A Study of Human Gaze Behavior During Visual Crowd Counting

Raji Annadi\textsuperscript{1}, Yupei Chen\textsuperscript{2}, Viresh Ranjan\textsuperscript{1}, Dimitris Samaras\textsuperscript{1}, Gregory Zelinsky\textsuperscript{2}, Minh Hoai\textsuperscript{1}
\textsuperscript{1}Department of Computer Science, \textsuperscript{2}Department of Psychology
Stony Brook University, Stony Brook, NY 11790

Abstract

In this paper, we describe our study on how humans allocate their attention during visual crowd counting. Using an eye tracker, we collect gaze behavior of human participants who are tasked with counting the number of people in crowd images. Analyzing the collected gaze behavior of ten human participants on thirty crowd images, we observe some common approaches for visual counting. For an image of a small crowd, the approach is to enumerate over all people or groups of people in the crowd, and this explains the high level of similarity between the fixation density maps of different human participants. For an image of a large crowd, our participants tend to focus on one section of the image, count the number of people in that section, and then extrapolate to the other sections. In terms of count accuracy, our human participants are not as good at the counting task, compared to the performance of the current state-of-the-art computer algorithms. Interestingly, there is a tendency to under count the number of people in all crowd images. Gaze behavior data and images can be downloaded from https://www3.cs.stonybrook.edu/~cvl/projects/crowd_counting_gaze/

1 Introduction

Large crowd gathering is commonplace, and estimating the size of a crowd is an important problem for different purposes ranging from journalism to public safety. It is known that humans are good at subitizing \cite{6}, i.e., predicting fast and accurate counts for small number of items. For an image of a dense crowd, humans can also provide a very accurate count given enough time. In fact, one common approach to produce ‘gold standard’ annotations \cite{4, 20, 25, 3} for training computer algorithms \cite{7, 16, 4, 1, 15, 21, 12, 17, 13, 11, 22, 2, 10, 23, 26, 24, 5, 19, 8, 9, 14, 18} for crowd counting is to ask human annotators to enumerate and place a dot on each person in the crowd images. The things that remain unclear are: 1) how well can humans count a large number of objects in dense images under limited time, and 2) what are the approaches that humans typically use for counting when there is not enough time to attend to and enumerate over all objects. In this paper, we seek the answers to these two questions by collecting and analyzing the gaze behavior of human subjects that were asked to count under a time constraint.

2 Data Collection

The gaze-behavior dataset we collected consists of 23,036 fixations from 10 participants with the task of counting the number of people in 30 images for 30 seconds per image. The 10 participants were Stony Brook University undergraduate and graduate students (6 male, 4 female, age range 18–28, normal or corrected-to-normal vision). A total of 30 diverse images were selected from the UCF-QNRF dataset \cite{4} and used in this data collection. These images contained as few as 65 people and as many as 12672 people and included both images taken in daylight and at night.

Each participant was instructed to freely view each image and count “how many people were there in the image?” Each trial started with a central fixation dot and the participant had to fixate at the dot and press a button to start the trial. When the image appeared, they had 30s to freely inspect the image and count
the number of people in that image. The image disappeared after 30 seconds and they were asked to type in a number that indicated the number of people in the image. The 30 images were randomly interleaved and it took \( \sim 20 \) minutes to finish the experiment. Eye movements were recorded using an EyeLink 1000 eye-tracker in desktop configuration (SR research) with a sample rate at 1000Hz. Eye-tracker calibration was conducted at the beginning of the experiment. All participants completed the experiment in the same controlled laboratory condition.

3 Gaze Behavior Analysis

This section describes several notable patterns in the gaze behaviors of our ten participants, obtained by visualizing their fixation density maps. All fixation density maps of the 10 participants on 30 crowd images are shown in the last 30 figures of this paper.

