DOI: https://doi.org/10.15276/hait.07.2024.17
UDC 004.932.72

Accurate crowd counting for intelligent video surveillance systems

Ruslan Y. Dobryshev[1)]
ORCID: https://orcid.org/0009-0007-8639-3157, rdobrishev@gmail.com
Maksym V. Maksymov[1)]
ORCID: https://orcid.org/0000-0002-3292-3112, prof.maksimov@gmail.com, Scopus Author ID: 7005088554
[1)] Odesa Polytechnic National University, 1, Shevchenko Av. Odesa, 65044, Ukraine

ABSTRACT

The paper presents a novel deep learning approach for crowd counting in intelligent video surveillance systems, addressing the growing need for accurate monitoring of public spaces in urban environments. The demand for precise crowd estimation arises from challenges related to security, public safety, and efficiency in urban areas, particularly during large public events. Existing crowd counting techniques, including feature-based object detection and regression-based methods, face limitations in high-density environments due to occlusions, lighting variations, and diverse human figures. To overcome these challenges, the authors propose a new deep encoder-decoder architecture based on VGG16, which incorporates hierarchical feature extraction with spatial and channel attention mechanisms. This architecture enhances the model's ability to manage variations in crowd density, leveraging adaptive pooling and dilated convolutions to extract meaningful features from dense crowds. The model's decoder is further refined to handle sparse and crowded scenes through separate density maps, improving its adaptability and accuracy. Evaluations of the proposed model on benchmark datasets, including Shanghai Tech and UCF CC 50, demonstrate superior performance over state-of-the-art methods, with significant improvements in mean absolute error and mean squared error metrics. The paper emphasizes the importance of addressing environmental variability and scale differences in crowded environments and shows that the proposed model is effective in both sparse and dense crowd conditions. This research contributes to the advancement of intelligent video surveillance systems by providing a more accurate and efficient method for crowd counting, with potential applications in public safety, transportation management, and urban planning.

Keywords: Crowd counting; intelligent video surveillance; deep learning; encoder-decoder architecture; density map estimation; hierarchical feature extraction; convolutional neural networks; public safety monitoring

***For citation*: Dobryshev R. Y., Maksymov M. V. "Accurate crowd counting for intelligent video surveillance systems". *Herald of Advanced Information Technology.* 2024; Vol.7 No.3: 253–261. DOI: https://doi.org/10.15276/hait.07.2024.17**

INTRODUCTION, FORMULATION OF THE PROBLEM

Intelligent video surveillance systems (IVS) have rapidly advanced and become a crucial component in modern security frameworks. In recent years, with the growth of urban areas, increasing foot traffic in public spaces, and rising security threats, the demand for solutions that can autonomously monitor large crowds has significantly increased. Intelligent video surveillance systems are now an integral part of smart cities, access control systems, and public safety enforcement. As urbanization continues to grow, the automation of video stream analysis for monitoring human crowds becomes critical. This opens the door to solving key problems, such as anomaly detection, public event safety management, and optimizing public transportation efficiency.

One of the most important tasks faced by modern intelligent surveillance systems is accurate crowd counting. This task plays a fundamental role in IVS functionality as it not only enables the estimation of crowd density but also helps identify abnormal situations, such as excessive crowding in confined spaces, which could indicate potential threats like evacuation risks. Additionally, accurate people counting facilitates better crowd management in densely populated areas such as stadiums, train stations, airports, and shopping malls. In the context of pandemics and other mass public health threats, the ability to precisely estimate the number of individuals in restricted areas allows for more effective application of distancing measures and other control protocols (Fig. 1).

Despite its relevance and importance, crowd counting is a highly complex task from a technical standpoint. Various factors complicate the process: variations in people's postures and positions, occlusions caused by overlapping individuals, changes in lighting, dynamic crowd movements, and the diversity of human figures. All of these factors present technical challenges for the development of

reliable and accurate counting algorithms. Particular difficulties arise when dealing with dense crowds, where individual figures overlap, making the visual separation of people extremely challenging. Addressing these issues requires sophisticated computer vision algorithms, machine learning, and deep learning techniques that can handle non-standard filming conditions.

