Pedestrian Crowd Detection and Segmentation using Multi-Source Feature Descriptors

Saleh Basalamah[1], Sultan Daud Khan[2]
Umm Al-Qura University, Saudi Arabia[1]
National University of Technology, Pakistan[2]

***Abstract*—Crowd analysis is receiving much attention from research community due to its widespread importance in public safety and security. In order to automatically understand crowd dynamics, it is imperative to detect and segment crowd from the background. Crowd detection and segmentation serve as preprocessing step in most crowd analysis applications, for example, crowd tracking, behavior understanding and anomaly detection. Intuitively, the crowd regions can be extracted using background modeling or using motion cues. However, these model accumulate many false positives when the crowd is static. In this paper, we propose a novel framework that automatically detects and segments crowd by integrating appearance features from multiple sources. We evaluate our proposed framework using challenging images with varying crowd densities, camera viewpoints and pedestrian appearances. From qualitative analysis, we observe that the proposed framework work perform well by precisely segmenting crowd in complex scenes.**

Keywords—Crowd detection; Fourier analysis; crowd analysis; crowd segmentation

I. Introduction

With the growing population of the world and with the increased urbanization, scientific community focus on developing tools and techniques to ensure crowd safety. Events like sports, festivals, concerts, and carnivals, where the participants count in thousands, may lead to crowd disaster. Therefore, event organizers and security personnel must adopt adequate safety measures to ensure crowd safety. Crowd disasters still occur very frequently despite strict security and safety measures. One of the main reason of crowd disaster is the thousands of people gathered in a constrained environment which results in critically increased densities.

In order to ensure public safety in such high density situations, surveillance cameras are installed in multiple locations providing coverage of whole crowd scene. Generally, this is the job of security personnel to detect abnormal activities by watching over the TV. This kind of manual surveillance is a tiresome job and due to limited human capacity prone to human errors. Therefore, as solution an automatic analysis of the crowd is required which can reliably analyze the crowd dynamics. Designing such virtual analyst has attracted the interest of computer science community. Despite the recent advancements in computer vision technology, research community still did achieve desired result for understanding crowd dynamics. This attributes to following reasons: (1) Most of existing methods are based on assumptions often violated in real world scenarios. (2) Due to limited data availability, it is hard to train network effectively. Therefore as a solution, crowd simulation models have been used to provide synthetic data for training and also for validation of the computer vision algorithms. In order to automatically analyze the crowd dynamics, several computer vision tool sets [1], [2], [3], [4], [5], [6] are proposed. Using live video stream, these tools compute important measurements that are of significant importance to crowd managers and security personnel. These measurements include but not limited, crowd counting, density estimation, anomaly detection, crowd tracking. Although these tool sets computes important measurements that are useful in understanding crowd dynamics yet these tools could not detect detect crowds in the scene.

For understanding crowd dynamics, detection and tracking of pedestrians are the important steps. However, before starting any crowd analysis, crowd detection and segmentation is the preprocessing step. Intuitively , Crowd detection and segmentation can be achieve by motion segmentation techniques. But we observe that in real videos, large portion of crowd remain stationary and these stationary groups can be captured using motion segmentation techniques. Another shortcoming of motion segmentation is that it accumulates many false positives by detection motion of objects belong to other categories.

Crowd detection and segmentation serve as pre-processing step in many applications of crowd surveillance. However, crowd detection and segmentation is challenging task due to the following reasons. (1) In high density crowds, pedestrians generally stand close to each other due to constrained and limited space and environment. This cause severe occlusions among pedestrians. (2) Severe clutter in the scene usually confuse detector to distinguish between background and crowd.

In order to address above challenges, we proposed a framework that integrate appearance features and train a linear Support Vector Machine (SVM) classifier. Our proposed framework takes input of arbitrary size and divide into multiple cells in a grid from. Then for each cell we compute three descriptors, i.e, Local Binary Pattern (LBP), Fourier Analysis and Gray Level Co-occurrence Matrix (GLCM). The corresponding appearance features are then concatenated in a lineary fashion and SVM classifier is trained to classify each cell into crowd or non-crowd patch. Later on, we employ 11 x 11 2D-gaussian kernel to smooth the final output.

Our contributions can be summarized as follows:

- Our method does not require detection and tracking instead rely on low level features that works well in all crowd scenes.
- Our approach reduce the computational cost by detecting crowd from a single image instead of using whole video sequence.

