GrCS: Granular Computing-Based Crowd Segmentation

Ven Jyn Kok and Chee Seng Chan, *Senior Member, IEEE*

***Abstract*—Crowd segmentation is important in serving as the basis for a wide range of crowd analysis tasks such as density estimation and behavior understanding. However, due to interocclusions, perspective distortion, clutter background, and random crowd distribution, localizing crowd segments is technically a very challenging task. This paper proposes a novel crowd segmentation framework-based on granular computing (GrCS) to enable the problem of crowd segmentation to be conceptualized at different levels of granularity, and to map problems into computationally tractable subproblems. It shows that by exploiting the correlation among pixel granules, we are able to aggregate structurally similar pixels into meaningful atomic structure granules. This is useful in outlining natural boundaries between crowd and background (i.e., noncrowd) regions. From the structure granules, we infer the crowd and background regions by granular information classification. GrCS is scene-independent and can be applied effectively to crowd scenes with a variety of physical layout and crowdedness. Extensive experiments have been conducted on hundreds of real and synthetic crowd scenes. The results demonstrate that by exploiting the correlation among granules, we can outline the natural boundaries of structurally similar crowd and background regions necessary for crowd segmentation.**

***Index Terms*—Crowd analysis, crowd segmentation, granular computing (GrC).**

I. Introduction

VISUAL crowd surveillance at large public events such as concerts, parades, and rallies are common in cities worldwide. The mere existence of crowd has the prospect of progressing into a hazardous scene, for instance, the recent stampede in the Shanghai 2014 New Year's Eve revelry which claimed 36 innocent lives.[1] Alarmingly, with rapid urbanization around the world, the formation of crowd by chance is becoming a norm, e.g., crowds in train stations during rush hour. Consequently, crowd analysis has emerged as a crucial focus in visual surveillance for a proactive crowd management to anticipate disasters and provide support in good time.

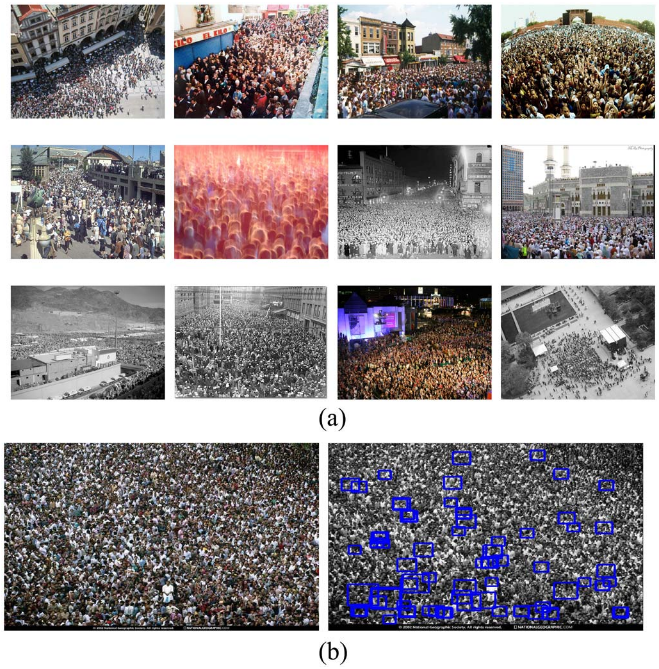

Fig. 1. (a) Crowd images with large variations in context of crowd, crowdedness, illumination conditions, background appearance, and physical layout of environment. (b) Left: crowd image. Right: person detection result using deformable part model [1]. The blue bounding boxes signify the detections result. False positive and fail detections are evident in the image.

Generally in crowd analysis, crowd segmentation serves as one of the fundamental steps for further analysis, such as crowd density estimation [2], crowd behavior analysis [3], and person tracking in crowd [4]. This is also stated in [5] and [6] that the localization of crowd segments is required prior to visual tasks such as tracking or behavior understanding.

As illustrated in Fig. 1(a), we show that crowd segmentation is a challenging task in computer vision due to the following.

1) *Context Variations of Crowd:* Crowd across all scenes varies drastically because of different crowdedness, illumination conditions, interocclusions and variations of clothing and poses. At the same time, perspective distortions due to camera orientation and position implicate changes of scales of individuals within a crowd.
2) *Cluttered Background in a Crowded Scene:* In all these images, it is observed that the background (noncrowd) regions (i.e., trees, grasses, and buildings) are cluttered in such a way that they resemble crowd region,

Manuscript received May 19, 2015; revised September 15, 2015; accepted March 1, 2016. Date of publication March 16, 2016; date of current version April 13, 2017. This work was supported by the High Impact MoHE through the Ministry of Higher Education Malaysia under Grant UM.C/625/1/HIR/MoE/FCSIT/08 and Grant H-22001-00-B00008. This paper was recommended by Associate Editor Y. Yang.

The authors are with the Centre of Image and Signal Processing, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia (e-mail: venjyn.kok@siswa.um.edu.my; cs.chan@um.edu.my).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2016.2538765

[1] BBC News: http://www.bbc.com/news/world-asia-china-30646918.

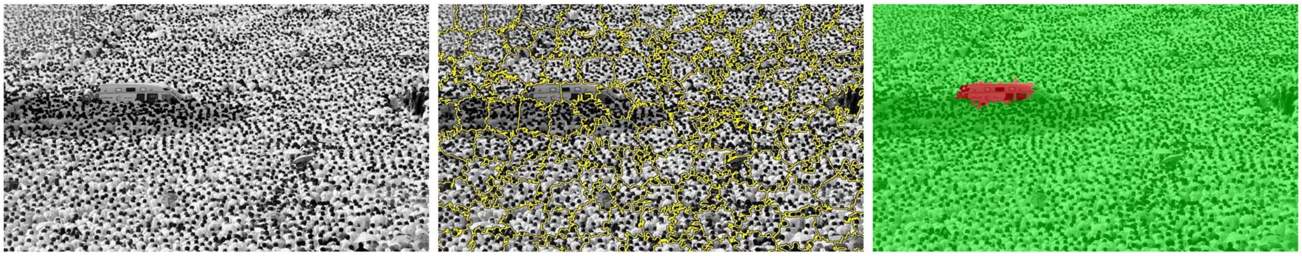

Fig. 2. GrCS. Left: dense crowd scene image. Middle: image segmented into structurally similar atomic clusters (structure granules), shown as regions within yellow outline. Perimeters of crowd and background are distinctively separated. Right: crowd and background regions segmentation achieved via classification of structure granules. A vehicle is outlined and classified as background region (shown as red overlay). Best viewed in color.

where finding a good separation between crowd and background regions is a daunting task.

3) *Unconstrained Physical Layout of the Environment:* The formation of crowd across different scenes is inherently dependent on constraints imposed by the environmental layout.

Due to the aforementioned issues, current crowd segmentation methods that detect each person in crowd are still in their infancy stage. This is an ongoing research problem [5]. We show an example in Fig. 1(b) of a dense crowd scene where individuals in crowd are severely interoccluded and mostly cannot be detected. Hence, other methods [2], [7]–[9] infer crowd segments by learning the textures of crowd scenes either using regular pixel-grid or overlapping multiscale pixel-grid (i.e., numerous range of neighboring pixels) at each pixel. In spite of the promising results, the use of pixel-grid imposes some constraints on inferring crowd segments. In the former, crowd images are divided into regular pixel grids where an optimized boundary adherence of crowd segments across different scene is difficult to achieve. In the latter, an antecedent version of the regular pixel-grid, namely, multiscale pixel grid is proposed to cope with crowd variation across different scenes. Since it is leveraging on its antecedent, conformation to varying crowd segments remains unresolved. With a smaller pixel-grid, localization accuracy is better with less probability of patch consisting both crowd and background regions; whereas a larger pixel-grid covers wider regions for analysis of structure [6], [7].

In this paper, we strive to exploit the correlation among image granules at different levels of granularity with the hope that granulation can alleviate the aforementioned constraints. The dichotomy articulated by Moravec [10] between humans and machines regarding the easiness and complexity in solving different problems remains valid today. Specifically, machines perform poorly in tasks that are seemingly effortless and natural for humans (i.e., recognizing crowd regions), but can easily solve problems that humans find challenging (i.e., numerical computation). One key advantage of the human mind has over a machine in cognition is the ability to segment visual information into meaningful units of analysis effortlessly [11]. More remarkably, this is achieved in vivid detail; disregarding the orientation, color intensity, and deformation present. We seek to transfer this structured problem solving ability of human cognition into crowd segmentation system, with the aim of alleviating the complexity to infer crowd segments. Interestingly, granular computing (GrC), an emerging computing paradigm of information processing [12], simulates human cognitive process by enabling abstraction on the essential details at different granularities. That is, correlations among granules are explored to solve various research problems in computer. So unlike conventional approaches [7], [9], the concept of GrC is incorporated in our approach in the form of granules, thereby, honoring the correlations of structures in crowd scenes from pixel level to crowd and background level. This is to mitigate the effects of the aforementioned issues (i.e., context variations of crowd, cluttered background, and unconstrained physical layout of the environment) for an effective crowd segmentation. The utilization of granules obviates the difficulty to segregate individuals in crowd due to context variations of crowd by enabling inference of crowd and background regions based on local structures. To circumvent the effects of cluttered background and unconstrained physical layout of the environment, we believe the key is to study the correlations among granules to represent structurally similar regions in crowd scene images.