Traditional methods, such as feature-based object detection or motion analysis, have long been employed in crowd counting tasks. However, these methods suffer from several limitations. First, they are often ineffective in high-density crowd environments, where the visual features of individuals may be obscured. Second, existing methods are often sensitive to environmental conditions: changes in lighting or the presence of dynamic objects in the background can significantly reduce their accuracy. Additionally, many methods require complex image preprocessing or the use of additional sensors, increasing the complexity and cost of such systems.

Modern approaches based on deep neural networks have dramatically improved crowd counting performance, but even they are not without limitations. One of the major issues with deep learning-based solutions is the need for large datasets for training, which is not always feasible in real-world scenarios.

Moreover, deep learning models are prone to overfitting and can be sensitive to changes in environmental parameters, such as camera angle or crowd density. Another significant drawback is the demand for extensive computational resources, which limits the deployment of these methods on low-power devices.

Thus, despite the impressive progress achieved in the field of crowd counting, existing methods still face numerous challenges. Insufficient accuracy, particularly under complex conditions, high computational costs, and dependence on data quality leave room for further research and development of more efficient and versatile solutions. The advancement of intelligent video surveillance systems demands continuous improvement in crowd counting methods, making this a highly active area of research and technical innovation.

Thus, **the purpose of this study** is to provide a technique for accurate deep-learning based density-aware crowd counting for intelligent video surveillance systems.

1. LITERATURE REVIEW

Crowd counting may be executed using two main methodologies: object detection and density map estimation. The first approach involves a picture as input, yielding a numerical result that denotes the total count of individuals

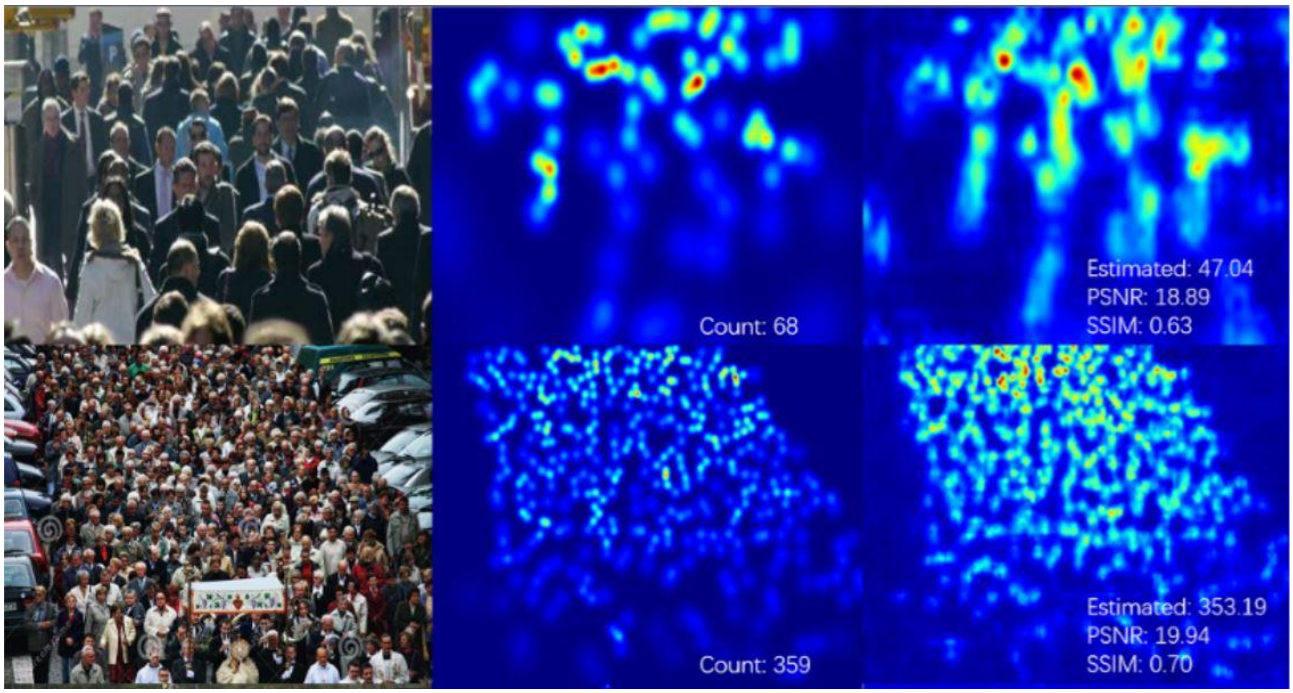


Fig. 1. **Illustration of the crowd counting task**
Source: **compiled by the [1]**

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Onlin)e

inside the frame. In the second technique, a model produces a crowd density map, which is then merged to ascertain the overall headcount.

