

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES A HYBRID INTELLIGENT MODEL FOR CROWD DENSITY ESTIMATION

Ali Salem Ali Bin Sama^{*1} & Hussein Salem Ali Bin Sama²

^{*1}Department of Computer Science and Information, College of Sharia and Islamic Studies at Al-Ahsa, Imam Muhammad bin Saud Islamic University, Saudi Arabia

^{*1}Department of Geology Engineering, Faculty of Oil & Minerals, Aden University, Yemen

²Department of Computer, Faculty of Education - Shabowah, University of Aden, Yemen

ABSTRACT

This work is aiming at the development of a hybrid intelligent model to tackle the problem of crowd density estimation. The presented hybrid model comprises Extreme Learning Machine (ELM) for pattern recognition, Differential Evolution (DE) for model construction, as well as texture feature extraction techniques for input pattern encoding. In this work, DE is adopted to design an efficient recognition model by performing training instances selection as well as ELM topology selection. To assess the performances of the proposed model, a three popular crowd density benchmark dataset are used in this study including PETS_2009, Chunxi_Road, and Mall dataset.

Keywords: *Extreme Learning Machine, Differential Evolution, Crowd Density Estimation, Hybrid Models.*

I. INTRODUCTION

Automatic crowd density estimation models play a vital role by giving an early alarm before crowd disasters occurred. In the past, many tragedies and death cases happened due to crowd crush. As an example, Water Festival disaster (2010) in Colombia where more than 380 persons died[1]. Another example of disaster happened during pilgrimage season in Mena (2006) where more than 360 persons were died[1]. As such, the development of a crowd density estimation model could assist in making appropriate decisions for emergency and safety control.

In the literature, numerous crowd density estimation methods have been presented and they could be classified into three categories namely holistic-based, detection-based, as well as localization-based techniques. The first category utilise global image feature such as texture[2], foreground pixels[3], as well as edge features[4] for input pattern classification. In order to classify the extracted global features, different classifiers have been used such as linear regression [3], neural networks[2]and Gaussian process regression [5].A texture-based holistic approach that adopt the grey level co-occurrence matrix (GLCM) was discussed by Marana [2]. Another work that uses 2D discrete wavelet transform (DWT) was presented by Xiaohua et al.[6]. However, Rahmalanet. al.[7] proposed Translation Invariant Orthonormal Chebyshev Moments (TIOCM) texture extraction technique.

Hussain et al.[8] developed a pixel-based crowd density estimation model . The developed model was applied for images which have been taken from Al Masjid al-Haram with aim of estimating the number of peoples. The first step of the developed was segmentation of foreground objects based on a reference image. Then, ANN was applied to recognize the present of peoples from those segmented foreground objects. The systems showed high accuracy results (i.e. 100%) for low crowd density cases, but the performance was dropped for high crowded cases due to the overlapping of foreground objects. Xiaohua et al[6] proposed a texture-based model for crowd density estimation that comprise 2D discrete wavelet transform (DWT) and ELM classifier. In their work, the input feature vector was classified by ELM into four crowd density classes i.e. low, moderate-low, moderate-high, and high density. Nevertheless, this model shows low performances with non-uniform crowds. Additional work that adopt neural network regression model was proposed by Hou et al.[9].The proposed model relay on weighted foreground pixel to estimate the crowd size.

Indoor crowd estimation model was developed by Huang et al.[10]. In their model, the static background was used to segment foreground pixels. Similar work was proposed by Ma[11] where a density map is calculated using the quasi-calibration model, and then a weight is assigned to each pixel to compensate for the effects of perspective. The weighted sum of pixels in the foreground mask was used to detect excessive crowding above a threshold in [11].

The second category of crowd density estimation is detection-based methods. These method utilise head, face, or human body detectors to obtain the location of each pedestrian within the crowd scene. After that, crowd counting and estimation could be performed directly. In fact, these methods are useful in sparse environments in which the detected object is fully visible. Walking pedestrian detection model was proposed by JonesandSnow[12]. In their work, Haar-like filter and appearance features are integrated with AdaBoost classifier to classify the input pattern. The main limitation of the proposed model in [12] it can detect only moving pedestrian because it depend on their motion characteristics.

Another work proposed by Lin et al.[13]. In their, work, the synergy of Haar wavelet transform(HWT) with support vector machines(SVM). HWT was used to is applied for feature extraction of the head-like contour, then ELM was applied to classify it as the contour of a head or not. The outcomes of this method have the limitations of inability to extract head contour when it is invisible. A recent work was proposed by Gall and Lempitsky[14]where Hough forests method was employed for pedestrian detection.

The last category is localization-based methods where the input image is divided into cells, and then a classifier is applied to predict the present of people. For example Conte et al.[15] proposed approach where it split the entire scene into small horizontal and vertical cells. The reported results indicate that the proposed method were able to achieve a good crowd density estimation accuracy. Another SVM-based scheme was developed by Wu et al.[16]. The proposed scheme uses GLCM for cell texture feature extraction and SVM to predict the density distribution inside the presented cell. The model was evaluated on real crowd videos and the outcomes assert the effectiveness of the proposed system.

