Multimed Tools Appl
https://doi.org/10.1007/s11042-017-5418-y

Granular-based dense crowd density estimation

Ven Jyn Kok[1] · **Chee Seng Chan**[2]

Received: 23 April 2017 / Revised: 26 September 2017 / Accepted: 12 November 2017

Abstract Dense crowd density estimation is one of the fundamental tasks in crowd analysis. While tremendous progress has been made to understand crowd scenes along with the rise of Convolutional Neural Networks (CNNs), research work on dense crowd density estimation is still an ongoing process. In this paper, we propose a novel approach to learn discriminative crowd features from granules, that conforms to the outline between crowd and background (i.e. non-crowd) regions, for density estimation. It shows that by studying the inner statistics of granules for density estimation, this approach is adaptive to arbitrary distribution of crowd (i.e. scene independent). Multiple features fusion is proposed to learn discriminative crowd features from granules. This is to be used as description of the crowd where a direct mapping between the features and crowd density is learned. Extensive experiments on public benchmark datasets demonstrate the effectiveness of our novel approach for scene independent dense crowd density estimation.

Keywords Dense crowd analysis · Density estimation · Texture features · Visual surveillance

✉ Ven Jyn Kok
vj.kok@ukm.edu.my

Chee Seng Chan
cs.chan@um.edu.my

[1] Faculty of Information Science and Technology, National University of Malaysia, 43600 Bangi, Selangor, Malaysia

[2] Center of Image and Signal Processing, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

Published online: 05 December 2017

1 Introduction

Dense crowd density estimation is a fundamental task in crowd analysis with wide spectrum of applications. For example, the number of individuals in a crowd can be an indicator of the comfort level in crowded scenes. It can also be a cue for imminent crowd disasters, e.g. crowd crush due to overcrowding. Crowd disasters often occur when density in crowd become so great that individuals are crammed together. Physical forces from various directions cause individuals to fall, thus creating a domino effect that forces individuals to either step on each other or fall as well [15]. Hence, any mass gathering is at high risk of turning fatal given physical stress (e.g. overcrowding). This is evident with the recurrent of lethal crowd disasters, such as the Shanghai New Years Eve revelry disaster 2014 (36 death) [20] and Saudi Arabia Hajj disaster 2015 (2,110 death) [18].

As this is the 21st century, with advancing technology, one would think that crowd disaster, specifically due to overcrowding, is a problem of the past. However, owing to the worldwide population growth [45], coupled with the continuing urbanization [46], the occurrence of crowded environment is a growing norm. To make matters worse, according to the population estimates and projections from the United Nations Population Division [45], the world population is expected to reach 9.6 billion by the year of 2050. The present of large crowd in any environment can, thus, disrupt and challenge the effectiveness of crowd management, safety and security. The Love Parade music festival in Germany is a well-known crowd disaster incident of the 21st century due to overcrowding [14], resulting in mass casualties. In response to the dynamic and degenerating risk of densely crowded environments, dense crowd density estimation has emerged as an increasingly crucial application in visual surveillance for a proactive crowd management. This endeavor is also further motivated by the need for a sophisticated crowd surveillance system [40].

Although substantial efforts have been made toward understanding crowd for density estimation [17, 49, 50], accurate and precise estimation of individuals in dense crowd scenes is still an open problem. This is because it is difficult to discern individuals in dense crowd since they are in close proximity with each other [39]. The complexity often manifests itself in the frequent, partial or complete occlusion between individuals [2]. An individual in a dense crowd may only be occupying a few pixels per individual, thus it is infeasible to discern individuals and one's body parts, as shown in Fig. 1. The problem is further hampered by the perspective distortions and ambiguities caused by varying physical layout of crowd environments. The formation of crowd across different scenes is inherently dependent on the constraints imposed by the environmental layout. Generally, perspective maps or multi-scale pixel grid is essential in these studies to cope with the aforementioned challenges for dense crowd density estimation in different scenes.

Fig. 1 Examples of dense crowd scenes with perspective distortion. Individuals who are closer to the camera view appear larger than those who are further away from the camera

The common problem of the aforementioned methods is the susceptibility to the pixel-grid constrain. That is, conformations to the natural outline between crowd and background (i.e non-crowd) are difficult to achieve (see Fig. 2a). This can lead to the problem where features extracted may not offer sufficient discrimination, and thus inevitably lead to incorrect density estimation. One can observe in Fig. 2d that imprecise delineation of crowd and non-crowd regions, as well as assumption of dependency between pixel-grids can lead to inaccurate person count estimation. This is because extracted features are not characterizing either crowd or background only. It is also worth noting that the notion of assuming dependency between pixel grids is impractical since fundamentally crowd density and distribution varies from regions to regions in unconstrained public scenes. For instance, background elements can be randomly positioned within dense crowds, as shown in Fig. 2a. Moreover, fixed pixel grids are not able to adapt to the large variation of individuals size due to perspective distortion [22].

Motivated by these shortcomings, we propose to learn discriminative crowd features from meaningful atomic granules for dense crowd density estimation. Intuitively, a meaningful atomic region, consisting of crowd or background only, allows discriminative features extraction of the respective regions, and thus enables accurate density estimation. Unlike pixel-grid, an atomic granule adheres to the natural outline of crowd and background regions. This is to alleviate the constrains of using pixel-grid. This kind of representation has been studied in various fields, such as image segmentation [19] and object detection [27] for applications like face recognition and vehicle tracking. Previous works demonstrate the effectiveness of exploiting the inner statistics of the patches in an image to provide a meaningful primitive from which image features are computed. Nevertheless, to the best of our knowledge, there is no attempt to explore such representation for dense crowd density estimation. Importantly, such granules representation is scene independent, thus enables density estimation of dense crowd at different locations.

In addition, the proposed approach is motivated by the fact that no single feature can provide sufficient information for density estimation in dense crowd scenes. As noted by Idrees et al. [17], this is predominantly due to low resolutions imagery, perspective distortion and severe occlusions. One can, however, observe that dense crowds portray textures which

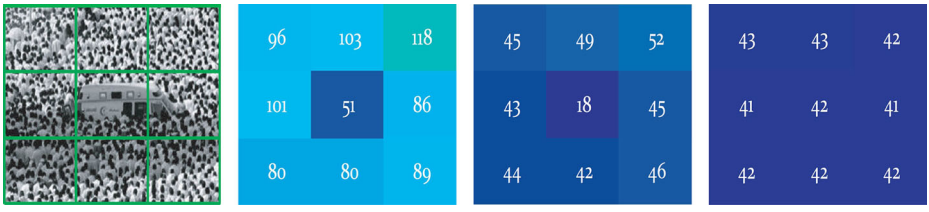


Fig. 2 Dense crowd density estimation by [17]. The dependency between pixel-grids is modeled by multi-scale Markov Random Field to enhance density estimation. Green outline indicates the partitions between pixel-grids. Crowd density for pixel-grids consisting of crowd and background (i.e. non-crowd) regions has been estimated to have similar density with crowd-only pixel-grids after dependency modeling. Best viewed in color

can be employed to infer crowd density. There is a relationship between low-level features and crowd density that is expected to facilitate dense crowd density estimation [31].

