosine similarity, θ_c is a threshold. In our experiments, k_p, k_f, and θ_c are set to be 0.2, 0.8, 0.7, which are decided from the experimental statistics.

From Eq. 2, we see that the diffused thermal energy $\mathbf{E}_{\mathbf{P}}$ is the summation from all other particles, which encodes the correlation among $\mathbf{P}$ and all other particles in the frame. Furthermore, in Eq. 4, the first term preserves the motion pattern of the heat source. The second term considers the spatial correlation between source and target particles. The third term guarantees that particles along the motion direction of the heat source receives more thermal energies. Furthermore, the cosine similarity $\cos(\mathbf{F}_{\mathbf{P}}, \mathbf{F}_{\mathbf{Q}})$ is introduced in Eq. 4 such that particle $\mathbf{P}$ will not accept energy from $\mathbf{Q}$ if their input motion vectors are far different (or less-coherent) from each other. That is, Eq. 4 successfully satisfies all the intuitions.

Fig. 4 shows one example of the thermal diffusion process, which reveals that:

1) Comparing Figs 4b and 4a, the original motion information is indeed preseved in the TEF. Moreover, TEF further strengthens particle motion coherency by thermal diffusion, which integrates the influence among particles. Coherent motions become more recognizable, thus more accurate coherent motion extraction can be achieved.
2) From Fig. 4c, we can see that the thermal energy for each heat source particle is propagated in a sector shape. Particles along the motion direction of the heat source (C and D) receive more energies than particles outside the motion direction (such as E). In Fig. 4d, since particles on the lower side of the heat source B have small (cosine) motion similarities with B, they do not accept thermal energies.

B. The Coarse-to-Fine Scheme

Although Eqs 2 and 4 can effectively strengthen the coherency among particles, it is based on a single input

Algorithm 1 Coarse-to-Fine Thermal Diffusion Process

1: $T = T_{max}$
2: Calculate the input motion vector field $\mathbf{F_P}(T)$ with T-frame intervals
3: $\mathbf{U_P} = \mathbf{F_P}(T)$
4: **for** $n = 0$ to Num_{itr} **do**
5:     Use Eq. 2 to create the new thermal energy field $\mathbf{E_P^n}$ based on $\mathbf{F_P}(T)$ and $\mathbf{U_P}$
6:     Normalize the vector magnitudes in $\mathbf{E_P^n}$
7:     $\mathbf{U_P} = \mathbf{E_P^n}$
8:     $T = T - T_{step}$
9:     **if** $T > 0$ **then**
10:         Calculate $\mathbf{F_P}(T)$ with the new T
11:     **end if**
12: **end for**
13: Output $\mathbf{E_P^n}$

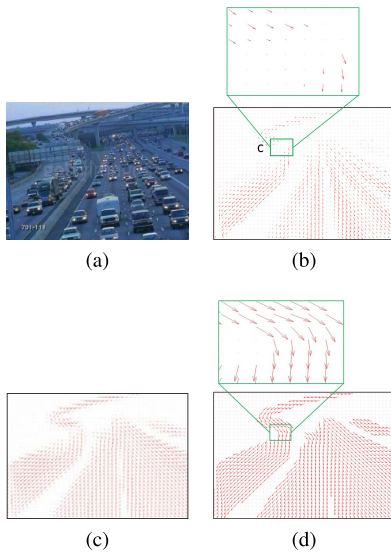(a)   (b)

(c)   (d)

Fig. 5.   (a), (b): An input video frame and its input motion vector field; (c), (d): TEF results of Algorithm 1 after 1 and 3 iterations, respectively ($T_{max} = 5$ and $T_{step} = 1$).

motion field, and only short-term motion information is considered, which is volatile and noisy. Thus, we propose a coarse-to-fine scheme to include long-term motion information. The entire coarse-to-fine thermal diffusion process is described in Algorithm 1.

The long-term motion vector field with a large frame interval T_{max} is first calculated and used to create the thermal energy field. Then, the TEF is iteratively updated with shorter-term motion vector fields, i.e., $\mathbf{F_P}(T)$ with smaller T. Figs 5a to 5d show the TEF results after different iteration numbers. When more iterations are performed, more motion information with different intervals will be included in the thermal diffusion process. Thus, more precise results can be achieved in the TEF, as in Fig. 5d. Fig. 1c shows another TEF result after the entire coarse-to-fine thermal diffusion scheme. We find that:

1) TEF is an enhanced version of the input motion where particles' energy directions in the TEF are similar to their original motion directions. Besides, since TEF include both the motion correlation among particles and

the short-/long-term motion information among frames, coherent motions are effectively strengthened and highlighted in TEF.

2) As mentioned, input motion vectors may be disordered, e.g., region A in Fig 1b. However, the thermal energies from other particles can help recognize these disordered motion vectors and make them coherent, e.g., Fig. 1c.

3) Input motion vectors may be extremely small due to slow motion or occlusion by other objects (region B in Fig. 1b and region C in Fig. 5b). It is very difficult to include these particles into the coherent region by traditional methods [7]–[10] because they are close to the background motion vector. However, TEF can strengthen these small motion vectors by diffusing thermal energies from distant particles with larger motions.

C. Finding Coherent Motions Through Triangulation

Coherent motion regions can be achieved by performing segmentation on the TEF. We propose a triangulation-based scheme as follows:

Step 1 (Triangulation): In this step, we randomly sample particles from the entire scene and apply the triangulation process [34] to link the sampled particles. The block labeled as "triangulation" in Fig. 3 shows one triangulation result, where red dots are the sampled particles and the lines are links created by the triangulation process [34].

Step 2 (Boundary Detection): We first obtain each triangulation link weight by:

$$\omega(\mathbf{P}, \mathbf{Q}) = \frac{||\mathbf{E_P} - \mathbf{E_Q}||}{||\mathbf{P} - \mathbf{Q}||} \tag{5}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are two connected particles, $\mathbf{E_P}$ and $\mathbf{E_Q}$ are the thermal energy vectors of $\mathbf{P}$ and $\mathbf{Q}$ in the TEF. A large weight will be assigned if the connected particles are from different coherent motion regions (i.e., they have different thermal energy vectors). Thus, by thresholding on the link weights, we can find links crossing the boundaries. The block labeled as "detected region boundary" in Fig. 3 shows one boundary detection result after step 2.

Step 3 (Coherent Motion Segmentation): Then, coherent motions can be easily segmented and we use the watershed algorithm [35]. The final coherent motions are shown in the block named "detected coherent motions" in Fig. 3.

V. CONSTRUCTING SEMANTIC REGIONS

With the extracted coherent motions, accurate motion information in a frame can be achieved. However, since coherent motions vary over time, it is essential to construct semantic regions from time-varying coherent motions to catch stable semantic patterns inside a scene. For this purpose, we propose a *two-step clustering scheme*. Assuming that in total M coherent motions ($\mathbf{C}_m$, $m = 1, ..., M$) from N TEFs extracted at N times, the two-step clustering scheme is:

Step 1 (Cluster Coherent Motion Regions): The similarity between two coherent motions $\mathbf{C}_m$ and $\mathbf{C}_k$ is computed as:

$$S_C(\mathbf{C}_m, \mathbf{C}_k) = \#\{(\mathbf{P}, \mathbf{Q}) | \mathbf{P} \in \mathbf{L}_m, \mathbf{Q} \in \mathbf{L}_k,$$
$$\cos(\mathbf{E_P}, \mathbf{E_Q}) \cdot e^{-k_p ||\mathbf{P} - \mathbf{Q}||^2} > \theta_{bp}\} \tag{6}$$

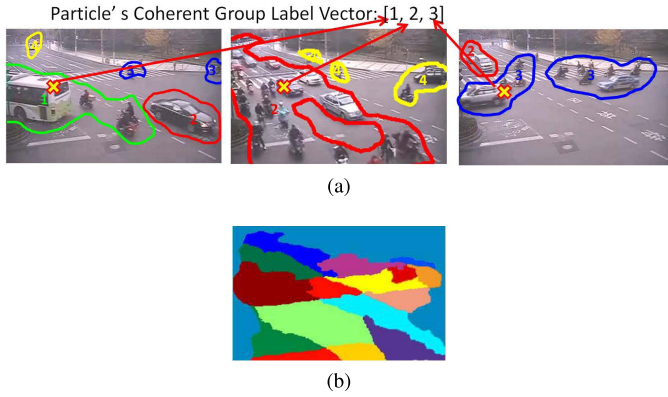Particle's Coherent Group Label Vector: [1, 2, 3]