In summary, Table 1 presents all the studied crowd density models with their dataset used, achieved accuracy, advantages, and disadvantage

Table 1. Summary of the proposed crowd density estimation methods

Method	Ref	Dataset	The proposed Model	Accuracy	Advantages	Disadvantage
Holistic-based	[2]	299 images captured from an area of Liverpool Street Railway Station, London, UK.	GLCM + Neural Network	81.88%	Fast in estimating the crowd scene.	Sensitive to the complexity of scene background.
	[7]	225 frames from video recorded outdoor reception	TIOCM + Self-Organizing Map	85.5%		
	[8]	58 images for Phase I, and 138 image for phase II	ROI features + Neural Network	81.67%		

	[6]	300 images	Wavelet features + SVM	89.30%		
	[10]	157 images	Image features + Radial basis function	89.00%		
Detection-based	[12]	21 video sequences with 83,152 frames	Harr filters+AdaBoost	93.00%	Accurate in computing the number of pedestrian presented in the scene.	Computational expensive and it is inappropriate for occluded scene.
	[13]	-	Harr filters+ SVM	90% - 95%.		
Localization-based	[15]	PETS 2009 dataset	Motion + Linear regression	82.00%	Combines the advantages of holistic-based with the accuracy of detection-based methods in predicting the location of the crowd region.	Sensitive to the cell size and it needs training for each dataset.
	[16]	70 images from real videos taken both in Hong Kong and Beijing	GLCM Texture + SVM	88.00%		

II. METHOD & MATERIAL

The general architecture of the proposed model is shown in Figure 1. As can be seen that the proposed hybrid model consists of three stages namely input stage, texture feature extraction stage, as well as pattern recognition stage. In the first stage, the input image is fed for further feature extraction operation. The second stage is responsible about encoding the input image using a three texture descriptors including the Histogram Of Gradient (HOG)[17], Local Binary Pattern (LBP)[18], and Gabor wavelet[19]. These features are computed from the input image and combined together as a single vector that will be fed to the last stage (ELM) for recognition purposes. The last stage is concern about the classification of the input pattern. ELM is used in this study to perform the classification task. The output decision of ELM will be one of five classes including very low image, low density image, medium density image high density image, or very high density image as shown in Figure 1.