The main technical contribution of this paper is the proposal of a novel granular-based density estimation model to learn low-level crowd features and effectively combine them for dense crowd density estimation. Instead of using perspective map or multi-scale pixel grid to cope with perspective distortion in different crowd scenes as most existing work [17, 49, 50], we exploit the correlation among pixels (i.e. inner statistics) in an image to form natural representation of crowd and background regions as primitive regions to learn discriminative crowd features. This enables adaptation to different dense crowd structure in scenes due to severe inter-occlusion, perspective distortion and varying crowdedness, for density estimation. In addition, we also conduct extensive evaluation with multiple deep learning structures [49, 50] and various features fusion to learn discriminative crowd feature for density estimation.

A preliminary version of this work was presented earlier in [22], which focuses on outlining the natural boundaries between crowd and background (i.e., noncrowd) regions for crowd segmentation. To achieve adequate separation between crowd and background regions, dense crowd analysis is conceptualized at different levels of granularity with the aim to map problems into computationally tractable subproblems. In comparison to our earlier work in [22], the present work introduces a novel granular-based dense crowd density estimation framework (Section 3) and proposes complementary features to characterize crowd for density estimation (Section 3.2). Moreover, extensive technical details, experimental evaluations and analyzes are provided (Section 4). Specifically, Section 4.4 provides experimental evaluations and analyzes with state-of-the-art approaches [17, 25, 38, 49, 50] on several challenging dense crowd datasets [17, 50]. Sections 4.5 and 4.6 provide analyzes regarding different texture features and number of granules for dense crowd density estimation, respectively.

2 Related work

Conventional approaches are mostly object-centric from which the behavior in a scene is learnt in three steps: object detection, tracking and compilation of tracked results for crowd segmentation [1, 51], behavior analysis [26, 43] and density estimation [11, 35, 37].[1] For instance, framework that performs clustering of coherent trajectories to represent a moving entity, and inferring the number of individuals in the scene by Rabaud and Belongi [37]. This approach is limited to crowd scenes with sparse crowd where continuous sets of image frames are accessible. The results presented in their work have shown promising performance when individuals are disconnected from each other. However when individuals in crowd scenes are closely positioned with each other, trajectories are incorrectly merged together. This is due to the phenomenon of collective motion occurring between moving interacting entities. Hence, as noted by Zhan et al. [48], conventional computer vision methods work well on sparse scenes (i.e. approximately 5-20 individuals [2]), but are inadequate to analyze dense crowd scenes. Correspondingly, a straightforward extension of these methods is neither suitable nor capable of analyzing dense crowd. This is because a crowd is

[1] Given the broad and expending nature of research in crowd analysis, this paper will narrow down the scope by focusing only on crowd density estimation. Interested readers are referred to [13, 23, 48] for detailed discussions on applications and advances in crowd analysis.

beyond a simple sum of individuals, where it can assume different complex behaviors. The difficulty of analyzing crowd increases disproportionately in relation to the number of individuals in a crowd.

Since delineating individuals in dense crowd scenes is difficult (because of the spatial overlaps), most existing density estimation approaches [4–6, 17, 32] obviate the steps to detect and / or track individuals. They put emphasis on extracting a set of low-level image features. Marana et al. [32] presented a method based on texture analysis to estimate crowd density, where the estimation was given in terms of discrete ranges (i.e. very low, low, moderate, high and very high). Their objective was to challenge scenes of dense crowd where each individual is greatly occluded. They assumed that crowd scene of high density tend to illustrate fine textures, whereas crowd scenes of low density are mostly made up of coarse patterns. Crowd density estimation by Davies et al. [8] is one of the earliest works that uses regression approach to learn the relationship between global features (e.g. number of edge pixels) and density of individuals. Similarly, works by Chan et al. [5] and Chan and Vasconcelos[4] propose to extract dynamic texture from homogeneous motion crowd segments and focus on learning mapping between large set of feature responses and density. However, a problem commonly encountered in regression based approach is perspective distortion in which individuals who are closer to the camera view appear larger than those who are further away from the camera (as illustrated in Fig. 1). The problem is exacerbated when single regression function is used for the whole image space. To address this problem, perspective normalization plays a key role by bringing the perceived size of individuals at different depths to the same scale [30]. Another commmonly used approach is to divide the image space into different pixel-grids (as illustrated in Fig. 2a) and each pixel-grid is modeled by a regression function to mitigate the influence of perspective distortion. Such approaches rely on local features modeling through the analysis of pixel-grids [6, 17].

In recent years, the computer vision field has witness a great leap forward through the adaptation of deep convolutional neural network (CNNs) in crowd analysis, such as crowd scene understanding [41] and crowd segmentation [21]. To the best of our knowledge, the exploration of CNNs on crowd density estimation is still new. Recently, Zhang et al. [49] and Zhang et al. [50] proposed a CNN based method for crowd density estimation. Although the work by Zhang et al. [49] shows good performance on most of the datasets, their approach requires perspective maps to bring the perceived size of individuals at different depths to the same scale. Such scene specific perspective information is not readily available in many practical applications of density estimation. To obviate the need of perspective maps, Zhang et al. [50] adopted a multi-column CNNs architecture. Each column corresponds to filters with receptive fields of different sizes (i.e. small, medium, large) to cope with large variation in people/head size due to perspective distortions. Similar to conventional approaches, generally, these studies require perspective map or multi-scale CNN filters to cope with perspective distortion in dense crowd scenes.

3 Proposed granular-based density estimation model

Given a dense crowd image, the aim is to estimate the number of individuals based on discriminative crowd features. In a public scene, the density of individuals can vary from region to region. As shown in Fig. 1, this density variation is mainly due to the effects of perspective distortion or constraints imposed by the environment layout. In this section, we introduce a novel dense crowd density estimation model using meaningful atomic granules. The key steps of granular-based dense crowd density estimation framework are illustrated in Fig. 3.

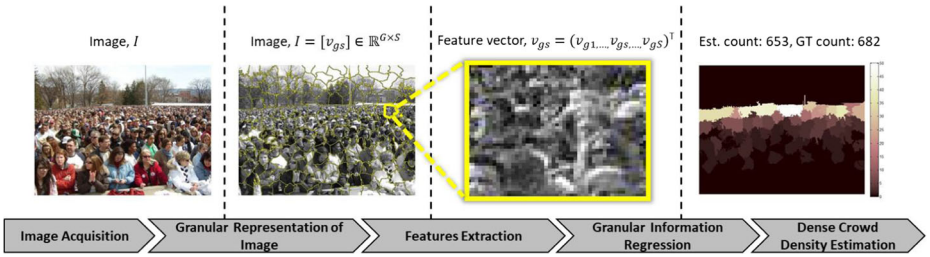


Fig. 3 An illustration of the key steps in granular-based dense crowd density estimation. Column 4 shows the estimated count (Est. count) and ground truth count (GT count) using the proposed framework. The heat map shows the dense crowd density where brighter color indicates higher density

The proposed framework represents dense crowd image using granules. This is to alleviate pixel-grid constrain, while enabling discriminative feature extraction of crowd regions for density estimation. The granules are formed from the aggregation of pixels with similar feature vector, adapting the pixel clustering approach [22]. This is to facilitate in distinguishing between crowd and background (i.e. non-crowd) regions for density estimation.

