Robust Identification of Dense or Sparse Crowd Based on Classifier Fusion

Saikat Dutta[1](✉), Soumya Kanti Naskar[1], Sanjoy Kumar Saha[1],
and Bhabatosh Chanda[2]

[1] Jadavpur University, Kolkata, India
saikat.dutta779@gmail.com, rijunaskar@gmail.com, sks_ju@yahoo.co.in
[2] Indian Statistical Institute, Kolkata, India
chanda@isical.ac.in

Abstract. For a video surveillance system, crowd behavior analysis and crowd managing are important tasks. Along with the event in which crowd participates, its volume and density are also important in managing the crowd. Hence, characterizing the crowd as dense or sparse is an essential component of a crowd handling system. In this context, most of the existing methods try to estimate the headcount. Unlike those, the proposed method exploits the domain-knowledge based low-level features to classify the crowd image as dense or sparse. We present three simple systems working with three different feature sets. These are all free from the burden of background estimation. Experiments are carried on a dataset formed by taking the images from UCF-CC50 and SanghaiTech. Performance of all three feature sets are satisfactory, and Corner-Point based methodology provides the best result.

Keywords: Crowd density · Crowd classification · Dense or sparse crowd

1 Introduction

Crowd management has become an important task and video surveillance systems can be of great help in this context. In daily life people may gather at various public places like railway station and market place, and also for different activities or events like sports and cultural. To ensure safety and proper management, crowd behavior analysis is crucial. The behavioral anomaly of the crowd depends not only on the nature of the participating group but also on the crowd volume and *density*. Hence, estimating these parameters through video surveillance system is an important step towards crowd behavior analysis and management. In this paper, we present three novel methods for classifying the crowd image as dense or sparse using domain knowledge based low level features. Finally, classifiers are fused to develop a robust system.

The paper is organized as follows. This brief introduction is followed by a review of past work presented in Sect. 2. Proposed methodology is elaborated in Sect. 3. Section 4 presents the experimental results and discussion. Concluding remarks are sited in Sect. 5.

B. Deka et al. (Eds.): PReMI 2019, LNCS 11941, pp. 131–139, 2019.
https://doi.org/10.1007/978-3-030-34869-4_15

2 Past Work

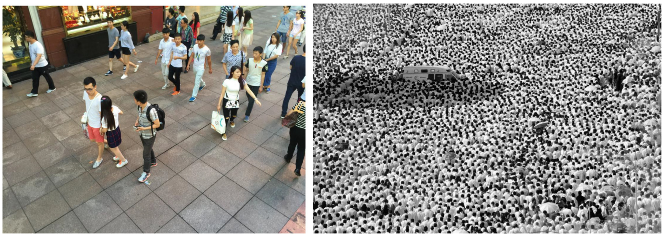
A large variety of methods exists in the literature. Some works are based on still images and some are on videos. Some works focus only on dense crowd images. One of the main approaches towards crowd density estimation is to count the population. This approach [1,2] can be sub-grouped as *human detection based* and *motion based*. In human detection based approach [3], the challenge lies in designing the human detector. and subsequent counting is straight forward. In motion based approach, the number of components with independent motion is taken as the count [4,5].

Marana et al. [6] used texture features in the form of Gray-level Dependence Matrices (GLDM) and applied Self Organizing Map (SOM) to classify crowd images to different density categories ranging from very low to very high. Li et al. [7] applied head-detector on the segmented foreground to obtain the count. Cheriyadat et al. [4] worked on image sequence with moving crowd, where low-level feature points are tracked, and regions with coherent motion are detected as objects for counting. SIFT features are also used for crowd detection in [8]. Corner points based methods are widely used to count the number of moving people [5,9]. Subburaman et al. [3] used gradient orientation features at interest points and Adaboost classifier. Jiang [10] proposed an improvisation on the regression based crowd counting mechanism. Idrees et al. [11] proposed a hybrid approach for highly dense crowd image, where head detector and interest point based count were combined with Fourier analysis. Hafeezallah et al. [12] introduced the curvelet frame change detection which enhances the statistical features for counting the individuals in the crowd.

In recent times convolutional neural network (CNN) is being used for crowd density estimation [13,14]. The network is trained with known crowd patches and then adapt it for target scenario. It is well known that obtaining a meaningful result from deep learning based method requires a huge training set whose distribution should be good representative of the population from which test (target) data would be drawn. Such a training set may not always be available. Second, it is observed that though a considerable variety of methods exists, there is not a single method that can handle all sorts of crowds. Moreover, some methods can handles image(s) of dense crowd only. Thus, characterizing a crowd as dense or sparse at the onset is essential in choosing an optimal strategy. In this work, we attempt to develop a robust system that can classify crowd image(s) into dense or sparse based on a small training set.