The notion of simplifying an image scene into structurally meaningful atomic regions (i.e., granules) is generally unprecedented in the existing crowd segmentation studies. It is important to have granulation that is able to adapt in different crowd structures in scenes due to varying crowdedness, perspective distortion, severe interocclusion, and cluttered background for a better crowd segmentation. As an example, Fig. 2 illustrates a crowd scene with severe interocclusion between individuals and the scale of individuals vary drastically due to the perspective and position of camera. Even so humans are able to distinguish the vehicle within the crowd with ease. Similarly, using the proposed method, it is observed in Fig. 2 (middle) that each granule (i.e., regions within the yellow outline) encompasses only a single context (i.e., crowd or background). This serves as a meaningful primitive region to infer the corresponding context [as shown in Fig. 2 (right)]. Accordingly, the vehicle (red overlay) surrounded by a swarm of crowd (green overlay) is effectively singled out despite severe occlusion and highly textured scene. In addition, crowd regions segmentation is illumination invariant.

The contributions of this paper are summarized into three main aspects.

1) *GrC-Based Crowd Segmentation:* We introduce a novel crowd segmentation framework using the concept and principles of GrC. GrC is incorporated in this paper to conceptualize crowd segmentation problems on different granularity similar to human cognition in problem

solving, with the intention of mapping it into computationally tractable subproblems.

2) *Adaptive Crowd Scene Granulation*: Contrary to regular-grid representation, we study the correlation among granules to represent structurally similar regions in crowd scene images to infer the crowd and background regions. This is required because occasionally structures of background in a scene image resemble crowd regions, which lead to vague outline between the crowd and the background. The constructed granules are scene-independent, conforming to the boundaries of crowd segments.

3) *Dataset With Ground Truth Annotations*: In order to facilitate this paper, over a hundred real crowd scenes are carefully annotated at pixel-level (i.e., each pixel is assigned a class label), with careful labeling around complex boundaries of crowd and background to provide precise boundaries and localization information. The dataset and ground truth shall be made public to support future crowd segmentation analysis.

The rest of the paper is organized as follows. In Section II, we provide literature relevant to this paper in crowd segmentation. Section III describes the proposed framework of crowd segmentation by modeling crowd scenes with GrC. The experimental results are presented and discussed in Section IV, followed by the possible extensions of this paper and the conclusion in Section V.

II. Related Work

Existing work on dense crowd analysis tends to exploit the collective coordination of crowd by analyzing crowd through analogies with studies in fluid dynamic [13]–[15] or treating a crowd as a collective entity [4], [16]–[19]. A number of approaches have been proposed for crowd segmentation [20], [21]. These studies lean toward analyzing dynamic crowd segmentation for crowd flow segmentation [14], [22], crowd behavior understanding [3], person tracking [4], [23], anomaly segmentation [18], [24], and crowd counting [25]. Crowd is generally studied with emphasis given on the evolution of its motions in an environment. Commonly, the approaches perform background subtraction [26], [27] or estimate collective crowd motion along temporal axis [4], [14] to identify the crowd segments. Such approaches are susceptible to false segmentation in cluttered environment with other moving entities (e.g., moving vehicles and waving trees), as well as limited to localizing crowd with variations in collective motion. Observations by Helbing *et al.* [28] highlighted that stationary crowd (e.g., spectators of a speech) implicitly influenced the motion flow of dynamic crowd, where crowd maneuver around stationary crowd to avoid collisions. Thus, it is of equal importance to detect stationary crowd segments for a complete crowd surveillance system. In this paper, we use spatial cues that are generally available in dynamic and stationary crowd for segmentation.

There is another branch of crowd segmentation research that utilizes the collectiveness of crowd as well. It exploits the texture patterns of collective crowd regions to detect crowd, regardless of the motion variations. Due to severe interocclusions and perspective distortion in dense crowd scene [as illustrated in Fig. 1(b)], appearance-based approaches which include head and shoulder segmentation are not feasible [2]. To alleviate the need of person detection in a crowd, imagery of crowd scenes is partitioned into regular pixel-grid with the purpose of achieving local texture consistency and is treated as a texture analysis problem. A study by [29] shows that crowd regions carry strong cue of texture variations. Arandjelovic [7] proposed an image-based crowd segmentation method using low-level local feature from single crowd image. Each pixel response is defined using multiscale pixel-grid, where the computation of the probability of a pixel-grid being a crowd region is based on a predefined average number of SIFT word segmentation per image area. Using similar approach, Idrees *et al.* [2] partition crowd scene into pixel-grid to construct a confidence map of crowd regions. In another study, Fagette *et al.* [9] performed crowd segmentation by retrieving multiscale pixel-grid texture features from crowd scene. Binary classification is conducted to infer crowd regions in image. Since these methods use regular pixel-grid, the representation is not adaptive to the random distribution of crowd perimeters and extent of crowdedness in real-world scene. Also, it is unclear how well they can be generalized to arbitrary crowd scenes. The number of layers in multiscale pixel grid is scene-dependent; it has to be empirically defined for each public crowd scene to optimize adherence to the arbitrary crowd distribution. In contrast to the aforementioned studies, we propose using the concept and principles of GrC to conceptualize crowd segmentation problem on different granularity to explore the affinity of structures in crowd scenes. The correlations of structures are utilized to define explicitly atomic regions with outlines that conform to the arbitrary perimeter between crowd and background regions. This method can be well generalized into real-world crowd scenes with varying geometric structures of environment and crowdedness. In addition, regions of crowd with different scales of individuals due to varying camera positions are segregated into meaningful structurally coherent regions.

Although not explicitly for the purpose of crowd segmentation, work that uses the human-centricity of GrC for image context understanding includes [30] and [31]. Pal *et al.* [30] applied GrC together with rough sets to perform grayscale image segmentation. Their method defines nonoverlapping pixel-grid of different sizes as granules to quantify the object-background regions in images. Rizzi and Del Vescovo [31] proposed to decompose each image into segments (i.e., granules) and map the correlation among image segments for image classification. The method performs abstractions to cope with a wide set of problem instances of image classification. In this paper, we propose to transfer the concept of GrC that is able to simulate human cognitive process to explore the correlation of structures from pixel level in crowd scenes specifically to infer crowd and background segments. The use of this structured problem-solving method is to alleviate the difficulty of defining the natural boundaries between crowd and background due to varying crowdedness, perspective distortion, severe interocclusion, and cluttered background.

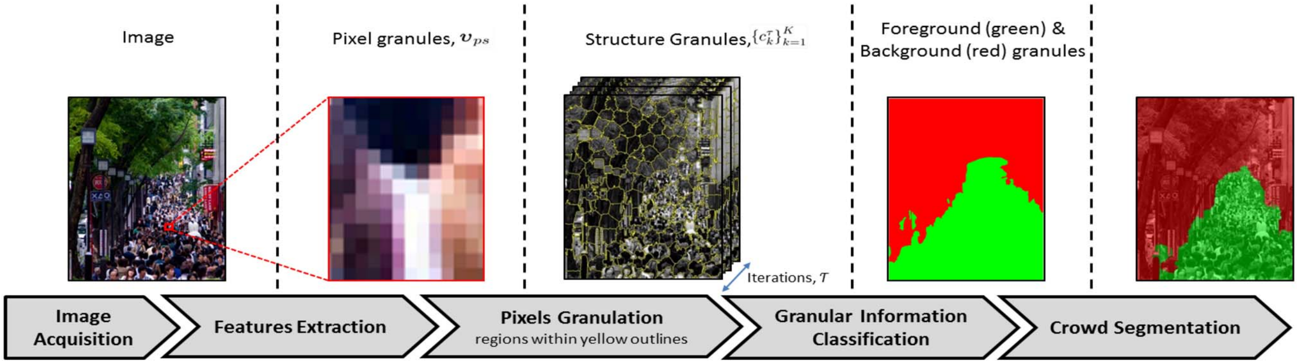
Image
Pixel granules, $\boldsymbol{v}_{ps}$
Structure Granules, $\{c_k^\tau\}_{k=1}^K$
Foreground (green) & Background (red) granules
Iterations, τ
Image Acquisition
Features Extraction
Pixels Granulation
regions within yellow outlines
Granular Information Classification
Crowd Segmentation

Fig. 3. GrCS framework. An illustration of the key steps and the different levels of granularity of image in granular computing-based crowd segmentation.

III. Proposed Framework

The key steps of GrC-based crowd segmentation (GrCS) framework are illustrated in Fig. 3, where granules are the basic elements and each level represents different levels of granularity. This is to simulate the ability of humans to conceptualize at different granularity levels with the intention of mapping problems into computationally tractable subproblems.

