Finding Coherent Motions and Semantic Regions in Crowd Scenes: A Diffusion and Clustering Approach

Weiyue Wang[1], Weiyao Lin[1,⋆], Yuanzhe Chen[1], Jianxin Wu[2], Jingdong Wang[3], and Bin Sheng[4]

[1] Dept. Electronic Engr., Shanghai Jiao Tong Univ., China
[2] National Key Laboratory for Novel Software Technology, Nanjing Univ., China
[3] Microsoft Research, Beijing, China
[4] Dept. Computer Science & Engr., Shanghai Jiao Tong Univ., China

Abstract. This paper addresses the problem of detecting coherent motions in crowd scenes and subsequently constructing semantic regions for activity recognition. We first introduce a coarse-to-fine thermal-diffusion-based approach. It processes input motion fields (e.g., optical flow fields) and produces a coherent motion filed, named as 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. Finally, these semantic regions are used to recognize activities in crowded scenes. Experiments on various videos demonstrate the effectiveness of our approach.

1 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. Fig. 1). 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.

In this paper, we focus on: (1) constructing an accurate coherent motion field to find coherent motions, and (2) finding stable semantic regions based on the detected coherent motions and recognizing activities in a crowd scene.

First, constructing an accurate coherent motion field is crucial to coherent motion detection. In Fig. 1, (c) is the input motion field and (d) is the coherent motion field which is constructed from (c) using the proposed approach. In (c), 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

⋆ Corresponding author.

D. Fleet et al. (Eds.): ECCV 2014, Part I, LNCS 8689, pp. 756–771, 2014.

coherent detection results directly from (c). However, by transferring (c) into a coherent motion field where the coherent motions among particles are suitably highlighted (i.e., (d)), coherent motion detection is greatly facilitated. However, although many algorithms have been proposed for coherent motion detection [2,21,26,27,12], 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 for describing the activity patterns in a scene is another important issue. Coherent motions at different times may vary widely, e.g. in Fig 1(a), changing of traffic lights will lead to different coherent motions. Coherent motions alone may not effectively describe the overall semantic patterns in a scene. Therefore, semantic regions need to be extracted from these time-varying coherent motions to achieve stable and meaningful semantic patterns. 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 custering process for this purpose.

Our contributions to crowd scene understand and activity recognition are:

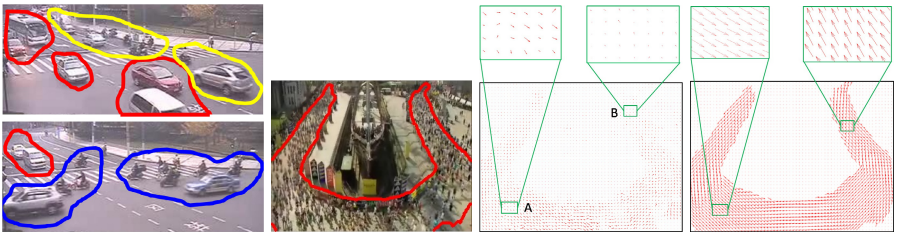
(1) We propose a coarse-to-fine thermal diffusion process to transfer the input motion field into a thermal energy field (TEF), i.e., a more accurate coherent motion field. TEF effectively encodes both motion correlation among particles and motion trends of individual particles. To our knowledge, this 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 further propose 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 crowd activity recognition based on these semantic regions can achieve good performance. Besides, the proposed clustering scheme can also effectively handle disconnectedness, which is caused by occlusion or low density regions in the crowd (cf. Fig. 1 (a), the yellow regions).

The remainder of this paper is organized as follows. Section 2 reviews related works. Section 3 describes the framework of the proposed approach. Sections 4-6 describe the details of our proposed thermal diffusion process, triangulation scheme, and two-step clustering scheme. Section 7 shows the experimental results and Section 8 concludes the paper.