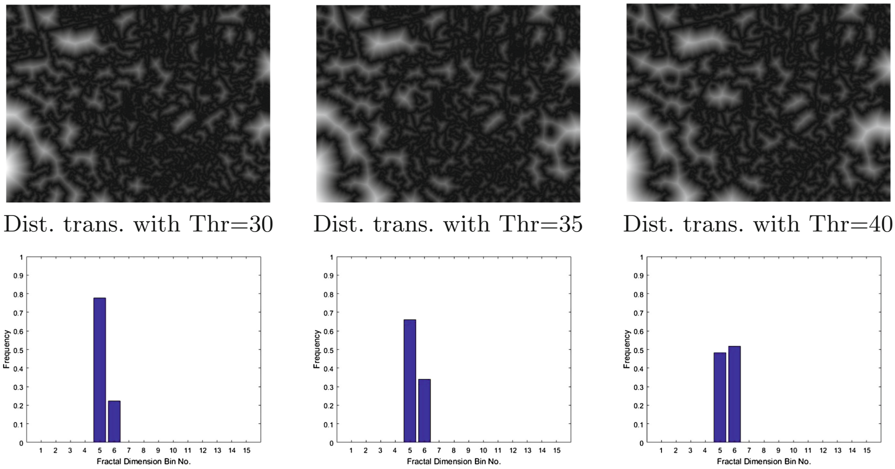
3 Proposed Methodology

In this work, we try to determine whether a crowd seen in an image is dense or sparse. Here, crowd image is conceived as texture image, and dense crowd image appears to be fine (micro) texture, while sparse crowd mimics coarse (macro) texture. Thus, sparse/dense crowd classification degenerates to fine/coarse texture classification. This motivates us to look for a variety of texture descriptors suitable for the task. Here we consider three different texture descriptors. First two try to rely on fractal dimension; whereas, the last one is based on count of interest (corner) points over. Feature extraction processes are detailed as follows.



Example of sparse crowd Example of dense crowd

Fig. 1. Sample images from the dataset

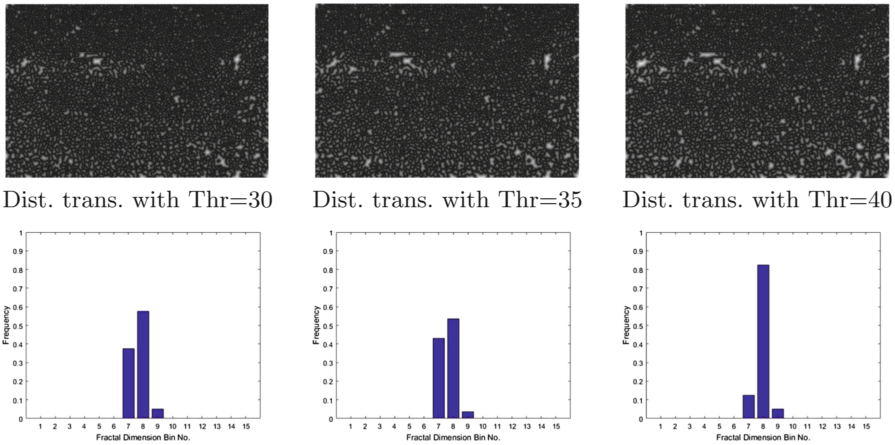


Dist. trans. with Thr=30 Dist. trans. with Thr=35 Dist. trans. with Thr=40

Histo. with Thr=30 Histo. with Thr=35 Histo. with Thr=40

Fig. 2. Distance transform based descriptor for example sparse crowd in Fig. 1

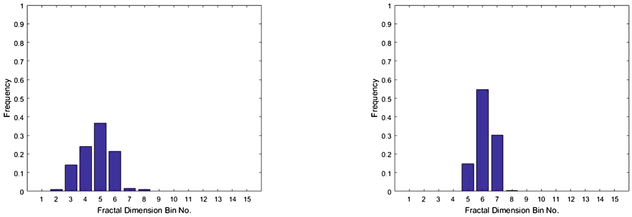
Descriptor Based on Distance Transform and Fractal Dimension: First, color image is converted into gray-scale image and segmented using morphological watershed algorithm [15,16], where gray-scale value of a pixel represents altitude at that location. The watershed line surrounds each region depicting a uniform surface feature. For Dense crowd images, a large number of small segments are obtained; while for sparse crowd, segments are large and small in number. Watershed algorithm produces a binary image with distinct regions with watershed line in-between. It may noted that one could have used any other segmentation scheme that generates closed contour.



