International Journal on Semantic Web and Information Systems
Volume 18 • Issue 1

Information Entropy Augmented High Density Crowd Counting Network

Yu Hao, Xi'an University of Posts and Telecommunications, China*
Lingzhe Wang, Xi'an University of Posts and Telecommunications, China
Ying Liu, Xi'an University of Posts and Telecommunications, China
Jiulun Fan, Xi'an University of Posts and Telecommunications, China

ABSTRACT

The research proposes an innovated structure of the density map-based crowd counting network augmented by information entropy. The network comprises of a front-end network to extract features and a back-end network to generate density maps. In order to validate the assumption that the entropy can boost the accuracy of density map generation, a multi-scale entropy map extraction process is imported into the front-end network along with a fine-tuned convolutional feature extraction process. In the back-end network, extracted features are decoded into the density map with a multi-column dilated convolution network. Finally, the decoded density map can be mapped as the estimated counting number. Experimental results indicate that the devised network is capable of accurately estimating the count in extremely high crowd density. Compared to similar structured networks which don't adapt entropy feature, the proposed network exhibits higher performance. This result proves the feature of information entropy is capable of enhancing the efficiency of density map-based crowd counting approaches.

KEYWORDS

CNN, Crowd Count, Density Map, Image Regression, Information Entropy

1. INTRODUCTION

The analysis of crowd in extremely-high density is essential to public safety. By predicting or alarming the potential hazardous incidents such as panic, casualties can be reduced or avoided. Crowd counting techniques can provide the real-time number of pedestrians within the footage, which is a crucial information to prevent stampede. The strategy of conventional computer vision-based techniques for crowd counting is to extract features such as HOG (Xu et al., 2016), contour (Dong et al., 2007; Weikert et al., 2020) and spatial-temporal information (Wang, 2019) from image patches obtained with a sliding window, and feed these features to classifiers such as SVM (Xu et al., 2016; Tu et al., 2013; Zhao et al., 2017), random forest (Li & Zhou, 2016; Pham et al., 2015) and Markov Model (Jalal et al., 2020) to determine if a pedestrian exists in the patch. Once the detection for the entire footage is completed, the total number of detected pedestrians can be obtained. The major defect of conventional approaches is the low performance on high crowd density. When the density increased to a high-level, pixel-wise information for each pedestrian decreases drastically, and more occlusions will occur. In this case, the accurate detection of individual becomes difficult. and it will cause a

DOI: 10.4018/IJSWIS.297144 *Corresponding Author

significant performance degradation. In order to tackle these issues, regression-based approaches attempt to fabric relations between the crowd distribution and certain global features of the entire footage, and estimate the total crowd count. Arteta et al. (2014) firstly introduced the concept of density map by convolving the pedestrian's spatial positions in training data with a Gaussian kernel. In the training phase, extracted features and density maps are exploited to train the decoding model. In the testing phase, features are feed to the well-trained model to decode the density map, which will be used to estimate the crowd count. This technique effectively addressed the problem of occlusions in high density, and inspired the deep-learning based crowd counting techniques.

The structure of deep learning-based techniques usually comprises a front-end (feature extraction) network and a back-end (density map generation) network (Cao et al., 2018; Li et al., 2018; Liu et al., 2019; Karthika, 2021; Ranjan et al., 2018; Sindagi & Patel, 2017; Zhang et al., 2016). The front-end network extracts multi-scale features from image data, while the back-end network decodes the features into a density map. Instead of extracting patches with a sliding window, deep learning-based approaches use entire image to fulfill the end-to-end training. Therefore, the processing speed is often much faster than the conventional. Also, the counting accuracy of deep learning-based approaches outperforms the conventional in most of cases.

In the training phase, by convolving the pedestrian's head position with convolution kernels, the ground-truth density map can be generated. However, due to the camera's perspective, head sizes of pedestrians often vary in the scene. In order to percept the different head size, convolution kernels in various scales are adapted to extract features in multiple scale. Zhang et al. (2016) introduced a novel structure namely Multi-Column Neural Network (MCNN). MCNN includes 3 independent feature extraction paths, where each path adapts kernels with different scales. To achieve the most accurate map generation in the training phase, the average distance between each pedestrian's neighbors is calculated to get the self-adapted variance of the kernel. Since the MCNN only exploits the global features, its ability of perception on local features is relatively weak. In order to further enhance the counting accuracy, Sindagi & Patel (2017) introduced the Context Pyramid CNN (CP-CNN) by adapting a triple-stream network to extract the context relation, including Global Context encoder, Density Map Encoder and Local Context Encoder. Encoded context features are then concatenated with generated density map to achieve a better count estimation. However, multi-path feature extraction approaches like MCNN and CP-CNN have a high time-consumption. In order to optimize the balance between processing time and the quality of density map, Congested Scene Recognition Net (CSRNet) (Li et al., 2018) exploits the initial ten layers of the VGG-16 network (Chen et al., 2020; Simonyan & Zisserman, 2014) to achieve the fast convolutional feature extraction. Next, the dilated convolution network (Chen, Papandreou, Kokkinos, Murphy & Yuille, 2017; Chen, Papandreou, Schroff & Adam, 2017; Yu & Koltun, 2016) with different dilate rates is utilized to percept different head sizes and prevent information lose in pooling layer. Context-Aware Network (CAN) (Liu et al., 2019) further enhanced CSRNet by applying spatial-pyramid pooling (He et al., 2014) between VGG-16 and dilated network to obtain the contrast context feature. More accurate density map can be generated by using the context feature to augment the raw feature.