Formally, a dense crowd image, $I = [v_{gs}] \in \mathbb{R}^{G \times S}$, where G is the number of granules in an image and S is the number of features. Each granule, g, in a dense crowd image, I, is represented as a feature vector, $v_{gs} = \left(v_{g1}, \ldots, v_{gs}, \ldots, v_{gS}\right)^{\top} \in \mathbb{R}^{G \times S}$, where $g = \{1, \ldots, G\}$ and $s = \{1, \ldots, S\}$. The feature vector, v_{gs} is formed by the mean of feature descriptor of pixel, p, within the respective granule, i.e. $\frac{1}{N} \sum_{p=1}^{N} v_{ps}$, such that N is the number of pixels within the respective granule.[2] The feature descriptor for each pixel, $v_{ps} = \left(v_{p1}, \ldots, v_{ps}, \ldots, v_{pS}\right)^{\top}$, is the concatenation of S different and complementary features. The texture features used in the proposed approach to represent pixels are discussed in Section 3.2. Dense crowd density estimation problem is subsequently formulated as a regression problem. In particular, a mapping function between feature vectors input and a scalar-valued crowd density output is learned.

3.1 Granular representation of dense crowd images

Although dense crowd can be irregular at a coarse level, the texture of crowd tend to correspond to a harmonic pattern (i.e. regular texture) at a finer scale patches [17], such as pixel-grids or granules. Moreover, crowd regions tend to present large number of texture features. As one can observe from Fig. 1, this is because of the appearance variations of crowd. These texture features carry strong cues regarding the number of people in a scene [30]. Thus, crowd regions in these granules can be treated as texture for processing.

In this work, dense crowd images are represented as granules for density estimation. It is the basic aspect of dense crowd scenes in this work, characterizing structurally meaningful atomic regions that distinguish between crowd and background regions for low-level feature extraction. Specifically, given a dense crowd scene image, pixels in the image are aggregated based on feature similarity [22] to form granules. These granules are structurally coherent atomic regions in the image that conform to the natural boundaries between crowd and background (as shown in Fig. 4). The key idea of these atomic regions is to have a pixel

[2]Note that the number of pixels, N, within each granule varies [22].

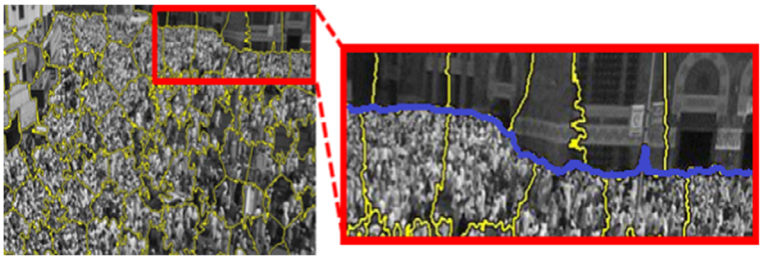


Fig. 4 Examples of granules on a dense crowd images. The yellow outlines indicate the partitions between granules. The blue outline indicates the boundary between crowd and background. The red box indicates clear separation of granules between crowd and background. Best viewed in color

aggregation process versatile to different crowd scenes, and so this will best categorize the diverse structures in the scene for robust density estimation.

The texture feature vector for each granule is the mean of texture features of pixels within the respective granule. The texture feature vector from each granule is used as description of the crowd, where a direct mapping between the features and crowd density is learned.

3.2 Crowd texture features

In this section, we introduce a complementary set of texture features for granules in dense crowd images, to facilitate the learning of discriminative crowd features for density estimation. The set of texture features includes the Local Standard Deviation (LSD), Dense Scale-Invariant Feature Transform (DSIFT) and Phase Congruency (PC). These features explicitly convey meaningful spatial content (i.e. texture) of the granules in dense crowd images for density estimation. 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 dense crowd analysis researches.

Local Standard Deviation (LSD) The LSD feature is inspired by the fact that dense crowd regions with different density tend to generate distinct local texture patterns, as shown in Fig. 5. That is, highly dense crowd regions (as shown in the first column of Fig. 5) comprise of fine patterns, whereas moderately dense crowd regions (as shown in the third column of Fig. 5) mostly contain coarse pattern. As related in [8] and [32], there is a correlation between crowd density and edge feature of crowd. Accordingly, this proposed approach is motivated to use edge feature to characterize crowd regions.

To this end, Local Standard Deviation (LSD) is employed to capture the local image structure, i.e. edges, formed by mass of crowd in dense crowd images. This is because LSD is a computationally simple and practical edge detection mechanism [28]. The output of LSD is a measure of the local average contrast. Specifically, calculating the LSD of pixels in a neighborhood can indicate the degree of variability of pixels intensities in that local region. Strong intensity contrast / variability of pixels characterizes edges in images.

Given a dense crowd image, LSD calculates the standard deviation of pixel intensities in a 5×5 neighborhood centering each pixel of interest (i.e. all the pixels in the image). The output of LSD is assigned to the respective pixel of interest. One of the main advantages of using LSD in the proposed approach is that edge sharpness of crowd images can be quantified. This is essential to delineate the various texture features in dense crowd images for density estimation.

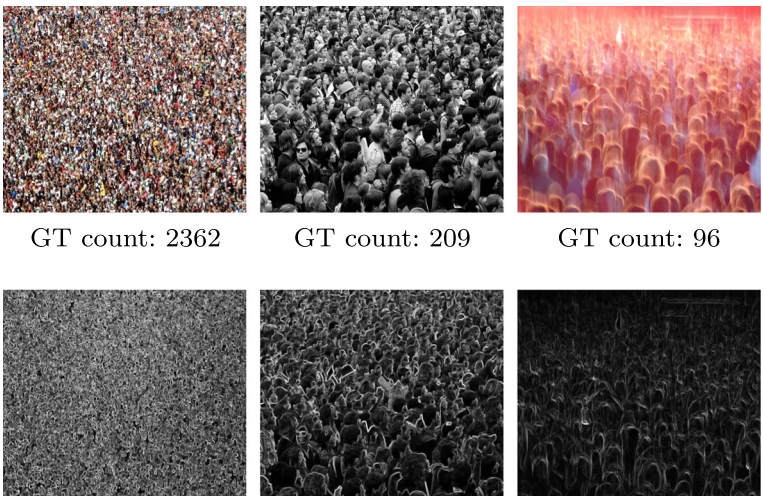


Fig. 5 (Top row) Examples of dense crowd scene images. (Bottom row) Images of local standard deviation (LSD) using 5×5 neighborhood. Note that the crowd density (ground truth (GT) count) decreases when viewed from left to right. Best viewed in color

Dense Scale-Invariant Feature Transform (DSIFT) DSIFT [47] is a variation of the SIFT algorithm [29], which is a state-of-the-art keypoint based approach to characterize local gradient information. By using SIFT, the number of interest points extracted from an image varies based on the image content, making the information incorporation on spatial configuration complicated [44]. Conversely, DSIFT extracts SIFT histogram for all pixels with overlapping patches. Compared to sparse features (e.g. SIFT [29], interest points [33]), dense features result in a good coverage of the entire scene [44]. This produces a constant amount of features per image area that contain essential information of the image content.

