2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)

Human Count Estimation in High Density Crowd Images and Videos

Rohit, Vandit Chauhan, Santosh Kumar, Sanjay Kumar Singh
Department of Computer Science & Engineering,
Department of Electronics Engineering,
Indian Institute of Technology (B.H.U.), Varanasi-221005, India
rohit.rohit.ece14@iitbhu.ac.in
vandit.chauhan.ece14@iitbhu.ac.in
santosh.rs.cse12@iitbhu.ac.in
sks.cse@iitbhu.ac.in

***Abstract*— This paper addresses the issue of detection and counting of people in exceedingly swarmed crowd images and video scenes. The identification of individual objects has observed enormous advancements over the late years, but crowd scenes still remain predominantly challenging for detection and counting purposes because of substantial impediments and occlusions, high group densities and large disparities in individuals' sizes and appearances. To address these difficulties, we propose epsilon-Support Value Regression (SVR) fusion-based approach to clout information on the global construction of the crowd scenes and to identify the people in these scenes. We prepared and tested our approach on a new dataset of head pose images and fifty densely crowd images containing over 64000 human head annotations, with counts varying in between 94 and 4543. Our dataset contains jammed crowd scenes with individuals in large numbers, as opposed to other datasets, which encompass very few individuals. The experimental results establish the robustness and viability of the proposed approach by careful assessment of the counting process in terms of Absolute and Normalized Absolute Errors. We also developed a prototype for evaluating the accuracy of proposed system.**

Keywords- Human Detection; Crowd; Recognition; Computer vision; High Density; SVR; HOG; Fourier Analysis; SIFT

I. INTRODUCTION

Human detection in crowd video scenes is getting more proliferation due to the variety of applications in crowd monitoring and tracking [2, 11, 23, 24, 26]. For crowd analysis, automated and semi-automated solutions for density estimation and counting exist in field of computer vision. There is currently a great interest in vision technology for monitoring all types of environments [1, 3]. This could have many aims and goals, e.g. security, resource management, or advertising EASE OF USE [4, 5]. However, highly dense crowd scenes (Fig. 1) pose more complications. Recognizing and monitoring in areas where there are chances of heavy occlusions, is still a big problem [2, 24, 26].

A huge loss of life and property has been recorded due to stampedes and crowd disasters in the recent years, as per Table 1. With the ever-increasing population, more such incidents are bound to happen without proper crowd management systems. These events involving large gatherings of people, need to be controlled and monitored, which can be done by combining image processing techniques and the field of computer vision. In this paper, we address the problem: How to recognize humans in high density crowded scenes?

Table 1. illustrates the loss of life in stampedes in recent years
(Source: https://en.wikipedia.org/wiki/List_of_human_crushes)

Place	Year	Casualties
Hindu festival, Datia District, Madhya Pradesh, India	2013	**>200**
Water Festival, Phnom Phen, Cambodia	2010	**592**
Pilgrimage, Mena, Saudi Arabia	2006	**554**
Religious Procession, Baghda, Iraq	2005	**1000**

To solve this problem, we propose a solution to count the number of human occurrences in these crowded video scenes. In this paper, we propose epsilon-Support Value Regression (SVR) fusion-based approach to leverage information on the global structure of the crowded scene and to identify people in the crowded scenes [15, 20, 23, 24, 25, 26].
The proposed method counts the approximate number of people in images, which would act as an indicator of the chances of any potential crowd disaster [1, 2, 11, 16, 23, 25, 26].

Fig. 1: *High-density crowd scenes pose major problems in detection and monitoring, due to heavy occlusions and significant variations in people's sizes and appearances.*

- Research Contribution

Following are the major contributions in this research:

1. In this paper, we propose a system to estimate the crowd density and counts of people in crowded scenes. SIFT features and descriptors are extracted, along with calculated Fourier analysis of the image.