- Our approach do not use background subtraction and do not make use of motion information.
- Our approach do not rely on pedestrian detection and tracking yet utilized appearance feature, therefore, can be applicable in both low and high density crowds.
- Our approach ease the process of crowd analysis by detecting only Region of Interest (crowded area).
- We evaluate our method on different scenes. The experiments results shows that our proposed method can precisely localize the crowd.

The rest of paper is organized as follows: we discuss related work in Section II, proposed methodology is discussed in Section III, Section V discusses experiment results and Section VI concludes the paper.

II. Related Work

There is inadequate work reported in literature on crowd detection segmentation. Most of crowd related literature focus on crowd counting, density estimation, tracking and anomaly detection. Automated analysis of crowd behaviors has large number of applications applications ranging from prediction of congestion to the discovery of abnormal behaviors or flows. Most of the research is focused on detecting anomalies in videos [7], [8], [9], [10], [11], [12], [13], counting people in crowds [14], [15], [16], [17], [18], [19], [20], [21], characterizing different motion flows and segmentation [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32]. Other works aimed at detecting/tracking individuals or group of individuals in crowd scene [33], [34], [35], [36], [37], [38], [39], [40] rather than identifying crowd behaviours. As there is growing interest in crowd dynamics understanding in general, identifying crowd behaviors specifically has not been studied in depth, and very few papers have explicitly concentrated on identifying or modelling crowd behaviour. Berkan et al. [41] identify five crowd behaviours, i.e. blocking, lane, bottleneck, ring/arch, fountainhead. A similar framework is proposed by [42], where the same five crowd behavior are identified. Khan et al. [43] proposed a method that can identify crowd behaviors by utilizing source and sink information of trajectories. Recognizing the recent success of convolution neural network in filed of object detection and segmentation, [44] proposed CNN based on two stream network fusion network [45], originally designed for video action recognition to identify multiple crowd behaviors in crowd scene. However, Convolution neural network (CNN) has not gained adequate performance that have been achieved in image classification and object detection. Part of reason is the lack of data set for training or the data sets are too small and noisy. Compared to the image classification, classification of crowd behaviors has the additional challenge of variations in motion, viewpoint and scales. Due to these challenges, we require more training examples. Another reason is that CNN are not able to take full advantage of temporal information existing between consecutive frames of the video.

III. Proposed Methodology

The overall methodology of our proposed framework is shown in Fig. 1. During training phase, given a set of images, we first divide each image into grid of cells. Then for each cell we compute low-level appearance features. including local-binary pattern, Fourier analysis and Gray Level Co-occurrence Matrix (GLCM). We then construct a long feature vector by concatenating all features. Then a linear classifier is trained using long feature vector. The size of feature vector is 128. During testing phase, each input image passed through the same steps and extracted long feature vector is then mapped to learned classifier for generating the confidence score for each cell. Later on, Gaussian kernel is applied to smooth the final output.

A. Local Binary Pattern

For texture and appearance base classification, Local Binary Pattern (LBP) is a perfect choice. For each pixel p, we compute LBP with angular quantization of 8 pixels and with spatial resolution of 1. We then compare pixel p with its 8 neighbourhood pixels in such a way that the output is 0 if the intensity of pixel p is less than its neighbourhood and 1 if it is greater. We then process all pixels of input image in the same way and generate normalized histogram. With this unique representation, we can only capture the appearance information of the whole image. For capturing the texture information, we compute Gray-Level Co-occurrence matrix from the intensity values. We then compute Entropy (**P**), Energy (**E**), Contrast (**C**) and Homogeneity (**H**) as in the following equations.

$$P = -\sum_x \sum_y p[x,y] \log p[x,y] \tag{1}$$

$$E = -\sum_x \sum_y p^2[x,y] \tag{2}$$

$$C = -\sum_x \sum_y (x-y)^2 p[x,y] \tag{3}$$

$$H = -\sum_x \sum_y p[x,y]/1 + |x-y| \tag{4}$$

where x and y represent horizontal and vertical components of the input image.

B. Fourier Analysis

Fourier Analysis is a unique way of extracting appearance information from the image. We observed that in high density crowds, where large number of people gather in a constrained environment, far away pedestrians cover only few pixels due to perspective distortions. In this way, pedestrian detection and histogram of gradient can can impart useful information. We further observed that high density crowds have repetitive structures, as all pedestrians appear same from a distant camera view point. This unique appearance of crowd can be easily represented by Fourier Transform F_t. When we convert input image to Fourier domain, these repetitive structure of pedestrian heads can be easily detected using the peaks of frequency domain. In order to tackle scale problem that is caused by perceptive distortions, we divide the image into patches. Here, we assume that the density in the patch is same.