As one can observe in crowd regions with highly irregular repetitive grain (as shown in Fig. 5 (Top row)), it is likely to have similar texture element around different regions of crowd, formed by parts of people [17]. The local intensity gradient can reveal local individual appearance, such as head and shoulder, which is informative for density estimation [30]. Therefore, in addition to edge feature of crowds, DSIFT is used in the proposed approach to model the appearance cue of crowd. DSIFT algorithm is implemented to extract feature descriptor for each pixel in a dense crowd image. The DSIFT feature descriptor corresponds to the spatial coordinate of image pixels, forming a dense description of the image.

Given a dense crowd image, the feature descriptor of each pixel of interest is constructed by overlying a window centering the pixel of interest. Each local window is further divided into smaller sub-windows (e.g. 4×4) where gradient orientation and magnitude are quantized into an 8 bin histogram in each sub-window. The feature descriptor of the pixel of interest is formed by concatenating the histogram of sub-windows, obtaining a $4 \times 4 \times 8 = 128$ dimensional vector as the SIFT representation.

Phase Congruency (PC) The gradient-based texture features, i.e. LSD and DSIFT, are sensitive to image illumination variations [24]. Hence, these extracted features can be image dependent. To compensate and complement the set of features used to represent textures of granules, a dimensionless measure of feature significance that is invariant to image

illumination is desired. Such measure can provide absolute quantifications of feature significance that is applicable to any dense crowd scene images.

Studies by Oppenheim and Lim [36] have shown that phase information of images can retain the important features of image context. Interestingly, the Local Energy Model developed by Morrone and Owens [34] postulates that features can be perceived at spatial positions of maximum phase congruency within an image in the frequency domain. Hence, the advantage of this model is that it is not based on local intensity gradient for feature detection. These texture features detected include edges and lines.

To construct a dimensionless measure of phase congruency of dense crowd images that is invariant to image illumination, the method introduced by Kovesi [24] is used. The [24] scheme calculates the phase congruency with Log-Gabor wavelet filters [10], which work as bandpass filters. It allows arbitrary large bandwidth filters to be constructed while maintaining a zero direct current (DC) component in the even-symmetric filter. Hence, the phase congruency of a pixel p in a dense crowd image, I, is expressed as the summation over orientation o and scale n:

$$PC(p) = \frac{\sum_o \sum_n W_o(p) \lfloor A_{no}(p) \Delta\Phi_{no}(p) - T_o \rfloor}{\sum_o \sum_n A_{no}(p) + \varepsilon} \tag{1}$$

where

$$\Delta\Phi_{no}(p) = cos(\phi_{no}(p) - \overline{\phi}_o(p)) - sin(\phi_{no}(p) - \overline{\phi}_o(p)) \tag{2}$$

such that $\lfloor \cdot \rfloor$ is a floor function which denotes that the enclosed quantity is not permitted to be negative; $W_o(p)$ is a weighting factor based on frequency spread; $A_{no}(p)$ is the local amplitude of pixel p on scale n and orientation o; T_o is introduced to compensate for noise influence. A small denominator $\varepsilon = 0.0001$ is added to avoid division by zero [24]. $\Delta\Phi_{no}(p)$ is a sensitive phase deviation measure, where $\overline{\phi}_o(p)$ is the mean phase angle for pixel p.

The output of the phase congruency takes on the values between [0, 1], providing an illumination invariant measure of texture features in dense crowd images. Figure 6 shows sample phase congruency outputs of dense crowd images.

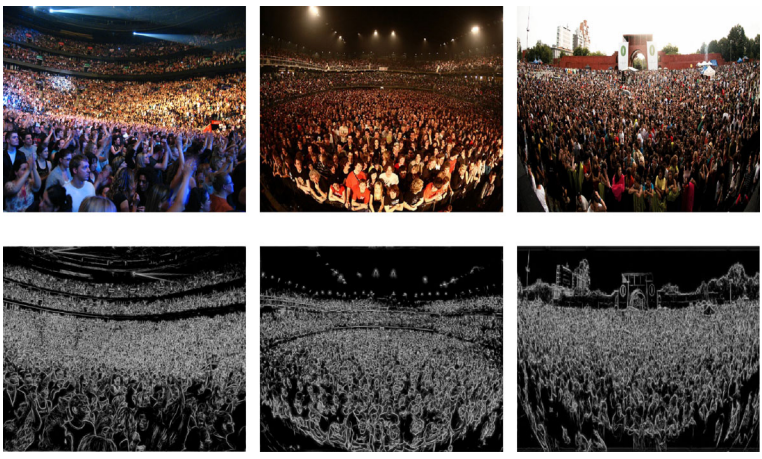


Fig. 6 (Top row) Examples of dense crowd scene images. (Bottom row) Images of the corresponding phase congruency (PC). Note that the texture features are invariant to changes in illumination. Best viewed in color

3.3 Density estimation by regression

The texture feature vector, υ_{gs}, of a granule in a dense crowd image is the mean of feature descriptor of each pixel, p, within the respective granule. The feature descriptor, υ_{ps}, of each pixel, p, is the concatenation of the Local Standard Deviation (LSD), Dense Scale-Invariant Feature Transform (DSIFT) and Phase Congruency (PC) texture features in this work. Given the granules of dense crowd images, dense crowd density estimation task is posed as a regression problem. The aim is to learn the relationship between the texture features and the crowd density, for dense crowd density estimation of new scenes.

For sparse crowd scenes (i.e. approximately 5-20 individuals [2]) where lower crowd density and fewer occlusions among individuals are observed, linear regressor (e.g. ridge regression [16]) may suffice. This is because the mapping between the features and people count typically presents a linear relationship [30]. Nonetheless, given a dense crowd environment, where there are severe partial and complete occlusions among individuals, a nonlinear regressor is required to capture the nonlinear trend in the feature space [3].

Formally, given M training data, which is represented as $\{\mathbf{x}_i, y_i\}_{i=1}^{M}$, $\mathbf{x}_i$ is a feature vector of granule in the training data, and y_i is the corresponding ground truth crowd density of the respective granule. The ground truth crowd density in a granule is the sum of ground truth annotation within the respective granule. The objective of regression is to predict the value of y given a new value of $\mathbf{x}$. In the proposed approach, the mapping between the texture features and the crowd density is estimated by learning a nonlinear function, in particular, a Kernel Ridge Regression (KRR). KRR with radial basis functions kernel is employed owing to its promising performance in the literature for crowd density estimation [6, 7]. In its simplest form, a ridge regression function (i.e. $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \phi(\mathbf{x})$) is a linear regressor with a cost function as follows:

$$C(\mathbf{w}) = \frac{1}{2}\sum_i (y_i - \mathbf{w}^T \phi(\mathbf{x}_i))^2 + \frac{1}{2}\lambda \|\mathbf{w}\|^2 \tag{3}$$

where $\frac{1}{2}\lambda \|\mathbf{w}\|^2$ is a regularization term to avoid over-fitting of the training data. The parameter $\lambda > 0$ is determined via cross-validation. The model parameter $\mathbf{w}$ is determined by minimizing the cost function $C(\mathbf{w})$.