(a)

(b)

Fig. 6. (a) Step 1: Coherent regions in the three TEFs have been assigned different cluster labels by Step 1 and are displayed in different colors); (b) Find semantic regions by clustering the cluster label vectors of the particles (best viewed in color).

where #{·} is the number of elements in a set. θ_{bp} is a threshold which is set to be the same as θ_c in Eq. 4 in our experiments. Furthermore, $\mathbf{L}_m$ and $\mathbf{L}_k$ are the sets of "indicative particles" for $\mathbf{C}_m$ and $\mathbf{C}_k$:

$$\mathbf{L}_m = \{\mathbf{P}|\cos(\mathbf{E_P}, \mathbf{V_P}) > \theta_c, \mathbf{P} \text{ is on the boundary of } \mathbf{C}_m\}$$
$$\mathbf{L}_k = \{\mathbf{Q}|\cos(\mathbf{E_Q}, \mathbf{V_Q}) > \theta_c, \mathbf{Q} \text{ is on the boundary of } \mathbf{C}_k\} \quad (7)$$

where $\mathbf{V_P} = [v_{\mathbf{P}}^x, v_{\mathbf{P}}^y]$ is the outer normal vector at $\mathbf{P}$, i.e., perpendicular to the boundary and pointing outward the coherent motion region, θ_c is the same threshold as in the condition for Eq. 4. That is, only particles which are on the boundaries of the coherent motion region and whose thermal energy vectors sharply point outward the region are selected as the indicative particles. Thus, we can avoid noisy particles and substantially reduce the required computations.

From Eq. 6, we can see that we first extract the indicative particles, then only utilize those high-correlation pairs, and the total number of such pairs are the similarity value between two coherent motions. It should be noted that the similarity will be calculated between any coherent motion pairs even if they belong to different TEFs.

Then, we construct a similarity graph for the M coherent motions, and perform clustering [36] on this similarity graph with the optimal number of clusters being determined automatically, the cluster results are grouped coherent regions.

Step 2 (Cluster to Find Semantic Regions): Each coherent motion is assigned a cluster label in Step 1, as illustrated in Fig. 6a. However, due to the variation of coherent motions at different times, there exist many ambiguous particles. For example, in Fig. 6a, the yellow cross particle belongs to different coherent motion clusters in different TEFs. This makes it difficult to directly use the clustered coherent motion results to construct reliable semantic regions. In order to address this problem, we further propose to encode particles in each TEF by the cluster labels of the particles' affiliated coherent motions. And by concatenating the cluster labels over different TEFs, we can construct a "cluster label" vector for each particle, as in Fig. 6a And with these label vectors, the same spectral clustering process as Step 1 can be performed on the particles to achieve the final semantic regions, as in Fig. 6b.
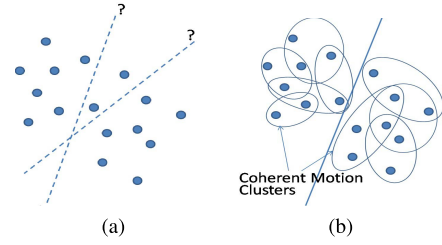

(a)   (b)

Fig. 7. (a) Directly segmenting semantic regions according to the particles' local features. (b) Segmenting semantic regions with the guidance of coherent motion clusters.

Comparing with previous semantic region segmentation methods [4], [31] which perform clustering using local similarity among particles, our scheme utilizes the guidance from the global coherent motion clustering results to strengthen the correlations among particles. For example, in Fig. 7a, when directly segmenting the particles by their local features, its accuracy may be limited due to similar distances among particles. However, by utilizing cluster labels to encode the particles, similarities among particles can be suitably enhanced by the global coherent cluster information, as in Fig. 7b. Thus, more precise segmentation results can be achieved.

A. Recognizing Pre-Defined Activities

Based on the constructed semantic regions, we are able to recognize pre-defined activities (i.e., activities with labeled training data) in the scene. In this paper, we simply average the TEF vectors in each semantic region and concatenate these averaged TEF vectors as the final feature vector for describing the activity patterns in a TEF. Then, a linear support vector machine (SVM) [37] is utilized to train and recognize pre-defined activities. Experimental results show that with accurate TEF and precise semantic regions, we can achieve satisfying results using this simple method.

B. Merging Disconnected Coherent Motions

Since TEF also includes long-distance correlations between distant particles, by performing our clustering scheme, we also have the advantage of effectively merging disconnected coherent motions, which may be caused by the occlusion from other objects or low density of the crowd. For examples, the two disconnected blue regions in the right-most figure in Fig. 6a are merged into the same cluster by our approach. Note that this issue is not well studied in the existing coherent motion research.

VI. MINING RECURRENT ACTIVITIES

With the extracted coherent motions and constructed semantic regions, crowd activities can be recognized by constructing and pre-labeling training data, as in Section V-A. However, since pre-defining or labeling crowd activities take lots of human labors, it is also desirable to automatically mine recurrent activity patterns in a crowd scene without human intervention. For this purpose, we propose a *cluster-and-marge process* which includes three steps: frame-level clustering, coherent motion merging, and flow curve extraction.
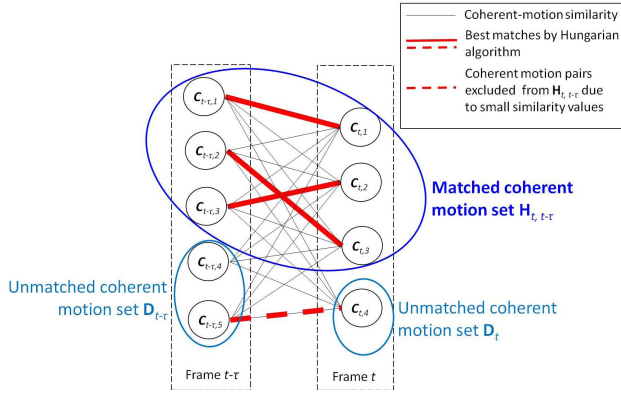
Fig. 8. Example of matched and unmatched coherent motion region sets (best viewed in color).

A. Frame-Level Clustering

The frame-level clustering step clusters frames according to the extracted coherent motions and semantic regions, such that frames with the same recurrent activity pattern can be organized into the same group. In this paper, we first calculate inter-frame similarities for all frame pairs and then utilize spectral clustering [36] to cluster frames according to these inter-frame similarities.

In order to calculate the inter-frame similarity between frames t and $t - \tau$, the similarities between all coherent motions from frames t and $t - \tau$ are first calculated using Eq. 6. Then, the inter-frame similarity $S_F (t, t - \tau)$ can be achieved from these coherent motion similarities and the segmented semantic regions. More specifically, we define $S_F (t, t - \tau)$ as

$$S_F (t, t - \tau) = S_{FU} (t, t - \tau) \cdot S_{FM} (t, t - \tau) \quad (8)$$

where $S_{FM} (t, t - \tau)$ is the similarity for the matched coherent motion pairs between t and $t - \tau$, $S_{FU} (t, t - \tau)$ is the similarity for the unmatched coherent motion regions in frames t and $t - \tau$. $S_{FM} (t, t - \tau)$ and $S_{FU} (t, t - \tau)$ can be calculated by Eqs 9 and 10.