For each patch P that belongs to input image I, we perform the following steps:

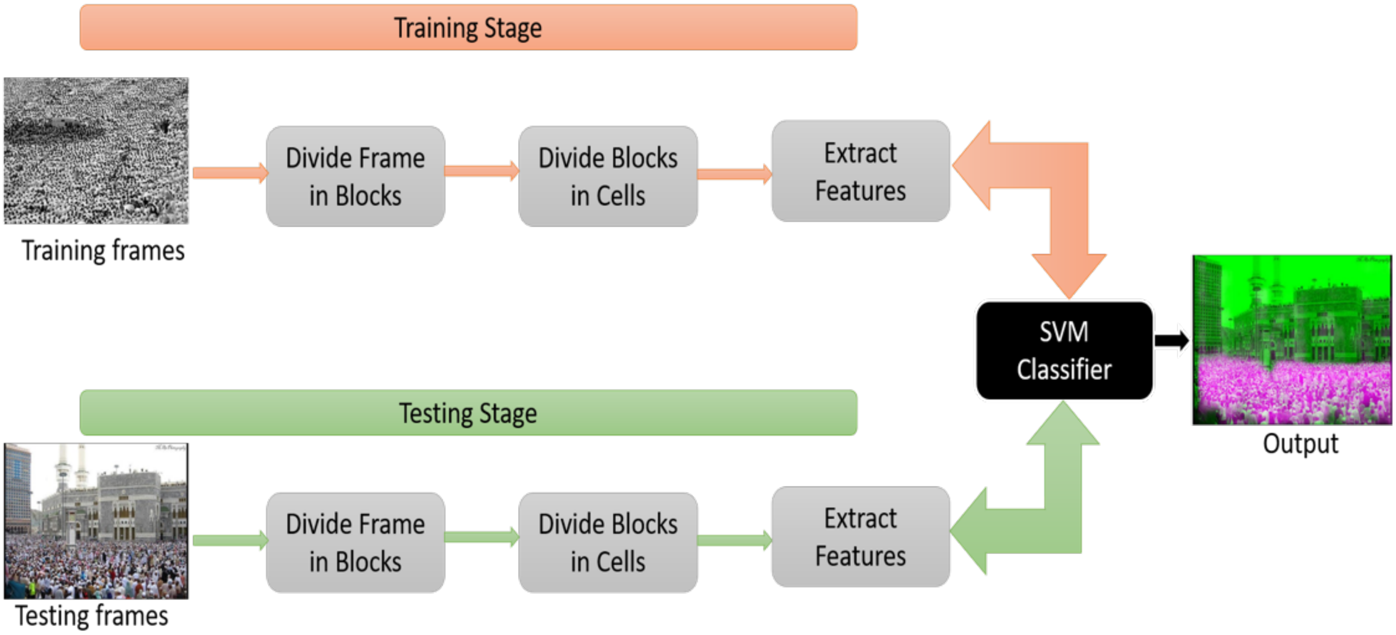

Fig. 1. Pipeline of proposed framework both during testing and training phase. Input image is divided into block and the into cells. Features are then extracted from each cell and train SVM classifier.

- Convert the patch into gradient Δ_P.
- Apply Fourier Transform on Δ_P and then apply low pass filter. This step is important to remove high frequency components from the signal as it contains information about the edges.
- Remove low amplitude component by applying a threshold υ. We set the value of υ to 0.4.
- Reconstruct the image P_r by applying inverse Fourier Transform and apply non-maxima suppression.

After reconstruction, we compute the following statistical features, mean (**M**), Variance (**V**), Skewness (**S**), and Kurtosis (**K**)

$$M = \frac{1}{xy} \sum_{x,y \in P_r} P_r(x,y) \tag{5}$$

$$V = \frac{1}{xy-1} \sum_{x,y \in P_r} \left(P_r(x,y) - \frac{1}{xy-1} \sum_{x,y \in P_r} P_r(x,y) \right)^2 \tag{6}$$

$$S = \frac{\frac{1}{xy-1} \sum_{x,y \in P_r} \left(P_r(x,y) - \frac{1}{xy-1} \sum_{x,y \in P_r} P_r(x,y) \right)^3}{\left(\frac{1}{xy-1} \sum_{x,y \in P_r} \left(P_r(x,y) - \frac{1}{xy-1} \sum_{x,y \in P_r} P_r(x,y) \right)^2 \right)^{\frac{3}{2}}} \tag{7}$$