The nonlinear version of the ridge regression, i.e. KRR, can be achieved via kernel trick [42]. That is, constructing the ridge regression model in higher dimensional feature space induced by a kernel function. In this work, the radial basis functions is used:

$$k(\mathbf{x}, \mathbf{x}') = exp(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}) \tag{4}$$

where the kernel width parameter σ is determined via cross-validation. The KRR functions is given by:

$$f(\mathbf{x}, \alpha) = \sum_i \alpha_i k(\mathbf{x}, \mathbf{x}_i) \tag{5}$$

where $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_i\}$ are Lagrange multipliers used to solve the KRR minimization problem [30].

The estimated density of an unseen dense crowd image is the summation of the estimation obtained for all granules in the corresponding image.

4 Experimental results and discussion

4.1 Dataset

Evaluations on the granular density estimation model for dense crowd scenes are conducted on benchmark datasets of real scenes obtained from [17] and [50]. The ground truth count for each image in the datasets [17, 50] has been provided by the respective authors. The annotations were done manually marking the positions or the center of the head of every individual in the scenes. The crowd in these datasets varies in terms of illuminations, crowdedness and perspectives. One of the biggest challenges in these datasets is the large range of crowd count between images.

UCF_CC_50 dataset [17] consists of 50 dense crowd images collected mainly from Flickr[3]. The number of individuals in these images ranges between 96 and 4628, with an average of 1280 individuals per image. The scenes in these images are diverse, depicting dense crowd in various settings, such as concerts, train stations and stadiums. The ground truths provided are manually annotated.

Shanghaitech dataset [50] is a new large-scale crowd counting dataset consists of 1198 annotated crowd images. A total of 330,165 people are manually labeled at the centers of their heads. There are two parts in this dataset: Part A (482 images) and Part B (716 images). Images in Part A are crawled from the Internet, whereas images in Part B are taken from a busy street of metropolitan areas in Shanghai. The number of individuals in Part A ranges between 33 and 3139, while Part B ranges between 9 and 578. Since this present work focuses on dense crowd scenes, only Part A of this dataset will be used for evaluation.

4.2 Evaluation metric

The performance of density estimation approach can be assessed by the similarity between the actual count and the estimated count of individuals in a scene [17]. Similar with existing approaches [17, 49, 50], we use mean and deviation of Absolute Difference (AD) to quantify the accuracy of density estimation and robustness of the proposed algorithm, respectively.

$$AD_i = \left| \zeta_i - \hat{\zeta}_i \right| \tag{6}$$

where i denotes the ith image, ζ_i is the actual (i.e. ground truth) count in each image, and $\hat{\zeta}_i$ is the estimated count.

4.3 Experiment settings

Consistent with [17, 49, 50], the UCF_CC_50 dataset is randomly divided into sets of 10 to perform 5-fold cross-validation to avoid bias. The Shanghaitech dataset is divided into training set and testing set, such that 300 images are used for training and 182 images for testing. Granules are randomly selected from the Shanghaitech training images to train the proposed granular density estimation model.[4] Note that there are no overlaps between training and

[3]Flickr - Photo Sharing!: https://www.flickr.com/

[4]The selection of those random granules can be repeated to train the proposed granular density estimation model. In this study, the mean and deviation of AD is the average of 100 independent evaluations. Evaluation with varying granules generated consistent mean and deviation of AD.

testing dense crowd images. In all the experiments, we set the number of granules, $G = 200$, which enables high localization of granules with adequate separation between crowd and background regions for density estimation.

4.4 Comparison with state-of-the-art methods

The effectiveness of the proposed model is demonstrated in the application of dense crowd density estimation on public scenes. Evaluations are conduction by benchmarking with [25, 38], as well as multi-scale pixel grid and CNNs approaches [17, 49, 50]. The methods by Lempitsky and Zisserman [25] and Rodriguez et al. [38] are among the few conventional approaches that are suitable for dense crowd density estimation and, therefore, are used for comparison. Most existing methods [12, 37] require person detection, hence they are more suitable for sparse crowd scenes analysis.

The comparisons on the UCF_CC_50 dataset are presented in Table 1. The method by Rodriguez et al. [38] has the highest mean and deviation of AD per image. This is because the method relies on head detections for density estimation. For dense crowd scenes with few pixels per individual, severe occlusions and appearance variations, it is challenging to determine one's head from another. Comparing methods by Lempitsky et al. [25] with Idrees et al. [17], method by Idrees et al. [17] uses three sources (i.e. head detection, SIFT and frequency domain analysis) and multi-scale MRF to model dependency among pixel grid, whereas Lempitsky et al. [25] uses only DSIFT feature for dense crowd density estimation. Accordingly, the work by Idrees et al. [17] has lower AD per image than the work by Lempitsky et al. [25]. This shows that to enhance density estimation in dense crowd scenes, multiple features is required to compensate and complement the insufficient of other features. By using granules which conform to the natural boundaries between crowd and background (i.e. non-crowd) regions, the proposed approach is able to extract crowd texture features essential for segmentation [22] and density estimation. This is not the case for [17] that uses pixel-grids. We show that when compared with CNNs approaches [49, 50], our proposed granular density estimation approach achieves comparable mean AD (i.e. 407.8) and lowest deviation of AD (i.e. 484.0). This essentially shows the robustness of granular-based representation of dense crowd images that exploits the correlation among pixels in an image to form natural representation of crowd and background regions. Such representation is able to adapt to different dense crowd scenes providing perceptually meaningful atomic

Table 1 Comparative results of dense crowd density estimation in UCF_CC_50 dataset [17]

Method	Absolute difference (AD)	
	mean	deviation
Rodriguez et al. [38]	655.7	697.8
Lempitsky et al. [25]	493.4	487.1
Idrees et al. [17] - before MRF	468.0	590.3
Idrees et al. [17] - after MRF	419.5	541.6
Zhang et al. [49]	467.0	498.5
Zhang et al. [50] - MCNN	**377.6**	509.1
Proposed	407.8	**484.0**

The lowest mean and deviation of Absolute Difference (AD) are in bold

Table 2 Comparative results of dense crowd density estimation in Shanghaitech dataset [50]

Method	Absolute difference (AD)	
	mean	deviation
LBP + RR	303.2	371.0
Zhang et al. [49]	181.8	277.7
Zhang et al. [50] - MCNN(CCR)	245.0	336.1
Zhang et al. [50] - MCNN	**110.2**	173.2
Proposed	251.2	**171.8**

The lowest mean and deviation of Absolute Difference (AD) are in bold

regions to learn discriminative features for density estimation. This is not the case for [49, 50], where perspective maps or scene geometry is required.