First, $S_{FM} (t, t - \tau)$ is defined as

$$S_{FM} (t, t - \tau) = \frac{\sum\limits_{(\mathbf{C}_{t,i}, \mathbf{C}_{t-\tau,j}) \in \mathbf{H}_{t,t-\tau}} \lambda_{i,j} S_C (\mathbf{C}_{t,i}, \mathbf{C}_{t-\tau,j})}{\max\{n_t, n_{t-\tau}\}} \quad (9)$$

where $S_C (\mathbf{C}_{t,i}, \mathbf{C}_{t-\tau,j})$ is the similarity between coherent motion regions $\mathbf{C}_{t,i}$ and $\mathbf{C}_{t-\tau,j}$, $\lambda_{i,j}$ is the corresponding weight. n_t and $n_{t-\tau}$ are the total number of coherent motion regions in frames t and $t - \tau$, respectively. $\mathbf{H}_{t,t-\tau}$ is the set of all matched coherent region pairs. In this paper, $\mathbf{H}_{t,t-\tau}$ and $\lambda_{i,j}$ are calculated by the Hungarian algorithm [38] which can achieve optimal coherent motion matching results based on the input coherent motion similarities. Furthermore, in order to exclude dissimilar coherent motion pairs from the matching result, coherent motion pairs $(\mathbf{C}_{t,i}, \mathbf{C}_{t-\tau,j})$ with small similarity values $S_C (\mathbf{C}_{t,i}, \mathbf{C}_{t-\tau,j})$ will be deleted from $\mathbf{H}_{t,t-\tau}$. Fig. 8 shows an example of the matched coherent motion pairs.
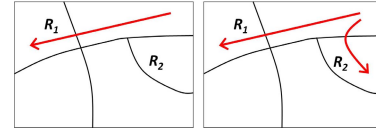

Fig. 9. Motion flows for two recurrent activities displayed over semantic regions.

The next term $S_{FU} (t, t - \tau)$ is defined as

$$S_{FU} (t, t - \tau) = \prod_{\mathbf{C}_{t-\tau,j} \in \mathbf{D}_{t-\tau}} \varepsilon (\mathbf{C}_{t-\tau,j}) \cdot \prod_{\mathbf{C}_{t,i} \in \mathbf{D}_t} \varepsilon (\mathbf{C}_{t,i}) \quad (10)$$

where $\mathbf{D}_{t-\tau}$ and $\mathbf{D}_t$ are the sets of unmatched coherent regions in frames $t - \tau$ and t, as shown in Fig. 8. $\varepsilon (\mathbf{C})$ is the unmatching cost for coherent motion region $\mathbf{C}$:

$$\varepsilon (\mathbf{C}) = \frac{\sum_{\mathbf{R}_k, \mathbf{R}_k \cap \mathbf{C} \neq \emptyset} \frac{1}{\rho (\mathbf{R}_k)}}{\#\{\mathbf{R}_k | \mathbf{R}_k \cap \mathbf{C} \neq \emptyset\}} \quad (11)$$

where $\mathbf{R}_k$ is the k-th semantic region of the scene, the term $\#\{\mathbf{R}_k | \mathbf{R}_k \cap \mathbf{C} \neq \emptyset\}$ represents the total number of semantic regions that have overlap with the coherent motion region $\mathbf{C}$. $\rho (\mathbf{R}_k)$ is the importance cost measuring whether semantic region $\mathbf{R}_k$ is important in distinguishing different recurrent activities. For example, assuming that a scene includes two recurrent activities, as in Fig. 9, it is obvious that the semantic region $\mathbf{R}_2$ on the right should have larger importance cost since the two recurrent activity patterns have different motion flows in $\mathbf{R}_2$. Comparatively, the semantic region $\mathbf{R}_1$ on the left should have smaller importance cost since both recurrent activity patterns have similar flows in $\mathbf{R}_1$. Therefore, when calculating the similarity between frames t and $t - \tau$, if there exists an unmatched coherent region $\mathbf{C}$ in $\mathbf{R}_2$, a large importance cost $\rho (\mathbf{R}_2)$ will be applied to reduce the inter-frame similarity, indicating that frames t and $t - \tau$ have different recurrent activity patterns. On the contrary, if there exists an unmatched coherent region $\mathbf{C}$ in $\mathbf{R}_1$, the inter-frame similarity will be less affected since a coherent region in $\mathbf{R}_1$ is less indicative of the differences between recurrent activities.

For $\rho (\mathbf{R}_k)$, we first perform a pre-clustering according to the matched coherent motion similarities $S_{FM} (t, t - \tau)$ which roughly clusters frames into different recurrent activity groups. Then a vector is constructed for each semantic region $\mathbf{R}_k$: $\left[Num_{\mathbf{R}_k, \mathbf{G}_1}, Num_{\mathbf{R}_k, \mathbf{G}_2}, ..., Num_{\mathbf{R}_k, \mathbf{G}_Z} \right]$ where $Num_{\mathbf{R}_k, \mathbf{G}_i}$ is the total number of coherent motions located in $\mathbf{R}_k$ in the i-th pre-clustered recurrent activity group $\mathbf{G}_i$, Z is the total number of pre-clustered recurrent activity groups. Finally, $\rho (\mathbf{R}_k)$ can be calculated by:

$$\rho (\mathbf{R}_k) = e^{k_s \cdot var\{Num_{\mathbf{R}_k, \mathbf{G}_1}, Num_{\mathbf{R}_k, \mathbf{G}_2}, ..., Num_{\mathbf{R}_k, \mathbf{G}_Z}\}} \quad (12)$$

where $var\{\cdot\}$ is the variance operation, $k_s = \frac{1}{Num_f^2}$ where Num_f is the total number of frames to be clustered. According to Eq. 12, if coherent motions appear evenly in $\mathbf{R}_k$ for different recurrent activities, i.e., the variance is smaller, it implies that $\mathbf{R}_k$ is less important in distinguishing different recurrent activities. On the contrary, if the appearance time of coherent motions in $\mathbf{R}_k$ has larger variation over different pre-clustered

Algorithm 2 Frame-Level Clustering Process

Input: Coherent regions $\mathbf{C}_{t,i}$ extracted for each frame t, and semantic regions $\mathbf{R}_k$ of the scene

Output: Recurrent activity groups including frames with similar recurrent activity patterns

1: Calculate similarities $S_C(\mathbf{C}_{t,i}, \mathbf{C}_{t-\tau,j})$ between all coherent motion regions from different frames
2: Calculate $S_{FM}(t, t-\tau)$ for all frame pairs based on $S_C(\mathbf{C}_{t,i}, \mathbf{C}_{t-\tau,j})$
3: Pre-cluster frames based on $S_{FM}(t, t-\tau)$
4: Calculate importance cost $\rho(\mathbf{R}_k)$ for all semantic regions based on the pre-clustering result
5: Calculate $S_{FU}(t, t-\tau)$ for all frame pairs according to the unmatched coherent regions and $\rho(\mathbf{R}_k)$
6: Calculate inter-frame similarities $S_F(t, t-\tau)$ for all frame pairs using $S_{FM}(t, t-\tau)$ and $S_{FU}(t, t-\tau)$
7: With $S_F(t, t-\tau)$, cluster frames into recurrent activity groups
8: Output the clustering result in line 9

recurrent activity groups, a large $\rho(\mathbf{R}_k)$ will be achieved to increase the importance of $\mathbf{R}_k$. The complete process of frame-level clustering is illustrated in Algorithm 2.

B. Coherent Motion Merging

After frame-level clustering, frames are clustered into different recurrent activity groups. Thus, by parsing frames in each recurrent activity group, complete motion patterns for each recurrent activity can be estimated. In this paper, we introduce a coherent motion merging step to merge similar coherent motions from the same recurrent activity group for achieving motion pattern regions. More specifically, we first apply the same operation as Step 1 in the two-step clustering scheme (Section V) to cluster coherent motion regions inside the same recurrent activity group. Then, coherent motions of the same cluster are merged together to form a motion pattern region. The merging process can be described by Eq. 13 and Fig. 10.