Our first observation is that our participants use different approaches for images of sparse and dense crowds. For an image with a sparse crowd where the total number of people is small (less than approximately 100 people) such as the one in Figure 1a, the fixation density map covers the entire crowd area as can be seen in Figure 1b. This suggests that the participants pay attention to every area of the crowd, and they probably aim to account for every single person or group of people in the image. For an image with a denser crowd such as the one in Figure 2a, the fixation density map covers only a small part of the crowd. This is understandable given the size of the crowd and the impossibility of enumerating over all people in the crowd. For an image of a dense crowd, the fixation density maps reveal that our participants tend to focus on one section of the crowd, count or estimate the number of people in that section, and extrapolate to the other sections.

![A sparse crowd image](image1.png) ![Fixation density map](image2.png)

Figure 1: A sparse crowd image and the corresponding fixation density map

Our second observation is that the number of fixations tends to decrease as the size of the crowd increases, as can be seen in Figure 3. For a sparse crowd, our participants were fast to change their eye fixations, and this indicates an enumeration behavior. Meanwhile for a dense crowd, there is some region of the image where our participants attended to for a longer period of time to estimate the total count in the region.

Our third observation is that there is high degree of similarity between the fixation density maps of different subjects for an input image of a sparse crowd, as shown in Figure 4. This further indicates the behavior to account for all people or group of people in a sparse crowd.

We also find that our participants are more successful in estimating the size of a crowd with a uniform spread. Their common approach is to count the number of people in a selected row and selected column and multiple the resulting numbers to obtain the final count. This behavior can be observed in Figure 5 where each fixation density map spreads over one row and one column.
Figure 2: A dense crowd image and the corresponding fixation density map

Figure 3: The mean and median number of fixations as a function of ground truth count. The number of fixations tends to decrease as the size of the crowd increases.

4 Count Accuracy

Table 1 lists the estimated counts from the 10 subjects on all 30 images. The images in the table are listed based on the sizes of the crowds, from smallest to largest. The table also reports the counting accuracy of the human subjects under several performance metrics. Following previous works on automated crowd counting (e.g., [16, 4, 1, 3, 25]), we evaluate the accuracy of human counting based on: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Normalized Absolute Error (MeanNAE), and Median Normalized Absolute Error (MedNAE). Let $C_i$ denote the ground truth count for Image $i$, and $\hat{C}_i$ is the estimated count.
Figure 4: Fixation density maps of two different subjects on the same image with a sparse crowd. These density maps are highly similar to each other.