In our context, a dense crowd image, $I = [\boldsymbol{v}_{ps}] \in \mathbb{R}^{N\times S}$, where N is the number of pixels in an image and S is the number of features for each pixel, p. Each pixel, p, in an image is the basic granule (i.e., pixel granule), represented as a feature vector, $\boldsymbol{v}_{ps} = (v_{p1}, \dots, v_{ps}, \dots, v_{pS})^\top \in \mathbb{R}^{N\times S}$, where $p = \{1, \dots, N\}$ and $s = \{1, \dots, S\}$. The feature vector, $\boldsymbol{v}_{ps}$ is formed by the concatenation of S features. Aggregation of the pixel granules (granulation process) with similarity of feature vector, $\boldsymbol{v}_{ps}$, will form a higher level set of granules (i.e., structure granules). We anticipate that these structure granules are structurally coherent atomic regions in the image that conform to the natural boundaries between different structures of crowd and background. The key idea of the atomic regions is to have a pixel aggregation process versatile to different crowd scenes, and so this will best categorize the diverse structures in the scene for robust crowd segmentation. From the structure granules, we pose crowd segmentation task as a classification problem to construct granulated view of foreground (i.e., crowd in the context of this paper) and background (i.e., sky, buildings, grasses, etc.) granules.

A. Pixel Granules

The finest level of granules represents the most basic aspect of crowd scenes, which is the pixel information: pixel intensity and spatial position in image plane. However, due to the complexity of discerning cluttered background from crowd, texture features are introduced in our proposed framework to increase the discriminative ability for texture differentiation. This is because background region such as carpet grass, can be easily misinterpreted as crowd region. Co-occurrence of multiple features, v_{ps}, is thus essential to complement the insufficiencies of other features. Similar strategy is used by humans where one's cognition uses existing information to understand a new subject matter.

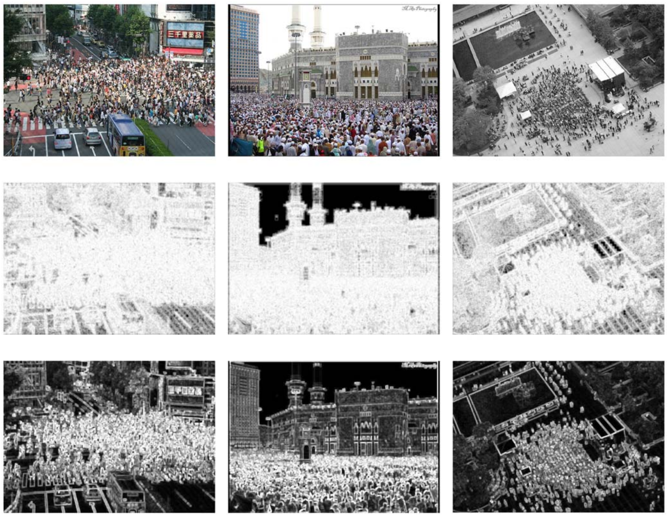

Fig. 4. Top row: example crowd scene images. Middle row: entropy images using 5 × 5 neighborhood. Bottom row: images of LRI using 5 × 5 neighborhood.

In this paper, the texture features are represented by the widely used local binary pattern (LBP) [32] and local range of intensity (LRI). Nevertheless, the proposed framework is not restricted to these sets of features employed in this paper. Diverse sets of features can be exploited to enhance and adapt to various image segmentation task.

1) Local Binary Patterns: LBP is adopted to capture the microstructure of local region by which we analyze the raw low-level spatial pattern of crowd. LBP is computationally simple yet a practical gray-level invariant approach to summarize local gray-level structure. Employing LBP to capture the dense microstructures in crowd regions, such as lines and edges formed by a mass of crowd can serve as a good indicator of the presence of crowd. In this paper, we implement an extended version of LBP operator known as uniform patterns [33] to cope with variance in rotation of captured microstructures.

Given pixels within a crowd image, I, we use a 3×3 circularly symmetric local neighborhood, i.e., eight sampling points centering each pixel of interest. The neighboring pixels are thresholded against the value of the corresponding pixel of interest and the value associated with the local neighborhood is concatenated to form a binary pattern. Texture descriptors of LBP uniform pattern correspond to the histogram formed by uniform and nonuniform binary pattern labels.

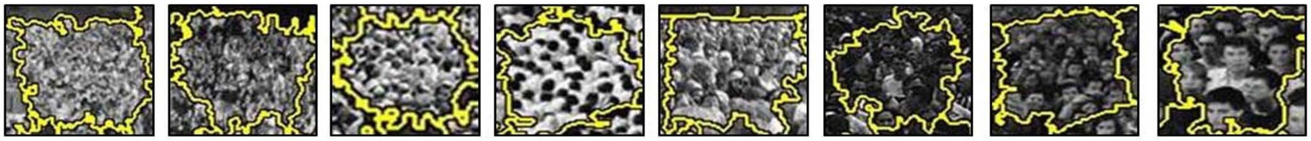

Fig. 5. Sample crowd structure granules with variabilities in terms of illumination, scale of persons per area, perspective, and interocclusion. Note that the scale of person per image area increases when view from left to right.

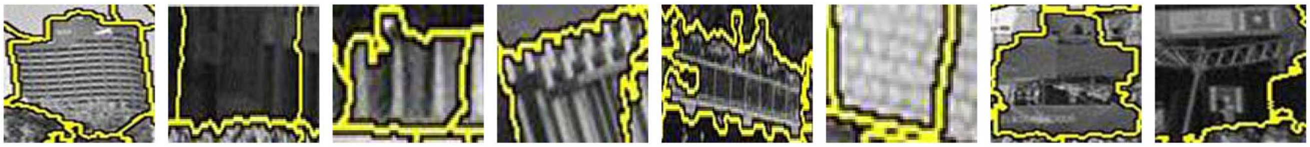

Fig. 6. Sample background structure granules with variabilities in terms of illumination and texture patterns.

2) Local Range of Intensity: The LRI is defined as the difference between the extrema (maximum and minimum) intensity values of a local neighborhood centering each pixel of interest. The notion of using local intensity variation to solve visual analysis problem in computer vision has been used by several researchers, such as [34]. We observed that crowd segments tend to exhibit larger range of intensity variation in comparison to background (i.e., noncrowd) regions, mainly due to varying individual appearances. Instead of using the conventional entropy measure [35], LRI is deemed more effective in quantifying the information content (statistical randomness) of local regions based on intensity variation in crowd scenes. As illustrated in Fig. 4 (middle row), entropy is susceptible to image noise and background clutters such as grass, trees, and buildings which produce similar entropy variation. We anticipate that by adopting LRI feature, the common hurdle of discriminating crowd regions from textured background in existing literatures can be eliminated.

B. Structure Granulars

Crowdedness and the distribution of crowd in crowd scenes are rarely uniform due to the physical layout of the environment and/or the viewpoint of the scene captured. Worse still, the textures of background (e.g., building structures and trees) and crowd (as a result of gait, clothing, and shape of person) vary drastically, as illustrated in Figs. 5 and 6, respectively. It, thus, can lead to vague boundaries between crowd and background [as shown in Fig. 4 (top row)]. On a finer scale, the variability of crowd region corresponds to a unison structure [3]. The structures can be intimately governed by the structure granules to outline the perimeters of coherent crowd structure and background. We explore the correlation among pixel granules for granulation, with the aim of forming structurally uniform structure granules adhering to the natural edges of crowd scenes for analysis. This is analogous to how human brains perceive and process visual information; one does not focus on individual pixels, instead, grouping them into semantically meaningful forms to understand the image. In GrC, granulation process is the aggregation of smaller and lower level granules into a larger and higher level granules according to their similar characteristics [36]. In terms of coarse and fine relationship [37], [38], pixel granules are the refinement of the structure granules where every pixel granule is contained in the structure granular level.

Structure granules are constructed by aggregating pixels (i.e., pixel granules) with similar structure feature vector, adapting the pixels clustering approach [39] with refinement. The refinement is necessary in this paper to enable auto-adaptability of structure granules to conform to the structure of local atomic regions. This is different from the existing cluster analysis solutions [40]–[43] that use distance measures such as the similarity between two granules defined as an average distance between subgranules. More precisely, we commence by initializing the number of structure granules, K, in an image, I. The greater the value of K, the finer is the crowd image partitioned, generating more structure granules. The initial structure granule centers, $\{c_k\}_{k=1}^{K}$, for an image, I, with N pixels is regularly seeded at a grid interval $G = \sqrt{(N/K)}$. Each c_k is represented by a feature vector, $\boldsymbol{\upsilon}_{c_k s} = (v_{c_k 1}, \ldots, v_{c_k s}, \ldots, v_{c_k S})^{\top}$. Within the search region $(2G \times 2G)$ for each structure granule center, c_k, similarity of each feature, $v_{c_k s} \in \boldsymbol{\upsilon}_{c_k s}$ of structure granule center, c_k, with pixel, p, within the respective search region is defined as

$$d_{ps}^{\tau} = \left\| v_{c_k s} - v_{ps} \right\|_2. \tag{1}$$

Anchor pixels for a structure granule are the pixels (i.e., pixel granules) that are associated with a specific structure granule center. The anchor pixels for each structure granule center, c_k are obtained by iteratively associating pixels in the image, I, to the nearest structure granule center using the shortest pairwise distance. The pairwise distance measure, D^{τ}, is formulated as

$$D^{\tau} = \sum_{s=1}^{S} \frac{d_{ps}^{\tau}}{m_s^{\tau-1}}\ ,\ \tau \in \{1, 2, 3, \ldots\} \tag{2}$$

$$\text{where } m_s^{\tau-1} = \max\left(m_s^{\tau-2}, d_{mps}^{\tau-1}\right) \tag{3}$$

$$d_{mps}^{\tau-1} = \max\left\{d_{ps}^{\tau-1}, \forall\ p \in 2G \times 2G\right\} \tag{4}$$

such that $d_{mps}^{\tau-1}$ is the maximum distance of a structure granule center, c_k, with the pixels within the respective search region at iteration $\tau - 1$. The anchor pixels together with its respective structure granule center will form a structure granule [i.e., a region within yellow outlines as shown in Fig. 2 (middle)].