2 Related Works

Although many works [2,21,26,27,12,17,25,9,5,10] have been proposed on coherent motion detection, due to the complex nature of crowd scenes, they are not



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

Fig. 1. (a) Example time-varying coherent motions; (b) Example frame of a Marathon video sequence, the red curve is the ground truth coherent motion region; (c) Input motion vector field of (b); (d) Coherent motion field from (c) using the proposed approach (Best viewed in color)

yet mature for the accurate detection of coherent motion fields. Cremers and Soatto [9] and Brox et al. [5] model the intensity variation of optical flow by an objective functional minimization scheme. However, these methods are only suitable for motions with simple patterns and cannot effectively analyze complex crowd patterns such as the circular flow in Fig. 1 (b). Other works introduce external spatial-temporal correlation traits to model the motion coherency among particles [21,26,27]. 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 will 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. 1 (c).

Besides the works on coherent motion detection, 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 [23,14]. However, the focus of our approach is different from these methods. In our approach, we focus on enhancing the correlation among coherent particles to facilitate coherent motion detection. Thus, the motion vectors of coherent particles will be enhanced even if their actual motions are small, such as the region B in Fig. 1 (c) and (d). In contrast, advanced optical flow estimation methods focus on estimating the *actual* motion of particles. Thus, they are still less capable of creating precise results when applied on coherent motion detection.

Another thread of related works is the anisotropic-diffusion-based methods [18,22,20] used in image segmentation. However, our approach also 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

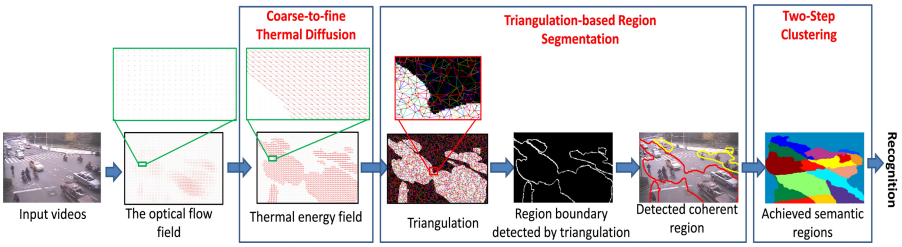


Fig. 2. The flowchart of the proposed approach (best viewed in color)

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, many coherent region boundaries are vague, subtle and unrecognizable. Simply applying the anisotropic-diffusion methods [18,22,20] 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, another important issue is the utilization of coherent motions to recognize 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 [2], 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 [19,24]. 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 [15,13] which recognize crowd activities by segmenting scenes into semantic regions, our approach differs from them in that: 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 in this paper, information from the coherent motions can effectively enhance the correlation among particles, resulting in more meaningful semantic regions to facilitate activity recognition.

3 Overview of the Approach

Fig. 2 shows framework of the proposed approach. The input motion fields are first extracted from input videos. In this paper, optical flow fields [2,6] are extracted, and each pixel in the frame is viewed as a particle. Then, 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)). After that, the triangulation-based scheme is applied to identify coherent motions. Finally, the

two-step clustering scheme is performed to cluster the coherent motions from multiple TEFs and construct semantic regions for the target scene. With these semantic regions, we can extract effective features to describe crowd activities in the scene and perform recognition accordingly. In the following, we will describe the details of the proposed coarse-to-fine thermal diffusion process, the triangulation-based scheme, and the two-step clustering scheme, respectively.

4 Coarse-to-Fine Thermal Diffusion

In order to facilitate coherency detection, 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 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. Besides, we also argue that the following intuitions should be satisfied:

(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.

4.1 Thermal Diffusion Process

Based on the above discussions, we borrow the idea from physical thermal propagation [7] and model the thermal diffusion process by Eqn. (1):

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

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

The first term in Eqn. (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 $\boldsymbol{F}_{\boldsymbol{P}}$ 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 our approach and the anisotropic-diffusion methods [20]. Without the $\boldsymbol{F}_{\boldsymbol{P}}$ term, Eqn. (1) can be solved by:

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

where $\boldsymbol{E}_{\boldsymbol{P},l}$ is the final diffused thermal energy for particle $\boldsymbol{P}$ after l seconds, $\boldsymbol{I}$ is the set of all particles in the frame, w and h are width height of the frame.