Evaluations have also been conducted on Shanghaitech dataset. We compare our proposed approach with two state-of-the-art CNNs approaches [49, 50].The MCNN approach by Zhang et al. [50] is based on density map estimation, whereas the MCNN-CCR is based on mapping the images to their total head counts. The regression approach using ridge regression (RR) and Local Binary Pattern (LBP) feature, denoted as the LBP + RR, serves as the baseline for comparison. The mean and deviation of AD in comparison are shown in Table 2. Similar to UCF_CC_50 dataset, evaluation on Shanghaitech dataset shows that multiple complementary texture features are essential to compensate the insufficient of other features. Hence, the proposed approach achieves lower mean and deviation of AD than baseline approach (i.e. LBP+RR). Despite the high level abstraction of CNNs based approach [49, 50], our proposed approach based on granules and complementary texture features achieves comparable mean of AD. However, even without perspective information of the scenes, the proposed approach achieves the lowest deviation of AD (i.e. 171.8) in comparison with CNNs approaches [49, 50]. This indicates that our proposed approach is scene-independent, and can be effectively applied to public dense crowd scenes with varieties of crowd distributions. Arbitrary distribution of crowd is effectively outlined using granules, providing a meaningful atomic region for discriminative features extraction.

4.5 Evaluation of different texture features

The qualitative results of the proposed approach using different features on UCF_CC_50 dataset are presented in Table 3. The first row in Table 3 shows the results of using Local

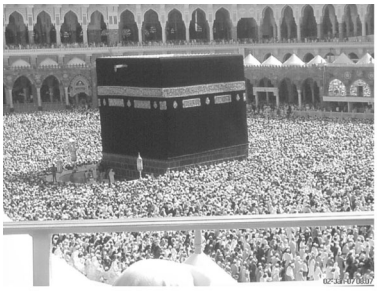
Table 3 Quantitative results of the proposed approach on UCF_CC_50 dataset using different texture features, i.e. Local Standard Deviation (LSD), Dense Scale-Invariant Feature Transform (DSIFT) and Phase congruency (PC)

Method	Absolute difference (AD)			
	per granule		per image	
	mean	deviation	mean	deviation
LSD	**5.9**	8.2	621.6	679.7
LSD + DSIFT	6.7	7.2	481.2	523.7
LSD + DSIFT + PC	6.4	**6.6**	**407.8**	**484.0**

The lowest mean and deviation of Absolute Difference (AD) are in bold

Standard Deviation (LSD) feature only, giving mean and deviation of AD 5.9 and 8.2 respectively, per granule, as well as 621.6 and 679.7 respectively, per image. Supplementing the proposed approach with DSIFT feature which captures the appearance cue in dense crowd scenes, improves the mean of AD per image by 140.4. To compensate and complement the gradient-based features (LSD and DSIFT) that are sensitive to image illumination variations, PC feature that is based on phase information in frequency domain is included. This improves the mean and deviation of AD per image to 407.8 and 484.0, respectively. Although the mean of AD per granule increases marginally, the deviation of AD reduces by 1.6.

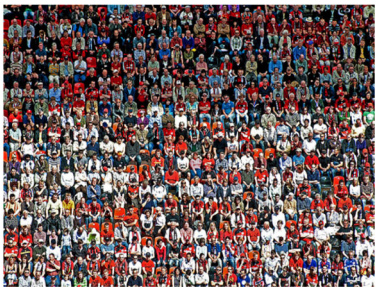
Figure 7 illustrates the estimated counts of dense crowd images with severe perspective distortion and varying illumination conditions. As shown in the first column of Fig. 7, our proposed approach is able to cope with varying scales of individuals in crowd for dense crowd density estimation. This is because DSIFT provides appearance cue features that remain invariant to changes in scale. Moreover, dense crowd images are represented using meaningful atomic granules that conform to the outline of crowd and background, as well as the variability of crowd structures. To comprehend the influence of PC feature on



LSD: 1168
LSD + DSIFT: 1526
LSD + DSIFT + PC: **2087**
GT count: **2104**

LSD: 678
LSD + DSIFT: 768
LSD + DSIFT + PC: **898**
GT count: **1046**

LSD: 257
LSD + DSIFT: 674
LSD + DSIFT + PC: **591**
GT count: **484**



LSD: 1595
LSD + DSIFT: 1209
LSD + DSIFT + PC: **588**
GT count: **440**

Fig. 7 Dense crowd images are shown with their respective ground truth count (GT count) and estimated count using multiple complementary texture features (i.e. Local Standard Deviation (LSD), Dense Scale-Invariant Feature Transform (DSIFT) and Phase Congruency (PC)). The improvements of the estimated count allude to the complementary nature of these features

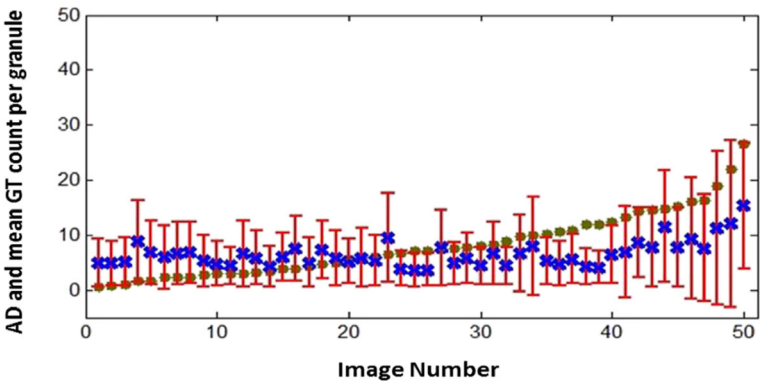


Fig. 8 Analysis of per granule estimates in terms of absolute difference (AD). The x-axis shows image numbers sorted with respect to mean ground truth (GT) count per granule. Olive dots: GT count per granule. Blue crosses: mean of absolute difference. Red bars: standard deviation of absolute difference. Best viewed in color

dense crowd density estimation, second column of Fig. 7 provides visualization of dense crowd scenes with varying illumination conditions and the corresponding estimated counts. The results show that PC feature is significant in improving dense crowd density estimation performance. The intuition is that PC feature is illumination invariant, hence it can provides quantifications of feature significance in different dense crowd scenes (i.e. scene independent) that is essential for density estimation.

Figure 8 shows the AD for granules in each dense crowd image. The images are sorted according to their respective mean ground truth count per granule to ease analysis. The mean

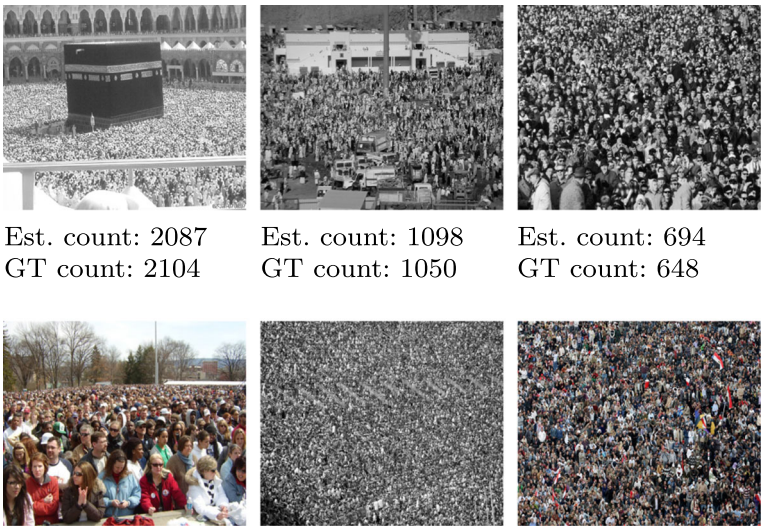


Fig. 9 Several dense crowd images from the UCF_CC_50 dataset with their respective ground truth count (GT count) and estimated count (Est. count) using the proposed approach