$$K = \frac{\frac{1}{xy-1} \sum_{x,y \in P_r} \left(P_r(x,y) - \frac{1}{xy-1} \sum_{x,y \in P_r} P_r(x,y) \right)^4}{\left(\frac{1}{xy-1} \sum_{x,y \in P_r} \left(P_r(x,y) - \frac{1}{xy-1} \sum_{x,y \in P_r} P_r(x,y) \right)^2 \right)^2} \tag{8}$$

where P_r is image patch and x and y are the horizontal and vertical coordinates of the patch.

C. Gray Level Co-occurrence Matrix

Gray Level Co-occurrence Matrix commonly used for texture extraction and detection. GLCM uses distribution of gray-level of neighboring pixels in relation with center pixel. Marana et al. proposed a method of GLCM for utilizing texture information for crowd density estimation. In our case, we adapt it to train a binary classifier that distinguish background from the crowded patches. The Gray-Level Co-Occurrence Matrix (GLCM) $P[x, y]$ is computed by taking the sum of all pixel pairs having gray value x and y separated by distance parameter d and at an orientation $theta$ of 0sidegree 45sidegree, 90sidegree and 135sidegree, respectively. The counting is converted to joint conditional probability $P(x, y/\theta, d)$.

After computing GLCM for input image, we then extract features, for example, Entropy (**P**), Energy (**E**), Contrast (**C**) and Homogeneity (**H**) using equations 1, 2, 3, and 4.

IV. Features Fusion

In this section, we discuss the fusion of different features from different sources as shown in Fig. 2. For training, we use N number of frame. In order to ensure patch wise training, we divide each image into dense overlapping patches. We then extract appearance features using above mentioned sources and

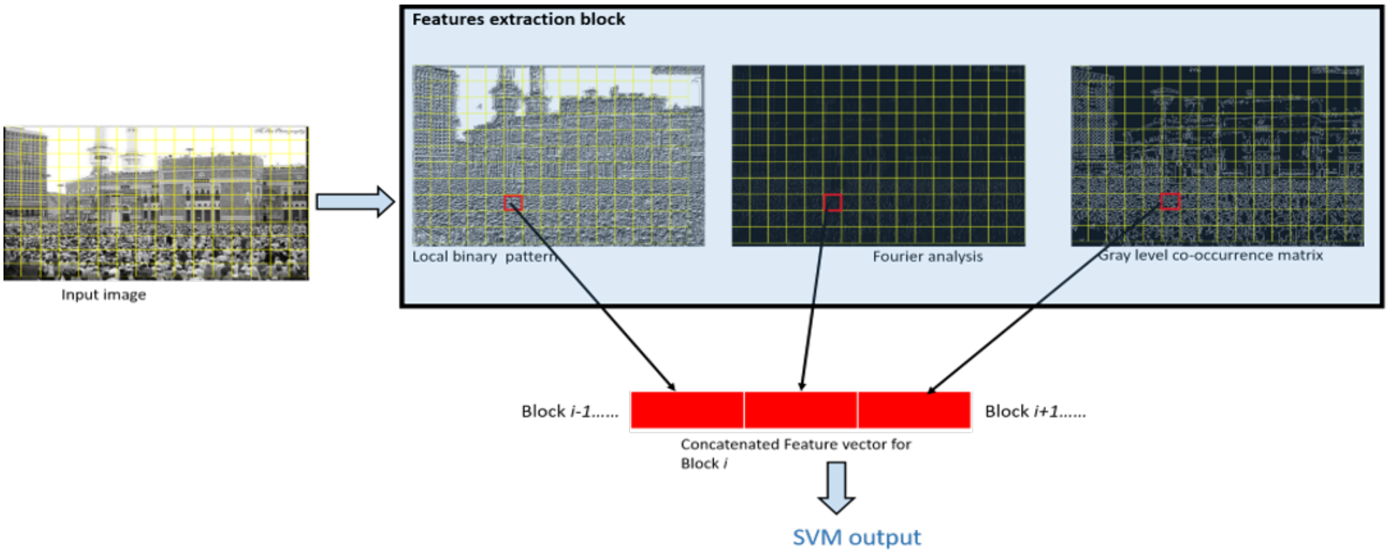
Features extraction block
Input image
Local binary pattern
Fourier analysis
Gray level co-occurrence matrix
Block i-1……
Block i+1……
Concatenated Feature vector for Block i
SVM output

Fig. 2. Sample input frame is divided into multiple blocks. Feature are extracted for block *i*(red). Features are concatenated and feed to the SVM classifier to assign a score.

combine the resultant features into a long feature vector. Let $F_i = \{m_i^1, m_i^2 \ldots, x_i^M\}$ is the final feature vector for input image i and M represent the number of patches in the image.