Conventional techniques for crowd counting mostly depended on detection-based methodologies. These technologies used image processing techniques to identify pre-engineered elements, such as body shapes or components, subsequently using machine learning models. Examples of these models include linear regression, ridge regression, Gaussian processes, support vector machines (SVMs), random forests, gradient boosting, and fundamental neural networks. Nevertheless, the precision of these approaches markedly diminished when addressing photographs of dense crowds owing to many problems, including object occlusion, poor resolution, and complications related to perspective and angles.

To address the shortcomings of detection-based techniques, regression-based methods were developed to estimate the total population inside a whole picture or its portions. In contrast to detection models, these approaches do not seek to identify particular body parts but rather use global picture properties such as texture, foreground contrast, and gradients.

These methods mitigate several issues associated with poor resolution and object occlusion; yet, they exhibit limited efficacy when used on pictures characterized by high crowd density.

Recent studies illustrate the superior efficacy of convolutional neural networks (CNNs) in crowd counting tasks, attributable to their capacity for automated extraction of intricate information. Analogous to other computer vision tasks, including image classification, object recognition, and segmentation, convolutional neural networks (CNNs) have emerged as the preeminent method for crowd counting, markedly surpassing conventional techniques.

In contrast to traditional methods that just forecast total headcount, convolutional neural networks (CNNs) are often used for crowd density estimates. This method entails forecasting a density map of the scene, which encompasses both the overall number of individuals and their spatial distribution inside the picture, therefore significantly augmenting scene analysis skills.

Further studies have extensively embraced the density estimate technique using convolutional neural networks (CNN) as a pivotal strategy for addressing the crowd counting issue. The design of these models has undergone significant improvements to achieve optimal accuracy.

Conventionally, we assess the efficacy of any deep learning model using benchmark datasets, and over time, we have introduced numerous specific datasets for crowd counting tasks.

The datasets have significantly increased the complexity of the issue by including elements such as elevated crowd density, size differences, scene variety, fluctuations in lighting, unequal crowd distribution, severe occlusions, and perspective distortions.

Over time, researchers have created more sophisticated CNN architectures, novel learning techniques, and enhanced assessment criteria to successfully tackle these issues and achieve high accuracy on complicated data.

In recent years, many deep neural network (DNN) models have been introduced for crowd counting, varying in size and design:

1. Single-column models, despite being small, have a lower performance in processing high-density pictures and encounter scaling variation challenges;
2. Multi-column models can manage scale variation in objects, but the number of columns limits their adaptability to diverse item scales. Moreover, multi-column models incur significant computational costs due to the need to train many columns concurrently, hence escalating resource requirements.;
3. Single-column models with multi-scale modules have been created to address scale changes more efficiently in terms of computing. These methodologies are derived from the Inception architecture, with some design alterations;
4. People often use encoder-decoder models when maintaining spatial resolution is crucial, particularly for producing high-quality density maps. These approaches provide multi-tiered supervision, enhancing control at different phases of the network;
5. Nonetheless, both single-column and multi-column models developed from the ground up exhibit constrained accuracy when evaluated on extensive datasets, particularly those with very packed pictures. Pre-trained models, like VGG, Inception, and ResNet, are often used to enhance counting precision. Models using pre-trained backbone neural networks (frontends) experience expedited training. Nonetheless, this results in augmented model size and execution duration, rendering them less appropriate for real-time applications.

2. PROPOSED MODEL

Typically, methodologies using deep neural networks (DNNs) use conventional and dilated convolutions as fundamental components to discern

local patterns and density maps. The majority use identical filters, pooling matrices, and configurations across the entire picture, implicitly assuming uniform congestion levels. This assumption often proves inaccurate in practice.

Due to the dynamic variability in congested environments, it is critical to employ a variety of features and branches in order to effectively react to and collect information at varying degrees of density. Despite the fact, that different proposed approaches demonstrating high efficiency via various strategies, there is significant potential for improvement in developing highly efficient convolutional layer architectures capable of properly addressing crowd scenarios with substantial density variations. Usually, a size factor 3 for kernel of a convolution filter is more efficient than bigger sizes for extracting significant features since it captures more details with lesser complexity, facilitating easier network training. Reduced receptive fields yielded enhanced performance.