2. Due to geometric alterations brought on by perspective issues and the foreshortening in these scenes, density may greatly vary across the field of vision. To counter this issue, images are divided in small patches of equal dimensions, and counts for these patches are evaluated. As long as the size of the patch is small, it is safe to assume constant density across the patch.
3. Counts based on head detection are not very accurate in high-density crowds, with heads being as small as tens of pixels, but to improve on this approach, we employed cascade training of head images, with manual selection of bounding boxes covering all possible positions and orientations of the human head.
4. The approach of Fourier analysis proved quite accurate for crowded patches, but since this approach is indifferent to the presence of people, it provided inaccurate results in the case of the non-crowd patches. To solve this issue, we calculated confidences for these patches, and combined them with the results obtained from the Fourier model.
5. Finally, we fused results from all the three approaches (i.e. head detection, Fourier analysis and interest points based method) to train our ep-SVR model, to provide better accuracy and consistency.

The rest of the paper is organized as follows: Section II presents related work in the direction of human detection and estimation of crowd density. Section III describes the various databases we used for our training and testing. Section IV demonstrates the proposed approach in detail. Section V shows the experimental results and analysis. Finally, we conclude the paper in section VI.

II. Related Work

Proposed methodologies and applications related to crowd estimation are briefly reviewed in this section. Singular identification for counting people in images/videos is implemented in [1, 2]. These sorts of strategies however are not viable for the sort of crowds we deal with, because severe occlusion and cluttering, low resolution, and few pixels per individuals due to foreshortening, result in serious difficulties in human, or even face detection.

Different operator-based models of crowds have been established since the social force model [9], that employ and utilize mathematical modeling of the forces acting on an individual, and simulating their interactions, and stemming macroscopic laws on group properties like the stream flow, see for example [4] and [5].

It also cultivated the multi-scale approach concept [6]. The interest value changes in accordance with an imperfect dynamical model and is revised at every progression with the observed instances. Multiple efforts have been made to apply this approach in non-linear models [3], [7] and [8].

Brostow et al. [10] and Rabaud et al. [11] proposed systems to count the moving objects in video frames by estimating immediate regions of coherent motion of objects. Computation of such patterns of motion were also proposed in [12-14], but not with a causal solution to the question of crowd estimation.

These algorithms necessitate video frames as the raw input data, with reasonably high frame rates for consistent motion estimation, but are not practical for still images of crowds, or even videos if the individuals in the crowd show insignificant or no motion, e.g., conferences, concerts or meetings.

Approaches such as [16] recommend dividing images into smaller blocks and regress on each of them individually. These methods [15, 16] try to counter the problems generally associated with foreshortening, and local geometric alteration due to the perspective. One major problem with this approach however is that it disregards the constraints on localized dependence and spatial consistency as information leveraged across the blocks is not pooled.

Another group of techniques proposed for counting humans in crowds rely upon the construct of direct relationships between localized features and counts, through regression. Global functions, where the parameters are estimated for the entire image or video, are envisioned in [17, 18, 19, 20]. These methods implicitly presume that the feature density is roughly uniform throughout the image, regardless of the location where the feature is computed. This assumption is largely unsound in most real-world situations due to perspective, changes in outlook, and variations in crowd density.

III. Description of Database

We collected our databases from the online available free web images for crowd ([21], [22]). The crowd image dataset comprises of 50 images including people in concerts, stadiums, pilgrimage etc. It includes an average of 1190 people per image, ranging from 94 to 4543 people. We obtained 63705 annotations, which the (x, y) co-ordinates of each individual present there. The database is shown in Figs. 2 and 3, respectively.

This dataset is used for training the SIFT interest point-based model. For training of the Cascade Head Detector, we created our database:

- By extracting frames from various videos of people on railway stations or pedestrians walking on busy roads.
- By using images of all possible head positions varying from -90 degree to +90 degree [22].
- By using 6684 negative images, including those of empty roads, malls, buildings, gardens and stadiums, etc.