The individual thermal energy $\boldsymbol{e}_{\boldsymbol{P},l}(\boldsymbol{Q}) = [e^x_{\boldsymbol{P},l}(\boldsymbol{Q}), e^y_{\boldsymbol{P},l}(\boldsymbol{Q})]$ is diffused from the heat source particle $\boldsymbol{Q} = (q^x, q^y)$ to particle $\boldsymbol{P}$ after l seconds, as:

$$e^{\gamma}_{\boldsymbol{P},l}(\boldsymbol{Q}) = u^{\gamma}_{\boldsymbol{Q}} \cdot e^{\frac{-k_p}{l}||\boldsymbol{P}-\boldsymbol{Q}||^2} \tag{3}$$

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

However, when $\boldsymbol{F}$ in Eqn. (1) is non-zero, it is difficult to get the exact solution for Eqn. (1). So we introduce an additional term $e^{-k_f|\boldsymbol{F}_{\boldsymbol{Q}}\cdot(\boldsymbol{P}-\boldsymbol{Q})|}$ to approximate the influence of $\boldsymbol{F}_{\boldsymbol{Q}}$ where k_f is a force propagation factor. Moreover, in order to prevent unrelated particles from accepting too much heat from $\boldsymbol{Q}$, we restrict that only highly correlated particles will propagate energies to each other. The final individual thermal energy from $\boldsymbol{Q}$ to $\boldsymbol{P}$ is:

$$e^{\gamma}_{\boldsymbol{P},l}(\boldsymbol{Q}) = \begin{cases} u^{\gamma}_{\boldsymbol{Q}} \times e^{-k_p||\boldsymbol{P}-\boldsymbol{Q}||^2} \times e^{-k_f|\boldsymbol{F}_{\boldsymbol{Q}}\cdot(\boldsymbol{P}-\boldsymbol{Q})|} & \text{if } \cos(\boldsymbol{F}_{\boldsymbol{P}}, \boldsymbol{F}_{\boldsymbol{Q}}) \geq \theta_c \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\boldsymbol{F}_{\boldsymbol{P}}$ and $\boldsymbol{F}_{\boldsymbol{Q}}$ are the input motion vectors of the current particle $\boldsymbol{P}$ and the heat source particle $\boldsymbol{Q}$, and $\cos(\boldsymbol{F}_{\boldsymbol{P}}, \boldsymbol{F}_{\boldsymbol{Q}})$ is the cosine similarity, θ_c is a threshold.

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

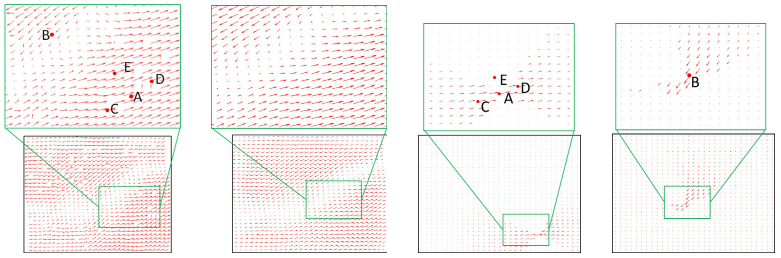
Fig. 3 shows one example of the thermal diffusion process, which reveals that:

(1) Comparing Fig. 3 (b) and (a), 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. 3 (c), 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. 3 (d), since particles on the lower side of the heat source B have small (cosine) motion similarities with B, they do not accept thermal energies.

4.2 The Coarse-to-Fine Scheme

Although Eqn. (2) can effectively strengthen the coherency among particles, it is based on a single input motion field, and only short-term motion information is



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

Fig. 3. (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

considered, which is volatile and noisy. Thus, we propose a coarse-to-fine scheme to include long-term motion information.

Algorithm 1: Coarse-to-Fine Thermal Diffusion Process

$T = T_{max}$.
calculate the input motion vector field $\boldsymbol{F_P}(T)$ with T-frame intervals.
$\boldsymbol{U_P} = \boldsymbol{F_P}(T)$.
for $n = 0$ to Num_{itr} // Num_{itr} is the total iteration time
  use Eqn. (2) to create the new thermal energy field $\boldsymbol{E_P^n}$ based on $\boldsymbol{F_P}(T)$ and $\boldsymbol{U_P}$.
  normalize the vector magnitudes in $\boldsymbol{E_P^n}$.
  $\boldsymbol{U_P} = \boldsymbol{E_P^n}$.
  $T = T - T_{step}$.
  if $T > 0$
    calculate $\boldsymbol{F_P}(T)$ with the new T.
  end if
end for
output $\boldsymbol{E_P^n}$

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., $\boldsymbol{F_P}(T)$ with smaller T. Fig. 4 (c)-(d) 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. 4 (d). Fig. 1 (d) 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. 1 (c). However, the thermal energies from other particles can help recognize these disordered motion vectors and make them coherent, e.g., Fig. 1 (d).