$$\overline{\mathbf{E}}_{\mathbf{P},\boldsymbol{\Psi}_j} = \sum_{\mathbf{C}_m \in \boldsymbol{\Psi}_j} \frac{\mathbf{E}_{\mathbf{P},\mathbf{C}_m}}{\#\{\boldsymbol{\Psi}_j\}} \tag{13}$$

if $\dfrac{\#\left\{\mathbf{E}_{\mathbf{P},\mathbf{C}_m}|\mathbf{E}_{\mathbf{P},\mathbf{C}_m} \neq [0,0], \mathbf{C}_m \in \boldsymbol{\Psi}_j\right\}}{\#\{\boldsymbol{\Psi}_j\}} > \theta_{mf}$, and it is $[0,0]$ if otherwise, where $\boldsymbol{\Psi}_j$ the j-th coherent motion cluster. $\overline{\mathbf{E}}_{\mathbf{P},\boldsymbol{\Psi}_j} = \left[\overline{E}^x_{\mathbf{P},\boldsymbol{\Psi}_j}, \overline{E}^y_{\mathbf{P},\boldsymbol{\Psi}_j}\right]$ is the merged motion vector result for $\boldsymbol{\Psi}_j$ at particle $\mathbf{P}$. $\mathbf{C}_m$ is a coherent motion region belonging to coherent motion cluster $\boldsymbol{\Psi}_j$. $\#\{\boldsymbol{\Psi}_j\}$ is the total number of coherent motion regions in cluster $\boldsymbol{\Psi}_j$. θ_{mf} is a threshold which is set as 0.4 in our experiments. $\mathbf{E}_{\mathbf{P},\mathbf{C}_m}$ is the TEF thermal energy for $\mathbf{C}_m$ at particle $\mathbf{P}$. Note that $\mathbf{E}_{\mathbf{P},\mathbf{C}_m}$ is set to $[0,0]$ if $\mathbf{P}$ is outside the region of $\mathbf{C}_m$. And $\#\left\{\mathbf{E}_{\mathbf{P},\mathbf{C}_m}|\mathbf{E}_{\mathbf{P},\mathbf{C}_m} \neq [0,0], \mathbf{C}_m \in \boldsymbol{\Psi}_j\right\}$ is the total number of non-zero TEF thermal energies at particle $\mathbf{P}$ and belonging to $\boldsymbol{\Psi}_j$.

According to Eq. 13, the merged motion pattern region $\overline{\mathbf{R}}_{\boldsymbol{\Psi}_j} = \{\overline{\mathbf{E}}_{\mathbf{P},\boldsymbol{\Psi}_j}\}$ for a coherent motion cluster $\boldsymbol{\Psi}_j$ is basically the normalized summation over all coherent regions in $\boldsymbol{\Psi}_j$. Besides, we further introduce a threshold θ_{mf} to filter out noisy
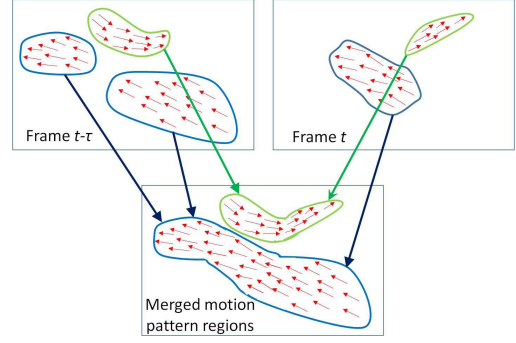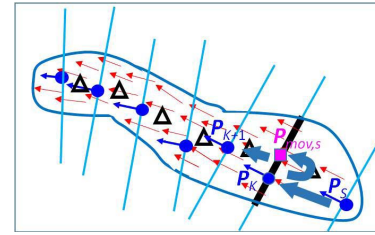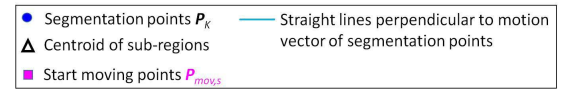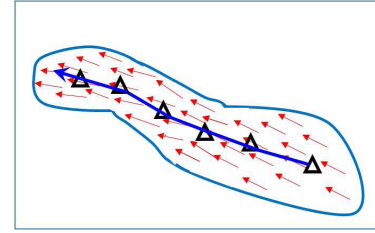

Fig. 10. Process of similar coherent motion merging. Frames t and $t - \tau$ are from the same recurrent activity group, the green coherent motion regions in frames t and $t - \tau$ belong to one coherent motion cluster, and the blue coherent motion regions in frames t and $t - \tau$ belong to another coherent motion cluster. (Best viewed in color).

Fig. 11. Flow curve extraction process. Finding segmentation points, draw straight lines to achieve sub-regions, and calculate centroids of each sub-region; (b) Link centroids to achieve the extracted flow curve. (Best viewed in color.)

or isolated particles which have low frequent motions in the coherent motion cluster $\boldsymbol{\Psi}_j$. An example of merged motion pattern regions is shown in Fig. 10.

C. Flow Curve Extraction

The motion pattern regions achieved in the previous step can represent the complete motion information for each recurrent activity. However, since motion pattern regions may overlap with each other and the contours of motion pattern regions may also be irregular, it is necessary to extract flow curves from these motion pattern regions such that recurrent activities can be more clearly described and visualized.

Our proposed flow curve extraction process can be described by Algorithm 3 and Fig. 11. According to
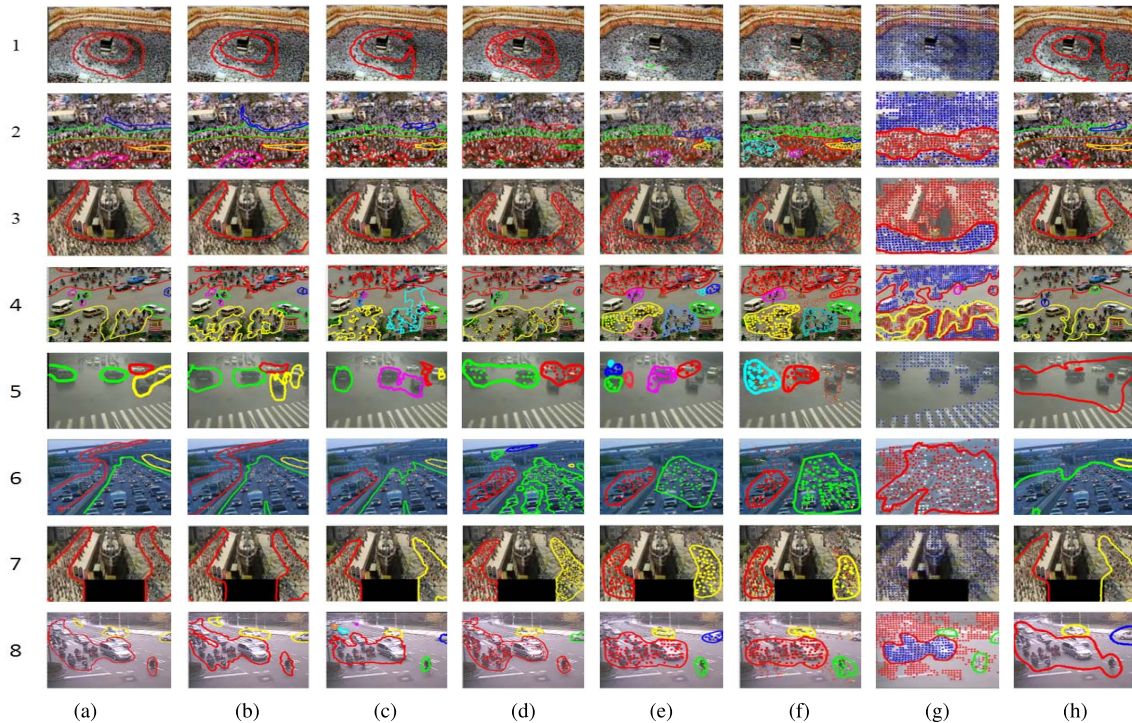
Fig. 12.   Coherent motion extraction results. (a): Ground Truth, (b): Results of our approach, (c): Results of [7], (d): Results of [8], (e): Results of [9], (f): Results of [10], (g): Results of [40], (h): Results of [28]. (Best viewed in color).