Dist. trans. with Thr=30 Dist. trans. with Thr=35 Dist. trans. with Thr=40

Histo. with Thr=30 Histo. with Thr=35 Histo. with Thr=40

Fig. 3. Distance transform based descriptor for example dense crowd in Fig. 1

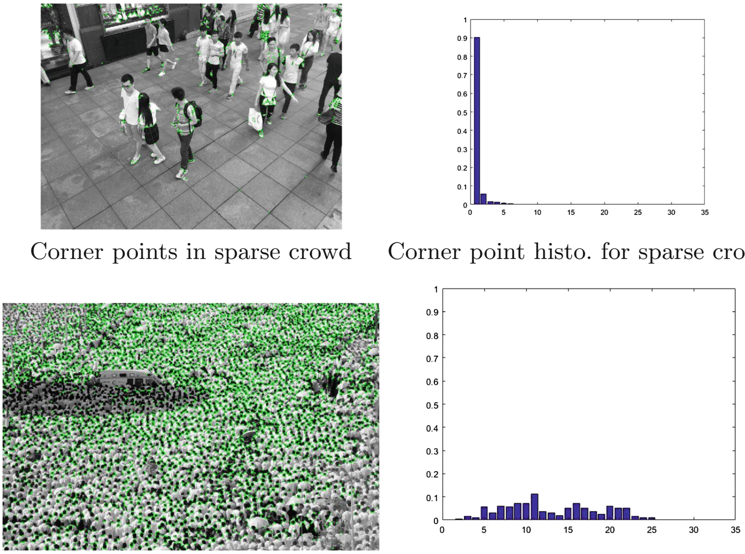


Histogram of Fractal Dimension for Sparse Crowd in Figure 1 Histogram of Fractal Dimension for Dense Crowd in Figure 1

Fig. 4. Fractal dimension descriptor for example sparse and dense crowds in Fig. 1

To extract texture feature from the said binary image, we apply distance transform [17]. The result of the transform is a two-dimensional matrix (say, T) of the same size as the image and a matrix element denotes the distance of the corresponding pixel from nearest watershed line. Hence, it reveals a kind (fine or coarse) of texture. Finally, texture feature is extracted from distance matrix in terms of fractal dimension. Note that, fractal dimension has already been used for texture segmentation [18,19]. It indicates roughness and self-similarity in the image. For a dense image, more self-similarity is expected compared to a sparse one. T is divided into $K \times K$ patches with a stride of K/p. Fractal dimension is computed over each patch. A normalized histogram of these fractal dimensions is taken as feature vector. Here, we empirically decide $K = 100$ and $p = 2$.

Watershed algorithm has a parameter that controls the segmentation process, and its selection is data-dependent and is a non-trivial task. Impact of different threshold values on segmentation will vary depending on the crowd density and the variation pattern can be an indicator of density. In our work, we take three



Corner points in sparse crowd Corner point histo. for sparse crowd

Corner points in dense crowd Corner point histo. for dense crowd

Fig. 5. Corner point based descriptor for example sparse and dense crowd in Fig. 1

threshold values: 40, 35 and 30 which are chosen empirically and applied to all the images in the dataset. Corresponding histograms are concatenated to form image texture descriptor. Figure 1 show sample sparse and dense images. Corresponding distance transform matrices and histograms are shown in Figs. 2 and 3. It is evident that the fractal dimension distribution is different the two types of crowd.

Descriptor Based on Fractal Dimension: Above algorithm is intuitively very promising, but it gets affected by certain issues. For example, we expect large and less number of segments in the binarized sparse crowd image. But the assumption fails in case of textured background. In order to get rid of it we drop the segmentation step. Fractal dimension is computed over each patch of gray-level image and these values are summarized into a normalized histogram of fractal dimension. The histograms of fractal dimension for sample sparse and dense crowd are shown in Fig. 4.

Descriptor Based on Corner Point: Fractal dimension based feature is global in nature and bears impact of background texture. To reduce such influence and to incorporate local character we focus on corner point based descriptor. Number of such points in a small patch of a dense crowd image is usually higher than that of sparse crowd image.

We extract corner points using Harris-Stephens algorithm [20]. Sensitivity factor is taken as 0.05. Then image is divided into patches as before. For each patch, corner points are counted. Histogram of normalized count is taken as the descriptor. The histograms of example crowd images are shown in Fig. 5.