Another strategy to enhance the counting performance is by enhancing the density map's resolution. Iterative-Counting CNN (IC-CNN) (Ranjan et al., 2018) attempts to generate high-resolution density map by merging original image, low-resolution density map and feature map extracted with MCNN together. Unlike the multi-path structured MCNN, Scale Aggregation Net (SANet) (Cao et al., 2018) adapted the single-stream multi-layered network structure based on inception model (Szegedy et al., 2015). For each layer, kernels with various scales are exploited to extract and aggregate features, and send them to the next layer. After 3 iterations of extraction and aggregation, transposed convolutions are implemented to generate the density map in high-resolution. SANet outperforms CP-CNN and CSRNet. However, due to the multi-scale extraction in each layer, the computational burden of SANet is relatively high. The selection of loss function in back-end network can also be crucial. The work

of Wan (2021) proved the pixel-wise L2 loss and Bayesian loss can be exploited as the generalized loss function to outperform others for accurate density map estimation.

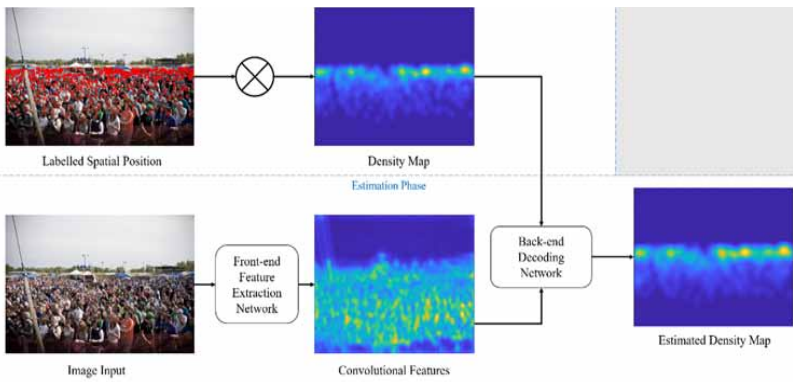
As statistical feature, information entropy is widely exploited for the detection of abnormal crowd behaviors (Hao et al., 2019; Zhang et al., 2019). In order to explore the entropy's capability to enhance the efficiency of deep learning-based crowd counting approaches, this paper introduced an entropy augmented crowd counting neural network. Same to the main-stream structure, the proposed network contains the front-end feature extraction network and back-end decoding/regression network as well. Inspired by CSRNet, the front-end adapted a fine-tuned VGG-16 network to achieve fast extraction of features. On the other hand, multi-scale entropy maps are obtained with an extraction model from the input image, and merged with convolutional features. Finally, a back-end decoding network based on the dilated convolution utilizes merged features to generate the density map. This paper attempts to prove the information entropy's effectiveness to enhance the accuracy of density map modeling and pedestrian counting in extremely high density.

The remainder of contents are distributed as below. Section 2 explains the principle of crowd counting techniques in high-density with density map estimation as well as relevant concepts, including the ground-truth density map generation and the self-adapted Gaussian kernel. Section 3 introduces the architecture of the devised network. In the front-end network, two feature extraction processes including a fine-tuned VGG16 network and a multi-scale entropy map extraction model are specifically described. In the decoding network, a multi-path dilated convolution network is introduced to decode the feature map into density map. Section 4 introduces adapted datasets for the measurement of the proposed technique, the evaluation criteria and comparative results with main-stream crowd counting approaches. Section 5 concludes the research and discusses the potential optimizing strategy of the proposed approach.

2. COUNTING PRINCIPLE OF DENSITY MAP-BASED APPROACHES

As illustrated in Figure 1, the counting strategy based on deep learning comprises two independent phases - training and estimation. In the former phase, labelled spatial positions of pedestrian's heads are convolved with a fixed gaussian kernel to generate the ground-truth density map. Next, the front-end network is used to extract features from the original image. The map generation network trains the decoding model by regressing these features into the density map. In the estimation phase, the well-trained decoding model is capable of transforming modelled features into density map. Since a mapping relation exists between density map and actual crowd number, a count can finally be estimated.

Figure 1. Principles of deep-learning based crowd counting approach



The density map is utilized to train the back-end decoding network, and estimate the count number. Therefore, the density map's quality ultimately determines the accuracy of counting result. In order to model the density map in training phase, spatial positions $\left(x_N, y_N\right)$ of N pedestrians are firstly collected from the manually labelled dataset. For each pedestrian i, a Gaussian kernel $G\left(x,y\right)$ is convolved with the value of 1 at position $\left(x_i, y_i\right)$. Next, convolution results of all pedestrians are accumulated to get the density map D of current footage. The global density map $D\left(x,y\right)$ can be expressed as Equation 1.