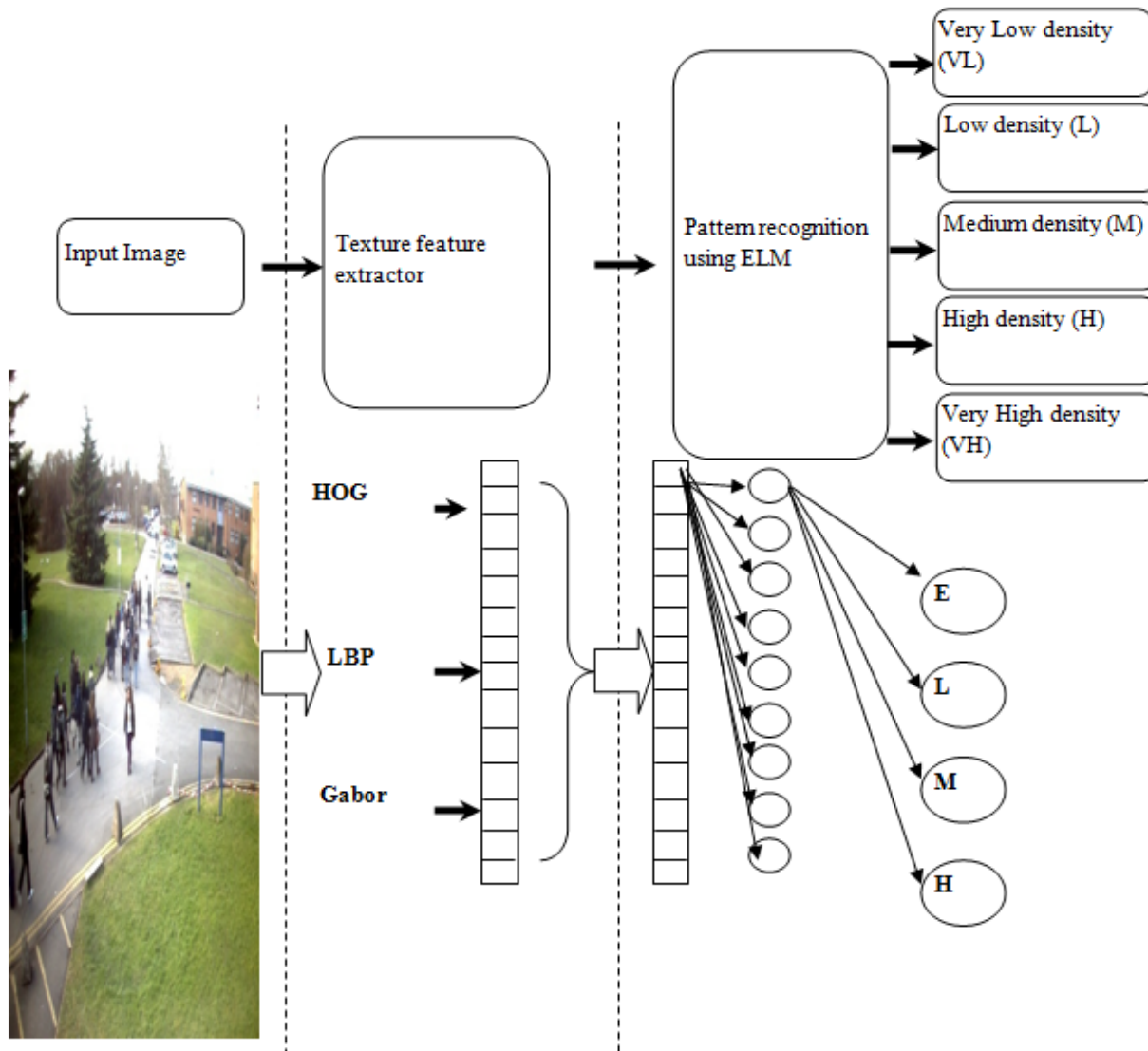


Figure 1: The structure of the proposed hybrid model

1. Texture feature extractor

In this study, a three types of texture feature extraction techniques are implemented which are HOG, Gabor filters, and LBP. Basically, HOG feature extractor identified by two parameters, i.e., block size and cell size. Each cell in HOG is mapped to 9-bin feature vector and the whole feature vector is generated by concatenating the computed vector from each block as indicated in Figure 2.



Figure 2: HOG feature vector

The main idea of Gabor filters is to generate a bank of filters at different scales and orientations using the following formula:

$$G(x, y) = \left(\frac{1}{2\pi\sigma_x\sigma_y} \right) \exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right) \exp(i2\pi w(x \cos \theta + y \sin \theta)) \quad (1)$$

where w lie in $[0, 0.5]$ and it is the radial frequency of the Gabor filter. θ lie in $[0, \pi]$ and it is the orientation variable which controls the angle of the filter, and σ lie in $[0, 2\pi]$ and it is the scale variable that controls the shape of the Gaussian function. As recommended in the previous study [19], a total of 40 Gabor filters are used for the preprocessing stage. These filters are generated at eight orientations $(0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi, \frac{5\pi}{4}, \frac{3\pi}{2}, \frac{7\pi}{4})$, and five scales $(0, \frac{\pi}{4}, \frac{\pi}{2}, \pi, \frac{3\pi}{4})$ as shown in Figure 3.

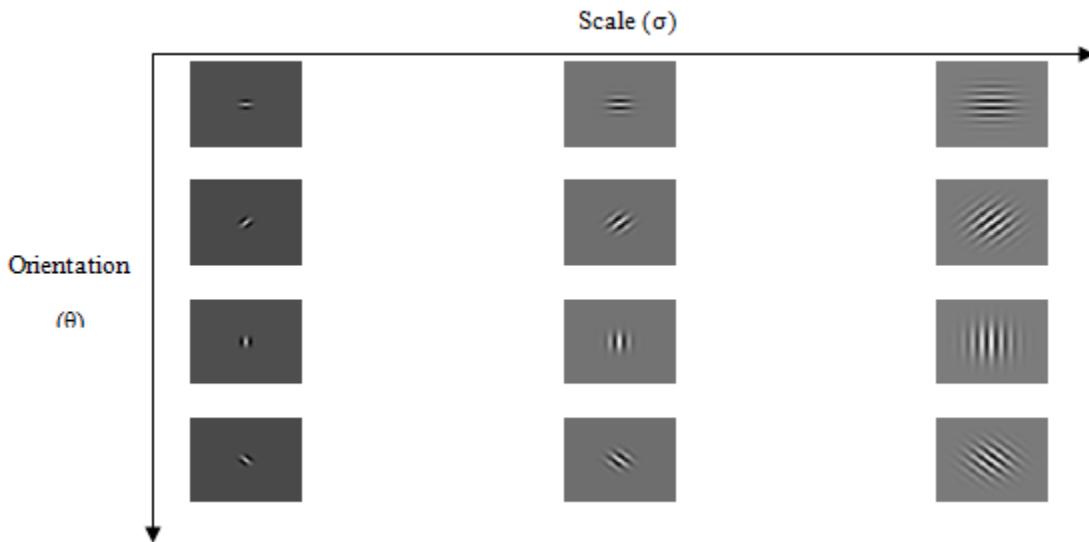


Figure 3: Gabor wavelet filters

LBP feature extraction technique has been widely used for pattern recognition [18][20]. Basically, LBP encodes input image texture features by thresholding all neighbor pixels P^{th} on a circle of radius R as indicated in Figure 4. Basically, LBP is computed as follows:

$$BP_{p,R}(X_c, Y_c) = \sum_{p=0}^{p-1} f(g_p - g_c)2^p, \text{ and } f(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where g_p and g_c are pixels intensity values of the center pixel X_c and its neighbor pixel Y_c .

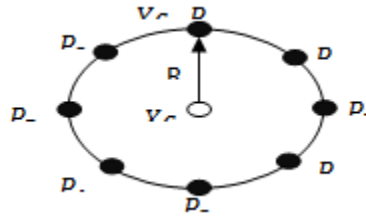


Figure 4: LBP feature extraction

2. Extreme learning machine

ELM was introduced by Huang et.al [21].The main advantage of ELM is that it provides good generalization performance at extremely fast learning speed. The general structure of ELM is shown in Figure 5 where it comprises three layers namely the input layer, hidden layer, and output layer.