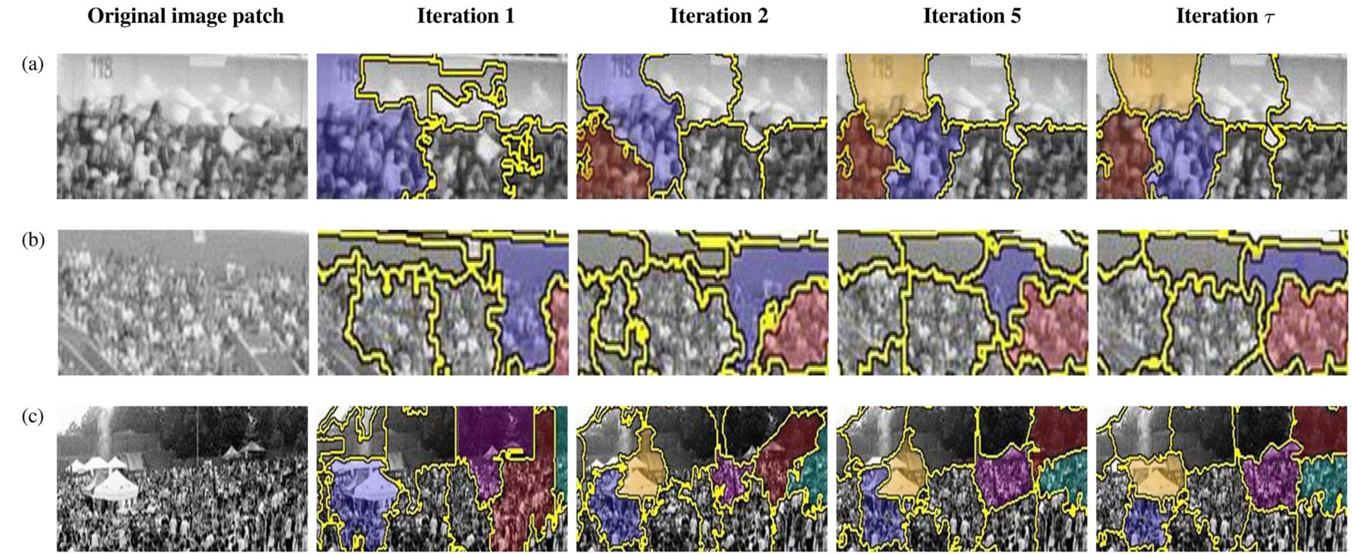

Fig. 7. Transition of structure granules at each iteration. Structure granules with significant localization improvement are overlaid with different colors (i.e., purple, red, yellow, pink, and green) to enhance the visualization of the improved separation between crowd and background regions over the iterations. (a) At iteration 1, it can be observed that the structure granule with purple overlay consists of crowd and background regions. After several iterations, at iteration τ, high localization of structure granules is achieved where crowd and background regions are well separated. That is, the structure granule with yellow overlay consists of background region only, whereas the structure granules with purple and red overlay consist of crowd region only. Similarly, (b) and (c) show the localization improvement of structure granules on two different crowd scenes.

Algorithm 1 Construction of Structure Granules

```
Require: An initial set of structure granule centers, $\{c_k\}_{k=1}^K \in I$, regularly seeded at a grid interval G and number of iterations, $\tau$, where $\tau \in \mathbb{N}$
Ensure: A set of new structure granule centers, $\left\{c_k^\tau\right\}_{k=1}^K \in I$
  repeat
    for each structure granule center, $c_k$ do
      for each feature, $v_{ps}$ do
        Compute $d_{ps}^\tau$ as to Eq. 1;
      end for
      Compute $D^\tau$ as to Eq. 2;
    end for
    Associate pixels to the nearest structure granule center, $c_k$, by $D^\tau$;
    Update set of structure granule centers, $\left\{c_k^\tau\right\}_{k=1}^K \in I$;
  until Separation between crowd and background regions is optimized
```

Note that, $m_s^{\tau-1}$ is a novel adaptive varying scaling parameter in the GrCS. This is in contrast to the constant scaling parameter scheme employed in [39]. Due to complex texture variations in crowd scenes, compactness of structure granules in terms of crowd and background boundary adherences is essential to provide an informative granulated view to comprehend scene context. Inspired by [44], in this paper, at each iteration, τ, the selection of our scaling parameter, $m_s^{\tau-1}$ for each d_{ps}^τ is computed by studying the local structure of the anchor pixels with structure granule center, c_k from previous iterations (3). Using a scaling parameter that honors the local structures of structure granule enables self-tuning of the pixel-to-granule center distances according to the local statistic of different features of the granule. The adaptive varying scaling parameters automatically find, at each iteration, the scales that enable high structure affinity of pixels within each structure granule and low structure affinity across neighboring granules for each structure feature, v_{ps} [as shown in (2)]. We demonstrate in Section IV-C that this in turn facilitates distinct separation adhere to the natural boundaries between crowd and background regions in crowd images.

A set of new structure granule centers, $\{c_k^\tau\}_{k=1}^K \in I$ is defined at each iteration, where each c_k^τ is represented by the average of feature vector, $\boldsymbol{\upsilon}_{ps}$ of anchor pixels within the respective clusters. The optimized clusters constructed at this stage form a vocabulary of structure granules providing the granular description of the crowd image. Fig. 7 shows examples of the transition of structure granules at each iteration. As the number of iterations, τ, increases, the localization of structure granules improves with optimized separation between crowd and background regions and eventually converges.[2] The pseudo code in Algorithm 1 describes the iterative process to construct structure granules given crowd scene image.

C. Crowd Segmentation

Given the structure granular, our objective is to achieve robust crowd regions inference, and so we pose crowd segmentation as a classification problem. We wish to take into

[2] The value of d_{ps}^τ and $m_s^{\tau-1}$ in (2) is always positive, thus, the D^τ is a series of positive terms. Since the number of features, S is finite, the number of terms in the series is also finite. Consequently, the sum of the partial terms of D^τ is a monotonically increasing sequence. In order to prove the sequence of partial terms of D^τ converges, we have to verify that the sequence of partial terms of D^τ is bounded or not. Being the sum of the partial terms of D^τ is a monotonically increasing sequence, it is bounded below. At the same time, due to the finite number of features, S, the largest sum of partial terms can be bounded by any real number, which proves that the sum of partial terms of D^τ is bounded above. This indicates that the sum of the partial terms of D^τ is bounded. Hence, the series D^τ converges.

consideration of the variability (as shown in Figs. 5 and 6) to infer class label (i.e., crowd or background) of input structure granules. Random forest algorithm is implemented due to the high generalization power yet able to avoid model overfitting, and being fast during training and testing [45], [46]. Each random decision tree is generated by a random subset, $\mathbf{E}'$ of the labeled training structure granules with replacement. At a specific leaf node, the labeled training structure granules, $\mathbf{E}'_{\text{node}} = \{\mathbf{c}_i, l_i\}_{i=1}^{A}$ are recursively split into left, $\mathbf{E}'_{\text{left}}$ and right, $\mathbf{E}'_{\text{right}}$ node subsets, where $\mathbf{c}_i$ is a feature vector of structure granule, l_i is the corresponding class label (i.e., crowd or background) and A is the number of training samples. The splitting is done given a set of thresholds, $\mathbf{T}$ and splitting function, f as

$$\mathbf{E}'_{\text{left}} = \left\{\mathbf{c}_i \in \mathbf{E}'_{\text{node}} | f(\mathbf{c}_i) < t\right\} \tag{5}$$

$$\mathbf{E}'_{\text{right}} = \mathbf{E}'_{\text{node}} \setminus \mathbf{E}'_{\text{left}}. \tag{6}$$

At each leaf node, the threshold, $t \in \mathbf{T}$ that best split the training granules with maximized gain, ΔG is selected

$$\Delta G = -\frac{\left|\mathbf{E}'_{\text{left}}\right|}{\left|\mathbf{E}'_{\text{left}}\right| + \left|\mathbf{E}'_{\text{right}}\right|} \cdot J_{\text{left}} - \frac{\left|\mathbf{E}'_{\text{right}}\right|}{\left|\mathbf{E}'_{\text{left}}\right| + \left|\mathbf{E}'_{\text{right}}\right|} \cdot J_{\text{right}} \tag{7}$$

where $J = -\sum_l p(l_i) \cdot (1 - p(l_i))$ is the Gini index and $p(l_i)$ is the class probability for l_i. Class labels of Q unseen structure granules, $\{\mathbf{c}_j\}_{j=1}^{Q}$ are inferred by traversing down all β decision trees. Each leaf node of a decision tree returns a prediction of the class label, l_j with class probability distribution $p(l_j|\mathbf{c}_j)$. The final class label (i.e., crowd or background) of structure granule is equated by averaging the probability estimate from each decision tree, defined as

$$l_j^* = \arg\max_{l_j} \frac{1}{\beta} \sum^{\beta} p_\beta\left(l_j|\mathbf{c}_j\right). \tag{8}$$

The class labels of structure granules in an unseen image computed are used to infer the foreground (i.e., crowd) and background granules in the crowd scene image. The construction of foreground and background granules is a process of granulation. Such granulation process provides a granulated view of the image which is intended to be on par with the way a human would annotate crowd and background regions in a crowd scene.