Figure 5: Fixation density maps from three different subjects on a crowd with a uniform density. The common behavior is to count the number of people in one row and count the number of rows.
| Image    | Count | GT Subj1 Subj2 Subj3 Subj4 Subj5 Subj6 Subj7 Subj8 Subj9 Subj10 | Mean NAE-I | Median NAE-I |
|----------|-------|---------------------------------------------------------------|-------------|--------------|
| Image 07 | 65    | 59  80  55  60  65  65  76  60  62  60                       | 0.09        | 0.08         |
| Image 08 | 82    | 90  70  90  45  55  40  65  50  62  90                       | 0.26        | 0.23         |
| Image 03 | 100   | 130  60  90  50  95  50  800 35 100 200                    | 1.65        | 0.45         |
| Image 06 | 101   | 68  80  110  60  160  50  150 60 65 170                    | 0.40        | 0.41         |
| Image 17 | 122   | 75  65  70  60  50  70  135 50 75 140                    | 0.40        | 0.43         |
| Image 09 | 233   | 230  100 110 180 300 150 320 120 120 220                | 0.34        | 0.36         |
| Image 15 | 291   | 220  300 120 100 70 120 350 70 160 350                | 0.45        | 0.52         |
| Image 10 | 353   | 350  250 175 200 350 180 350 150 200 500                | 0.32        | 0.42         |
| Image 12 | 405   | 550  250 200 200 350 200 370 60 120 500                | 0.43        | 0.44         |
| Image 22 | 443   | 250  350 200 400 300 400 500 200 150 500                | 0.32        | 0.27         |
| Image 27 | 703   | 430  250 150 300 350 200 350 150 160 300                | 0.62        | 0.61         |
| Image 14 | 720   | 350  600 250 400 500 500 600 120 400 700                | 0.95        | 0.44         |
| Image 30 | 858   | 700  200 225 350 250 300 400 150 130 600                | 0.61        | 0.68         |
| Image 11 | 915   | 750  500 375 150 1000 400 1000 150 200 500                | 1.47        | 0.58         |
| Image 02 | 936   | 580  400 300 500 3000 1500 5000 150 480 5000               | 1.49        | 0.64         |
| Image 20 | 1237  | 800  1000 275 700 600 1800 1800 400 500 3500              | 0.63        | 0.49         |
| Image 29 | 1349  | 2000 550 400 1000 5000 1200 1900 500 500 900              | 0.69        | 0.54         |
| Image 05 | 1442  | 1600 2000 375 800 7000 1000 800 1000 400 4000             | 1.38        | 0.73         |
| Image 01 | 1515  | 900  1000 350 1300 700 1000 800 500 900 1500             | 0.41        | 0.41         |
| Image 28 | 1852  | 1400 500 375 800 17000 600 700 600 250 4000             | 1.67        | 0.77         |
| Image 13 | 1929  | 1200 600 500 600 14 1500 9000 100 500 20000            | 1.84        | 0.74         |
| Image 23 | 2555  | 1300 1300 400 800 15500 1500 1800 300 1200 15000         | 2.03        | 0.77         |
| Image 16 | 3002  | 2000 4000 600 3000 10000 1600 8000 700 1500 2000         | 0.75        | 0.48         |
| Image 26 | 4282  | 1300 5000 650 1500 12000 2500 18000 800 1200 130000       | 3.87        | 0.77         |
| Image 18 | 4609  | 4800 3000 700 2000 18000 1900 25000 700 3000 15000       | 1.32        | 0.72         |
| Image 04 | 5520  | 3500 4000 1200 1000 10000 3000 15000 5000 1300 25000      | 0.96        | 0.77         |
| Image 24 | 6052  | 2100 6000 700 5000 20000 3000 12400 1000 1000 12000      | 0.82        | 0.83         |
| Image 25 | 7695  | 11000 10000 1400 1000 15000 5000 40000 2500 5000 28000   | 1.16        | 0.75         |
| Image 21 | 10274 | 6500 6000 900 1500 12500 5000 80000 2000 4500 1000000     | 10.78       | 0.68         |
| Image 19 | 12672 | 1600 1500 1200 4000 18000 9000 100000 2000 25000 1600000  | 13.79       | 0.86         |

Table 1: Ground truth and estimated counts from 10 different subjects on the 30 images. The images in the table are listed based on the sizes of the crowds. The table reports the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) of individual subjects. This table also reports the Mean and Median Normalized Absolute Error (NAE) calculated for individual subjects (i.e., aggregating over images, referred as MeanNAE and MedNAE in the table) or individual images (i.e., aggregating over subjects, referred to as MeanNAE-I and MedNAE-I in the table).

The MAE, RMSE, MeanNAE and MedNAE for Subject $j$ are defined as follows.

$$
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |C_i - \hat{C}_i^j|; \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (C_i - \hat{C}_i^j)^2};
$$

$$
\text{MeanNAE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|C_i - \hat{C}_i^j|}{C_i}; \quad \text{MedNAE} = \text{Median}_{i=1}^{n} \left\{ \frac{|C_i - \hat{C}_i^j|}{C_i} \right\},
$$

(1)
where $n = 30$ the number of images. We also compute the Mean NAE and Median NAE for each individual image by taking the mean or median over the 10 subjects. For Image $i$, the Mean NAE (MeanNAE-I) and Median NAE (MedNAE-I) are defined as follows:

$$\text{MeanNAE-I} = \frac{1}{m} \sum_{j=1}^{m} \frac{|C_i - \hat{C}_i^j|}{C_i}; \quad \text{MedNAE-I} = \text{Median}_{j=1,m} \left\{ \frac{|C_i - \hat{C}_i^j|}{C_i} \right\},$$

(3)

where $m = 10$ is the number of subjects.