The mathematical formula of ELM is defined as follows

$$\min_{\mathbf{B}} \|\mathbf{HB} - \mathbf{T}\| \tag{3}$$

$$\mathbf{B} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{T} \tag{4}$$

$$\mathbf{H} = \begin{bmatrix} g(\mathbf{w}_1 \mathbf{x}_1 + b_1) & \dots & g(\mathbf{w}_j \mathbf{x}_1 + b_j) \\ \vdots & \ddots & \vdots \\ g(\mathbf{w}_1 \mathbf{x}_N + b_1) & \dots & g(\mathbf{w}_j \mathbf{x}_N + b_j) \end{bmatrix} \tag{5}$$

Where H is the output of the hidden layer, B is the weight matrix, and T is the target class.

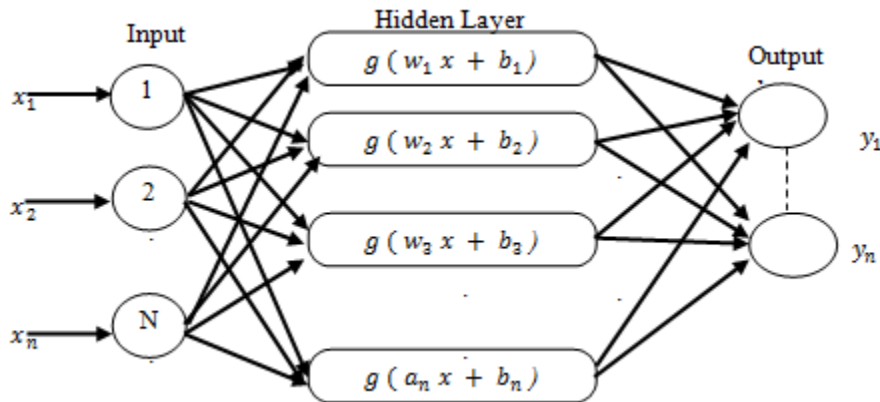


Figure 5: Extreme learning machine

3. Differential Evolution Algorithm

DE has been introduced in [22]. Generally, DE consists of five main stages which are initialization, mutation, crossover, and selection, as shown in Figure 6. In the initialization stage the upper and lower bounds of the optimization problem is defined and a random X vectors of size NP are generated. The maximum number of iteration is set to G_{max}. In the mutation stage the aim is to add a random value to the best vector X_{best}, however the crossover and selection stage is concern about the hybridization between current vector X values with other vectors according to probability of crossover CR