V. Experimental Results

In this section, we evaluate the effectiveness of our proposed approach. In order to evaluate our proposed method, we publicly available UCF_CC_50 [16] dataset. UCF_CC_50 is challenging dataset that contains 50 images captured from 50 different scenes with significant variations in resolution, camera view points and densities. The density in images varies from 94 person / image to 4543 persons/image. We randomly divide the data set into training and testing samples using the same convention used in [16]. During training, we cropped multiple patches from the image and divide them into crowd and non-crowd patches. Fig. 3 clearly illustrates the input image with corresponding crowd and non-crowd patches. We feed these patches to train our classifier.

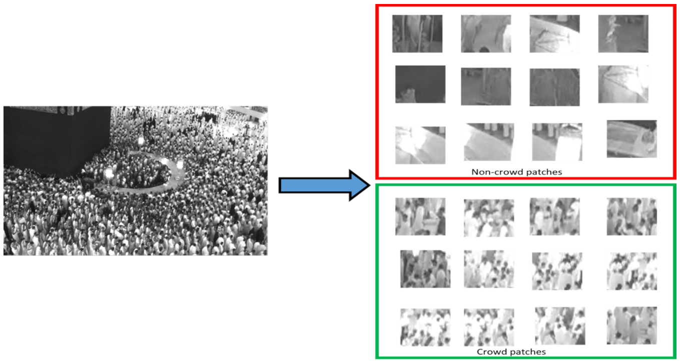

Fig. 3. Sample image used for training. The image is divided into positive and negative patches. We extract features from these patches and feed them for training SVM classifier.

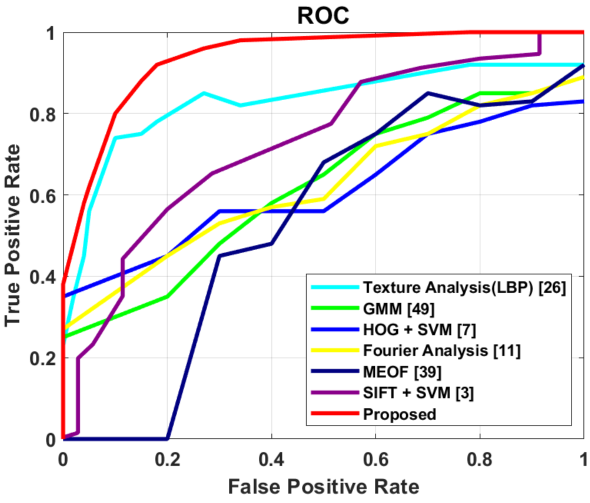

Fig. 4. ROC curves of different methods using UCF_CC_50 [16] dataset.

The results of our proposed framework is shown in Fig. 5. From the Figure, it is obvious that our proposed framework effectively discriminate crowd from non-crowd regions and precisely segment the crowded area. From the Fig. 5, it is obvious that our proposed method achieve impressive results with small number of false positives. These false positives attribute to the fact that our proposed framework also treat “leafy” areas as crowded regions.

We also compare our results with other baseline methods in a quantitative way. The first based line method is traditional Gaussian mixture model (GMM) for background subtraction. The second baseline model is based on motion extraction using optical flow (MEOF). Third method (HOG + SVM) is patch based model that utilizes Histogram of Oriented Gradient (HOG) features to train SVM classifier. In addition we also compare our results with SIFT + SVM [46], Fourier Analysis [16], and Texture [47]. We keep the size of patch to

TABLE I. COMPARATIVE ANALYSIS WITH OTHER TECHNIQUES ON UCF_CC_50 [16]

Method	**AUC**
GMM [48]	0.27
MEOF [27]	0.15
HOG + SVM [49]	0.45
SIFT + SVM [46]	0.56
Fourier analysis [16]	0.37
Texture analysis (LBP) [47]	0.58
Proposed	0.63

32 x 32 pixels in all our experiments. We use Area-under-curve (AUC) from ROC curves as evaluation metrics. We report the ROC and AUC results in Fig. 4 and Table I respectively. From the results, it is obvious that our proposed framework outperforms other state-of-the-art methods. We further observed that texture and appearance features work well in high density crowds since they capture regular and repetitive structure of the crowds.