Fig 2: (*p-r) Images from the Crowd Images Dataset.*

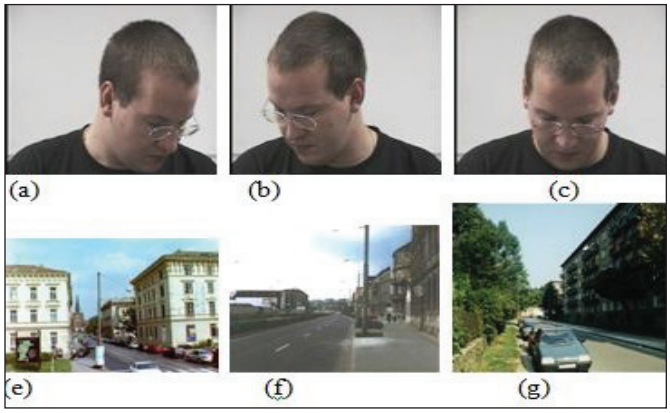

Fig 3: *Image from the Head Pose Dataset and ((e)-(g)) images from the Negative Images Datasets [21, 22]*

IV. PROPOSED SYSTEM

This section explains the proposed system in detail. The several components of the system are- (1) head detection, (2) Fourier analysis, (3) feature extraction, and (4) fusion of all three methods to form the final model. The block diagram of our proposed system is shown in Fig. 4.

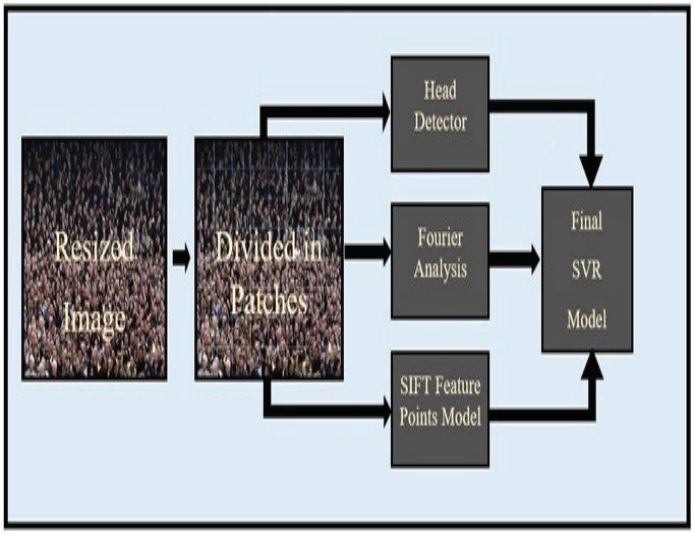

Fig.4: *Block diagram of proposed crowd detection and monitoring system*

The brief description of each component of proposed system is described as follows:

A. Head detection

Counts based on human heads appearing in densely crowded images are calculated by HOG-based feature descriptors [24, 26]. The central idea behind this approach is that the local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. The testing of head detection of humans in the crowded scene is shown in Fig. 5.

In practice, this is implemented by dividing the image window into small spatial regions "patch", for each patch accumulating a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. The combined histogram entries calculated from the representation [26].

We first collected images of all possible positions and orientations of the human head for our head pose dataset. Negative images are sampled in the ratio of 1:2. Negative images should not include portions of head, otherwise, the number of false negatives will increase. It is best to include images of empty roads, buildings, malls, stadiums, parks, etc. in order to achieve acceptable results. The Cascade Detector was trained for 35 stages with a false positive rate of 0.4 using these datasets. The target size was chosen as a trade-off between better results and increased time of detections [24, 26].

Fig.5: *Testing Head Detector*

B. Fourier Analysis

When density of crowd becomes too high, the HOG approach doesn't hold up well [2, 24], because in crowds of such large proportions, heads appear like small dots to the camera undetectable through the head detector. Fourier analysis provides a more elegant approach to counting humans in this case.

A crowd is inherently repetitive in nature, since all humans appear the same from a distance. Fourier Transform of an image can provide all its information i.e. positions where there are large changes in intensity values, endpoints of local extremes. The repetitions, as long as they occur consistently in space, can be resolved by Fourier analysis.