Algorithm 3 Flow Curve Extraction

Input: A motion pattern region $\overline{\mathbf{R}}_{\boldsymbol{\Psi}_j}$ merged from coherent region cluster $\boldsymbol{\Psi}_j$

Output: A flow curve extracted from $\overline{\mathbf{R}}_{\boldsymbol{\Psi}_j}$

1: Calculate the skeleton of $\overline{\mathbf{R}}_{\boldsymbol{\Psi}_j}$ [35]
2: Find the end point $\mathbf{P}_s$ of the skeleton which is on "backward" position to all other end points, where the "backward" direction is defined as the reversed direction of the motion flows in $\overline{\mathbf{R}}_{\boldsymbol{\Psi}_j}$
3: $\mathbf{P}_K = \mathbf{P}_s$, where $\mathbf{P}_K$ is the current segmentation point
4: **while** $\mathbf{P}_{K+1}$ is inside $\overline{\mathbf{R}}_{\boldsymbol{\Psi}_j}$ **do**
5:     $\mathbf{P}_{mov,s}$ as the middle point of the line perpendicular to the motion vector $\overline{\mathbf{E}}_{\mathbf{P}_K,\boldsymbol{\Psi}_j}$ at $\mathbf{P}_K$
6:     **for** n=0 to Num_{mov} $\{Num_{mov}$ is the number of movements$\}$ **do**
7:         Move from $\mathbf{P}_{mov,s}$ to $\mathbf{P}_{mov,e}$ by $\overline{\mathbf{E}}_{\mathbf{P}_{mov,s},\boldsymbol{\Psi}_j}$, where $\overline{\mathbf{E}}_{\mathbf{P}_{mov,s},\boldsymbol{\Psi}_j}$ is motion vector at $\mathbf{P}_{mov,s}$ in $\overline{\mathbf{R}}_{\boldsymbol{\Psi}_j}$
8:         $\mathbf{P}_{mov,s}$=$\mathbf{P}_{mov,e}$
9:     **end for**
10:    $\mathbf{P}_{K+1}$= $\mathbf{P}_{mov,s}$, where $\mathbf{P}_{K+1}$ is the next segmentation point

11:    Draw two straight lines perpendicular to the motion vectors of at $\mathbf{P}_K$ and $\mathbf{P}_{K+1}$, respectively
12:    Calculate the centroid of the sub-region segmented by the lines in line 13
13:    $\mathbf{P}_K = \mathbf{P}_{K+1}$
14: **end while**
15:    Sequentially link together all centroid points achieved by line 14

16:    Smooth the linked curve by line 17
17:    Output the curve by line 18

Algorithm 3 and Fig. 11, our approach first sequentially cuts a motion pattern region $\overline{\mathbf{R}}_{\boldsymbol{\Psi}_j}$ into sub-regions along the motion direction in $\overline{\mathbf{R}}_{\boldsymbol{\Psi}_j}$. Then the centroids of sub-regions are linked

together to achieve the output flow curve. With the above process, the extracted flow curve can accurately catch the major motion flow of a motion pattern region. Furthermore, it should be noted that in step 5 of Algorithm 3, if the line perpendicular to the motion vector $\overline{\mathbf{E}}_{\mathbf{P}_{K+1},\boldsymbol{\Psi}_j}$ at $\mathbf{P}_{K+1}$ is intersecting with a branched motion region (i.e., the motion region diverges around $\mathbf{P}_{K+1}$), multiple $\mathbf{P}_{mov,s}$ points will be achieved and the following flow curve extraction process will be performed on each $\mathbf{P}_{mov,s}$ respectively. In this way, we can properly achieve branched flow curves at the branch region.

VII. Experimental Results

Our approach is implemented by Matlab and the optical flow fields [32] are used as the input motion vector fields while each pixel in the frame is viewed as a particle. In order to achieve motion vector fields with T-frame intervals ($T = 10$ in our experiments), the particle advection method [7] is used which tracks the movement of each particle over T frames.

A. Results for Coherent Motion Detection

We perform experiments on a dataset including 30 different crowd videos collected from the UCF dataset [7], the UCSD dataset [39], the CUHK dataset [10], and our own collected set. This dataset covers various real-world crowd scene scenarios with both low- and high-density crowds and both rapid and slow motion flows. Some example frames of the dataset is shown in Fig. 12.

We compare our approach with four state-of-the-art coherent motion detection algorithms: The Lagrangian particle dynamics approach [7], the local-translation domain segmentation approach [8], the coherent-filtering approach [9], and the

collectiveness measuring-based approach [10]. In order to further demonstrate the effectiveness of our approach, we also include the results of a general motion segmentation method [40] and an anisotropic-diffusion-based image segmentation method [28].

1) Qualitative Comparison on Coherent Motion Detection: Fig. 12 compares the coherent motion detection results for different methods. We include the manually labeled ground truth results in the first column. From Fig. 12, we can see that our approach can achieve better coherent motion extraction than the compared methods. For example, in sequence 1, our approach can effectively extract the circle-shape coherent motion. Comparatively, the method in [7] can only detect part of the circle while the methods in [9] and [10] fail to work since few reliable key points are extracted from this over-crowded scene. For sequences 2 and 4 where multiple complex motion flows exist, our approach can still precisely detect the small and less differentiable coherent motions, such as the pink region on the bottom and the blue region on the top in sequence 2 (a). The compared methods have low effectiveness in identifying these regions due to the interference from the neighboring motion regions. In sequences 3 and 6, since motions on the top of the frame are extremely small and close to the background, the compared methods fail to include these particles into the coherent motion region. However, in our approach, these small motions can be suitably strengthened and included through the thermal diffusion process. Furthermore, the methods in [28] and [40] do not show satisfying results, e.g., in sequences 5 and 6. This is because: (1) the crowd scenes are extremely complicated such that the extracted particle flows or trajectories become unreliable, thus making the general motion segmentation methods [40] difficult to create precise results; (2) Since many coherent region boundaries in the crowd motion fields are rather vague and unrecognizable, good boundaries cannot be easily achieved without suitably utilizing the characteristics of the motion vector fields. Thus, simply applying the existing anisotropic-diffusion segmentation methods [28] cannot achieve satisfying results.

2) Capability to Handle Disconnected Coherent Motions: Sequences 5-8 in Fig. 12 compare the algorithms' capability in handling disconnected coherent motions. In sequence 7, we manually block one part of the coherent motion region while in sequences 5, 6, and 8, the red or green coherent motion regions are disconnected due to occlusion by other objects or low density. Since the disconnected regions are separated far from each other, most compared methods wrongly segment them into different coherent motion regions. However, with our thermal diffusion process and two-step clustering scheme, these regions can be successfully merged into one coherent region.

3) Quantitative Comparison: Table I compares the quantitative results for different methods. In Table I, the average Particle Error Rates (PERs) and the average Coherent Number Error (CNE) for all the sequences in our dataset are compared to measure the overall accuracy of coherent motion detection. PER is calculated by PER = # of Wrong Particles/Total # of Particles. CNE is calculated by

TABLE I

AVERAGE PER AND CNE FOR ALL SEQUENCES IN THE DATASET

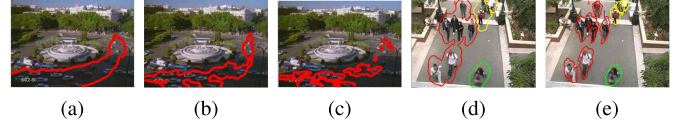|	Proposed	[7]	[8]	[9]	[10]	[40]	[28]
PER (%)	**7.8**	32.5	19.5	25.6	24.1	66.4	21.4
CNE	**0.14**	1.24	0.93	1.05	0.96	1.78	0.84

(a)   (b)   (c)   (d)   (e)

Fig. 13. The coherent motion detection results of our approach under different k_p and k_f values. (a) $k_p = 0.2$. (b) $k_p = 0.5$. (c) $k_p = 0.7$. (d) $k_f = 0.6$. (e) $k_f = 0.9$.

$$CNE = \frac{\sum_i |Num_d(i) - Num_{gt}(i)|}{\sum_i 1}$$ where $Num_d(i)$ and $Num_{gt}(i)$ are the numbers of detected and ground-truth coherent regions for sequence i, respectively, $\sum_i 1$ is the total number of sequences.

Table I further demonstrates the effectiveness of our approach. In Table I, we can see that 1) Our approach can achieve smaller coherent detection error rates than the other methods, 2) Our approach can accurately obtain the coherent region numbers (close to the ground truth) while other methods often over-segment or under-segment the coherent regions.