VI. CONCLUSION

In this paper, we propose a novel approach to detect and segment crowd based on low-level appearance features. The proposed framework is tested on challenging images with large scale variations in densities, viewpoints and appearances of pedestrians in crowd. From experiment results, we showed that our proposed framework can precisely detect and segment crowded regions in the scene.

In future, we plan to integrate the proposed framework with various crowd tracking and crowd behavior understanding applications.

ACKNOWLEDGMENT

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU for this research.

REFERENCES

[1] S. D. Khan, M. Tayyab, M. K. Amin, A. Nour, A. Basalamah, S. Basalamah, and S. A. Khan, “Towards a crowd analytic framework for crowd management in majid-al-haram,” *arXiv preprint arXiv:1709.05952*, 2017.

[2] S. D. Khan, M. Saqib, and M. Blumenstein, “Towards a dedicated computer vision tool set for crowd simulation models,” *arXiv preprint arXiv:1709.02243*, 2017.

[3] S. D. Khan, G. Vizzari, and S. Bandini, “A computer vision tool set for innovative elder pedestrians aware crowd management support systems.” in *AI* AAL@ AI* IA*, 2016, pp. 75–91.

[4] ——, “Facing needs and requirements of crowd modelling: Towards a dedicated computer vision toolset,” in *Traffic and Granular Flow’15*. Springer, 2016, pp. 377–384.

[5] S. D. Khan, L. Crociani, and G. Vizzari, “Integrated analysis and synthesis of pedestrian dynamics: First results in a real world case study,” *From Objects to Agents*, 2013.

[6] ——, “Pedestrian and crowd studies: Towards the integration of automated analysis and synthesis.”

[7] H. Ullah, A. B. Altamimi, M. Uzair, and M. Ullah, “Anomalous entities detection and localization in pedestrian flows,” *Neurocomputing*, vol. 290, pp. 74–86, 2018.

[8] L. Kratz and K. Nishino, “Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models,” 2009.

[9] H. Ullah and N. Conci, “Crowd motion segmentation and anomaly detection via multi-label optimization,” in *ICPR workshop on Pattern Recognition and Crowd Analysis*, 2012.

[10] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1975–1981.

[11] H. Ullah, M. Ullah, and N. Conci, “Real-time anomaly detection in dense crowded scenes,” in *Video Surveillance and Transportation Imaging Applications 2014*, vol. 9026. International Society for Optics and Photonics, 2014, p. 902608.

[12] H. Ullah, L. Tenuti, and N. Conci, “Gaussian mixtures for anomaly detection in crowded scenes,” in *Video Surveillance and Transportation Imaging Applications*, vol. 8663. International Society for Optics and Photonics, 2013, p. 866303.

[13] H. Ullah, M. Ullah, H. Afridi, N. Conci, and F. G. De Natale, “Traffic accident detection through a hydrodynamic lens,” in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2470–2474.

[14] V. Rabaud and S. Belongie, “Counting crowded moving objects,” in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 705–711.

[15] C. Zhang, H. Li, X. Wang, and X. Yang, “Cross-scene crowd counting via deep convolutional neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 833–841.

[16] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, “Multi-source multi-scale counting in extremely dense crowd images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2547–2554.

[17] M. Arif, S. Daud, and S. Basalamah, “Counting of people in the extremely dense crowd using genetic algorithm and blobs counting,” *IAES International Journal of Artificial Intelligence*, vol. 2, no. 2, p. 51, 2013.

[18] ——, “People counting in extremely dense crowd using blob size optimization,” *Life Science Journal*, vol. 9, no. 3, pp. 1663–1673, 2012.

[19] M. Saqib, S. D. Khan, and M. Blumenstein, “Texture-based feature mining for crowd density estimation: A study,” in *Image and Vision Computing New Zealand (IVCNZ), 2016 International Conference on*. IEEE, 2016, pp. 1–6.

[20] S. Khan, G. Vizzari, S. Bandini, and S. Basalamah, “Detecting dominant motion flows and people counting in high density crowds,” 2014.

[21] M. Marsden, K. McGuinness, S. Little, and N. E. O’Connor, “Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification,” in *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*. IEEE, 2017, pp. 1–7.