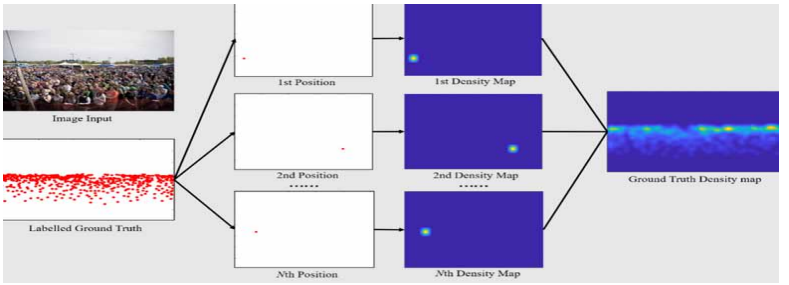
$$D\left(x,y\right) = \sum_{i=1}^{N} \delta\left(x - x_i, y - y_i\right) * G_{\sigma_i}\left(x,y\right) \tag{1}$$

Where $G_{\sigma_i}\left(x,y\right) = \frac{1}{2\pi\sigma_i} \exp\left\{-\frac{x^2 + y^2}{2\sigma_i^2}\right\}$, and $\delta\left(x - x_i, y - y_i\right)$ is the Dirac delta function representing pedestrian i's head position, σ is the standard deviation of $G\left(x,y\right)$. When modeling $D\left(x,y\right)$, the value of σ determines the scale of pedestrian's head. In real-life scenario, scales of heads may vary or distort due to the perspective of camera and occlusion between pedestrians. Therefore, a fixed σ can't guarantee the quality of $D\left(x,y\right)$ in extremely high crowd density, which can potentially impact the accuracy of count estimation. The self-adapted σ is a possible solution to tackle this issue. Based on the assumption that pedestrians are evenly distributed in extremely crowded scene, shorter Euclidean distance $\rho_{i,j}$ between 2 pedestrians i and j indicates they are farther from the camera, and the corresponding σ should be lower. Thus, the value of σ_i can be modelled based on distances with k nearest neighbors. The self-adapted σ_i can be expressed as Equation 2.

$$\sigma_i = \eta \sum_{j=1}^{k} d_j \tag{2}$$

Where d_j is the Euclidean distance between pedestrian j and current i. η indicates the scale parameter and empirically set to 0.1 based on experimental results. The process of ground-truth density map modelling can be illustrated as Figure 2.

Figure 2. The procedure of ground-truth density map modelling



In the estimation phase, features are obtained from the input image $I\left(x,y\right)$ with the extraction network, and sent to the decoding network to estimate the corresponding density map. The front-end network is often a multi-path network, in order to extract features in various types or scales. For example, MCNN (Zhang et al., 2016) adapts a triple-stream network with kernel size 5, 7 and 9 to handle different perceptive scales. CP-CNN (Sindagi & Patel, 2017) proposed a triple-stream network including global context, convolutional and local context feature streams. The adapted feature extraction network stream is expected to provide additional information for the density map estimation. However, since the multi-path network usually has higher time consumption when extracting features than single-stream, which can hamper the efficiency of network. Li et al. (2018) and Liu et al. (2019) replace the front-end part with pre-trained single-stream network to achieve fast feature extraction, and exploits multi-stream dilated convolution network to percept and decode the estimated density map. This paper adapted the concept of pre-trained front-end network as well as the additional feature extraction path. Like the import of contextual information in CAN (Liu et al., 2019), a multi-scale entropy feature extraction process is imported into the dual-stream front-end network.

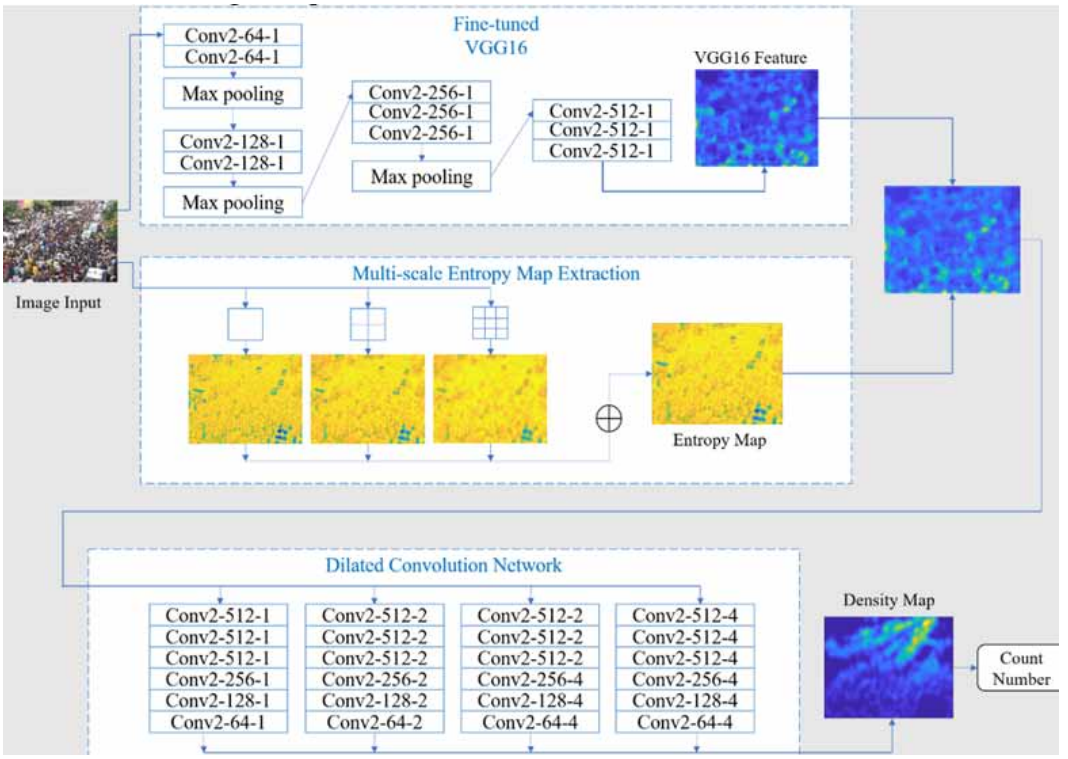
During the process of regression/decoding features into the estimated density map, the primary defect of utilizing ordinary convolution network is the loss of feature's inner structure and local patterns due to the down-sampling. This defect impacts the accuracy of the density map estimation. Using smaller convolution kernel can better percept local features, but the scale of perception field will decrease. Usually, the multi-stream dilated convolution structure is implemented to tackle the defect of using ordinary networks while increasing the perceptive field. The proposed back-end network applied a multi-stream hybrid dilated convolution structure to estimate the density map based on multi-scale perceptions. Once the final density map $\hat{D}$ is decoded, the estimated number of pedestrians $\hat{c}$ can be obtained from all pixels p within $\hat{D}$, as expressed in Equation 3.