4) Effect of Different Parameter Values: Finally, Fig. 13 shows the results of our approach under different parameter values, i.e., k_p and k_f in Eqs 3 and 4. From Figs 13a to 13c, we can see that k_p mainly governs the thermal diffusion distance. A small k_p will make the thermal energies to be diffused farther and thus can achieve larger coherent motion regions. When k_p increases, the extracted coherent motion region will shrink. Furthermore, k_f determines the directivity of thermal diffusion. When k_f increases, the diffused thermal energies will concentrate more along the motion direction of the source heat particles. On the contrary, when k_f decreases, the thermal energies will be propagated more uniformly to all directions around the heat source particle. Thus, the boundaries will shrink horizontally with larger k_f, as in Fig. 13e. However, note that in all examples in Fig. 13, our approach can always suitably merge coherent regions together even when they become disconnected when the parameter value changes.

B. Results for Semantic Region Construction and Pre-Defined Activity Recognition

We perform experiments on two crowd videos in our dataset, as the first and second rows in Fig. 14. 400 video clips are selected from each video with each clip including 20 frames. Four crowd activities are defined for each video and the example frames for the crowd activities are shown in Fig. 14. Note that these videos are challenging in that: (1) the crowd density in the scene varies frequently including both high density as Fig. 14d and low density clips as Fig. 14c; (2) The motion patterns are varying for different activities, making it difficult to construct meaningful and stable semantic regions; (3) There are large numbers of irregular motions that disturb
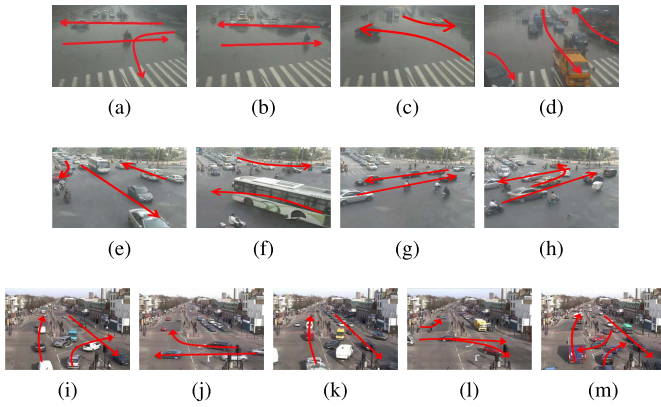
Fig. 14. Example frames of the ground-truth activities in different videos. First and second rows: videos in our dataset; Third row: video of QMUK Junction dataset [41]. HD: Horizontal pass and down turn; HP: Horizontal pass; BT: Both turn; VP: Vertical pass; VL: Vertical pass and left turn; HU: Horizontal pass and up turn; VR: Vertical pass and right turn; VB: Vertical pass and both turn. (a) HD. (b) HP. (c) BT. (d) VP. (e) VL. (f) BT. (g) HP. (h) HU. (i) VR. (j) HU. (k) VP. (l) HD. (m) VB.

Fig. 15. Constructed semantic regions of different methods for the videos in Fig. 14. The caption " [8]" denotes the method " [8]+Two step". (Best viewed in color). (a) Original. (b) Our. (c) Direct. (d) [8]. (e) [4]. (f) Original. (g) Our. (h) Direct. (i) [8]. (j) [4]. (k) Original. (l) Our. (m) Direct. (n) [8]. (o) [4].

the normal motion patterns (e.g., people running the red lights or bicycle following irregular paths); (4) The number of clips in the dataset is small, which increases the difficulty of constructing reliable semantic regions. Moreover, in order to further demonstrate the effectiveness of our approach, we also perform experiments on a public QMUK Junction dataset [41] where five crowd activities are defined, as shown in the third row of Fig. 14.

1) Accuracy on Semantic Region Construction: For each video in Fig. 14, we randomly select 200 video clips and use them to construct the corresponding semantic regions. Fig. 15 compares the results of four methods: (1) Our approach ("Our"), (2) Directly cluster regions based on the particles' TEF vectors ("Direct", note that our approach differs from this method by clustering over the cluster label vectors), (3) Use [8] to achieve coherent motion regions and then apply our two-step clustering scheme to construct semantic regions ("[8]+Two-Step", we show the results of [8] because in our experiments, [8] has the best semantic region construction results among the compared methods in Table I), (4) The activity-based scene segmentation method in [4] ("[4]"). We also show original scene images and plot all major activity flows to ease the comparison ("original scene").

Fig. 15 shows that the methods utilizing "coherent motion cluster label" information ("our" and " [8]+two-step") create more meaningful semantic regions than the other methods, e.g., successfully identifying the horizontal motion regions in the middle of the scene in Fig. 15b. This shows that our cluster label features can effectively strengthen the correlation among particles to facilitate semantic region construction. Furthermore, comparing our approach with the "[8]+Two-Step" method, it is obvious that the semantic regions by our approach are more accurate (e.g., more precise semantic region boundaries and more meaningful segmentations in the scene). This further shows that more precise coherent motion detection results can result in more accurate semantic region results.
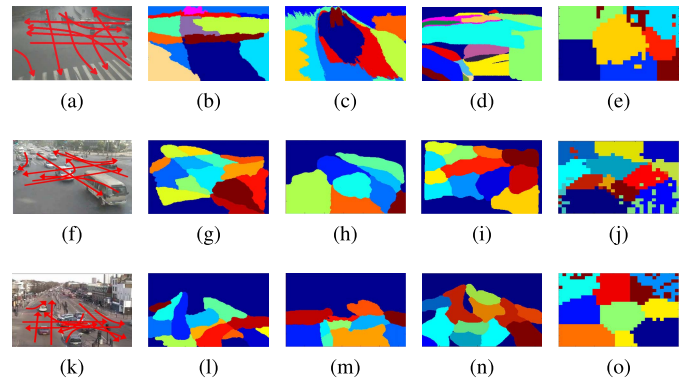
TABLE II
RECOGNITION ACCURACY OF DIFFERENT METHODS

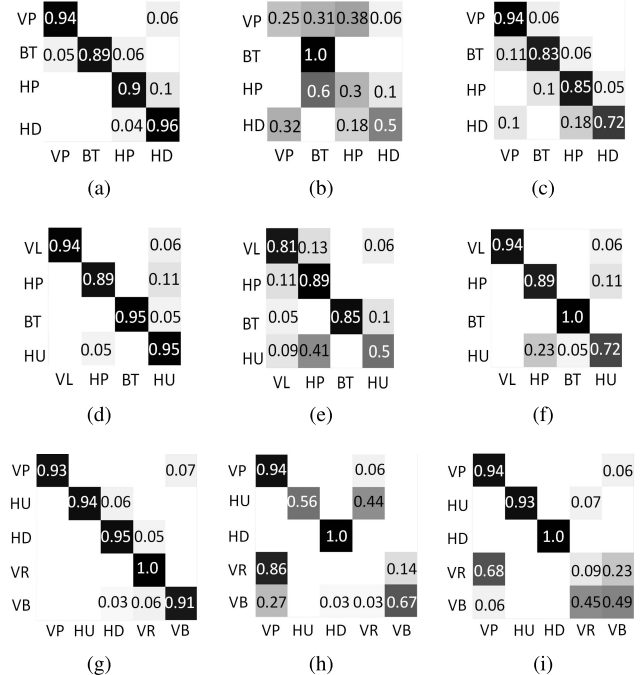|	Our (%)	Our+ OF (%)	Direct (%)	[8]+Two-Step (%)	[4] (%)	[3] (%)
Fig. 14a video	**92.2**	87.75	77.0	89.5	79.2	67.0
Fig. 14e video	**90.69**	83.83	73.53	81.76	72.35	69.80
Fig. 14i video	**93.58**	91.03	83.42	88.32	84.83	82.69

Fig. 16. The confusion matrix of clustering results. (The clustering results in the 1st, 2nd, and 3rd rows correspond to the videos in the 1st, 2nd, and 3rd rows in Fig. 14, respectively). (a) Our. (b) Direct Clustering. (c) Pre-clustering. (d) Our. (e) Direct Clustering. (f) Pre-clustering. (g) Our. (h) Direct Clustering. (i) Pre-clustering.