We take the Fourier Transform of all the patches of an image [15], selected as 64 by default. To eliminate high frequency content, we allow the gradient patches to pass through an appropriate Butterworth low pass filter. Noise which normally presents as low amplitude frequencies, is eliminated by careful inspection of the amplitude spectrum. The image is reconstructed using the Inverse Fourier Transform. To calculate peaks, we apply non-maximal suppression, employing the Canny Edge detector algorithm. The calculated peaks are an apt estimation of human count in the region, especially in patches representing crowds. But Fourier analysis can't inherently differentiate between crowd and non-crowd patches, and hence there is a need to discard counts from these non-crowd patches.

For this purpose, confidence scores of the patches are calculated as discussed in the next sub-section.

C. Feature extraction

Interest points in crowded images are detected using SIFT features. The descriptors of these points frame the base strategy that classifies between crowd and non-crowd objects [20, 23] and independently gives both the crowd counts and confidence scores to correct the Fourier analysis approach. Sky, building structures and trees are naturally recurring in outdoor images, and given the facts that head detection gives false positives in such regions, and Fourier analysis is practically crowd-blind, it is important to discard counts from such patches [23].

For obtaining counts, we obtain SIFT features and descriptors of the patches. K-means clustering technique is then applied to these SIFT descriptors. Descriptors are 128-dimension vectors, which give information about a point in image, which includes co-ordinates, width and height. Clustering gives us cluster centers for all points. The descriptors are then categorized in these formed clusters to construct a histogram of dimension [1×k]. Employing Support Vector Regression (SVR) on the ground truths of the patches, a model is constructed which estimates the crowd count in a patch using its clustered histogram bin.

$$N(I) = N(P_1 \cup P_2 \cup P_3 \cup +.....P_n) =$$
$$N(P_1) + N(P_2) + N(P_3) + ... + N(Pn)$$

Since we assumed independence among patches, the number of people in an image can be found simply by summation of people in its individual patches. SIFT features are intrinsically scant in nature, hence the frequency γ of a particular feature in a patch can also be modeled as a Poisson Random Variable [23],

$$p(\gamma_i / crowd) = \frac{e^{(-\lambda_i^+)}.(\lambda_i^+)^{\gamma_i}}{\gamma_i!}$$

with expected value, λ_i^+ . Given a set of positive (+) and negative (-) examples, the densities of the feature vary in positive and negative images, and can be used to identify crowd patches from non-crowd ones.

We calculate estimated histogram of positive and negative samples and store it in vectors, λ_i^+ and λ_i^- . Assuming independence among features, the log-likelihood of the ratio of patch containing crowd to non-crowd is:

$$\log(\gamma_1, \gamma_2, \gamma_3 ..... \gamma_k \mid crowd) - \log(\gamma_1, \gamma_2, \gamma_3 ....$$

$$..\gamma_k \mid -crowd) = \sum_i \lambda_i^- - \lambda_i^+ + \gamma_i (\log \lambda_i^+ - \log \lambda_i^-)$$

Above equation gives the confidence score for a patch [23].

D. Fusion

Once we have the counts for all the patches from our dataset, we fuse them to get our final model. We prepare an epsilon-SVR model by taking all these counts in a matrix and training them on their Ground Truths. The experimental results are summarized in Table 2.

Table 2: shows average patch and image errors for proposed approach

Methods	Per Patch		Per Image	
	Absolute Error	***Normalized Absolute Error***	***Absolute Error***	***Normalized Absolute Error***
Fourier Confidence	18.5±23.5	1.8±4.7	605.2±634.1	0.52±0.48
Head detection Method	19.1±18.1	3.1±4.7	695.4±558.8	1.22±1.93
SIFT feature	10.8±18.1	1.06±2.04	422.6±574.6	0.40±0.80
Rodriguez et al. [24]	-------	-------	655.7±697.8	0.71±1.02
Lempistsky et al. [25]	-------	-------	493.4±487.1	0.61±0.91
Proposed approach	**10.4±18.1**	**.99±2.35**	**415.2±519.5**	**0.36±0.42**

V. Experimental Results and Analysis

The final system was tested on our own dataset on a commercial DELL laptop (Inspiron 15, 3000 Series, 4th Gen Intel Core I5 Processor, 1.7 GHz, 4 GB DDR3 RAM) to check the accuracy and efficacy of the proposed approach.