Secondly, the use of patch-based and multi-patch processing is time-consuming since identical features must traverse several pathways and patches again. To leverage the advantages of multi-variant techniques, it is advisable to extract proximate characteristics from the network and thereafter direct them to other branches for refinement to identify more intricate features. To use a more complex network for crowd counting, it is essential to implement the previously discussed strategies inside a multi-branch framework to enhance performance.

This paper proposes a new deep encoder-decoder architecture that incorporates hierarchical feature extraction with focus models to give better features for estimating crowds of different sizes and densities.

The overall architecture of proposed technique is illustrated in Fig. 2.

This novel structure is composed of selective pooling as well as 1×1 and 3×3 convolutions, which are employed to enhance the feature matrices in order to effectively manage objects of varying sizes inside a scene using hierarchical feature extraction.

As previously mentioned, we formulate the problem of crowd counting by regressing the density map of individuals in relation to a scene. There are five primary components that make up this framework.

These components are as follows: convolutional network based on VGG16, a hierarchical feature extractor, a branch block, decoder block, and focus block. The total accuracy and efficiency of the model for counting the number of persons in a crowd is correlated with each of these blocks, and there is a connection between the aforementioned blocks.

The foundation of proposed model is based on VGG16, which is often utilized for the extraction of low-level characteristics. When we consider the compromise between time and accuracy, we delete the layers that are located between the last few pooling levels.

Next, we apply a focus block to highlight the most significant aspects. After that, these features are introduced into the hierarchical block, which is a mix of selective pooling and factor 1 and factor 3 convolutions. This block is responsible for creating features for the decoder block.

The next phase involves the use of a global average pooling and a fully connected layer in order to categorize the input scene as either very dense or sparse. After that, we transmit this information to the corresponding decoder using the same structure.

In the decoder, there are four layers of dilated convolution that are 3×3 in size. We place an focus module after each of these layers. Furthermore, to handle the disparities in congestion that occur in sparse and dense locations, we build two variants of the decoder module.

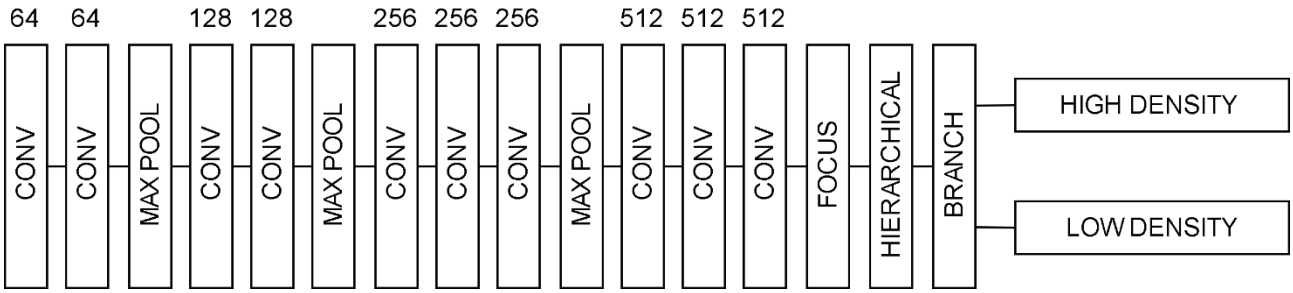


Fig. 2. **The overview of proposed architecture**
Source: **compiled by the authors**

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

These variants are responsible for generating low- and high-density maps inside the input scene, and then assigning these maps to the regression losses that correspond to them.

Using different features from the final layer of the decoder, we construct the resulting density map in the final stage.

Proposed framework completes the end-to-end training of the model by applying a classification loss alongside the same loss for the sparse, dense, and final output density maps.

As a result:

1. The focus block focuses its attention on the major characteristics, specifically areas that are congested;
2. the hierarchical block is able to create more productive features, which are better suited for the crowd counting job with different versions. Adaptive pooling techniques and dilated convolutions of varying sizes combine to accomplish this;
3. with the assistance of the branch block, the appropriate branch of the decoder may be located in accordance with the amount of congestion in the region;
4. we design the mid-branch decoder to handle any changes in congestion within the input picture.