2) Performances on Recognizing Pre-Defined Activities: In order to recognize the pre-defined activities in Fig. 14, for each video, we randomly select 200 video clips and construct semantic regions by the methods in Fig. 15. After that, we derive features from the TEF and train SVM classifiers by the method in Section V-A. Finally, we perform recognition

(a) Recur. act. mining results for video of Fig. 14a   (b) Recur. act. mining results for video of Fig. 14e   (c) Recur. act. mining results for video of Fig. 14i
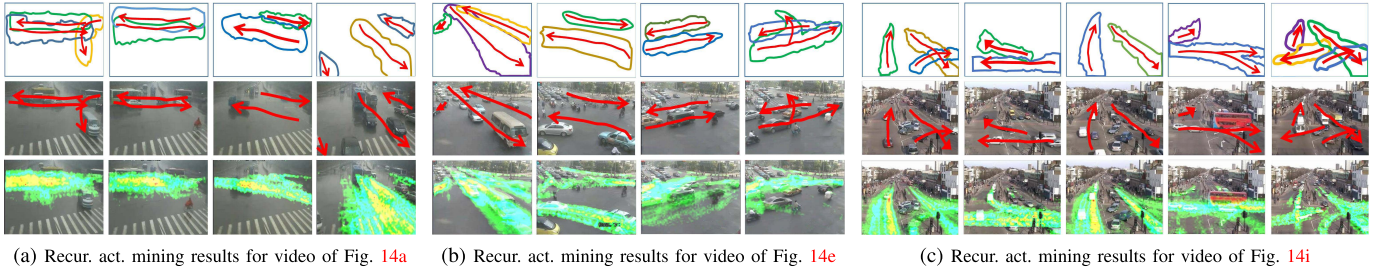
Fig. 17.   Recurrent activity mining results. First rows in (a), (b), and (c): the merged motion pattern regions and the extracted flow curves by our approach; Second rows in (a), (b), and (c): the extracted flow curves of our approach displayed over the video frame; Third rows in (a), (b), and (c): the recurrent activities extracted by [12].

on the other 200 video clips in the same video. Besides, we also include the results of two additional methods: (1) a state-of-the-art dense-trajectory-based recognition method [3] ("Dense-Traj"); (2) the method which uses our semantic regions but uses the input motion field (i.e., the optical flows) to derive the motion features in each semantic region ("Our+OF"). From the recognition accuracy shown in Table II, we observe that:

1) Methods using more meaningful semantic regions (i.e., "our", "our+OF", and " [8]+Two step") achieve better results than other methods. This shows that suitable semantic region construction can greatly facilitate activity recognition.

2) Approaches using TEF ("Our") achieve better results than those using the input motion field ("Our+OF"). This demonstrates that compared with the input motion filed, our TEF can effectively improve the effectiveness in representing the semantic regions' motion patterns.

3) The dense-trajectory method [3] which extracts global features does not achieve satisfying results. This is because the global features still have limitations in differentiating the subtle differences among activities. This further implies the usefulness of semantic region decomposition in analyzing crowd scenes.

C. Results for Recurrent Activity Mining

In this experiment, we use the same videos as in Fig. 14 for mining recurrent activities. For each video, we sample one frame per second, then calculate coherent motions for the sampled frames, and finally apply our cluster-and-merge process to achieve recurrent activity patterns. Note that the target for recurrent activity mining is to automatically discover recurrent activities from an input video without pre-defining activity types or pre-labeling training data. And ideally, good activity mining approaches should achieve similar activity patterns as the human-observed activity types in Fig. 14.

1) Performances on Frame-Level Clustering: For each video, we apply our frame-level clustering step to cluster the sampled frames into four recurrent activity groups. Our clustering results are compared with two methods: (1) *Direct clustering.* Directly clustering based on the TEF difference between two frames (i.e., use the summation of absolute thermal energy differences between the co-located particles in two TEFs as the inter-frame similarity). (2) *Pre-clustering.*

Using the matched-coherent-motion similarities $S_{FM}(t, t-\tau)$ in Eq. 9 as the inter-frame similarity for clustering.

Fig. 16 compares the clustering confusion matrixes of different methods. From Fig. 16, we can see that since frames of the same recurrent activity may contain different parts of a complete activity flow (e.g., Fig. 10), their TEFs may have large differences. Therefore, directly using TEF difference for clustering (i.e., direct TEF clustering) cannot achieve satisfying results. Comparatively, by including coherent motions to evaluate inter-frame similarities (i.e., "pre-clustering" and "our"), the clustering accuracy can be improved. However, the pre-clustering method still have limitations in differentiating similar recurrent activities, e.g., HP and HU in Figs 14g and 14h. Comparatively, by introducing the importance cost of semantic regions to measure the effects of unmatched coherent motions, our frame-level clustering approach can have stronger capability in differentiating similar recurrent activity patterns.

2) Performances on Coherent Motion Merging and Flow Curve Extraction: Fig. 17 shows the results of our coherent motion merging and flow curve extraction steps. Besides, we also compare our approach with a state-of-the-art activity mining method which utilizes a Probabilistic Latent Sequential Motif (PLSM) model to discover recurrent activities [12], which are shown as the last rows in Figs 17a, 17b, and 17c. From Fig. 17, we can have the following observations:

1) The recurrent activities mined by our approach is similar to the human-observed activity types in Fig. 14. This demonstrates that our proposed cluster-and-merge process can effectively discover desired activity types from an input video.

2) Note that although the clustering result in our frame-level clustering step is not 100 percentage accurate (as in Fig. 16), the extracted flow curves are less affected by the wrongly clustered frames because: (i) The noisy or isolated thermal energy vectors from the wrongly clustered frames will be filtered by the threshold θ_{mf} in Eq. 13. (ii) The flow curve extraction process will further reduce the effects of wrong frames by dividing sub-regions to derive flow curves, as in Fig. 11a.

3) Comparing our approach with the PLSM-based method [12], we can see that: (i) By introducing coherent regions to measure inter-frame similarities and derive motion pattern regions, our approach can achieve cleaner activity flows which are more coherent with the

human-observed activity types in Fig. 14. Comparatively, results of the PLSM-based method still include noisy motion patterns, e.g., the last column in Fig. 17b. (ii) Our approach can precisely differentiate motion flows inside a recurrent activity. However, the PLSM-based method has limitations in differentiate motion flows when they are located close to each other, e.g., the second column in Fig. 17a. (iii) The differences between similar recurrent activities are clearly differentiated and visualized by our approach, while they are less obvious in the results of the PLSM-based method, e.g., the third and fourth columns in Fig. 17b.
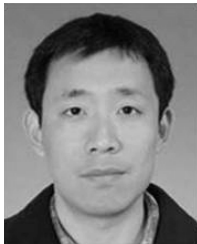
VIII. CONCLUSION

In this paper, we study the problem of coherent motion detection, semantic region construction, and recurrent activity mining in crowd scenes. A thermal-diffusion-based algorithm together with a two-step clustering scheme are introduced, which can achieve more meaningful coherent motion and semantic region results. Based on the extracted coherent motions and semantic regions, a cluster-and-merge process is further proposed which can effectively discover desirable activity patterns from a crowd video. Experiments on various videos show that our approach achieves the state-of-the-art performance.

REFERENCES

[1] W. Wang, W. Lin, Y. Chen, J. Wu, J. Wang, and B. Sheng, "Finding coherent motions and semantic regions in crowd scenes: A diffusion and clustering approach," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 756–771.

[2] M. Hu, S. Ali, and M. Shah, "Learning motion patterns in crowded scenes using motion flow field," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2008, pp. 1–5.

[3] J. Wu, Y. Zhang, and W. Lin, "Towards good practices for action video encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 2577–2584.

[4] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jan. 2009, pp. 1988–1995.

[5] T. Mei, Y. Rui, S. Li, and Q. Tian, "Multimedia search reranking: A literature survey," *ACM Comput. Surv.*, vol. 46, no. 3, pp. 1:1–1:36, 2012.

[6] H. Liu, T. Mei, J. Luo, H. Li, and S. Li, "Finding perfect rendezvous on the go: Accurate mobile visual localization and its applications to routing," in *Proc. 20th ACM Int. Conf. Multimedia (MM)*, 2012, pp. 9–18.

[7] S. Ali and M. Shah, "A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–6.

[8] S. Wu and H. Wong, "Crowd motion partitioning in a scattered motion field," *IEEE Trans. Syst., Man, and*, vol. 42, no. 5, pp. 1443–1454, Oct. 2012.