Careful fine-tuning of the model parameters such as the size of clusters and the thresholds for noise suppression produced convincing results. Absolute Error of around 10 people per patch and around 415 per image is an improvement compared to the previous proposed models. In Table 2 details of errors of all three methods and the proposed system is depicted along with the details of the previous models.

For the evaluation of the experimental results, we have compared various methods with the proposed approach. For the comparison, we applied the methods of Rodriguez et al. [24], and Lempistsky and Zisserman [25]. Since these models work on databases of images of crowd scenes, rather than videos, these are the most suitable methods for comparison with the proposed approach.

Figs. 6(a) and 6(b) illustrate the Absolute Errors and Normalized Absolute Errors for the estimation of patches in the individual crowd scenes, respectively. The mean value per patch is shown with the red circle, standard deviations with orange bars, and green triangles shows patch ground truths. For better analysis of the experimental results, the x-axis shows image number which is sorted in ascending order with respect to actual ground truth counts in both plots. It can be demonstrated that Absolute Errors per patch increases as the actual Patch Ground Truth count increases, except for the images in the range 22 – 44 with corresponding actual counts in the range of 900 – 2400 per image. This particular range shows better results on both the Absolute Error and Normalized Absolute Error fronts. Since the actual Ground Truths vary in a diverse range of 94 – 4543, the proposed system tends to underfit the data on the higher end of actual Ground Truths, while it overfits for the lower end of the spectra.

We also checked our system on videos of crowd (available on the free web). We extracted frames of these videos and used our system to estimate counts of people in the crowd scenes. On an average, our system requires 31 seconds for processing each frame. Since the chances of immediate and abrupt changes in the crowds of such enormous amounts are small, this time should be sufficient for crowd monitoring purposes. We also checked our model on Purirathyatra images, a Hindu festival associated with Lord Jagannath held at Puri in the state of Odisha, India, with high chances of a possible stampede-like situation every year, we calculated an average of 1190 people, which was reasonably accurate.

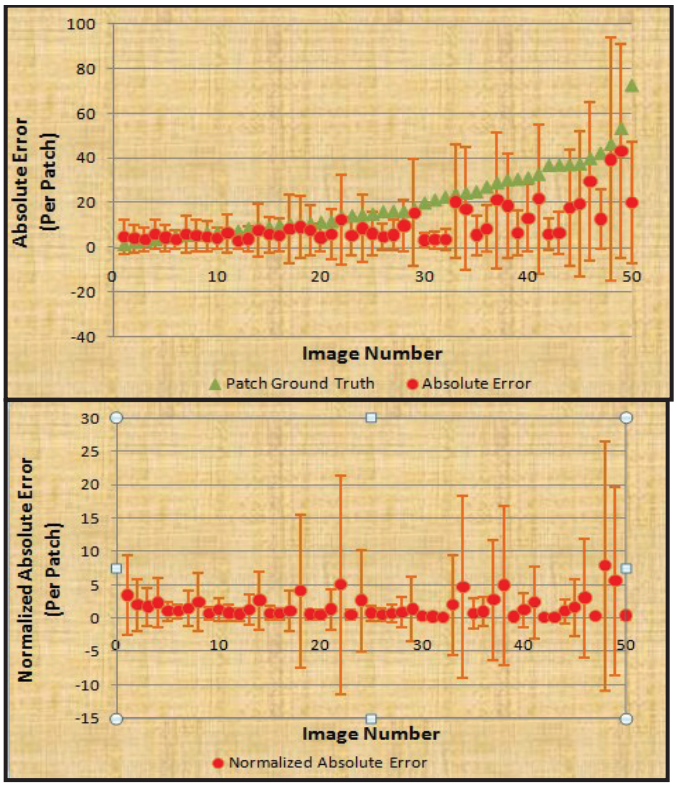

Fig.6: *This figure shows analysis of patch estimates in terms of absolute and normalized absolute differences. The x-axis shows image numbers sorted with respect to patch ground truth. Means are shown in red circles, standard deviation with orange bars and patch ground truths with green triangles.*

VI. Conclusion and Future Directions

In this paper, we propose a system for detecting and counting humans in highly crowded images and video scenes. We combine information from three different sources in terms of counts, confidences and absolute errors at the given patch level. We demonstrated that the proposed system performs better and consistent counting of individuals in densely crowded scenes and videos. The proposed system scales well to different densities, generating consistent Absolute and Normalized Error rates across images with diverse counts.