As can be seen from Table 1, our human subjects are not as good at counting under time constraint, relative to the current state-of-the-art automated crowd counting methods. The lowest MAE and RMSE values are 1025 and 1650, respectively. As a reference, the MAE and RMSE of the Bayesian+ method [15] on all test images of UCF-QNRF [4] are 88.7 and 154.8 respectively.

Comparing between images, we see that the absolute counting error is proportional to the size of crowd. However, the correlation between the absolute error and the crowd size is not linear. Considering the absolute error normalized by crowd size, the median value (MedNAE-I) is generally higher for a larger crowd, as can be seen from the last column of Table 1 and Figure 6.

Figure 7 is the scatter plot of the estimated counts and the ground truth counts. As can be seen, our human subjects tend to under-count the number of people in each image. On all images, the median and mean estimates are smaller than the ground truth. For images with smaller crowd sizes, the majority of the estimates are lower than the actual crowd sizes.

5 Conclusions

In this paper, we have presented a gaze-behavior dataset consisting of 23K fixations from 10 human participants on thirty crowd images. We have analyzed the gaze behavior of the human participants as well as the accuracy
Figure 7: Estimated counts versus ground truth counts in log-scale plot. Black dots are the individual estimates from our subjects. For each image, the figure only shows eight out of ten estimates, removing the minimum and maximum estimated values; this corresponds to removing the possible outliers outside the 10 to 90 percentiles range. The blue circles and red stars are the mean and median values of the estimated values respectively. The estimates above the diagonal line correspond to over counting, while the ones below the diagonal line correspond to under counting. For all images, both mean and median values are lower than the actual counts. This indicates the under-counting tendency of our subjects. For images with a smaller number of people, most estimated counts are lower than the actual number of people in the image.
of their count estimates. We found that our human participants used different approaches for sparse and dense crowds, and the number of fixations tends to decrease as the size of the crowd increases. For a sparse crowd, the common approach of our participants was to enumerate over all people or group of people in the crowd, leading to fixation density maps that were highly similar to one another. For a denser crowd, our participants tended to focus on counting the number of people in a small section and then extrapolate to other sections. In terms of count accuracy, our human participants performed poorly, compared to the performance of the state-of-the-art computer algorithms. We also observed a tendency to under count the number of people in each image.

References

[1] X. Cao, Z. Wang, Y. Zhao, and F. Su. Scale aggregation network for accurate and efficient crowd counting. In Proceedings of the European Conference on Computer Vision, 2018.
[2] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, and A. G. Hauptmann. Learning spatial awareness to improve crowd counting. In Proceedings of the International Conference on Computer Vision, 2019.
[3] 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.
[4] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah. Composition loss for counting, density map estimation and localization in dense crowds. In Proceedings of the European Conference on Computer Vision, 2018.
[5] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, and L. Shao. Crowd counting and density estimation by trellis encoder-decoder networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[6] E. L. Kaufman, M. W. Lord, T. W. Reese, and J. Volkmann. The discrimination of visual number. American Journal of Psychology, 62:498–525, 1949.
[7] Y. Li, X. Zhang, and D. Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[8] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao. Density map regression guided detection network for rgb-d crowd counting and localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[9] C. Liu, X. Weng, and Y. Mu. Recurrent attentive zooming for joint crowd counting and precise localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[10] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin. Crowd counting with deep structured scale integration network. In Proceedings of the International Conference on Computer Vision, 2019.
[11] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, and H. Wu. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[12] W. Liu, M. Salzmann, and P. Fua. Context-aware crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[13] Y. Liu, M. Shi, Q. Zhao, and X. Wang. Point in, box out: Beyond counting persons in crowds. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[14] E. Lu, W. Xie, and A. Zisserman. Class-agnostic counting. In Proceedings of the Asian Conference on Computer Vision, 2018.
[15] Z. Ma, X. Wei, X. Hong, and Y. Gong. Bayesian loss for crowd count estimation with point supervision. In Proceedings of the International Conference on Computer Vision, 2019.
[16] V. Ranjan, H. Le, and M. Hoai. Iterative crowd counting. In Proceedings of the European Conference on Computer Vision, 2018.
[17] M. Shi, Z. Yang, C. Xu, and Q. Chen. Revisiting perspective information for efficient crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[18] J. Wan and A. Chan. Adaptive density map generation for crowd counting. In Proceedings of the International Conference on Computer Vision, 2019.