[9] B. Zhou, X. Tang, and X. Wang, "Coherent filtering: Detecting coherent motions from crowd clutters," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 857–871.

[10] B. Zhou, X. Wang, and X. Tang, "Measuring crowd collectiveness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3049–3056.

[11] R. Emonet, J. Varadarajan, and J.-M. Odobez, "Temporal analysis of motif mixtures using Dirichlet processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 140–156, Jan. 2013.

[12] V. Jagannadan, R. Emonet, and J.-M. Odobez, "A sequential topic model for mining recurrent activities from long term video logs," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 100–126, 2013.

[13] B. Zhou, X. Wang, and X. Tang, "Random field topic model for semantic region analysis in crowded scenes from tracklets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3441–3448.

[14] X. Wang, K. T. Ma, G.-W. Ng, and W. E. L. Grimson, "Trajectory analysis and semantic region modeling using nonparametric Bayesian models," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 287–321, 2011.

[15] W. Hu, X. Li, G. Tian, S. Maybank, and Z. Zhang, "An incremental DPMM-based method for trajectory clustering, modeling, and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1051–1065, May 2013.

[16] B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2287–2301, Nov. 2011.

[17] J. C. Nascimento, M. A. T. Figueiredo, and J. S. Marques, "Trajectory classification using switched dynamical hidden Markov models," *IEEE Trans. Image Process.*, vol. 19, no. 5, pp. 1338–1348, May 2010.

[18] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 935–942.

[19] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu, "Crowd analysis: A survey," *Mach. Vis. Appl.*, vol. 19, nos. 5–6, pp. 345–357, 2008.

[20] D. Cremers and S. Soatto, "Motion competition: A variational approach to piecewise parametric motion segmentation," *Int. J. Comput. Vis.*, vol. 62, no. 3, pp. 249–265, 2005.

[21] T. Brox, M. Rousson, R. Deriche, and J. Weickert, "Colour, texture, and motion in level set based segmentation and tracking," *Image Vis. Comput.*, vol. 28, no. 3, pp. 376–390, 2010.

[22] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas, "Abnormal detection using interaction energy potentials," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3161–3167.

[23] L. Li and Y. Yang, "Optical flow estimation for a periodic image sequence," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 1–10, Jan. 2010.

[24] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1744–1757, Sep. 2012.

[25] D. Lin, E. Grimson, and J. Fisher, "Learning visual flows: A Lie algebraic approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 747–754.

[26] A. del Bimbo, P. Nesi, and J. L. C. Sanz, "Optical flow computation using extended constraints," *IEEE Trans. Image Process.*, vol. 5, no. 5, pp. 720–739, May 1996.

[27] M. J. Black, G. Sapiro, D. H. Marimont, and D. Heeger, "Robust anisotropic diffusion," *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 421–432, Mar. 1998.

[28] Y. Wu, Y. Wang, and Y. Jia, "Adaptive diffusion flow active contours for image segmentation," *Comput. Vis. Image Understand.*, vol. 117, no. 10, pp. 1421–1435, 2013.

[29] J. Weickert, *Anisotropic Diffusion in Image Processing*. Stuttgart, Germany: Teubner, 1998.

[30] T. Xu *et al.*, "Single and multiple view detection, tracking and video analysis in crowded environments," in *Proc. Int. Conf. Adv. Video Signal-Based Surveill. (AVSS)*, Sep. 2012, pp. 494–499.

[31] J. Li, S. Gong, and T. Xiang, "Scene segmentation for behaviour correlation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 383–395.

[32] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/Kanade meets horn/schunck: Combining local and global optic flow methods," *Int. J. Comput. Vis.*, vol. 61, no. 3, pp. 211–231, 2005.

[33] H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids*. Oxford, U.K.: Oxford Science, 1959.

[34] H. Edelsbrunner and N. R. Shah, "Incremental topological flipping works for regular triangulations," *Algorithmica*, vol. 15, no. 3, pp. 223–241, 1996.

[35] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2001.

[36] Z. Lu, X. Yang, W. Lin, H. Zha, and X. Chen, "Inferring user image-search goals under the implicit guidance of users," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 3, pp. 394–406, Mar. 2014.

[37] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[38] L. Lovász and M. D. Plummer, *Matching Theory*. Amsterdam, The Netherlands: North Holland, 1986.

[39] *UCSD Anomaly Detection Dataset*. [Online]. Available: http://www.svcl.ucsd.edu/projects/anomaly/, accessed Jan. 2014.

[40] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 282–295.

[41] *QMUL Junction Dataset*. [Online]. Available: http://www.eecs.qmul.ac.uk/~ccloy/downloads_qmul_junction.html, accessed Aug. 2015.

Jianxin Wu (M'12) received the B.S. and M.S. degrees in computer science from Nanjing University, and the Ph.D. degree in computer science from the Georgia Institute of Technology. He was an Assistant Professor with Nanyang Technological University, Singapore. He is currently a Professor with the Department of Computer Science and Technology, Nanjing University, China, and is associated with the National Key Laboratory for Novel Software Technology, China. His research interests are computer vision and machine learning. He served as an Area Chair of ICCV 2015 and a Senior PC Member for AAAI 2016.

Weiyao Lin received the B.E. and M.E. degrees from Shanghai Jiao Tong University, China, in 2003 and 2005, respectively, and the Ph.D. degree from the University of Washington, Seattle, USA, in 2010. He is currently an Associate Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University. He has authored or co-authored over 80 technical papers, including over 30 refereed journal papers and one book chapter. His research interests include image/video processing, video surveillance, and computer vision. He served as an Associate Editor of the *Journal of Visual Communication and Image Representation*, *Signal Processing: Image Communication*, the *Journal of Circuits, Systems, and Signal Processing*, and the IEEE ACCESS.
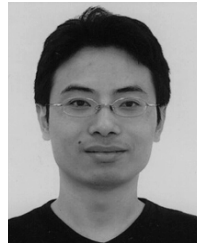
Jingdong Wang received the B.Eng. and M.Eng. degrees in automation from the Department of Automation, Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong. He is a Lead Researcher with the Visual Computing Group, Microsoft Research Asia. His areas of interest include computer vision, multimedia, and machine learning. He is mainly working on image search, including indexing and compact coding for large scale similarity search as well as interactive image search, and visual understanding including image recognition, salient object detection, person re-identification, and image segmentation. He served as an Area Chair in ACMMM 2015 and ICME 2015, a Track Chair in ICME 2012, and a Special Session Chair in ICMR 2014. He has made many contributions to Microsoft products, including Bing image search, Project Oxford, and XiaoIce.

Yang Mi received the B.S. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014, where he is currently pursuing the M.E. degree in electrical engineering. His research interests include image processing and computer vision.
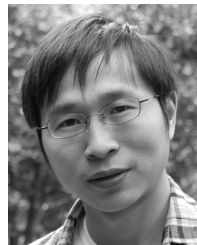
Tao Mei (M'07–SM'11) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively. He is a Lead Researcher with Microsoft Research Asia, Beijing, China.. He has authored or co-authored over 100 papers in journals and conferences, ten book chapters, and edited three books. He holds 13 U.S. granted patents and more than 20 in pending. His current research interests include multimedia information retrieval and computer vision. He is a Senior Member of the ACM. He was a recipient of several paper awards from prestigious multimedia journals and conferences, including the IEEE Circuits and Systems Society Circuits and Systems for Video Technology Best Paper Award in 2014, the IEEE TRANSACTIONS ON MULTIMEDIA Prize Paper Award in 2013, and the best paper awards at ACM Multimedia in 2009 and 2007. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, *ACM/Springer Multimedia*, and *Neurocomputing*, and a Guest Editor of six international journals. He was the General Co-Chair of ACM ICIMCS 2013, and the Program Co-Chair of the IEEE ICME 2015, the IEEE MMSP 2015, and MMM 2013.

Weiyue Wang received the B.S. degree in electrical engineering form Shanghai Jiao Tong University, China, in 2014, and the M.S. degree in electrical and computer engineering from The Ohio State University, USA, in 2015. She is currently pursuing the Ph.D. degree in computer science with the University of Southern California, USA. Her research interests include computer vision, pattern analysis, and computer graphics.