$$\hat{c} = \sum\nolimits_{p \in I} \hat{D}\left(p|I\right) \tag{3}$$

3. THE STRUCTURE OF INFORMATION ENTROPY AUGMENTED CROWD COUNTING NETWORK

The specified structure of the devised network is portraited as Figure 3. As previously mentioned, the network comprises a front-end and a back-end networks. The front-end network at the top part of Figure 3 extracts features, and the back-end network at the bottom part decodes the estimated density map. Firstly, the original image is sent into 2 independent streams within the front-end network, including a fine-tuned network with first 10 layers of VGG-16, and a multi-scale entropy map extraction model. Next, the extracted VGG-16 feature and entropy map are merged for density map estimation and sent to the back-end decoding network. The back-end is a multi-stream hybrid dilated convolution network, dilated factors are set to 1, 2, 2, 4 to percept features in various scales. Once the density map is modelled, a count number can be estimated according to Equation 3.

Figure 3. The structure of entropy augmented crowd count network



3.1 The Fine-tuned VGG16 Feature Extraction Network

The adapted VGG-16 in the front-end network includes ten 2-dimensional convolution layers and three max-pooling layers. The label Conv2-64-1 in Figure 3 indicates the layer's type is convolution 2D, there are 64 different filters, and the dilation factor is 1. Every convolution layer is attached with a ReLU layer. The introduced approach utilized the standard pre-trained model. Thus, the training process of VGG-16 can be skipped. Assuming the gray-scale image of original input is I, the extracted feature F^v can be expressed as Equation 4.

$$F^v = f\left(I\right) \tag{4}$$

Where $f\left(x\right)$ indicates the feature extraction process of the VGG-16 network. Note that the dimension of F^v is 512. Since it will be merged with the multi-scale entropy map in the following process, F^v is rescaled to the identical size of original image and then normalized. The normalization approach is adapted from the research of Zhang et al. (2019), which can be expressed as Equation 5.

$$F^{v'} = \frac{F^v - F^v_{min}}{F^v_{max} - F^v_{min}} F^v \tag{5}$$

Where $F^{v'}$ is the normalized feature, F^{v}_{max} is the maximum value, and F^{v}_{min} is the minimum values in F^{v}. When normalized, features are ready to be merged with the multi-scale entropy map for density map estimation.

3.2 Extracting the Multi-scale Entropy Map

Shannon Entropy is the quantitative measurement of the information's uncertainty. As visual feature, the entropy describes the irregularity of pixel's distribution in image. The entropy of local image part reveals details in the footage, which can provide additional information for the modelling of density map. The aim of the devised entropy feature extraction process is to obtain the multi-scale Entropy Map from the original image. To obtain the entropy map $H^{\varepsilon_i}\left(x,y\right)$ on scale ε_i from input image $I\left(x,y\right)$, the entropy $H_{x,y}$ of each pixel $q_{x,y} \in I$ will be calculated with pixels $q \in \left(x \pm \varepsilon_i / 2, y \pm \varepsilon_i / 2\right]$, which can be expressed as Equation 6.

$$H_{x,y} = -\sum_{j=1}^{\varepsilon_i^2} p_j \log p_j \tag{6}$$

Where p_j is the probability of q_j's gray scale level. By calculating $H_{x,y}$ of all pixels within $I\left(x,y\right)$, the raw entropy map H^{ε_i} on scale ε_i can be obtained. Similar with F^{v}, the entropy map will also be normalized into $H^{\varepsilon_i'}$ as Equation 7.

$$H^{\varepsilon_i'} = \frac{H^{\varepsilon_i} - H^{\varepsilon_i}_{min}}{H^{\varepsilon_i}_{max} - H^{\varepsilon_i}_{min}} H^{\varepsilon_i} \tag{7}$$