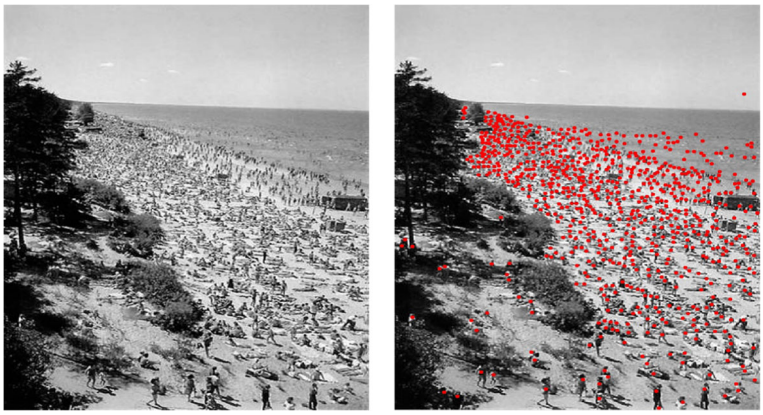


Fig. 10 Example of low-resolution dense crowd image where it is challenging to distinguish individuals from background. Left: Dense crowd image. Right: Image with ground truth annotations (red dots). This shows that manual annotations are prone to human mistakes. Best viewed in color

and deviation of AD per granule are shown with blue crosses and red bars respectively. The ground truth per granule for each image is shown as olive dots. As shown in Fig. 8, the AD per granule is consistent despite the increase of ground truth count, except for the images in the range of 46 to 50. The images from the range of 1 to 45 consist of 96 – 2704 ground truth count of individuals. This indicates that the proposed approach performs density estimation consistently for granules in this range. The reason for the increasing mean and standard deviation of AD for images in the range of 46 – 50 is because these images contain the highest ground truth count, with the largest ground truth count which is 4628 (i.e. a 4821% of the smallest ground truth count). Likewise, the ground truth count per granule also increases in contrast to the ground truth count per granule for other images. Figure 9 shows several dense crowd images from the dataset with their respective ground truth count and estimated count using the proposed approach.

From Fig. 8, it is observed that there are a few images with relatively higher AD per granule than the overall images within the range of 1 to 45. Upon scrutinizing the results, it is observed that some of these images correspond to low resolution images where informative texture features may have been diminished. It is also challenging for human to ascertain individuals from background in the scenes (as shown in Fig. 10). Since most ground truth provided [17, 50] are manually annotated, it is prone to human mistake.

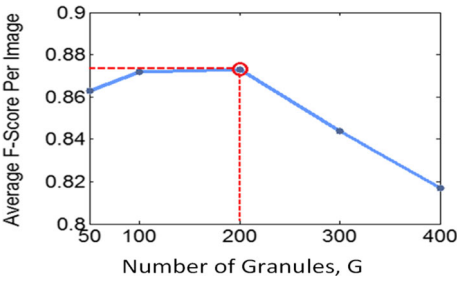


Fig. 11 This figure shows analysis of average f-score measure per crowd image in terms of number of granules, G. For $G = 200$, the average f-score per image is 0.873. (from [22])

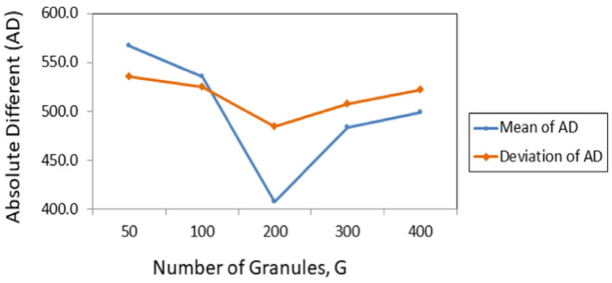


Fig. 12 Dense crowd density estimation evaluation results with different number of granules, G. Blue and red indicate the mean and deviation of Absolute Different (AD)

4.6 Evaluation of number of granules

The parameter G determines the number of granules in an image. The greater the G value, the more the granules are used to represent an image. Evaluation with different number of granules, G, and the corresponding localization accuracy is shown in Fig. 11. High localization accuracy of granules is sought after to enable characterization of discriminative texture features for dense crowd density estimation. The F-score measure is used according to the well-known PASCAL challenge [9] to evaluate the localization accuracy by overlapping granules with pixel basis ground truth annotation of crowd and background obtained from [22]. The result shows that the higher the G value, the less precise is the separation between crowd and background per image. This is as expected, because with respect to the image size, with a greater G value, the image is represented with smaller size granules, where each granule contains fewer number of pixels. Consequently, fewer texture features are present to characterize the content (i.e. crowd or background) of the corresponding granule for accurate localization. Likewise, the smaller the G value, the fewer the granules are used to represent an image, which in turn generate larger size granules. When the size of a granule becomes too large, it can no longer represent the characteristics of a local atomic region.

We further evaluate different number of granules, G, for dense crowd density estimation. The performance of dense crowd density estimation with different number of granules, G, is depicted in Fig. 12. The result is consistent to the corresponding localization result in Fig. 11. This shows that perceptually meaningful atomic region, consisting of crowd or background only, allows discriminative features extraction of the respective regions, and thus enables accurate density estimation. Hence, in this study, we empirically set $G = 200$, which forms compact granules that outlines the natural boundaries between crowd and background regions for dense crowd density estimation.

5 Conclusion

In this paper, we have proposed a novel approach for dense crowd density estimation by using granules that conform to the natural outline between crowd and background. The proposed density estimation approach allows the granules to adapt themselves to the arbitrary distribution of crowd in which the underlying texture features characterizing crowd and background regions can be extracted. Moreover, using a set of complementing texture features is essential to compensate the insufficiencies of other features. The experimental results on public dense crowd datasets demonstrate that the use of granules is effective

in improving density estimation in dense crowd scenes. Despite the importance of dense crowd density estimation research, it is acknowledged that one of the main challenges for this research is generating ground truth for evaluation. This is because manual annotation of ground truth is costly and prone to human error.

Acknowledgements This research is supported by the GGPM grant GGPM-2017-024, from the National University of Malaysia (UKM); and Chee Seng Chan is supported by the Fundamental Research Grant Scheme (FRGS) MoHE Grant FP070-2015A, from the Ministry of Education Malaysia.