3. EXPERIMENTAL RESULTS

Our approach will be evaluated for efficiency in this section. We conduct these tests on different datasets and compare the results with different popular approaches. Since their release, these methods have already been used to compare different methodologies.

Evaluation metrics. Each computer application necessitates the establishment of assessment metrics to assess the effectiveness of the solutions.

In crowd counting, many measures are used to evaluate model performance by juxtaposing anticipated outcomes with annotated ground facts.

The two predominant metrics in crowd counting are Mean Absolute Error (MAE) and Root Mean Squared Error (MSE), defined as follows:

$$MAE = \frac{1}{M}\sum_{m=1}^{M}\left|C_m^{est} - c_m^{gt}\right|, \quad (1)$$

$$MSE = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(C_m^{est} - C_m^{gt}\right)^2}. \quad (2)$$

where M refers to the quantity of training or testing data; C_m^{gt} indicates the precise count of individuals inside the region of interest of the m-th scenario, and C_m^{est} is the anticipated number of individuals in the crowd.

Data Augmentation. We use data augmentation to reduce the danger of overfitting to the minimal number of training photos. We supplement data with five forms of cropping and resizing. We crop each photograph to 25 % of its source size.

The cropped photos produce four non-overlapping segments from each size of the source photo. Furthermore, the other variant is randomly selected from the source image.

To resize, we simply resize the input picture to the dimensions 768x1024 or 1024x768, depending on the scale of the input data.

If an input image's height exceeds its width, we simply choose 1024x768, and otherwise, we scale it to 768x1024 size.

Results on the Shanghai Dataset. On the Shanghai Tech dataset, it is difficult to provide an accurate estimate of the number of pedestrians because the issue is generated by a variety of circumstances and variations in the amount of congestion.

The KNN technique is used to determine the mean path between each person and its three closest neighbors, and γ is equals to 0.25. This is done in order to set σ for the part A of dataset.

In the case of part B, we used a constant value of 15 for σ.A comparison is made between our approach and the most current state-of-the-art methods that have been published on this dataset (Table 1).

Table 1. **Experimental results for the first part of Shanghai dataset**

Model	MAE	MSE
Proposed	**60.5**	**93.2**
DRSAN	69.4	96.3
CSRNet	68.1	114.9
SFCN	65.0	107.7
TEDnet	63.9	108.8
CAN	62.4	99.8
SPN	61.8	99.7
ACSCP	75.9	102.9
ADCrowdNet	63.4	99.1

Source: **compiled by the authors**

We go through the original published papers of the other techniques and compile the findings of those approaches. In the experiment, it is clear that proposed model has obtained a mean absolute error (MAE) of 60.5 and a mean squared error (MSE) of 93.2. Other top-ranked approaches are not as

advantageous as our model, which demonstrates considerable benefits over these approaches.

As can be shown in Table 2, proposed model has obtained an MAE of 6.9 and an MSE of 11.0 on the second part of dataset.

Table 2. **Experimental results for the second part of Shanghai dataset**

Model	MAE	MSE
Proposed	**6.9**	**11.0**
DRSAN	10.9	17.9
CSRNet	10.7	15.9
SFCN	7.8	13.2
TEDnet	8.3	12.9
CAN	8.0	12.4
SPN	9.5	14.6
ACSCP	17.4	27.7
ADCrowdNet	7.9	13.0

Source: **compiled by the authors**

Both of these findings are superior to other popular crowd counting models. The combination of the hierarchical block and the two-variant decoder seems to be the key to our proposed model's ability to handle both sparse and crowded scenes, as shown by these findings.

Because of these factors, the model that we have presented is able to differentiate between the crowd levels of the source video and evaluate the crowd in accordance with the crowd level for improved estimate.

Results on the Wex dataset. The findings of the MAE metric are shown in Table 3, which is based on five distinct scenarios from the Wex dataset.