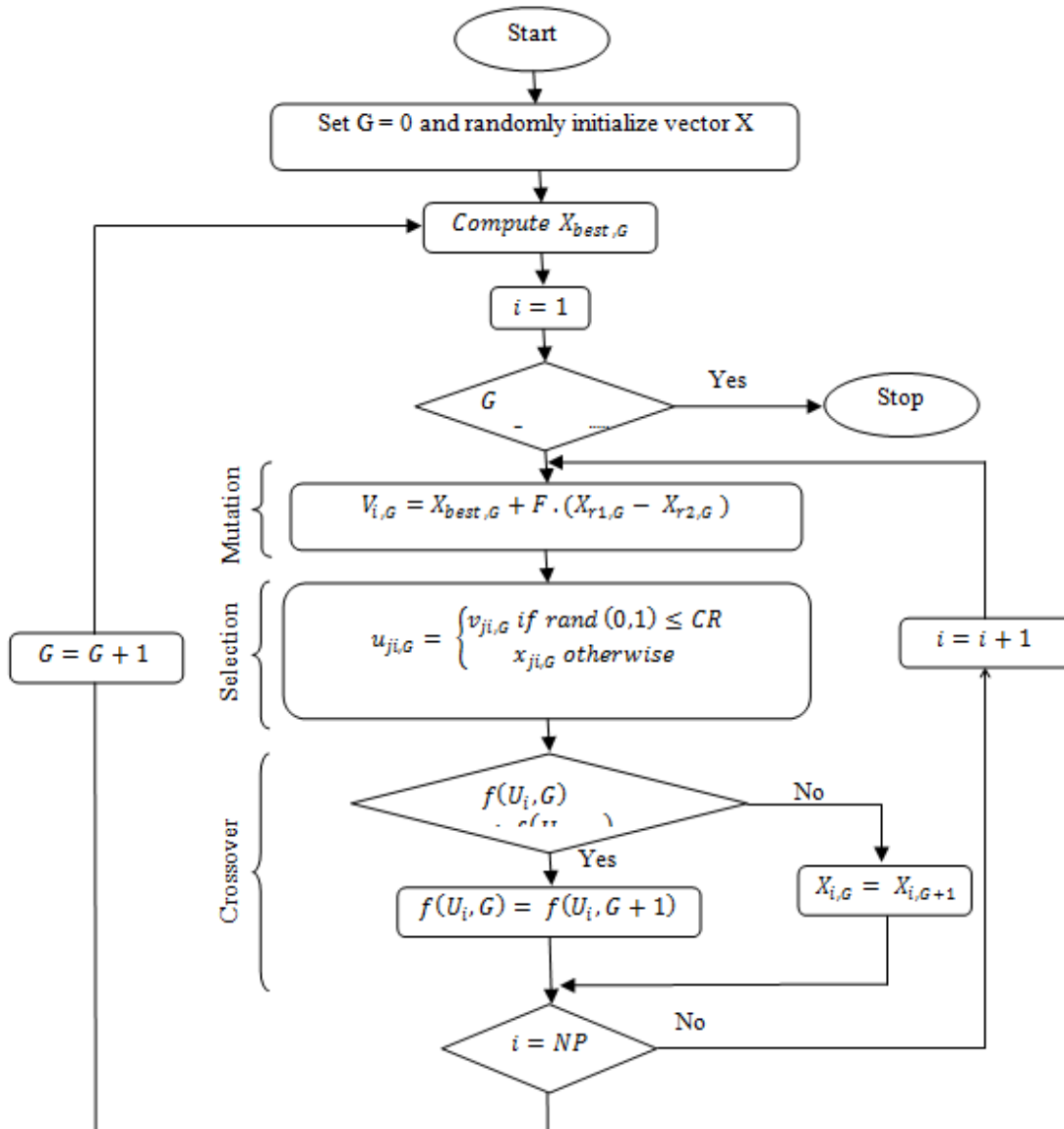


Figure 6: Differential Evolution Algorithm (DE)

4. Model construction steps

As mentioned earlier that ED is employed in this study to construct an efficient crowd intensity estimation system. The detailed steps of construction procedure are shown in Figure 8. Each step is explained as follows:

- 1- Initialize the DE vectors: In this step, each DE vector is initialized with a random position, X_1 .
- 2- Extract DE vector parameters: The current values of the executed DE vector are extracted to train ELM model. The encoding scheme of DE vector is shown in Figure 7. Mainly, the vector containing two components i.e. ELM parameters, and the selected training negative instance.
- 3- Train ELM: In this step, ELM classifier is trained according to the DE vector values.
- 4- Check the stopping criteria: Two stopping conditions for the DE algorithm are considered, i.e., either the maximum number of iterations is met, or a perfect (100%) fitness measure is achieved during the execution process.
- 5- Evaluate the fitness function: The trained ELM classifier is evaluated with the validation dataset and the

fitness function is computed as follows.

$$\text{Fitness} = G_mean = \sqrt{(\text{Sensitivity}_1 * \dots * \text{Sensitivity}_n) * (\text{Specificity}_1 * \dots * \text{Specificity}_n)} \quad (6)$$

where, TP is the true positive rate, and TN is the true negative rate. The penalty term here is introduced to give higher property of positive class over the negative class.

Execute DE operations: The mutation, crossover, and selection operations of DE are executed based on Figure 6.

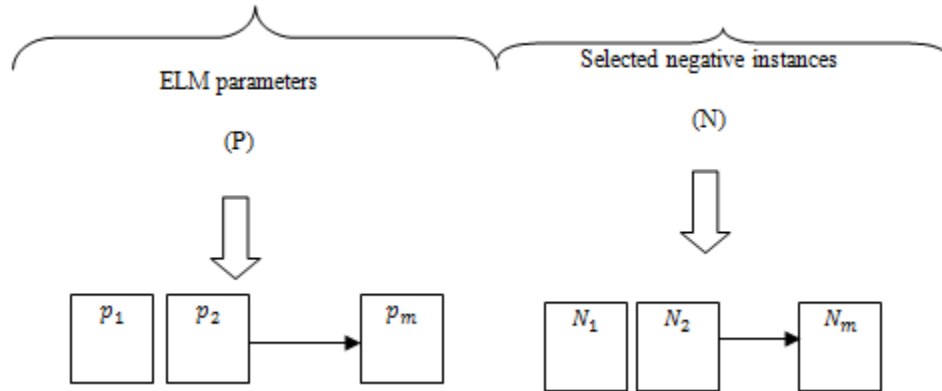


Figure 7: DE vector components

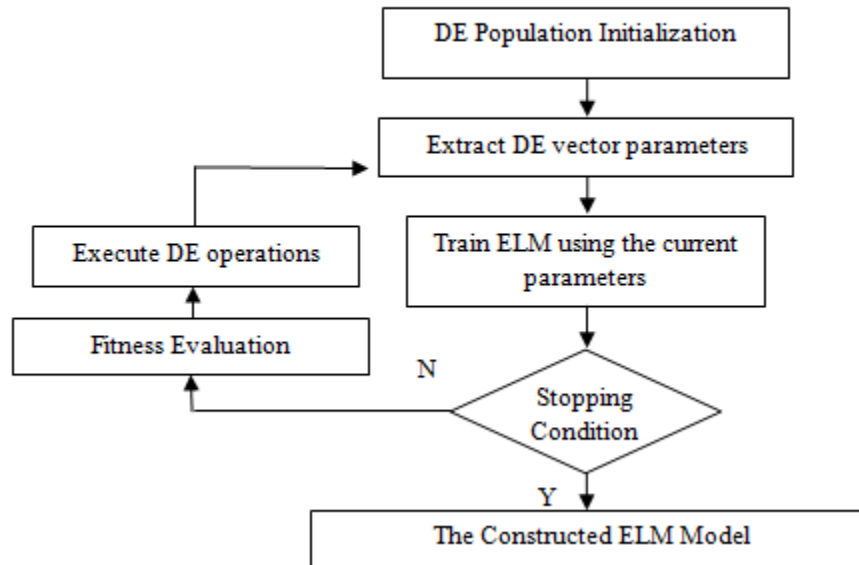


Figure 8: The flowchart of the model construction steps

III. RESULT & DISCUSSION

To evaluate the effectiveness of the proposed crowd estimation model, a total of three public databases are employed in this study including the shopping mall dataset (Mall) [23], Chunxi_Road[5], and PETS 2009[24]. Table 2 illustrates the details of each dataset in terms of the total numbers of frames, number of frames per second, as

well as the range of crowd number in a frame, as well as shown in Table 2. As in previous work[25], the frames has been classified into five class which are very low density class, low density class, medium density class, high density class, and very high density class. A number of sample images for each class are shown in Figure 9.