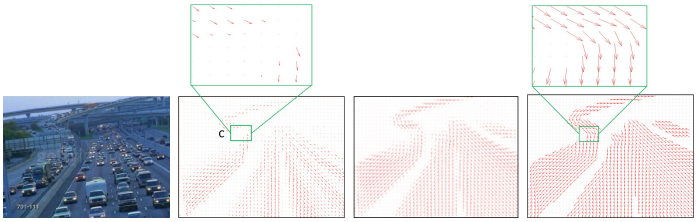


Fig. 4. (a),(b): An input video frame and its input motion vector field; (c),(d): TEF results of Algorithm 1 after 1 and 3 iterations, respectively (T_{max}=5 and T_{step}=1)

(3) Input motion vectors may be extremely small due to slow motion or occlusion by other objects (region B and C in Fig. 4 (b), respectively.) It is very difficult to include these particles into the coherent region by traditional methods [2,21,26,27] 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.

5 Coherent Motion Extraction 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 [11] to link the sampled particles. The block labeled as "triangulation" in Fig. 2 shows one triangulation result, where red dots are the sampled particles and the lines are links created by the triangulation process [11].

Step 2: Boundary detection. We first obtain each triangulation link weight by:

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

where $\boldsymbol{P}$ and $\boldsymbol{Q}$ are two connected particles, $\boldsymbol{E_P}$ and $\boldsymbol{E_Q}$ are the thermal energy vectors of $\boldsymbol{P}$ and $\boldsymbol{Q}$ 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. 2 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 [3]. The final coherent motions are shown in the block named "detected coherent motions" in Fig. 2.

6 Two-Step Clustering

Since coherent motions may vary over time, it is essential to construct semantic regions from time-varying coherent motions to catch the stable semantic patterns

inside a scene, for which we propose a two-step clustering scheme. Assuming that in total M coherent motions ($\boldsymbol{C}_m$, $m = 1, ..., M$) 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 $\boldsymbol{C}_m$ and $\boldsymbol{C}_k$ is computed as:

$$S(\boldsymbol{C}_m, \boldsymbol{C}_k) = \#\{(\boldsymbol{P}, \boldsymbol{Q}) | \boldsymbol{P} \in \boldsymbol{L}_m, \boldsymbol{Q} \in \boldsymbol{L}_k, \cos(\boldsymbol{E}_{\boldsymbol{P}}, \boldsymbol{E}_{\boldsymbol{Q}}) \cdot e^{-k_p ||\boldsymbol{P}-\boldsymbol{Q}||^2} > \theta_{bp}\} \tag{6}$$

where $\#\{\cdot\}$ is the number of elements in a set, and θ_{bp} is a threshold. Furthermore, $\boldsymbol{L}_m$ and $\boldsymbol{L}_k$ are the sets of "indicative particles" for $\boldsymbol{C}_m$ and $\boldsymbol{C}_k$:

$$\begin{cases} \boldsymbol{L}_m = \{\boldsymbol{P} | \cos(\boldsymbol{E}_{\boldsymbol{P}}, \boldsymbol{V}_{\boldsymbol{P}}) > \theta_c, \boldsymbol{P} \text{ is on the boundary of } \boldsymbol{C}_m\} \\ \boldsymbol{L}_k = \{\boldsymbol{Q} | \cos(\boldsymbol{E}_{\boldsymbol{Q}}, \boldsymbol{V}_{\boldsymbol{Q}}) > \theta_c, \boldsymbol{Q} \text{ is on the boundary of } \boldsymbol{C}_k\} \end{cases} \tag{7}$$

where $\boldsymbol{V}_{\boldsymbol{P}} = [v_{\boldsymbol{P}}^x, v_{\boldsymbol{P}}^y]$ is the outer normal vector at $\boldsymbol{P}$, i.e., perpendicular to the boundary and pointing outward the coherent motion region. θ_c is the same threshold as in Eqn. (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 Eqn. (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 [16] on this similarity graph with the optimal number of clusters being determined automatically, the cluster results are grouped coherent regions.