In future, we plan to enforce smoothness constraint on nearby patches to reduce dependency amongst neighboring patches and improve the overall performance of the proposed system. We also emphasize to bind the time constraint during estimation of patches across the used crowd scenes and videos.

References

[1] W. Ge and R. Collins. Marked point processes for crowd counting. In CVPR, 2009.
[2] M. Li, Z. Zhang, K. Huang, and T. Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In ICPR, 2008.
[3] A. Seiler, G. Evensen, J. Skjervheim, J. Hove, and J. Vabø. Using the enkf for history matching and uncertainty quantification of complex reservoir models. Computational Methods for Large-Scale Inverse Problems and Quantification of Uncertainty, edited by L. Biegler et al, pages 247–271, 2011.
[4] A. Treuille, S. Cooper, and Z. Popovi´c. Continuum crowds. ACM Transactions on Graphics (TOG).
[5] N. Bellomo and C. Dogbe. On the modelling crowd dynamics from scaling to hyperbolic macroscopic models. Mathematical Models and Methods in Applied Sciences, 18(supp01):1317–1345, 2008.
[6] E. Cristiani, B. Piccoli, and A. Tosin. Multiscale modeling of pedestrian dynamics, volume 12. Springer, 2014.
[7] A. Doucet, N. De Freitas, K. Murphy, and S. Russell. Rao-blackwellised particle filtering for dynamic bayesian networks. In Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, pages 176–183. Morgan Kaufmann Publishers Inc., 2000
[8] G. Evensen. Data assimilation: the ensemble Kalman filter. Springer Science & Business Media, 2009.
[9] D. Helbing, P. Molnar, I. J. Farkas, and K. Bolay. Self-organizing pedestrian movement. Environment and planning B: planning and design, 28(3):361–383, 2001.
[10] G. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In CVPR, 2006.
[11] V. Rabaud and S. Belongie. Counting crowded moving objects. In CVPR, 2006.
[12] X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception by hierarchical bayesian models. In CVPR, 2007.
[13] T. Xiang and S. Gong. Beyond tracking: Modelling activity and understanding behaviour. IJCV, 67(1):21–51, 2006.
[14] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern mod-els. In CVPR, 2009.
[15] W. Ma, L. Huang, and C. Liu. Crowd density analysis using co-occurrence texture features. In ICCIT, 2010.
[16] V. Lempitsky and A. Zisserman. Learning to count objects in images. In NIPS, 2010.
[17] A. Chan, Z. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In CVPR, 2008.
[18] S. Cho, T. Chow, and C. Leung. A neural-based crowd esti-mation by hybrid global learning algorithm. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 29(4):535–541, 1999.
[19] D. Kong, D. Gray, and H. Tao. Counting pedestrians in crowds using viewpoint invariant training. In BMVC, 2005.
[20] D. Ryan, S. Denman, C. Fookes, and S. Sridharan. Crowd counting using multiple local features. In Digital Image Computing: Techniques and Applications, 2009.
[21] www.prima.inrialpes.fr/perso/Gourier/Faces/HPDatabase.html accessed on 20-6-2016
[22] www.crcv.ucf.edu/data/crowd_counting.php accessed on 30-6-2016
[23] Idrees, Haroon, et al. "Multi-source multi-scale counting in extremely dense crowd images." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013
[24] Rodriguez, Mikel, et al. "Density-aware person detection and tracking in crowds." 2011 International Conference on Computer Vision. IEEE, 2011.
[25] Lempitsky, Victor, and Andrew Zisserman. "Learning to count objects in images." Advances in Neural Information Processing System.
[26] Xu, Tianchun, et al. "Crowd counting using accumulated HOG." Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016 12th International Conference on. IEEE, 2016.