IV. Experimental Results

Evaluations on the GrCS framework are conducted on benchmark datasets of real and synthetic crowd scenes obtained from [2], [7], [9], and [47]. These datasets consist of crowd scenes in various events, such as parades, concerts, and rallies. The crowd in these datasets varies in terms of illuminations, crowdedness, and perspectives. The resolutions of the images range from 240 × 320 to 1024 × 1024. To evaluate the efficiency of the proposed framework (i.e., conform precisely to the boundaries between crowd and background regions), we are persuaded to annotate manually the crowd and background regions as ground truth for real crowd scenes. Ground truth of each image is annotated at the pixel level, with careful labeling around complex boundaries of crowd. Examples of ground truth annotation are illustrated in the second row of Fig. 11. The ground truth for synthetic crowd images is generated by the Agoraset crowd generator [48]. Each ground truth segment is highly accurate, i.e., adhering to the precise outline between crowd and background, where it would be almost infeasible to achieve manually [49].

A. Experiment Settings

In all the experiments, we set the number of structure granules, $K = 200$ and the number of iterations, $\tau = 10$ which enables high localization of structure granules with adequate separation between crowd and background regions. The varying scaling parameter, $m_s^{\tau-1}$, for each d_{ps}^{τ} is initialized as $m_s^0 = 10$. Evaluation with different values of initialization constant generates consistent structure granules adhering to the boundaries of crowd. To construct granulated view of foreground (i.e., crowd) and background granules, we use random forest classifier with the number of random decision trees, $\beta = 2000$ and 100 randomly sampled variable at each split node. Crowd scene dataset is randomly divided into sets of 40 images to perform fivefold cross-validation to avoid bias. Each structure granule is represented by the mean of feature descriptor, $\boldsymbol{\upsilon}_{ps}$ from pixel granulation, with entropy measures and pixel-wise SIFT [50] features of anchor pixels and structure granule center, c_k. The feature responses of crowd and background structure granules are combined as input to train the random forest classifier.

B. Crowd Segmentation

We demonstrate the effectiveness and robustness of the GrCS for real and synthetic crowd scenes understanding in the application of crowd segmentation. Evaluations are conducted by benchmarking this paper with the multiscale pixel grid approaches [7], [9]. Each evaluation is compared against the benchmark dataset used in each respective approach.

Segmented crowd regions are shown as green overlay, whereas background regions with red overlay. For quantitative evaluation, the F-score measure is used according to the well-known PASCAL challenge [51] to evaluate the accuracy of crowd segmentation by overlapping it with ground truth annotation (as per pixel basis).

1) Synthetic Crowd Scenes: Evaluations on synthetic crowd scenes are conducted to gauge the applicability of GrCS. Crowd segmentation on synthetic crowd scenes is less taxing given the flat background texture. We show that when scales of person in crowd are uniform (as shown in row 1 of Fig. 8), GrCS achieves similar or better F-score than [9] in classifying crowd and background regions. However, on crowd scenes with perspective distortion and varying crowdedness, GrCS is more superior at discerning crowd and background regions, as illustrated in rows 2–4 in Fig. 8. This is not the case for [9], where their segmentation does not accurately highlight the person in crowd. GrCS framework achieves good segmentation of individuals in crowd, simply because novel adaptive varying

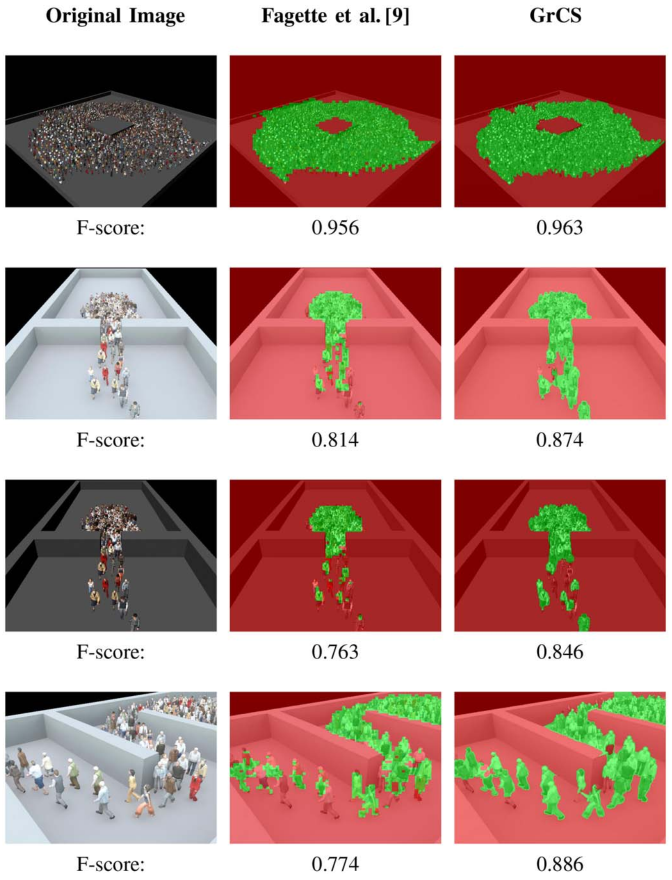

Fig. 8. Comparative results of crowd segmentation on synthetic crowd scenes with [9]. Best viewed in color.

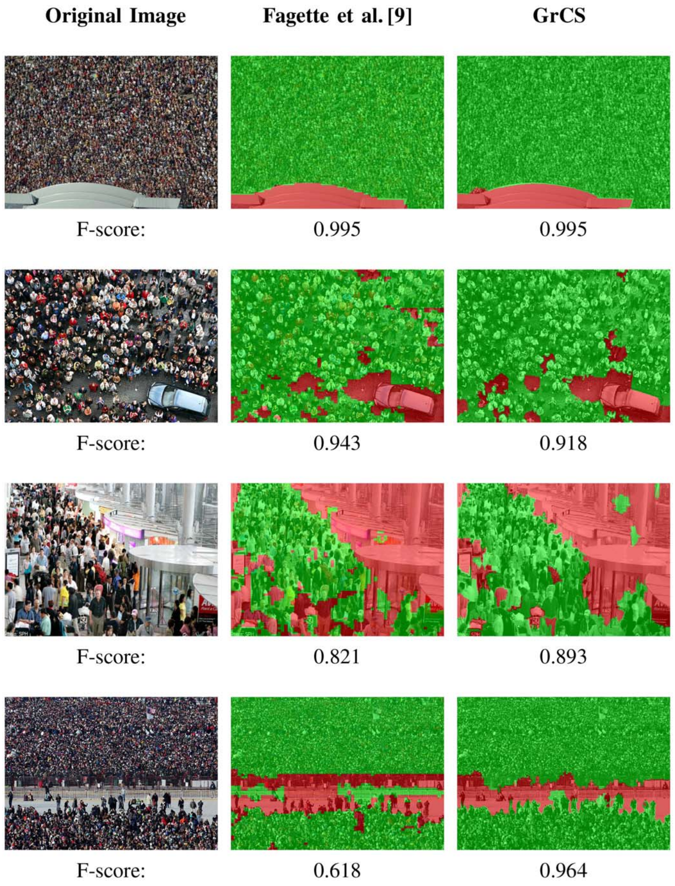

Fig. 9. Comparative results of crowd segmentation on real crowd scenes with [9]. Best viewed in color.

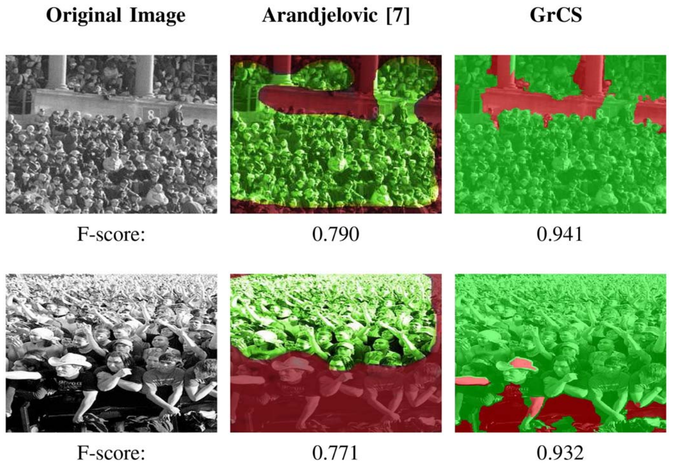

Fig. 10. Comparative results of crowd segmentation on real crowd scenes with [7]. Best viewed in color.

scaling parameter enables conformation of each structure granules adhering to the complex boundaries between crowd and background. With optimized structure granules, individuals in sparse crowd are adequately segmented.