Fig. 5. (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)

Step 2: Cluster to find semantic regions. Each coherent motion is assigned a cluster label in Step 1, as illustrated in Fig. 5 (a). However, due to the variation of coherent motions at different time, there exist many ambiguous particles. For example, in Fig. 5(a), 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

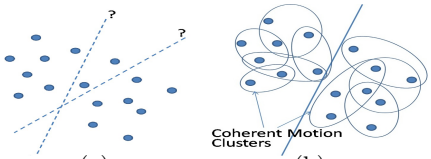


Fig. 6. (a) Directly segmenting semantic regions according to the particles' local features. (b) Segmenting semantic regions with the guidance of coherent motion clusters.

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. 5(a). And with these label vectors, the same spectral clustering process as Step 1 [16] can be performed on the particles to achieve the final semantic regions, as in Fig. 5 (b).

Comparing with previous semantic region segmentation methods [15,13] 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. 6 (a), 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. 6 (b). Thus, more precise segmentation results can be achieved.

6.1 Activity Recognition

Based on the constructed semantic regions, we are able to recognize activities 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) [8] is utilized to train and recognize activities. Experimental results show that with accurate TEF and precise semantic regions, we can achieve satisfying results using this simple method.

6.2 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. 5 (a) are merged into the same cluster by our approach. Note that this issue is not well studied in the existing coherent motion research.

7 Experimental Results

Our algorithm is implemented by Matlab and the optical flow fields [6] 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 [2] is used which tracks the movement of each particle over T frames. Furthermore, the parameters k_p, k_f, θ_c, and θ_{bp} in Eqns (4) and (6) are set to be 0.2, 0.8, 0.7, and 0.7, respectively. These values are decided from the experimental statistics.

7.1 Results for Coherent Motion Detection

We perform experiments on a dataset including 30 different crowd videos collected from the UCF dataset [2], the UCSD dataset [1], the CUHK dataset [27], 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. 7.

We compare our approach with four state-of-the-art coherent motion detection algorithms: The Lagrangian particle dynamics approach [2], the local-translation domain segmentation approach [21], the coherent-filtering approach [26], and the collectiveness measuring-based approach [27]. In order to further demonstrate the effectiveness of our approach, we also include the results of a general motion segmentation method [4] and an anisotropic-diffusion-based image segmentation method [22].

Qualitative Comparison on Coherent Motion Detection. Fig. 7 compares the coherent motion detection results for different methods. We include the manually labeled ground truth results in the first column. From Fig. 7, 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 [2] can only detect part of the circle while the methods in [26] and [27] 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 [4] and [22] 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 [4] difficult to create precise results;