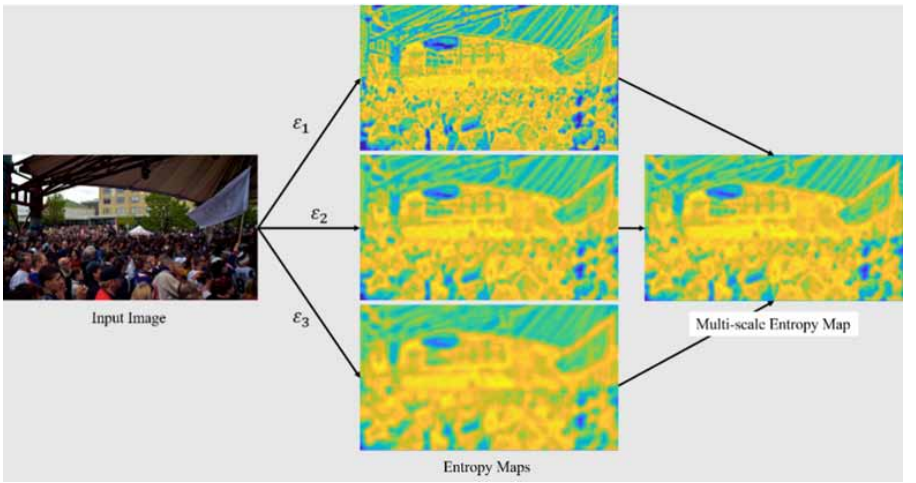
In order to percept different head sizes in the footage, multiple ε_i can be applied in the extraction process to obtain the multi-scale entropy map. The selection of ε is determined based on the actual environment of footage. Since the crowd distribution of the dataset exploited in this research is relatively stable, the variance of head sizes is small. Values of ε, 2ε and 4ε are selected as scales for the entropy map generation. Before merging with VGG16 features, entropy maps on multiple scales are integrated as H, which can be expressed as Equation 8.

$$H = \frac{1}{K}\sum_{i=1}^{K} H^{\varepsilon_i'} \tag{8}$$

Where K is the number of scales. The introduced multi-scale entropy map extraction process can be illustrated as Figure 4. Once H is obtained, it can be merged with extracted VGG16 features as the input F for the back-end network. The merging procedure can be simply expressed as Equation 9, where η is the weight factor of H, it determines how much the H can influence the feature map F.

$$F = F^{v} + \eta H \tag{9}$$

Figure 4. The extraction procedure of multi-scale entropy map



3.3 The Hybrid Dilated Convolution Network

The back-end network adapted a multi-stream hybrid dilated convolution network. For all four streams, hybrid dilated rates are set as 1, 2 and 4. The third stream applied a hybrid dilated rate of 2 and 4. By applying the dilated rate larger than 1, the down-sampling process can be avoided while expanding the perceptive field, and the internal structure of input data is preserved. Multiple dilated rates ensure features in various scales can be precepted. For each input F, the back-end network decodes an estimated density map $\hat{D}$, and the estimated count number $\hat{c}$ can be obtained with Equation 3.

4. EXPERIMENT RESULTS AND ANALYSIS

The devised experiments aim to assess the effectiveness of proposed approach by comparing performances with others, and prove the adaption of entropy is able to improve the density map-based counting approach. This section introduces datasets utilized in experiments, the evaluation metrics of performance and comparative results on various approaches.

4.1 Datasets

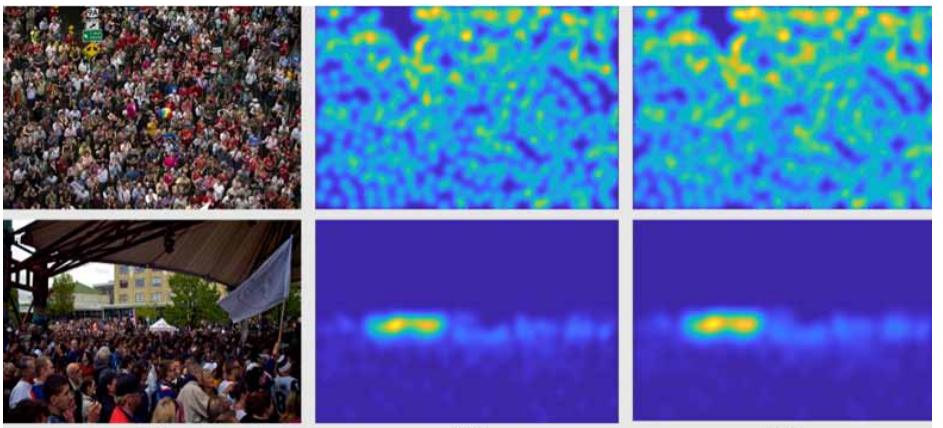
The ShanghaiTech dataset is firstly mentioned in the research of Zhang et al. (2016). It is most widely adapted for the analysis of deep-learning based crowd counting techniques (Cao et al., 2018; Li et al., 2018; Liu et al., 2019; Ranjan et al., 2018; Sindagi & Patel, 2017; Zhang et al., 2016). This dataset consists of 2 subsets, namely set A and set B. Set A comprises 300 images for training and 182 images for testing. Set B comprises 400 images for training and 316 images for testing. Each image contains more than hundreds of manually labeled pedestrians, and the variation trend of crowd number in each image is relatively stable.