Table 3. **Experimental results for the Wex dataset**

Model	Sc1	Sc2	Sc3	Sc4	Sc5	Avg
Proposed	**1.8**	**9.0**	9.7	**7.4**	**2.3**	**6.1**
DRSAN	2.7	12.0	10.4	10.5	3.9	7.9
CSRNet	3.0	11.6	**8.7**	16.5	3.5	8.7
SFCN	2.7	13.6	10.7	12.4	3.5	8.6
TEDnet	2.4	9.9	11.4	13.7	2.7	8.0
CAN	2.9	11.9	9.9	8.0	4.4	7.4
SPN	2.7	13.5	9.0	15.3	3.6	8.9
ACSCP	2.9	14.0	9.7	7.9	2.8	7.4
ADCrowdNet	2.1	14.3	11.7	8.0	3.0	7.8

Source: **compiled by the authors**

The suggested model has produced an average MAE of 6.1, as seen in the table. This represents a significant improvement over the results obtained by CAN, surpassing the state-of-the-art state by a margin of 1.3.

Additionally, the suggested model produces the lowest MAE of four out of all five scenarios, with MAE values equal to 1.8, 9.0, 7.4, and 2.3, respectively. This is the case for all five scenes. Based on established results, the suggested model outperforms state-of-the-art techniques in a variety of situations.

Results on the UCF dataset. For the purpose of creating ground truth density maps on the UCF CC 50 dataset, we choose a configuration that is analogous to the Shanghai Tech-A setting.

Table 4 demonstrates that the suggested model performs much better than the models that are considered to be state-of-the-art when applied to this dataset.

We are able to attain a mean absolute error of 120.1 with a mean squared error of 157, which surpasses the performance of previous benchmark models.

As a result of our trials, we have found that the suggested model is capable of providing an accurate estimation of the total number of individuals across all subgroups.

It is possible to draw the conclusion that the suggested model is capable of functioning well in both sparse and crowded circumstances.

Table 4. **Experimental results for the UCF dataset**

Model	MAE	MSE
Proposed	**120.1**	**157**
DRSAN	158.9	189.8
CSRNet	165.7	197.4
SFCN	173.9	201.9
TEDnet	149.3	174.7
CAN	169.7	192.5
SPN	159.1	186.0
ACSCP	181.0	209.2
ADCrowdNet	156.9	178.8

Source: **compiled by the authors**

CONCLUSIONS

This research presents a unique deep framework for crowd counting. A density-variant decoder has been integrated into the model in order to accommodate the significant density variance that exists within the crowded scenes.

We have also incorporated hierarchical features and focus blocks. In order to give more accurate crowd counting using two-scale density maps, the proposed model has made use of a branch module.

This module is responsible for transferring the hierarchical characteristics directly to the decoder variant that is the most appropriate.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

In order to aggregate these density maps, it makes use of the sigmoid function and generates a gating mask for the purpose of constructing the final density map.

The performance of proposed model in terms of its resilience, accuracy, and generalization has been proved by extensive tests conducted on a variety of benchmark datasets. In comparison to the approaches that are considered to be state-of-the-art, proposed model is able to obtain superior performance on virtually all of the main crowd counting datasets.

Throughout the course of this research, we have studied a variety of techniques for crowd counting and density estimation in order to come up with novel solutions that have the potential to beat the findings of the present state of the art by significant margins.

On the basis of our experiences, a number of aspects that need more investigation have been recognized and summarized as follows:

1) A suitable method for counting the number of people in a crowd ought to have a low level of complexity. In light of this rationale, we believe that future studies need to concentrate more on solutions that are based on a single column arrangement;

2) it may be a good idea to employ a form of zooming approach in the center of models if a congested location is recognized. This will allow us to focus on the high density zone and extract more helpful features from that area for an accurate density estimate. This will allows us to address the intra-dense area that exists inside a scene;

3) patch-based processing of average characteristics maps in the CNN-based model is an additional enhancement that may be made to the crowd counting framework. The results of our preliminary inquiry have shown that it has the potential to result in further enhancements to the precision of approaches for counting crowds.

REFERENCES

1. Wang, B., et al. “SD-Crowd: A shallow-deep CNN architecture for crowd counting”. *arXiv preprint.* 2019. DOI: https://doi.org/10.48550/arXiv.1904.00385.

2. Sindagi, V. A. & Patel, V. M. “Multi-level bottom-top and top-bottom feature fusion for crowd counting”. *Proceedings of the IEEE International Conference on Computer Vision*. 2019. p. 1002–1012. DOI: https://doi.org/10.1109/ICCV.2019.00109.

3. Zhang, Cong, et al. “Cross-scene crowd counting via deep convolution neural networks”. *Neurocomputing*, 2018; Vol. 314: 298–310. DOI: https://doi.org/10.1109/CVPR.2015.7298684.