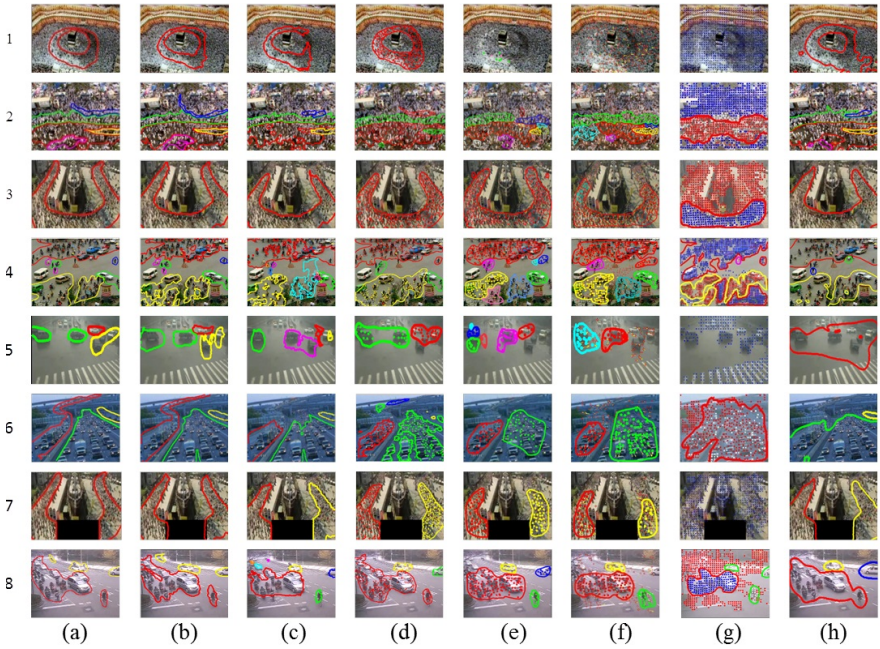


Fig. 7. Coherent motion extraction results. (a): Ground Truth, (b): Results of our approach, (c): Results of [2], (d): Results of [21], (e): Results of [26], (f): Results of [27], (g): Results of [4], (h): Results of [22]. (Best viewed in color).

(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 [22] cannot achieve satisfying results.

Table 1. Average PER and CNE for all sequences in the dataset

Methods	**Proposed**	[2]	[21]	[26]	[27]	[4]	[22]
Average PER (%)	**7.8**	32.5	19.5	25.6	24.1	66.4	21.4
Average CNE	**0.14**	1.24	0.93	1.05	0.96	1.78	0.84

Capability to Handle Disconnected Coherent Motions. Sequences 5-8 in Fig. 7 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.

(a) VP (b) BT (c) HR (d) HL

Fig. 8. Example frames of the activities in the crossroad dataset

Quantitative Comparison. Table 1 compares the quantitative results for different methods. In Table 1, 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 $CNE = \frac{\sum_i |Num_d(i) - Num_{gt}(i)|}{\Sigma_i 1}$ where $Num_d(i)$ and $Num_{gt}(i)$ are the number of detected and ground-truth coherent regions for sequence i, respectively. And $\Sigma_i 1$ is the total number of sequences.

Table 1 further demonstrate the effectiveness of our approach. In Table 1, 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.

7.2 Results for Semantic Region Construction and Activity Recognition

We perform experiments on a dataset of a crowd crossroad scene. This dataset includes 400 video clips with each clip includes 20 frames. There are totally four crowd activities in the dataset: vertical pass (VP), both turn (BT), horizontal pass and right turn (HR), and horizontal pass and left turn (HL), as in Fig. 8. This is a challenging dataset in that: (1) the crowd density in the scene varies frequently including both high density as Fig. 8 (a) and low density clips as Fig. 8 (b); (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 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.

Accuracy on Semantic Region Construction. We randomly select 200 video clips to construct semantic regions. Fig. 9 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 [21] to achieve coherent motion regions and then apply our two-step clustering scheme to construct semantic regions ("[21]+Two-Step", we choose to show the results of [21] because from our experiments, [21] has the best semantic region construction results among

the compared methods in Table 1), (4) The activity-based scene segmentation method in [15] ("[15]"). We also show original scene images and plot all major activity flows to ease the comparison ("original scene").

Fig. 9 shows that the methods utilizing "coherent motion cluster label" information ("our" and "[21]+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). 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 "[21]+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.

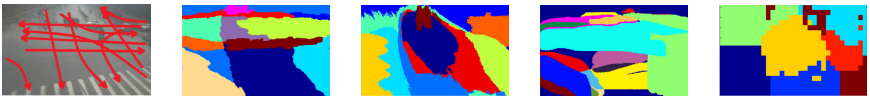


Fig. 9. Constructed semantic regions of different methods. (Best viewed in color).

Table 2. Recognition accuracy of different methods

Methods	Our	Our+OF	Direct	[21]+Two-Step	[15]	[19]
Accuracy	92.2%	87.75%	77.0%	89.5%	79.2%	67.0%

Performances on Activity Recognition. We randomly select 200 video clips and construct semantic regions by the methods in Fig. 9. After that, we derive features from the TEF and train SVM classifiers by the method in Section 6.1. Finally, we perform recognition on the other 200 video clips. Besides, we also include the results of two additional methods: (1) a state-of-the-art dense-trajectory-based recognition method [19] ("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 2, we observe that:

(1) Methods using more meaningful semantic regions ("our", "our+OF", and "[21]+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 [19] 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.