[22] S. D. Khan, G. Vizzari, and S. Bandini, “Identifying sources and sinks and detecting dominant motion patterns in crowds,” *Transportation Research Procedia*, vol. 2, pp. 195–200, 2014.

[23] M. Saqib, S. D. Khan, N. Sharma, and M. Blumenstein, “Extracting descriptive motion information from crowd scenes,” in *2017 International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE, 2017, pp. 1–6.

[24] M. Saqib, S. D. Khan, and M. Blumenstein, “Detecting dominant motion patterns in crowds of pedestrians,” in *Eighth International Conference on Graphic and Image Processing (ICGIP 2016)*, vol. 10225. International Society for Optics and Photonics, 2017, p. 102251L.

[25] H. Ullah, M. Ullah, and N. Conci, “Dominant motion analysis in regular and irregular crowd scenes,” in *International Workshop on Human Behavior Understanding*. Springer, 2014, pp. 62–72.

[26] S. Ali and M. Shah, “A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis,” in *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*. IEEE, 2007, pp. 1–6.

[27] H. Ullah and N. Conci, “Structured learning for crowd motion segmentation,” in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 824–828.

[28] H. Ullah, M. Ullah, and M. Uzair, “A hybrid social influence model for pedestrian motion segmentation,” *Neural Computing and Applications*, pp. 1–17, 2018.

[29] M. Gao, J. Jiang, J. Shen, G. Zou, and G. Fu, “Crowd motion segmentation and behavior recognition fusing streak flow and collectiveness,” *Optical Engineering*, vol. 57, no. 4, p. 043109, 2018.

[30] H. Ullah, M. Uzair, M. Ullah, A. Khan, A. Ahmad, and W. Khan, “Density independent hydrodynamics model for crowd coherency detection,” *Neurocomputing*, vol. 242, pp. 28–39, 2017.

[31] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu, “Unified crowd segmentation,” in *European conference on computer vision*. Springer, 2008, pp. 691–704.

[32] H. Mansour, C. Dicle, D. Tian, M. Benosman, and A. Vetro, “Method and system for segmenting pedestrian flows in videos,” Feb. 13 2018, uS Patent 9,892,520.

[33] G. J. Brostow and R. Cipolla, “Unsupervised bayesian detection of independent motion in crowds,” in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 594–601.

[34] L. Kratz and K. Nishino, “Tracking with local spatio-temporal motion patterns in extremely crowded scenes,” 2010.

[35] S. D. Khan, G. Vizzari, S. Bandini, and S. Basalamah, “Detection of social groups in pedestrian crowds using computer vision,” in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2015, pp. 249–260.

[36] S. Ali and M. Shah, “Floor fields for tracking in high density crowd scenes,” in *European conference on computer vision*. Springer, 2008, pp. 1–14.

[37] S. D. Khan, F. Porta, G. Vizzari, and S. Bandini, “Estimating speeds of pedestrians in real-world using computer vision,” in *International Conference on Cellular Automata*. Springer, 2014, pp. 526–535.

[38] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, “You’ll never walk alone: Modeling social behavior for multi-target tracking,” in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 261–268.

[39] S. D. Khan, “Estimating speeds and directions of pedestrians in real-time videos: A solution to road-safety problem,” in *CEUR Workshop Proceedings*, 2014, p. 1122.

[40] T. Zhao and R. Nevatia, “Tracking multiple humans in crowded environment,” in *null*. IEEE, 2004, pp. 406–413.

[41] B. Solmaz, B. E. Moore, and M. Shah, “Identifying behaviors in crowd scenes using stability analysis for dynamical systems,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 10, pp. 2064–2070, 2012.

[42] M. Ullah, H. Ullah, N. Conci, and F. G. De Natale, “Crowd behavior identification,” in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1195–1199.

[43] S. D. Khan, S. Bandini, S. Basalamah, and G. Vizzari, “Analyzing crowd behavior in naturalistic conditions: Identifying sources and sinks and characterizing main flows,” *Neurocomputing*, vol. 177, pp. 543–563, 2016.

[44] C. Dupont, L. Tobías, and B. Luvison, “Crowd-11: A dataset for fine grained crowd behaviour analysis,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 9–16.

[45] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.

[46] O. Arandjelovic, “Crowd detection from still images,” in *BMVC 2008: Proceedings of the British machine vision association conference 2008*. BMVA Press, 2008, pp. 1–10.