Table 2: The studied database sketches

Database	The total number of frames	Number of frames per second	The range of crowd number in a frame
Chunxi_Road[25]	1241	-	3-14
Mall[23]	2000	<2	13-53
PETS 2009[24]	663	7	8-33

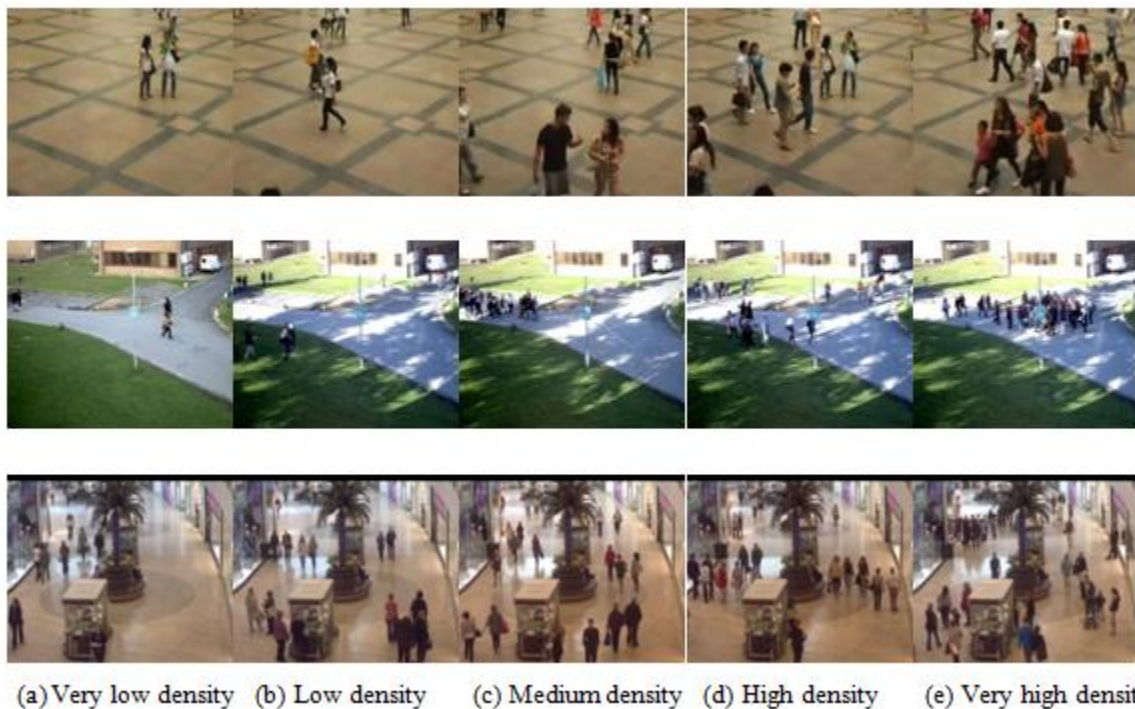


Figure 9: Examples of crowd PETS_2009(top line), Chunxi_Road(bottom line), andMall dataset (middle line)

Texture features analysis

This section assesses the contribution of each feature extraction technique (i.e. HOG, LBP, and Gabor) on model recognition performance. Each experiment was repeated 10 times, and then the mean recognition rate were computed and reported in Table 3 and Figure 10. As indicated from the result that the combination of texture feature extractors was able to achieve better performances as compared with individual extractor (i.e. HOG, LBP, and Gabor). It should be noted that HOG feature extractor outperform other feature extractors in terms of recognition performances. Particularly, the recognition rate on PETS_2009 images was 96% for HOG, 92 % for LBP, and 93% for Gabor filters respectively. The superior performances for HOG was due to the advantage of HOG in encoding and representing texture features that make it less sensitive to illumination pixels changes [17]. The lowest recognition rate was reported from LBP because its sensitivity to the variations in Gray-Scale values.

In order to compare the reported accuracy rate from the hybrid texture technique against individual extractor (i.e. HOG, LBP, and Gabor) statistically, the t-test statistical method [26, 27] was employed. The significance level (α) was set to 0.05 (i.e. 95% confidence level) and the p-values of the performance indicator as shown in Table 4. The null hypothesis imply that the performance of models X and Y was equivalent, whereas the alternative hypothesis

claimed that model X outperformed model Y. As can be seen in Table 4, all the reported p-values were lower than 0.05 which implies that the hybrid texture performed significantly better (at the 95% confidence level) than individual techniques. It is worth noting that integrating multiple texture features usually increases model complexity as well as the required cost of computational time. Nevertheless, when accuracy is the main issue, computational time could be ignored.