[19] J. Wan, W. Luo, B. Wu, A. B. Chan, and W. Liu. Residual regression with semantic prior for crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[20] Q. Wang, J. Gao, W. Lin, and X. Li. Nwpu-crowd: A large-scale benchmark for crowd counting. arXiv preprint arXiv:2001.03360, 2020.

[21] Q. Wang, J. Gao, W. Lin, and Y. Yuan. Learning from synthetic data for crowd counting in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[22] C. Xu, K. Qiu, J. Fu, S. Bai, Y. Xu, and X. Bai. Learn to scale: Generating multipolar normalized density maps for crowd counting. In Proceedings of the International Conference on Computer Vision, 2019.

[23] Z. Yan, Y. Yuan, W. Zuo, X. Tan, Y. Wang, S. Wen, and E. Ding. Perspective-guided convolution networks for crowd counting. In Proceedings of the International Conference on Computer Vision, 2019.

[24] Q. Zhang and A. B. Chan. Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[25] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[26] M. Zhao, J. Zhang, C. Zhang, and W. Zhang. Leveraging heterogeneous auxiliary tasks to assist crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
Figure 8: Image grid containing image 1, its annotation and fixation density maps from all subjects.

Figure 9: Image grid containing image 2, its annotation and fixation density maps from all subjects.
Figure 10: Image grid containing image 3, its annotation and fixation density maps from all subjects.

Figure 11: Image grid containing image 4, its annotation and fixation density maps from all subjects.
Figure 12: Image grid containing image 5, its annotation and fixation density maps from all subjects.

Figure 13: Image grid containing image 6, its annotation and fixation density maps from all subjects.
Figure 14: Image grid containing image 7, its annotation and fixation density maps from all subjects.

Figure 15: Image grid containing image 8, its annotation and fixation density maps from all subjects.
Figure 16: Image grid containing image 9, its annotation and fixation density maps from all subjects.

Figure 17: Image grid containing image 10, its annotation and fixation density maps from all subjects.
Figure 18: Image grid containing image 11, its annotation and fixation density maps from all subjects.

Figure 19: Image grid containing image 12, its annotation and fixation density maps from all subjects.
Figure 20: Image grid containing image 13, its annotation and fixation density maps from all subjects.

Figure 21: Image grid containing image 14, its annotation and fixation density maps from all subjects.
Figure 22: Image grid containing image 15, its annotation and fixation density maps from all subjects.

Figure 23: Image grid containing image 16, its annotation and fixation density maps from all subjects.
Figure 24: Image grid containing image 17, its annotation and fixation density maps from all subjects.

Figure 25: Image grid containing image 18, its annotation and fixation density maps from all subjects.
Figure 26: Image grid containing image 19, its annotation and fixation density maps from all subjects.

Figure 27: Image grid containing image 20, its annotation and fixation density maps from all subjects.
Figure 28: Image grid containing image 21, its annotation and fixation density maps from all subjects.

Figure 29: Image grid containing image 22, its annotation and fixation density maps from all subjects.
Figure 30: Image grid containing image 23, its annotation and fixation density maps from all subjects.

Figure 31: Image grid containing image 24, its annotation and fixation density maps from all subjects.
Figure 32: Image grid containing image 25, its annotation and fixation density maps from all subjects.

Figure 33: Image grid containing image 26, its annotation and fixation density maps from all subjects.
Figure 34: Image grid containing image 27, its annotation and fixation density maps from all subjects.

Figure 35: Image grid containing image 28, its annotation and fixation density maps from all subjects.
Figure 36: Image grid containing image 29, its annotation and fixation density maps from all subjects.

Figure 37: Image grid containing image 30, its annotation and fixation density maps from all subjects.