2) Real Crowd Scenes: Contrary to synthetic crowd scene, real crowd scenes are more challenging given the varying crowd context, cluttered background and unconstrained physical layout of environment. We further test the GrCS on real crowd scenes such as shown in Figs. 9 and 10. Analogous with synthetic crowd scene, evaluation on real crowd scenes shows that when the scale of a person in a crowd are uniform where each person occupies only few pixels, the GrCS is comparable with [9] (as shown in row 1 of Fig. 9). Evaluation on crowd scenes with perspective distortion and different crowdedness shows that our proposed method is able to cope better with varying scales of individuals in crowd to discern crowd and background regions in comparison to [9] and [7], as illustrated in row 3 of Fig. 9 and row 2 of Fig. 10. This is because the correlation among granules is exploited to represent structurally similar regions in crowd scenes and the variability of structures is taken into consideration during the granular information classification.

Background textures have significant influence on the crowd segmentation performance. For example in row 4 of Fig. 9, Fagette *et al.* [9] failed to segment crowd that has been overlaid by the steel barricades. Worst still, due to the crowd-like structure of steel barricade, it is mistakenly inferred as crowd segment. On the contrary, the GrCS is able to infer the actual crowd and background (i.e., steel barricade) segments. Arbitrary distribution of crowd and background regions is effectively outlined using GrCS (as shown in the fourth row of Fig. 9 and the first row of Fig. 10). It provides a more natural representation of crowd and background regions in comparison with [7] and [9]. This essentially illustrates the advantage of granulation process that is adaptive to different

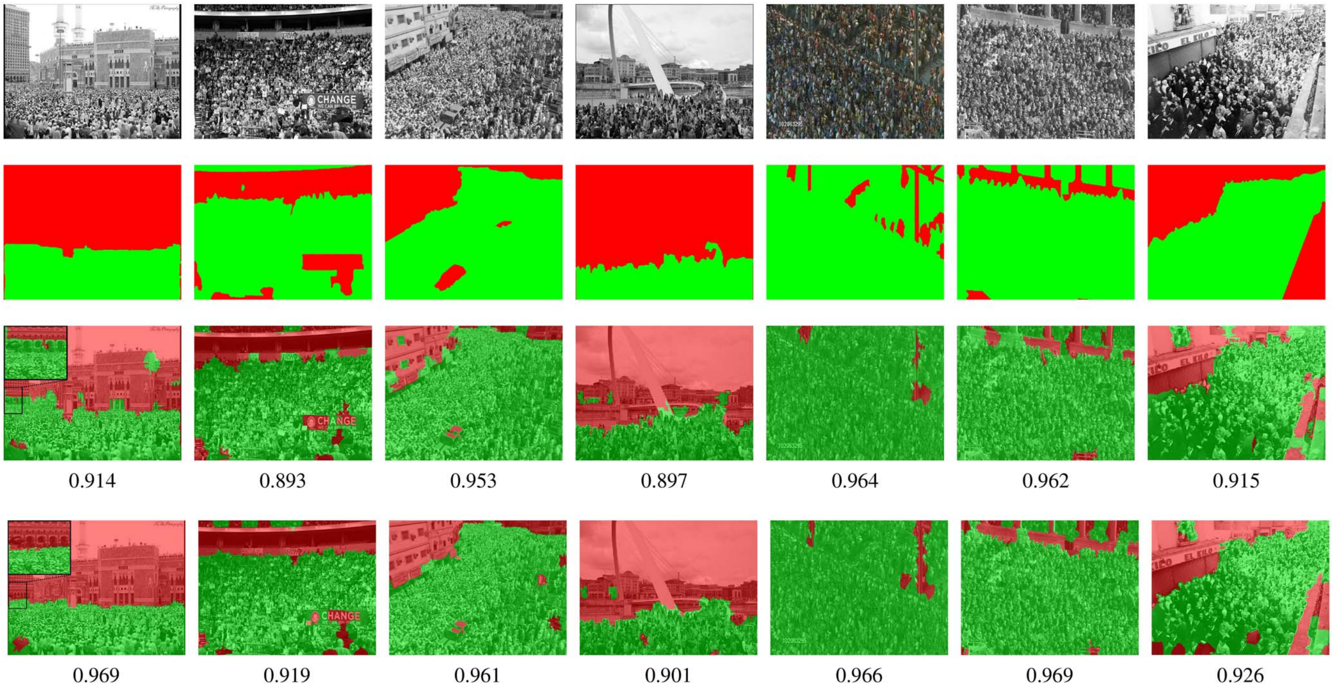

Fig. 11. Comparative results of crowd segmentation on real crowd scenes with SLIC [39]. First row: real crowd scenes. Second row: ground truth annotations. Third row: crowd segmentation using SLIC [39] with the respective F-score measures. Forth row: GrCS (adaptive varying scaling parameter) with the respective F-score measures. Best viewed in color.

crowd structure in scenes over pixel-grid. In addition, GrCS framework which utilizes LRI feature is less susceptible to false segmentation.

C. Adaptive Varying Scaling Parameters

We compare the GrCS approach using adaptive varying scaling parameter against constant scaling parameter [39] on real crowd scenes. Examples of the ground truth and the segmentations results in comparison are shown in Fig. 11. Using constant scaling parameter, we can well separate crowd from uncluttered background regions, but it performs poorly on complex and cluttered background. This is observed in the third row first column of Fig. 11, where the ambiguous perimeter between crowd and building structure is inaccurately outlined. Moreover, since some of the structure granules constructed using [39] contain both crowd and background texture (as shown in Fig. 12), it is understandable that the granular information is prone to classification error. As illustrated in the first and second column of Fig. 11, constant scaling parameter approach leads to textured regions of buildings inaccurately inferred as crowd, whereas the GrCS approach is able to define crowd and background regions corresponding to ground truth annotation.

To comprehend the influence of adaptive scaling parameter on crowd segmentation, Fig. 12 provides visualization of the ground truth and the comparative results of structure granules using the novel adaptive varying scaling parameters and the constant scaling parameter [39] (taken from random regions in crowd scenes from the first two columns in Fig. 11). The results show that using constant scaling parameter [39], the structure granules fail to adhere to the perimeters between different structures (particularly, crowd and background), in contrast to GrCS which uses adaptive varying scaling parameters. The main reason is, since each pixel, p, is represented by multiple structure features, v_{ps} that capture varying aspects of textures, using a constant scaling parameter for all d_{ps} throughout the iterations will not work well to capture the local affinity of each texture feature, v_{ps}, of pixels within the structure granule. Note that constant scaling parameter will act as normalization constant. Thus, any value of constant scaling parameter would generate similar structure granules.

D. Number of Structure Granules

The parameter K determines the number of structure granules in an image. The greater the K value, the more the structure granules constructed per image. Fig. 13 provides visualization of the influence of the parameter K on the crowd segmentation performance. The result shows that the higher the K value, the less precise is the segmentation per image. This is as expected, because with respect to the image size, with a greater K value, the image is decomposed into smaller size structure granules, where each granule contains fewer number of pixels. Consequently, fewer structures are present to infer the content (i.e., crowd or background) of the corresponding granule. Likewise, the smaller the K value, the fewer the structure granules constructed per image, which in turn generate larger size structure granules. When the size of a structure granule becomes too large, it can no longer represent the structure characteristics of a local region. In all the experiments in this paper, we empirically set

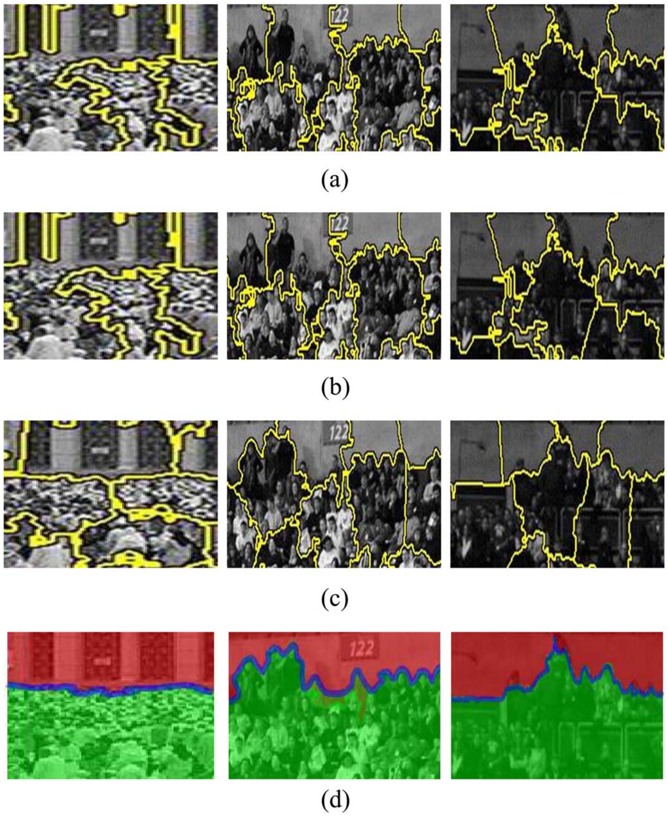

Fig. 12. Comparative results of structure granulation using constant value scaling parameter [39] and our proposed adaptive varying scaling parameters. In ideal segmentation results, crowd regions are shown as green overlay, background with red overlay, and blue line indicate ideal boundary between crowd and background. Boundaries between crowd and background of structure granules using adaptive varying scaling parameters are closer to the ground truth. Best viewed in color. (a) Constant scaling parameter = 10 [39]. (b) Constant scaling parameter = 20 [39]. (c) Our proposed, $m_s\tau - 1$. (d) Ground truth.