Table3: Performance comparison of different texture feature extractor on LFW

Model	Recognition rate		
	PETS_2009	Mall	Chunxi_Road
HOG	93.01%	73.40%	83.80%
LBP	90.24%	70.20%	81.51%
GABOR	91.64%	72.30%	82.73%
Hybrid	97.92%	75.70%	88.00%

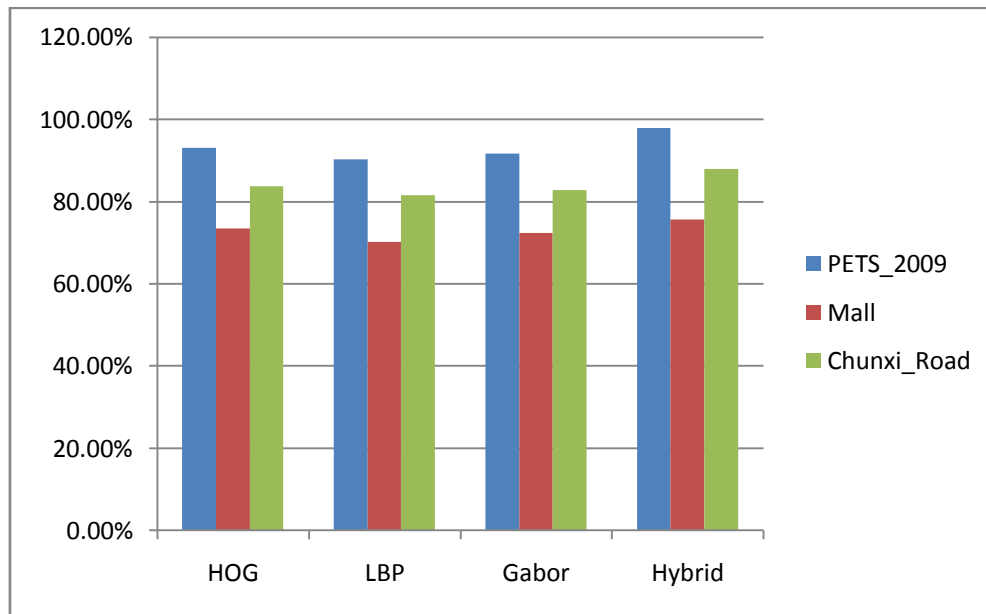





Figure 10: The recognition rate of HOG, LBP, Gabor, and Hybrid textures

Table 4: The p-values of the performance indicators

Model	Y	HOG	Gabor	LBP
	X			Hybrid (HOG + Gabor + LBP)
p-value		0	0	0

For illustration purposes, a number of misclassified cases from Mall dataset are shown in Table 5.

Table 5: Example of misclassified cases for Mall dataset (The column is true class but the row represents the misclassified to class).

Crowd density	Very Low	Low	Medium	High	Very High
Very Low					
Low					
Medium					
High					
Very High					

Further analysis was conducted by measuring the consumed time by each feature extraction technique. Specifically, the average time in seconds required to compute the texture features for single image from PETS_2009 was computed and reported in Table 6. It can be seen that LBP is the most expensive operation in terms of computational time cost because of the need to compute LBP for each pixel independently. On the other, Gabor wavelet feature extractor required the least time because that Gabor texture features is computed from a convolution operation between Gabor filters with the whole image where it saves time.

Table 6: Computational time analysis on PETS_2009

Texture feature	Time(sec)
HOG	0.64
LBP	0.93
Gabor	0.52

Comparison with related work

This section is aiming to compare the achieved outcomes from the proposed hybrid model against others reported results in the literature. Particularly, the reported from BP network[28], Cascade Optimized ConvNets[25], and SVM [29]are shown in Table 7 form comparison purposes. As indicated in Table 7 that the proposed model reports more than 15% accuracy as compared withBP network [28] and more than 9% accuracy against SVM [29] respectively. This is due to the benefits of integrating multiple features extraction techniques and employing DE to construct an efficient model.

Table 7: Comparison on PETS_2009

Approach	Accuracy%
BP network[28]	82.84
Cascade Optimized ConvNets[25]	96.80
SVM[29]	87.87
The proposed model	97.00

IV. CONCLUSION

In this work a hybrid intelligent model has been introduced for the problem of crowd density estimation. The hybrid model integrates DE algorithm with ELM to perform the recognition task. To assess the proposed model, a number of benchmark problems have been used including PETS_2009, Mall, and Chunxi_Road dataset images. The outcome of the hybrid model indicates its ability to achieve superior performances on PETS_2009 with 97.00% recognition rate. As compared with the reported results in the literature, the reported results show that it yields competitive performances. Further work could be conducted by applying deep learning models with optimization algorithms to handle the problem of density estimation.

V. ACKNOWLEDGEMENTS

This work is fully supported by Al-Imam Muhammad Ibn Saud Islamic University, Grant Scheme entitled ‘A Hybrid Intelligent Model for Crowd Density Estimation’, under Grant Nos. 370806.