Usually for a dense crowd, the non-zero histogram bins spread over large counts, whereas sparse crowd they are usually restricted to lower range of counts with strong peak. To reduce the effect of noise, an edge preserving smoothing [21] can be applied as pre-processing.

3.1 Classification

For all the three descriptors, we have used Decision tree as classifier [22]. During training, data is split at each decision node based on maximization of information gain at child nodes. During test, a simple condition is tested on feature at each node and corresponding branch is taken. This process goes on recursively and eventually a leaf node is reached based on which we predict the class-label.

Fusing the Classifier: It is understood from the description of features that some of them are supplementary and some are redundant too. Second, the classifier must be robust. That means, standard deviation of various test run must be as low as possible. So it may worth exploring the fusion of the classifiers based on these features. We have tried both kind of fusion: feature level fusion and decision level fusion. In the former case, three sets of features obtained based on (i) distance transform and fractal dimension, (ii) fractal dimension, and (iii) count of corner points are concatenated together to form a single feature vector, which is then fed to the classifier. In the latter case, output or decision obtained from each of the classifiers using three different feature sets as stated above are combined through an artificial neural network with three input nodes, two output nodes and a hidden layer. Results of fused classifiers are also reported.

4 Experimental Results and Discussion

We have performed the experiments on a machine with Intel®Core$^{\text{TM}}$i5-5200U CPU and 4 GB RAM. All the codes are written in MATLAB®.

Although there are many public datasets for crowd counting and tracking, dataset for crowd density based classification is not readily available, at least, to the best of our knowledge. Hence, we have created a dataset by collecting images from UCF-CC50 dataset [23] and SanghaiTech dataset [24]. The images are selected manually in a manner such that the pictures mostly contain the region of interest, *i.e.*, spaces where crowd is actually present. Multiple raters were employed to categorize these clearly as *dense* or *sparse*. Based on the raters opinion ground-truth is associated with each image as label. Final label is assigned to each image based on majority voting. The dataset thus prepared contains 64 dense and 64 sparse crowd images to avoid imbalance in dataset of either type.

To run the experiment with the given dataset, we have randomly partitioned the dataset of each category into two halves, trained the model, i.e., decision tree classifier with one half and test on the other half. This is done 50 times and an average score of accuracy is reported in Table 1 as a quantitative measure of performance of the proposed system.

For comparison among the descriptors, experiment is done for each descriptor separately and average classification accuracy is shown in the first three rows of Table 1. Table 1 reveals that accuracy due to corner point based descriptor (96.18%) is significantly higher than that of the fractal dimension based descriptors (80.06% and 88.59%). Second, lower standard deviation of the former indicates that this descriptor develops more robust descriptors compared to the other two. We tried to work with other widely used classifiers like neural network and SVM. But, the performance was poor and that can be attributed to limited dataset. For the same reason also we could not explore deep learning approach.

Table 1. Classification accuracy for different descriptors

	Sparce accuracy	Dense accuracy	Overall accuracy
Distance transform descriptor	82.56% ± 10.18%	77.56% ± 8.15%	80.06% ± 4.65%
Fractal dimension descriptor	88.06% ± 8.64%	89.12% ± 6.06%	88.59% ± 4.34%
Corner points descriptor	96.43% ± 2.60%	95.93% ± 3.64%	96.18% ± 1.89%
Feature level fusion	95.00% ± 3.68%	93.81% ± 4.21%	94.41% ± 2.21%
Decision level fusion	95.5% ± 4.77%	91.93% ± 7.66%	93.72% ± 4.31%
MCNN [24]	94.09% ± 2.42%	97.18% ± 2.18%	91.0% ± 4.78%

As suggested earlier, we have explored both feature level and decision level fusion of classifier.

Results are shown in 4th and 5th rows of Table 1. It is revealed that though in both cases robustness is improved, it cannot exceed the performance of corner point based descriptor. These indicates that fractal dimension based features are complementary to corner based descriptors and do not add any value while they are fused. Second, performance of decision level fusion and feature level fusion are same in terms of statistical significance.

We have compared the performance with Multi-column CNN (MCNN) used in [24]. The pretrained network is used to prepare the density map for the images of our dataset and that is used as input to neural network with one hidden layer. Results in Table 1 shows that accuracy of MCNN is less than corner point based descriptor and fused classifiers (both feature level and decision level).