Additionally, several datasets with unique patterns are frequently adapted in various researches as well. UCF-QNRF is introduced in the work of Haroon et al. (2018). Since this dataset collects images from multiple data sources, its image quality and pedestrian number have a drastic variation trend. Thus, UCF-QNRF is more appropriate for testing the system's adaptiveness than accuracy. The data scale of UCF_CC_50 (Haroon et al., 2013; Jiang et al., 2021) is very limited, it comprises totally 50 images. Therefore, UCF_CC_50 is mostly utilized in conventional approaches instead of deep-learning based approaches. WorldExpo'10 set (Cong et al., 2015) contains large amount of video data instead of static images, where a portion of video frames is manually annotated. Differed with

others, this set labels the Region of Interest in all footages, which can help reducing the computational burden during the analysis. All exploited datasets contain the manually labeled spatial position of each pedestrian and total head count.

This paper chooses ShanghaiTech as the primary dataset for the measurement of technical performance, and exploits some of above-mentioned main-stream sets to compare the effectiveness of different approaches. Figure 5 illustrates some random images of ShanghaiTech set, as well as the corresponding ground-truth and estimated density maps obtained with the devised approach. The image at the first row comprises an evenly distributed crowd from the over-head view, and the one at the bottom contains a crowd with large perspective rate. Despite the structure of footages varies dramatically, estimated density maps can be successfully generated. Results prove the devised approach is able to handle crowd observed by various camera views.

Figure 5. (a) Original images; (b) Ground truth maps; (c) Estimated maps



4.2 Evaluation Metrics

To assess the efficiency of devised approach, measurable criteria should be adapted for the analysis. The main-stream evaluating metrics of deep-learning based counting approaches are Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) of estimated and ground truth pedestrian counts in the entire test set. Definitions of MAE and RMSE can be expressed as Equation 10.

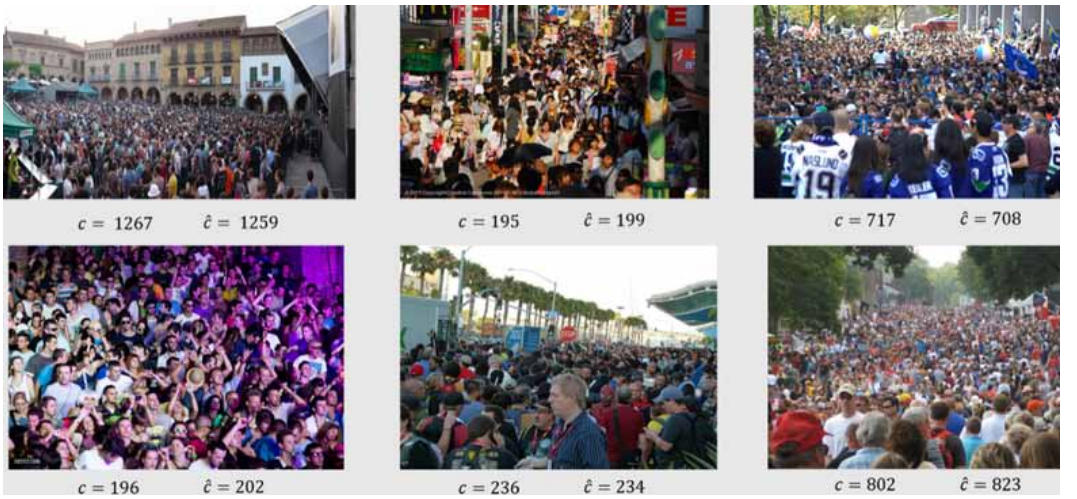
$$\begin{cases} MAE = \frac{1}{M}\sum_{i=1}^{M}\left|c_i - \widehat{c_i}\right| \\ RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(c_i - \widehat{c_i}\right)^2} \end{cases} \tag{10}$$

Where M is the total number of footages within the test set, c_i indicates the ground-truth of total head count, and $\widehat{c_i}$ is the estimated pedestrian number of image i. For the same dataset, the approach with smaller MAE and RMSE indicates better performance.

4.3 Experimental Results

In the first experiment, crowd count $\hat{c}$ is estimated for all images within the test sets of ShanghaiTech. The Figure 6 illustrates the comparative results between the $\hat{c}$ and ground-truth c of some images. The result indicates the proposed approach is capable of estimating the count on footages with various view perspectives and crowd density.

Figure 6. Sample Images of ShanghaiTech set, corresponding ground truth and estimated crowd count



The MAE and RMSE are calculated with the proposed estimation approach on both A and B subsets of ShanghaiTech. Main-stream networks including MCNN, CP-CNN, IC-CNN and CSRNet are implemented on the same dataset to compare the performance. As listed in Table 1, the proposed network obtained the lowest MAE and RMSE on both subsets. The results indicate the introduced approach outperformed others. Since the proposed approach has the similar VGG-16 feature extraction process and the back-end network as CSRNet, the better performance indicates the entropy map indeed improved the accuracy of estimation.