[47] R. Ma, L. Li, W. Huang, and Q. Tian, “On pixel count based crowd density estimation for visual surveillance,” in *IEEE Conference on Cybernetics and Intelligent Systems, 2004.*, vol. 1. IEEE, 2004, pp. 170–173.

[48] Z. Zivkovic, “Improved adaptive gaussian mixture model for background subtraction,” in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 2. IEEE, 2004, pp. 28–31.

[49] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” 2005.

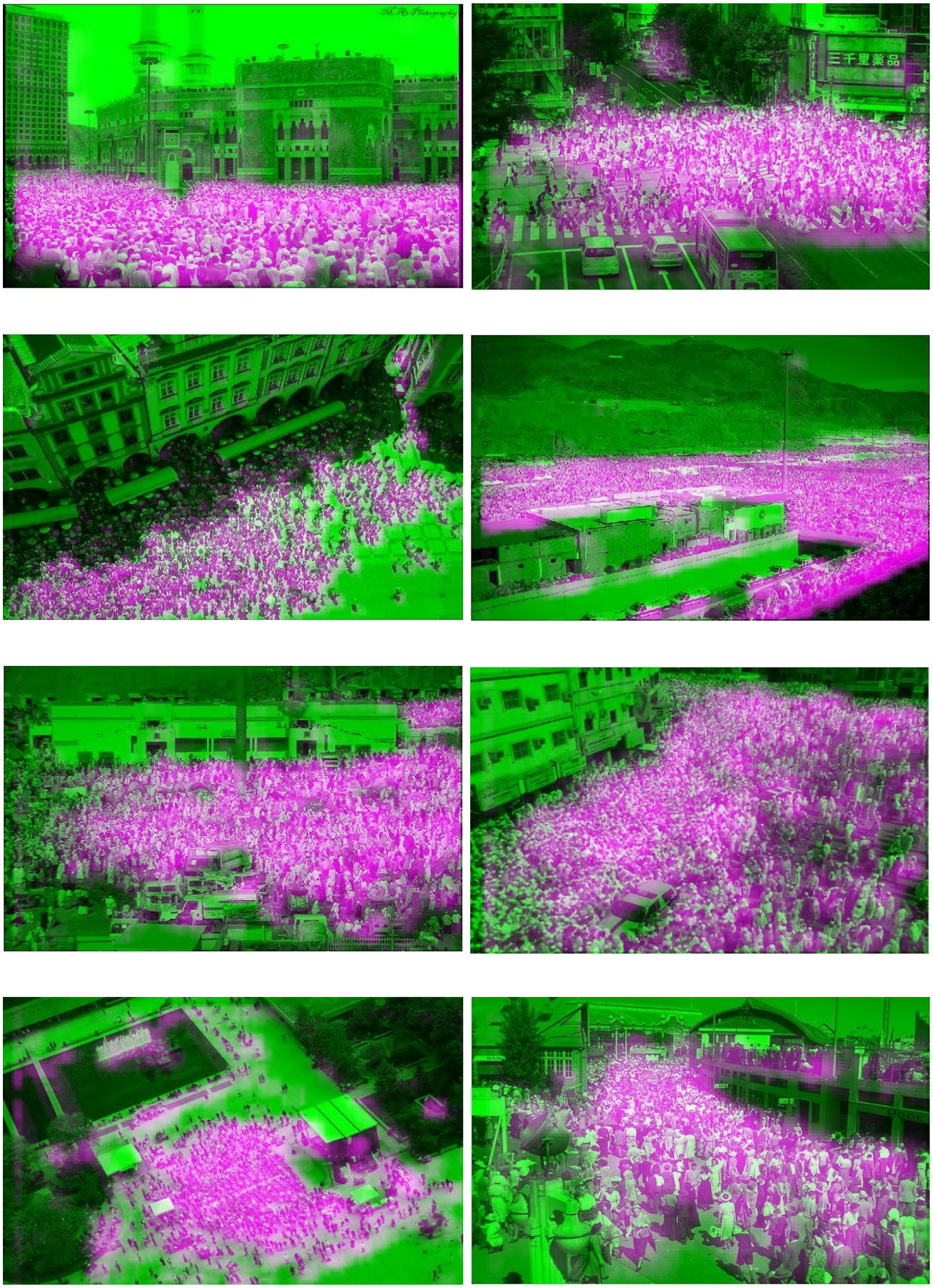

Fig. 5. Crowd segmentation results predicted by proposed approach in different crowd scenes: Segmentation mask is overlaid on image where the green color shows the background or non-crowd while pink color shows crowded areas).