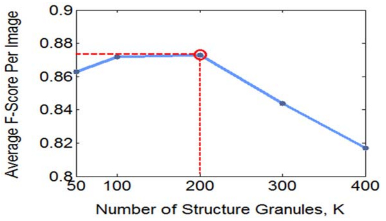

Fig. 13. Analysis of average f-score measure per crowd image in terms of number of structure granules, K. For $K = 200$, the average f-score per image is 0.873.

$K = 200$, which forms compact structure granules that outlines the natural boundaries between crowd and background regions.

E. Compactness of Structure Granules

Given the feature descriptor, $\boldsymbol{\upsilon}_{ps}$, of each pixel in a crowd scene, structure granules are formed by aggregating correlated pixel granules (detailed in Section III-B). The sought after characteristics of structure granules are as follows.

1) Boundaries between the structure granules of crowd and background regions are distinct, with each segregated into different structure granules.
2) Structure granules conform to the natural outline of arbitrary distribution of crowd.
3) Each structure granule contains structurally similar pixels of crowd scenes (i.e., high localization accuracy).

This is to cope with varying scales of individuals due to perspective distortion. The intuition is that each structure granule provides a compact and localized primitive characterizing the local structure for crowd segmentation.

An example of the structure granules (pixels granulation) on crowd scene constructed using GrCS is shown in Fig. 14 with yellow outlines indicating the partitions between granules. It is observed that this crowd scene has severe perspective distortion of crowd. Still, the GrCS is able to aggregate neighboring individuals of similar scale into structurally uniform atomic regions. Groups of individuals in crowds that appear much bigger in the images are segregated into different granules from those that appear smaller (regions in orange, green, and red box). At the same time, crowd regions with different crowdedness are observed to be grouped into separated granules. Despite complex background clutters (i.e., trees, building patterns, and image noise), the aggregation of correlated pixels enables precise segregation of crowd and background regions, as illustrated in blue box.

To evaluate the boundary adherences (compactness) of structure granules in the crowd scenes quantitatively, we consider local grouping of structurally similar pixels as a clustering problem, and use the widely adopted measurement in clustering evaluations (i.e., purity [52]). The purity measure of structure granules is utilized to quantify the quality of the granules against the pixel-level ground truth annotation labels (i.e., crowd or background). A structure granule is considered pure if it contains label from only one class, which is either crowd or background. Otherwise, a structure granule is considered as impure. In this context, an impure structure granule denotes that there is inaccurate separation between crowd and background regions. We quantify the accuracy of separation by using the purity measure, which is bounded within the [0, 1] range. A higher purity measure suggests a higher accuracy of boundaries between crowd and background regions.

Fig. 15 shows the comparison and relative improvement of our structure granules against varying scales of pixel-grid representation. Due to the aggregation of correlating pixel granules, structure granules are able to conform to the natural boundaries between different structures, in particular, crowd and background structure. Accordingly, the average purity measure of the proposed structure granules (0.854) outperforms the pixel-grid representation in all scales. Note that the proposed structure granules representation does not require manual intervention to achieve optimal boundaries adherence.

Furthermore, we show the purity measures of structure granule per crowd scene in Fig. 16. We observe that there are few crowd scenes with relatively lower purity measures. Upon scrutinizing our results, we observe that these images correspond to poorly illuminated crowd scenes, i.e., concerts and cinema, in which the lack of illumination may weaken informative textures structures and diminish scene details. Even so, the purity measures of the respective images are above 0.73.

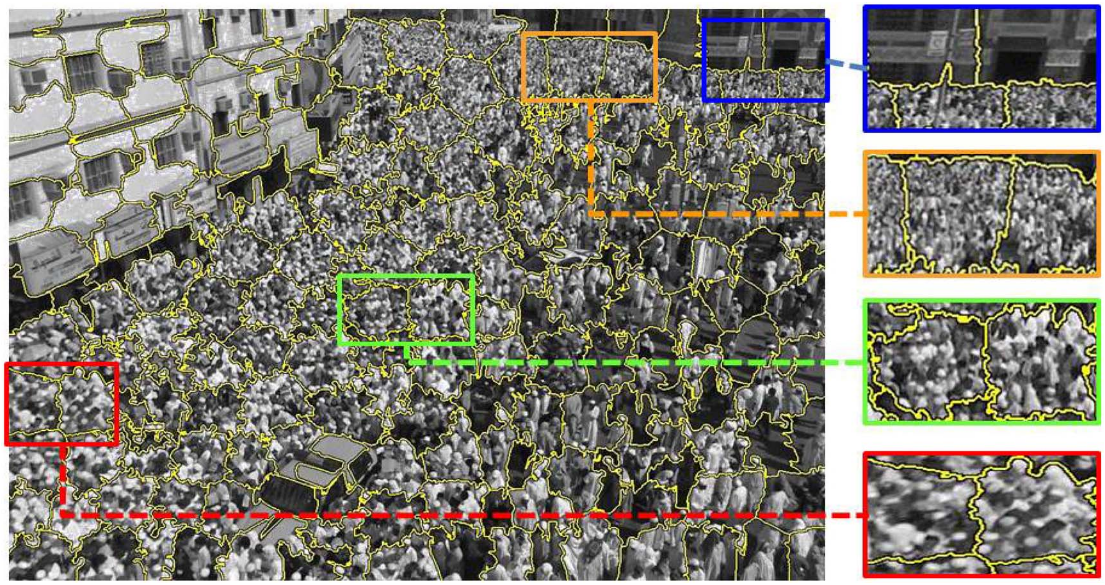

Fig. 14. Examples of structure granules on a dense crowd image. Yellow outline indicates the partitions between granules. Blue box: clear separation of structure granules between crowd and background. Orange, green, and red boxes: structure granules of crowd with significantly different crowdedness. Best viewed in color.

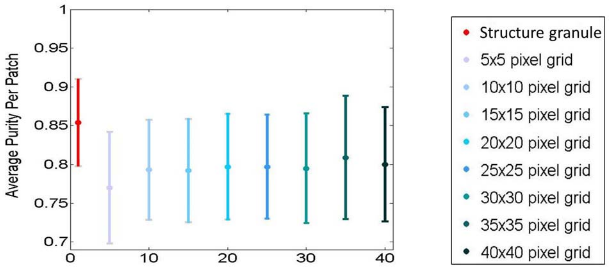

Fig. 15. Quantitative comparison of the boundary adherence (purity) measure of structure granules with different pixel-grid sizes. Means are shown in dots, standard deviations with bars.

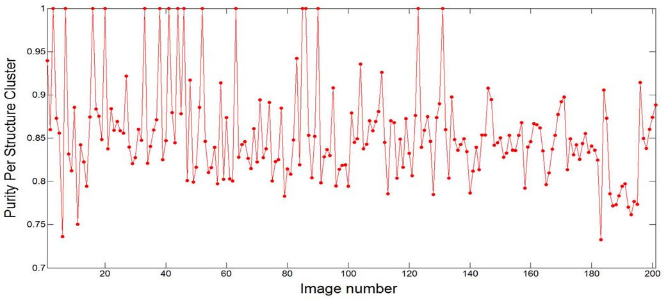

Fig. 16. Boundary adherence (purity) measure per structure granule with respect to image. The average purity is 0.854.

V. Conclusion

In this paper, we have explored a new research direction in dense crowd scene analysis using the theory and principles of GrC to conceptualize crowd segmentation problem at different levels of granularity. Structure granules constructed by aggregating similar neighboring pixel granules are served as primitive characterizing local textures instead of regular pixel-grid. Experimental results on public and synthetic crowd scenes have shown that the granulation approach is effective in grouping structurally similar pixels into clusters to cope with perspective distortion, varying crowdedness, and cluttered background for an effective interpretation of crowd and background regions. Though the structure granular is effective in outlining boundaries between multiscale crowd and background regions, the basis of granules for all granularity level are texture features. Thus, granulated view of different granularity level is limited when crowd scenes are poorly illuminated. Future investigation includes identifying texture features that are more robust toward characterizing poor illuminated crowd scenes.

References

[1] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” in *Proc. CVPR*, Anchorage, AK, USA, 2008, pp. 1–8.

[2] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, “Multi-source multi-scale counting in extremely dense crowd images,” in *Proc. CVPR*, Portland, OR, USA, 2013, pp. 2547–2554.