Table 1. Estimated Counting Results of multiple approaches

	ShanghaiTech Subset A		**ShanghaiTech Subset B**	
Approaches	MAE	RMSE	MAE	RMSE
MCNN	110.2	173.2	26.4	41.3
CP-CNN	73.6	106.4	20.1	30.1
IC-CNN	68.5	116.2	10.7	16.0
CSRNet	68.2	115.0	10.6	16.0
Proposed	**67.8**	**114.2**	**10.1**	**14.3**

The performance on other datasets is analyzed as well. For UCF_CC_50 set, the proposed approach has higher MAE and RMSE than CP-CNN and IC-CNN. Since UCF_CC_50 set has only 50 images, the result indicates the proposed approach doesn't have the highest efficiency on dataset with small scale. On the contrary, the devised approach has lower MAE and RMSE than CSRNet, which proves the adapted entropy map feature is capable of enhancing the counting performance on small scale training data. For the WorldExpo'10 set, only annotated data is exploited to evaluate the performance. The result indicates the devised approach has higher effectiveness than CSRNet and most approaches on the video data. Comparative experiments are further conducted between proposed approach and networks involved with contextual features such as SANet and CAN. The result shows the performance of proposed approach is slightly lower than SANet and CAN, which indicates the higher-level contextual features performs better than statistical features on density map-based approaches. It can be expected that by merging the entropy into contextual features, the counting accuracy can be further enhanced if the computational efficiency is ignored. For the model complexity, the extraction process of entropy did increase the computational time. However, the impact is primarily on the training phase. The detecting phase only suffered with an inferior increasement of processing time.

Table 2. Results on set UCF_CC_50 and set WorldExpo'10

	UCF_CC_50		**WorldExpo'10**	
Approaches	MAE	RMSE	MAE	RMSE
MCNN	377.6	509.1	11.6	16.3
CP-CNN	295.8	320.9	8.9	11.2
IC-CNN	260.9	365.5	10.3	14.5
CSRNet	266.1	397.5	8.6	10.8
SANet	258.4	334.9	8.2	10.1
CAN	212.2	243.7	7.4	9.8
Proposed	**262.8**	**385.6**	**8.3**	**10.2**

5. CONCLUSION

In this paper, an entropy feature augmented density map estimation network is proposed for the pedestrian counting in high crowd density. This paper attempts to explore the information entropy's capability of improving the accuracy of the density map's estimation. In the devised network, the input image is separately processed with a fine-tuned VGG16 network and information entropy calculation model to extract features in multiple scales. The merged features are decoded with a multi-stream hybrid dilated convolution network to produce the estimated map. As experimental results indicate, by adapting additional information entropy feature to the CSRNet-based network structure, the estimation performance is substantially increased on images and videos with extremely high crowd density. However, the statistical entropy feature doesn't outperform the contextual features. One potential optimizing strategy is to merge entropy feature with contextual features. This could further improve the accuracy of estimation by sacrificing a portion of processing speed. The future work will be concentrated on the integration of information entropy and other features while maintaining the computational efficiency.

ADDITIONAL FUNDING INFORMATION

The publisher has waived the Open Access Publication fee for this article.

CONFLICT OF INTEREST

We all declare that we have no conflict of interest in this paper.

ACKNOWLEDGMENT

This research was funded by National Science Foundation of China, grant number 62071378; Key Research and Development Program of Shaanxi, grant number 2019GY-054; Key Projects of Postgraduate Joint Cultivation Workstation of Xi'an University of Posts and Telecommunications, grant number YJGJ201902.

REFERENCES

Arteta, C., Lempitsky, V., Noble, J. A., & Zisserman. (2014). Interactive Object Counting. In *2014 European Conference on Computer Vision* (pp. 504-518). doi:10.1007/978-3-319-10578-9_33

Cao, X., Wang, Z., Zhao, Y., & Su, F. (2018). Scale Aggregation Network for Accurate and Efficient Crowd Counting. In *Proceedings of the European Conference on Computer Vision* (pp.734-750). doi:10.1007/978-3-030-01228-1_45

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *99*(1), 117–126. PMID:28463186

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). *Rethinking atrous convolution for semantic image segmentation*. ArXiv.

Chen, J. R., Chao, Y. P., Tsai, Y. W., Chan, H. J., Wan, Y. L., Tai, D. I., & Tsui, P. H. (2020). Clinical Value of Information Entropy Compared with Deep Learning for Ultrasound Grading of Hepatic Steatosis. *Entropy (Basel, Switzerland)*, *22*(1), 1006–1018.

Cong, Z., Hongsheng, L., Xiaogang, W., & Xiaokang, Y. (2015). Cross-Scene Crowd Counting via Deep Convolutional Neural Networks. In *2015 Conference on Computer Vision and Pattern Recognition* (pp.833–841). doi:10.1109/CVPR.2015.7298684

Dong, L., Parameswaran, V., Ramesh, V., & Zoghlami, I. (2007). Fast Crowd Segmentation Using Shape Indexing. *IEEE 11th International Conference on Computer Vision*, 1-8.

Hao, Y., Xu, Z. J., Liu, Y., Wang, J., & Fan, J. L. (2019). Effective Crowd Anomaly Detection Through Spatio-temporal Texture Analysis. *International Journal of Automation and Computing*, *16*(1), 27–39. doi:10.1007/s11633-018-1141-z

Haroon, I., Imran, S., Cody, S., & Mubarak, S. (2013). Multi-Source Multi-Scale Counting in Extremely Dense Crowd Images. *26th Conference on Computer Vision and Pattern Recognition*, 2547–2554.

Haroon, I., Muhmmad, T., Kishan, A., Dong, Z., Somaya, A. M., Nasir, R., & Mubarak, S. (2018). Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds. *European Conference on Computer Vision*, 532-546.

He, K., Zhang, X., Ren, S., & Sun, J. (2014). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *The 13th European Conference on Computer Vision*, 1904-1916.