8 Conclusion

In this paper, we study the problem of coherent motion detection and semantic region construction in crowd scenes, and introduce a thermal-diffusion-based algorithm together with a two-step clustering scheme, which can achieve more meaningful coherent motion and semantic region results. Experiments on various videos show that our approach achieves the state-of-the-art performance.

Acknowledgements. This work is supported in part by the following grants: National Science Foundation of China (No. 61001146, 61202154, 61025005, U1201255), Shanghai Pujiang Program (12PJ1404300), and Chinese National 973 Grants (2010CB731401).

References

1. http://www.svcl.ucsd.edu/projects/anomaly/
2. Ali, S., Shah, M.: A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: CVPR (2007)
3. Beucher, S., Meyer, F.: The morphological approach to segmentation: the watershed transformation. Optical Engineering (1992)
4. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 282–295. Springer, Heidelberg (2010)
5. Brox, T., Rousson, M., Deriche, R., Weickert, J.: Colour, texture, and motion in level set based segmentation and tracking. Image Vis. Comput. (2010)
6. Bruh, A., Weickert, J., Schnörr, C.: Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods. Int'l J. Computer Vision (2005)
7. Carslaw, H., Jaeger, J.: Conduction of Heat in Solids. IEEE Trans. Pattern Analysis and Machine Intelligence (1986)
8. Chang, C., Lin, C.: LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 1–27 (2011)
9. Cremers, D., Soatto, S.: Motion competition: A variational approach to piecewise parametric motion segmentation. Int. J. Comput. Vis. (2005)
10. Cui, X., Liu, Q., Gao, M., Metaxas, D.N.: Abnormal detection using interaction energy potentials. In: CVPR (2011)
11. Edelsbrunner, H., Shah, N.: Incremental topological flipping works for regular triangulations. Algorithmica (1996)
12. Hu, M., Ali, S., Shah, M.: Learning motion patterns in crowded scenes using motion flow field. In: ICPR (2008)
13. Li, J., Gong, S., Xiang, T.: Scene segmentation for behaviour correlation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 383–395. Springer, Heidelberg (2008)
14. Lin, D., Grimson, E., Fisher, J.: Learning visual flows: a lie algebraic approach. In: CVPR (2009)
15. Loy, C.C., Xiang, T., Gong, S.: Multi-camera activity correlation analysis. In: CVPR (2009)
16. Lu, Z., Yang, X., Lin, W., Zha, H., Chen, X.: Inferring user image search goals under the implicit guidance of users. IEEE Trans. Circuits and Systems for Video Technology (2014)

17. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: CVPR (2009)
18. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Analysis and Machine Intelligence 12(7), 629–639 (1990)
19. Wang, H., Klaser, A., Schmid, C., Liu, C.: Action recognition by dense trajectories. In: CVPR (2011)
20. Weickert, J.: Anisotropic diffusion in image processing. Teubner, Stuttgart (1998)
21. Wu, S., Wong, H.: Crowd motion partitioning in a scattered motion field. IEEE Trans. Systems, Man, and Cybernetics (2012)
22. Wu, Y., Wang, Y., Jia, Y.: Adaptive diffusion flow active contours for image segmentation. Computer Vision and Image Understanding, 1421–1435 (2013)
23. Xu, L., Jia, J., Matsushita, Y.: Motion detail preserving optical flow estimation. IEEE Trans. Pattern Analysis and Machine Intelligence 34(9), 1744–1757 (2012)
24. Xu, T., Peng, P., Fang, X., Su, C., Wang, Y., Tian, Y., Zeng, W., Huang, T.: Single and multiple view detection, tracking and video analysis in crowded environments. In: AVSS (2012)
25. Zhan, B., Monekosso, D., Remagnino, P., Velastin, S., Xu, L.: Crowd analysis: a survey. Machine Vision and Applications (2008)
26. Zhou, B., Tang, X., Wang, X.: Coherent filtering: Detecting coherent motions from crowd clutters. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 857–871. Springer, Heidelberg (2012)
27. Zhou, B., Tang, X., Wang, X.: Measuring crowd collectiveness. In: CVPR (2013)