[3] B. Solmaz, B. E. Moore, and M. Shah, “Identifying behaviors in crowd scenes using stability analysis for dynamical systems,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 2064–2070, Oct. 2012.

[4] S. Ali and M. Shah, “Floor fields for tracking in high density crowd scenes,” in *Proc. ECCV*, vol. 5303. Marseille, France, 2008, pp. 1–14.

[5] H. Idrees, K. Soomro, and M. Shah, “Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 1986–1998, Oct. 2015.

[6] K. Kang and X. Wang, “Fully convolutional neural networks for crowd segmentation,” *arXiv preprint arXiv:1411.4464*, 2014.

[7] O. Arandjelovic, “Crowd detection from still images,” in *Proc. BMVC*, Manchester, U.K., 2008, pp. 1–10.

[8] S. Ghidoni, G. Cielniak, and E. Menegatti, “Texture-based crowd detection and localisation,” in *Intelligent Autonomous Systems*. Heidelberg, Germany: Springer, 2013, pp. 725–736.

[9] A. Fagette, N. Courty, D. Racoceanu, and J.-Y. Dufour, “Unsupervised dense crowd detection by multiscale texture analysis,” *Pattern Recognit. Lett.*, vol. 44, pp. 126–133, Jul. 2014.

[10] H. P. Moravec, *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA, USA: Harvard Univ. Press, 1988.

[11] W. R. Hendee and P. N. T. Wells, *The Perception of Visual Information*. New York, NY, USA: Springer, 1997.

[12] W. Pedrycz, *Granular Computing: An Emerging Paradigm*, vol. 70. Heidelberg, Germany: Springer, 2001.

[13] B. E. Moore, S. Ali, R. Mehran, and M. Shah, "Visual crowd surveillance through a hydrodynamics lens," *Commun. ACM*, vol. 54, no. 12, pp. 64–73, 2011.

[14] S. Ali and M. Shah, "A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proc. CVPR*, Minneapolis, MN, USA, 2007, pp. 1–6.

[15] M. Shah, "Visual crowd surveillance is like hydrodynamics," in *Proc. ACM Multimedia*, Firenze, Italy, 2010, pp. 3–4.

[16] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents," in *Proc. CVPR*, Providence, RI, USA, 2012, pp. 2871–2878.

[17] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. CVPR*, Miami, FL, USA, 2009, pp. 935–942.

[18] M. K. Lim, V. J. Kok, C. C. Loy, and C. S. Chan, "Crowd saliency detection via global similarity structure," in *Proc. ICPR*, Stockholm, Sweden, 2014, pp. 3957–3962.

[19] Y.-L. Hou and G. K. H. Pang, "Multicue-based crowd segmentation using appearance and motion," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 2, pp. 356–369, Mar. 2013.

[20] T. Li *et al.*, "Crowded scene analysis: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 367–386, Mar. 2015.

[21] V. J. Kok, M. K. Lim, and C. S. Chan, "Crowd behavior analysis: A review where physics meets biology," *Neurocomputing*, vol. 177, pp. 342–362, Feb. 2016.

[22] S. Wu, Z. Yu, and H.-S. Wong, "Crowd flow segmentation using a novel region growing scheme," in *Proc. PCM*, Bangkok, Thailand, 2009, pp. 898–907.

[23] R. Mazzon, S. F. Tahir, and A. Cavallaro, "Person re-identification in crowd," *Pattern Recognit. Lett.*, vol. 33, no. 14, pp. 1828–1837, 2012.

[24] M. J. Leach, E. P. Sparks, and N. M. Robertson, "Contextual anomaly detection in crowded surveillance scenes," *Pattern Recognit. Lett.*, vol. 44, pp. 71–79, Jul. 2014.

[25] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. CVPR*, Anchorage, AK, USA, 2008, pp. 1–7.

[26] L. Dong, V. Parameswaran, V. Ramesh, and I. Zoghlami, "Fast crowd segmentation using shape indexing," in *Proc. ICCV*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[27] D. Kong, D. Gray, and H. Tao, "A viewpoint invariant approach for crowd counting," in *Proc. ICPR*, vol. 3. Hong Kong, 2006, pp. 1187–1190.

[28] D. Helbing, P. Molnár, I. J. Farkas, and K. Bolay, "Self-organizing pedestrian movement," *Environ. Plan. B*, vol. 28, no. 3, pp. 361–384, 2001.

[29] A. N. Marana, S. A. Velastin, L. D. F. Costa, and R. A. Lotufo, "Automatic estimation of crowd density using texture," *Safety Sci.*, vol. 28, no. 3, pp. 165–175, 1998.

[30] S. K. Pal, B. Uma Shankar, and P. Mitra, "Granular computing, rough entropy and object extraction," *Pattern Recognit. Lett.*, vol. 26, no. 16, pp. 2509–2517, 2005.

[31] A. Rizzi and G. Del Vescovo, "Automatic image classification by a granular computing approach," in *Proc. 16th IEEE Signal Process. Soc. Workshop Mach. Learn. Signal*, Arlington County, VA, USA, 2006, pp. 33–38.

[32] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.

[33] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[34] J. Chen, S. Shan, G. Zhao, X. Chen, W. Gao, and M. Pietikäinen, "A robust descriptor based on weber's law," in *Proc. CVPR*, Anchorage, AK, USA, 2008, pp. 1–7.

[35] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.

[36] J. T. Yao, A. V. Vasilakos, and W. Pedrycz, "Granular computing: Perspectives and challenges," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1977–1989, Dec. 2013.

[37] Y. Yao, "Perspectives of granular computing," in *Proc. IEEE Int. Conf. Granular Comput.*, vol. 1. Beijing, China, 2005, pp. 85–90.

[38] Y. Yao, "Interpreting concept learning in cognitive informatics and granular computing," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 4, pp. 855–866, Aug. 2009.

[39] R. Achanta *et al.*, "Slic superpixels," School Comput. Commun. Sci., Ecole Polytechnique Fédéral de Lausssanne (EPFL), Lausanne, Switzerland, Tech. Rep. 149300, 2010.

[40] W. Pedrycz and A. Bargiela, "An optimization of allocation of information granularity in the interpretation of data structures: Toward granular fuzzy clustering," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 582–590, Jun. 2012.

[41] A. Bargiela, W. Pedrycz, and K. Hirota, "Granular prototyping in fuzzy clustering," *IEEE Trans. Fuzzy Syst.*, vol. 12, no. 5, pp. 697–709, Oct. 2004.

[42] W. Pedrycz and A. Bargiela, "Granular clustering: A granular signature of data," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 32, no. 2, pp. 212–224, Apr. 2002.

[43] X.-Q. Tang and P. Zhu, "Hierarchical clustering problems and analysis of fuzzy proximity relation on granular space," *IEEE Trans. Fuzzy Syst.*, vol. 21, no. 5, pp. 814–824, Oct. 2013.

[44] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. NIPS*, 2004, pp. 1601–1608.

[45] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[46] W. L. Hoo, T.-K. Kim, Y. Pei, and C. S. Chan, "Enhanced random forest with image/patch-level learning for image understanding," in *Proc. ICPR*, Stockholm, Sweden, 2014, pp. 3434–3439.

[47] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, "Data-driven crowd analysis in videos," in *Proc. ICCV*, Barcelona, Spain, 2011, pp. 1235–1242.

[48] P. Allain, N. Courty, and T. Corpetti, "AGORASET: A dataset for crowd video analysis," in *Proc. 1st Int. Workshop Pattern Recognit. Crowd Anal.*, Tsukuba, Japan, Nov. 2012, pp. 1–6.

[49] N. Courty, P. Allain, C. Creusot, and T. Corpetti, "Using the AGORASET dataset: Assessing for the quality of crowd video analysis methods," *Pattern Recognit. Lett.*, vol. 44, pp. 161–170, Jul. 2014.

[50] C. Liu, J. Yuen, and A. Torralba, "SIFT Flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[51] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[52] C. C. Aggarwal, "A human-computer interactive method for projected clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 4, pp. 448–460, Apr. 2004.

Ven Jyn Kok received the bachelor's degree in electrical and electronics engineering from Universiti Tenaga Nasional, Bandar Baru Bangi, Malaysia, and the master's degree in data communications from the University of Sheffield, South Yorkshire, U.K. She is currently pursuing the Ph.D. degree with the University of Malaya, Kuala Lumpur, Malaysia.

Her current research interests include image processing, pattern recognition, and machine learning.

Chee Seng Chan (S'05–M'09–SM'14) received the Ph.D. degree from the University of Portsmouth, Portsmouth, U.K., in 2008.

He is currently a Senior Lecturer with the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. His current research interests include computer vision and fuzzy qualitative reasoning, image/video content analysis, and human–robot interaction.

Dr. Chan was a recipient of the Institution of Engineering and Technology (Malaysia) Young Engineer Award, in 2010, the Hitachi Research Fellowship, in 2013, and the Young Scientist Network-Academy of Sciences Malaysia, in 2015. He is the Founding Chair of the IEEE Computational Intelligence Society, Malaysia Chapter, and the Founder of Malaysian Image Analysis and Machine Intelligence Association. He is a Chartered Engineer and a member of the Institution of Engineering and Technology.