REFERENCES

1. B. Krausz, C. Bauckhage, *Loveparade 2010: Automatic video analysis of a crowd disaster*, *Computer Vision and Image Understanding*, 116 (2012) 307-319.
2. A. Marana, S. Velastin, L. Costa, R. Lotufo, *Estimation of crowd density using image processing*, *Image Processing for Security Applications (Digest No.: 1997/074)*, *IEE Colloquium on*, (IET1997), pp. 11/11-11/18.
3. A.C. Davies, J.H. Yin, S.A. Velastin, *Crowd monitoring using image processing*, *Electronics & Communication Engineering Journal*, 7 (1995) 37-47.
4. D. Kong, D. Gray, H. Tao, *A viewpoint invariant approach for crowd counting*, *Pattern Recognition*, 2006. *ICPR 2006. 18th International Conference on*, (IEEE2006), pp. 1187-1190.
5. A.B. Chan, Z.-S.J. Liang, N. Vasconcelos, *Privacy preserving crowd monitoring: Counting people without people models or tracking*, *Computer Vision and Pattern Recognition*, 2008. *CVPR 2008. IEEE Conference on*, (IEEE2008), pp. 1-7.
6. L. Xiaohua, S. Lansun, L. Huanqin, *Estimation of crowd density based on wavelet and support vector machine*, *Transactions of the Institute of Measurement and Control*, 28 (2006) 299-308.
7. H. Rahmalan, M.S. Nixon, J.N. Carter, *On crowd density estimation for surveillance*, (2006).
8. N. Hussain, H.S.M. Yatim, N.L. Hussain, J.L.S. Yan, F. Haron, *CDES: A pixel-based crowd density estimation system for Masjid al-Haram*, *Safety Science*, 49 (2011) 824-833.
9. Y.-L. Hou, G.K. Pang, *People counting and human detection in a challenging situation*, *IEEE transactions on systems, man, and cybernetics-part a: systems and humans*, 41 (2011) 24-33.
10. D. Huang, T.W. Chow, W. Chau, *Neural network based system for counting people*, *IECON 02 [Industrial Electronics Society, IEEE 2002 28th Annual Conference of the]*, (IEEE2002), pp. 2197-2201.
11. R. Ma, L. Li, W. Huang, Q. Tian, *On pixel count based crowd density estimation for visual surveillance*, *Cybernetics and Intelligent Systems*, 2004 *IEEE Conference on*, (IEEE2004), pp. 170-173.
12. M.J. Jones, D. Snow, *Pedestrian detection using boosted features over many frames*, *Pattern Recognition*, 2008. *ICPR 2008. 19th International Conference on*, (IEEE2008), pp. 1-4.
13. S.-F. Lin, J.-Y. Chen, H.-X. Chao, *Estimation of number of people in crowded scenes using perspective transformation*, *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 31 (2001) 645-654.

14. J. Gall, V. Lempitsky, *Class-specific hough forests for object detection, Decision forests for computer vision and medical image analysis*, (Springer, 2013), pp. 143-157.
15. D. Conte, P. Foggia, G. Percannella, M. Vento, *Counting moving persons in crowded scenes, Machine vision and applications*, 24 (2013) 1029-1042.
16. X. Wu, G. Liang, K.K. Lee, Y. Xu, *Crowd density estimation using texture analysis and learning, Robotics and Biomimetics, 2006. ROBIO'06. IEEE International Conference on, (IEEE2006)*, pp. 214-219.
17. O. Déniz, G. Bueno, J. Salido, F. De la Torre, *Face recognition using histograms of oriented gradients, Pattern Recognition Letters*, 32 (2011) 1598-1603.
18. T. Ahonen, A. Hadid, M. Pietikainen, *Face description with local binary patterns: Application to face recognition, IEEE transactions on pattern analysis and machine intelligence*, 28 (2006) 2037-2041.
19. C. Liu, H. Wechsler, *Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, IEEE Transactions on Image processing*, 11 (2002) 467-476.
20. H.S. Bhatt, S. Bharadwaj, R. Singh, M. Vatsa, *Memetically Optimized MCWLD for Matching Sketches With Digital Face Images, Information Forensics and Security, IEEE Transactions on*, 7 (2012) 1522-1535.
21. G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, *Extreme learning machine: theory and applications, Neurocomputing*, 70 (2006) 489-501.
22. J. Kennedy, R. Eberhart, *Particle swarm optimization, Neural Networks, 1995. Proceedings., IEEE International Conference on 1995*, pp. 1942-1948 vol.1944.
23. K. Chen, C.C. Loy, S. Gong, T. Xiang, *Feature Mining for Localised Crowd Counting, BMVC2012*, pp. 3.
24. A. Ellis, J. Ferryman, *PETS2010 and PETS2009 evaluation of results using individual ground truthed single views, Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on, (IEEE2010)*, pp. 135-142.
25. M. Fu, P. Xu, X. Li, Q. Liu, M. Ye, C. Zhu, *Fast crowd density estimation with convolutional neural networks, Engineering Applications of Artificial Intelligence*, 43 (2015) 81-88.
26. B. Efron, R.J. Tibshirani, *An Introduction to the Bootstrap (Chapman and Hall, 1993)*.
27. B. Efron, *Bootstrap Methods: Another Look at the Jackknife, The Annals of Statistics*, 7 (1979) 1-26.
28. G. Kim, T. An, M. Kim, *Estimation of Crowd Density in Public Areas Based on Neural Network, KSII Transactions on Internet & Information Systems*, 6 (2012).
29. H. Su, H. Yang, S. Zheng, *The large-scale crowd density estimation based on effective region feature extraction method, Asian Conference on Computer Vision, (Springer2010)*, pp. 302-313.