References

1. Ali S, Shah M (2007) A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp 1–6
2. Ali S, Nishino K, Manocha D, Shah M (2013) Modeling, simulation and visual analysis of crowds: A multidisciplinary perspective. In: Modeling, Simulation and Visual Analysis of Crowds, The International Series in Video Computing, vol 11. Springer, New York, pp 1-19
3. Chan AB, Dong D (2011) Generalized gaussian process models. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2681–2688
4. Chan AB, Vasconcelos N (2012) Counting people with low-level features and bayesian regression. IEEE Trans Image Process 21(4):2160–2177
5. Chan AB, Liang ZSJ, Vasconcelos N (2008) Privacy preserving crowd monitoring: Counting people without people models or tracking. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp 1–7
6. Chen K, Loy CC, Gong S, Xiang T (2012) Feature mining for localised crowd counting. In: Proceedings of the British Machine Vision Association Conference, vol 1, pp 21.1–21.11
7. Chen K, Gong S, Xiang T, Loy CC (2013) Cumulative attribute space for age and crowd density estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp 2467–2474
8. Davies AC, Yin JH, Velastin SA (1995) Crowd monitoring using image processing. Electron Commun Eng J 7(1):37–47
9. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
10. Field DJ (1987) Relations between the statistics of natural images and the response properties of cortical cells. J Opt Soc Amer A 4(12):2379–2394
11. Fu H, Ma H, Xiao H (2014) Scene-adaptive accurate and fast vertical crowd counting via joint using depth and color information. Multimed Tools Appl 73(1):273–289
12. Ge W, Collins RT (2009) Marked point processes for crowd counting. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp 2913–2920
13. Grant JM, Flynn PJ (2017) Crowd scene understanding from video: A survey. ACM Trans Multimed Comput Commun Appl 13(2):19
14. Helbing D, Mukerji P (2012) Crowd disasters as systemic failures: Analysis of the love parade disaster. EPJ Data Sci 1(1):1–40
15. Helbing D, Brockmann D, Chadefaux T, Donnay K, Blanke U, Woolley-Meza O, Moussaid M, Johansson A, Krause J, Schutte S, Perc M (2014) Saving human lives: What complexity science and information systems can contribute. J Stat Phys 158(3):735–781
16. Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1):55–67
17. Idrees H, Saleemi I, Seibert C, Shah M (2013) Multi-source multi-scale counting in extremely dense crowd images. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2547–2554
18. Jon G (2015) Saudi arabia hajj disaster death toll at least 2,110. Business Insider, http://www.businessinsider.com/ap-saudi-arabia-hajj-disaster-death -toll -at-least-2110-2015-10?IR=T&r=US&IR=T
19. Kae A, Sohn K, Lee H, Learned-Miller E (2013) Augmenting crfs with boltzmann machine shape priors for image labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2019–2026
20. Kaiman J (2015) Shanghai: dozens killed and injured in stampede at new year celebrations. The Guardian, http://www.theguardian.com/world/2014/dec/31/shanghai-35-people-killed-42-injured-new-year-crush

21. Kang K, Wang X (2014) Fully convolutional neural networks for crowd segmentation. arXiv:14114464
22. Kok VJ, Chan CS (2017) Grcs: Granular computing-based crowd segmentation. IEEE Trans Cybern 47(5):1157–1168
23. Kok VJ, Lim MK, Chan CS (2016) Crowd behavior analysis: A review where physics meets biology. Neurocomputing 177:342–362
24. Kovesi P (1999) Image features from phase congruency. Videre: J Comput Vis Res 1(3):1–26
25. Lempitsky V, Zisserman A (2010) Learning to count objects in images. In: Advances in Neural Information Processing Systems, pp 1324–1332
26. Lim MK, Kok VJ, Loy CC, Chan CS (2014) Crowd saliency detection via global similarity structure. In: International Conference on Pattern Recognition. IEEE, pp 3957–3962. lim and Kok contributed equally
27. Liu L, Xing J, Ai H, Lao S (2012) Semantic superpixel based vehicle tracking. In: International Conference on Pattern Recognition. IEEE, pp 2222–2225
28. Lloyd CD (2006) Local Models for Spatial Analysis. CRC Press, https://books.google.com.my/books?id=bIKToJ9en1UC
29. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
30. Loy CC, Chen K, Gong S, Xiang T (2013) Crowd counting and profiling: Methodology and evaluation. In: Modeling, Simulation and Visual Analysis of Crowds. Springer, pp 347-382
31. Marana A, Velastin S, Costa L, Lotufo R (1997) Estimation of crowd density using image processing. In: IEE Colloquium on Image Processing for Security Applications. IET, pp 11/1–11/8
32. Marana A, Velastin S, Costa L, Lotufo R (1998) Automatic estimation of crowd density using texture. Saf Sci 28(3):165–175
33. Mikolajczyk K, Schmid C (2004) Scale andamp; affine invariant interest point detectors. Int J Comput Vis 60(1):63–86
34. Morrone MC, Owens RA (1987) Feature detection from local energy. Pattern Recogn Lett 6(5):303–313
35. Mousse MA, Motamed C, Ezin EC (2017) People counting via multiple views using a fast information fusion approach. Multimed Tools Appl 76(5):6801–6819
36. Oppenheim AV, Lim JS (1981) The importance of phase in signals. Proc IEEE 69(5):529–541. https://doi.org/10.1109/PROC.1981.12022
37. Rabaud V, Belongie S (2006) Counting crowded moving objects. In: IEEE Conference on Computer Vision and Pattern Recognition, vol 1. IEEE, pp 705–711
38. Rodriguez M, Laptev I, Sivic J, Audibert JY (2011) Density-aware person detection and tracking in crowds. In: IEEE International Conference on Computer Vision. IEEE, pp 2423–2430
39. Rodriguez M, Sivic J, Laptev I (2013) Analysis of crowded scenes in video, pp 251–272. https://doi.org/10.1002/9781118577851.ch15
40. Ryan D, Denman S, Sridharan S, Fookes C (2015) An evaluation of crowd counting methods, features and regression models. Comput Vis Image Underst 130:1–17
41. Shao J, Kang K, Change Loy C, Wang X (2015) Deeply learned attributes for crowded scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 4657–4666
42. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge
43. Solmaz B, Moore BE, Shah M (2012) Identifying behaviors in crowd scenes using stability analysis for dynamical systems. IEEE Trans Pattern Anal Mach Intell 34(10):2064–2070
44. Tuytelaars T (2010) Dense interest points. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp 2281–2288
45. United Nations Department of Economic and Social Affairs, Population Division (2013) World Population Prospects: The 2012 Revision, vol 1. United Nations Publications, New York
46. United Nations Department of Economic and Social Affairs, Population Division (2014) World Urbanization Prospects: The 2014 Revision. Highlights United Nations Publications, New York
47. Vedaldi A, Fulkerson B (2010) Vlfeat: An open and portable library of computer vision algorithms. In: Proceedings of the International Conference on Multimedia. ACM, pp 1469–1472
48. Zhan B, Monekosso DN, Remagnino P, Velastin SA, Xu LQ (2008) Crowd analysis: a survey. Mach Vis Appl 19(5-6):345–357
49. Zhang C, Li H, Wang X, Yang X (2015) Cross-scene crowd counting via deep convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 833–841
50. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016) Single-image crowd counting via multi-column convolutional neural network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 589–597
51. Zhou B, Wang X, Tang X (2011) Random field topic model for semantic region analysis in crowded scenes from tracklets. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp 3441–3448