5 Conclusion

In this work we have presented a simple method to classify a crowd image as dense or sparse. Proposed method exploits three different descriptors based om domain knowledge. It is found that among those features, interest point based feature performs best because it includes local information. Most important part is that neither of the features require interest region segmentation nor background subtraction. It is also seen that classifier fusion leads to more robustness or less variation in performance. But as these methods rely on texture information, a texture-heavy sparse crowd image may be wrongly classified as dense one. This issue may be addressed in future. Moreover, dataset can be further enhanced to include more variety and also to utilize deep learning. However, the work shows proposed feature based methodology has good potential in classifying the crowd as dense or sparse.

References

1. Ali, S., Nishino, K., Manocha, D., Shah, M.: Modeling, simulation and visual analysis of crowds: a multidisciplinary perspective. In: Ali, S., Nishino, K., Manocha, D., Shah, M. (eds.) Modeling, Simulation and Visual Analysis of Crowds. TISVC, vol. 11, pp. 1–19. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-8483-7_1
2. Hashemzadeh, M., Pan, G., Yao, M.: Counting moving people in crowds using motion statistics of feature-points. Multimed. Tools Appl. **72**(1), 453–487 (2014)
3. Subburaman, V.B., Descamps, A., Carincotte, C.: Counting people in the crowd using a generic head detector. In: 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 470–475. IEEE (2012)
4. Cheriyadat, A. M., Bhaduri, B.L., Radke, R.J.: Detecting multiple moving objects in crowded environments with coherent motion regions. In: Computer Vision and Pattern Recognition Workshops (2008)
5. Albiol, A., Silla, M.J., Albiol, A., Mossi, J.M.: Video analysis using corner motion statistics. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, pp. 31–38 (2009)
6. Marana, A.N., Velastin, S.A., Costa, L.D.F., Lotufo, R.: Automatic estimation of crowd density using texture. Saf. Sci. **28**(3), 165–175 (1998)
7. Li, M., Zhang, Z., Huang, K., Tan, T.: Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In: International Conference on In Pattern Recognition (ICPR) (2008)
8. Arandjelovic, O.: Crowd detection from still images. In BMVC 2008: Proceedings of the British Machine Vision Association Conference, pp. 1–10. BMVA Press (2008)
9. Dittrich, F., Koerich, A., Oliveira, L.: People counting in crowded scenes using multiple cameras. In: 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 138–141. IEEE (2012)
10. Mei, J.: An improved method of crowd counting based on regression (2013)
11. Idrees, H., Saleemi, I., Seibert, C., Shah, M.: Multi-source multi-scale counting in extremely dense crowd images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
12. Hafeezallah, A., Abu-Bakar, S.: Crowd counting using statistical features based on curvelet frame change detection. Multimed. Tools Appl. **76**(14), 15777–15799 (2017)
13. Zhang, C., Li, H., Wang, X., Yang, X.: Cross-scene crowd counting via deep convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 833–841. IEEE (2015)
14. Han, K., Wan, W., Yao, H., Hou, L.: Image crowd counting using convolutional neural network and Markov random field. arXiv preprint arXiv:1706.03686 (2017)
15. Couprie, M., Bertrand, G.: Topological gray-scale watershed transformation. In: Vision Geometry Vi, vol. 3168, pp. 136–147. International Society for Optics and Photonics (1997)

16. Bertrand, G.: On topological watersheds. J. Math. Imaging Vis. **22**(2–3), 217–230 (2005)
17. Rosenfeld, A., Pfaltz, J.L.: Sequential operations in digital picture processing. J. ACM (JACM) **13**(4), 471–494 (1966)
18. Keller, J.M., Chen, S., Crownover, R.M.: Texture description and segmentation through fractal geometry. Comput. Vis. Graph. Image Processing **45**(2), 150–166 (1989)
19. Chaudhuri, B.B., Sarkar, N.: Texture segmentation using fractal dimension. IEEE Trans. Pattern Anal. Mach. Intell. **17**(1), 72–77 (1995)
20. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, vol. 15, no. 50. Citeseer, pp. 10–5244 (1988)
21. Huang, T., Yang, G., Tang, G.: A fast two-dimensional median filtering algorithm. IEEE Trans. Acoust. Speech Signal Process. **27**(1), 13–18 (1979)
22. Breiman, L.: Classification and Regression Trees. Routledge, London (2017)
23. Idrees, H., Saleemi, I., Seibert, C., Shah, M.: Multi-source multi-scale counting in extremely dense crowd images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2547–2554 (2013)
24. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 589–597 (2016)