Jalal, A., Khalid, N., & Kim, K. (2020). Automatic Recognition of Human Interaction via Hybrid Descriptors and Maximum Entropy Markov Model Using Depth Sensors. *Entropy (Basel, Switzerland)*, *22*(1), 817–829. doi:10.3390/e22080817 PMID:33286588

Jiang, X., Zhang, L., Tianzhu, Z., Pei, L., Bing, Z., Yanwei, P., & Mingliang, X. (2021). Density-Aware Multi-Task Learning for Crowd Counting. *IEEE Transactions on Multimedia*, *23*(1), 443–453. doi:10.1109/TMM.2020.2980945

Karthika, R. (2021). Enhanced Learning Experiences Based on Regulatory Fit Theory Using Affective State Detection. *International Journal on Semantic Web and Information Systems*, *17*(4), 37–55. doi:10.4018/IJSWIS.2021100103

Li, T., & Zhou, M. (2016). ECG Classification Using Wavelet Packet Entropy and Random Forests. *Entropy (Basel, Switzerland)*, *18*(1), 285–298. doi:10.3390/e18080285

Li, Y., Zhang, X., & Chen, D. (2018). CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1091-1100. doi:10.1109/CVPR.2018.00120

Liu, W., Salzmann, M., & Fua, P. (2019). Context-Aware Crowd Counting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5099-5108.

Pham, V. Q., Kozakaya, T., Yamaguchi, O., & Okada, R. (2015). COUNT Forest: CO-Voting Uncertain Number of Targets Using Random Forest for Crowd Density Estimation. *2015 IEEE International Conference on Computer Vision*, 3253-3261. doi:10.1109/ICCV.2015.372

Ranjan, V., Le, H., & Hoai, M. (2018). Iterative Crowd Counting. *Proceedings of the European Conference on Computer Vision*, 270-285.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *2014 International Conference on Learning Representations*, 1230-1245.

Sindagi, V. A., & Patel, V. M. (2017). Generating High-Quality Crowd Density Maps using Contextual Pyramid CNNs, *Proceedings of the IEEE International Conference on Computer Vision*, 1861-1870. doi:10.1109/ICCV.2017.206

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going Deeper with Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Tu, J., Zhang, C., & Hao, P. (2013). Robust real-time attention-based head-shoulder detection for video surveillance. *The 20th IEEE International Conference on Image Processing*, 3340-3344.

Wan, J., Liu, Z., & Chan, A. B. (2021). A Generalized Loss Function for Crowd Counting and Localization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1974-1983. doi:10.1109/CVPR46437.2021.00201

Wang, M. (2019). FollowMe: A mobile crowd sensing platform for spatial-temporal data sharing. *International Journal of High Performance Computing and Networking*, *14*(4), 416–424. doi:10.1504/IJHPCN.2019.102347

Weikert, D., Mai, S., & Mostaghim, S. (2020). Particle Swarm Contour Search Algorithm. *Entropy (Basel, Switzerland)*, *22*(1), 407–420. doi:10.3390/e22040407 PMID:33286181

Xu, T., Chen, X., Wei, G., & Wang, W. (2016). Crowd counting using accumulated HOG. *The 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 1877-1881.

Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. *2016 International Conference on Learning Representations*, 1105-1113.

Zhang, X., Lin, D., Zheng, J., Tang, X., Fang, Y., & Yu, H. (2019). Detection of Salient Crowd Motion Based on Repulsive Force Network and Direction Entropy. *Entropy (Basel, Switzerland)*, *21*(1), 608–622. doi:10.3390/e21060608 PMID:33267322

Zhang, Y., Zhou, D., Chen, S., Gao, S., & Ma, Y. (2016). Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 589-597. doi:10.1109/CVPR.2016.70

Zhao, H., Sun, M., Deng, W., & Yang, X. (2017). A New Feature Extraction Method Based on EEMD and Multi-Scale Fuzzy Entropy for Motor Bearing. *Entropy (Basel, Switzerland)*, *19*(1), 14–28. doi:10.3390/e19010014

Yu Hao received Ph.D. degree in computing and engineering from the University of Huddersfield, UK in 2019 and the M. Sc. degree in computer science from the Wichita State University, USA in 2011. Currently, he is a lecturer in School of Communications and Information Engineering, Xi'an University of Posts and Telecommunications, China. His research focuses on the analysis of crowd abnormal behaviors.

Lingzhe Wang studied Information Engineering at Xi'an University of Posts and Telecommunications from 2016 to 2020. He was admitted to the University of Information and Communication Engineering in the same year and joined the video image processing team. The main research directions are artificial intelligence, abnormal behavior detection.

Ying Liu received the Ph.D. degree in computer vision from the Monash University, Australia in 2007. And she worked as a post doctor researcher at Nanyang Technological University, Singapore until 2010. She is the chief engineer of Shaanxi Forensic Science Digital Information Laboratory Research Center, China since 2012. She has published over 60 peer-reviewed journal and conference papers in the relevant fields. She was grant annual best paper of Pattern Recognition and Tier A paper from Australia Research Council. Her research interest includes pattern recognition, machine learning and forensic science.

Jiu-Lun Fan received the B.Sc. and M.Sc. degrees in mathematics from the Shaanxi Normal University, China in 1985 and 1988, respectively, and the Ph. D. degree in electronic engineering from the Xidian University, China in 1998. Currently, he is the president of Xi'an University of Posts and Telecommunications, China since 2015. He has published over 200 peer-reviewed journal and conference papers in the relevant fields. His research interests include signal processing, pattern recognition and communications security.