4. Lian, D., et al. “Density map regression guided detection network for RGB-D crowd counting and localization”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019. p. 1821–1830. DOI: https://doi.org/10.1109/CVPR.2019.00192.

5. Liu, C., Weng, X. & Mu, Y. “Recurrent attentive zooming for joint crowd counting and precise localization”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019. p. 1217–1226. DOI: https://doi.org/10.1109/CVPR.2019.00131.

6. Xu, Chenfeng, et al. “Autoscale: Learning to scale for crowd counting”. *Proceedings of the IEEE International Conference on Computer Vision.* 2019. DOI: https://doi.org/10.48550/arXiv.1912.09632.

7. Ma, Z., Wei, X., Hong, X. & Gong, Y. “Bayesian loss for crowd count estimation with point supervision”. *Proceedings of the IEEE International Conference on Computer Vision.* 2019. p. 6141–6150DOI: https://doi.org/10.1109/ICCV.2019.00624.

8. Ilyas, Naveed & Zaheer, Ahmad. “An effective modular approach for crowd counting in an image using convolutional neural networks”. *Scientific reports*, 2022; Vol. 5795: 154–162. DOI: https://doi.org/10.1038/s41598-022-09685-w.

9. Liu, Shunchang, et al. “Harnessing Perceptual Adversarial”. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2022: 2055–2069. DOI: https://doi.org/10.1145/3548606.3560566.

10. Liu, L., et al. “Crowd counting using deep recurrent spatial-aware network”. *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 2018. p. 849–855. DOI: https://doi.org/10.24963/ijcai.2018/118.

11. Idrees, Haroon, et al. “Composition loss for counting, density map estimation and localization in dense crowds”. *Lecture Notes in Computer Science.* 2018. DOI: https://doi.org/10.1007/978-3-030-01258-8_33.

12. Shi, Z., et al. “Crowd counting with deep negative correlation learning”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018. p. 5382–5390. DOI: https://doi.org/10.1109/CVPR.2018.00564.

13. Hua, Cheng, et al. “Crowd Counting with Dilated Inception Convolution”. *Proceedings of the 2021 7th International Conference on Computing and Artificial Intelligence*. 2021. p. 208–215. DOI: https://doi.org/10.1145/3467707.3467738.

14. Gao, J., Wang, Q. & Li, X. “PCC Net: Perspective crowd counting via spatial convolutional network”. *IEEE Transactions on Circuits and Systems for Video Technology*. 2019; 29: 3486–3498. DOI: https://doi.org/10.1109/TCSVT.2019.2919139.

15. Sindagi, V. A. & Patel, V. M. “CNN-Based cascaded multi-task learning of high-level prior and density estimation for crowd counting”. *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS).* 2017. DOI: https://doi.org/10.1109/AVSS.2017.8078491.

16. Xiong, F., et al. “From open set to closed set: Counting objects by spatial divide-and-conquer”. *Proceedings of the IEEE International Conference on Computer Vision*. 2019. DOI: https://doi.org/10.48550/arXiv.1908.06473.

17. Wan, J., et al. “Residual regression with semantic prior for crowd counting”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019. p. 4031–4040. DOI: https://doi.org/10.1109/CVPR.2019.00416.

18. Dong, Jingwei, et al. “Crowd Counting by Multi-Scale Dilated Convolution Networks”. *Electronics.* 2023; Vol. 12: 12. DOI: https://doi.org/10.3390/electronics12122624.

19. Weizhe, Liu, et al. “Context-aware crowd counting”. *CVPR 2019*. DOI: https://doi.org/10.48550/arXiv.1811.10452.

20. Sam, Deepak Babu, et al. “Locate, Size and Count: Accurately Resolving People in Dense Crowds via Detection”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. DOI: https://doi.org/10.48550/arXiv.1906.07538.

21. van Noord, Nanne, et al. “Learning scale-variant and scale-invariant features for deep image classification”. *Pattern Recognition*. 2017. DOI: https://doi.org/10.1016/j.patcog.2016.06.005.

22. Wang, Qi, et al. “NWPU-Crowd: A large-scale benchmark for crowd counting and localization”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. DOI: https://doi.org/10.1109/TPAMI.2020.3013269.

23. Liu, J., et al. “DecideNet: Counting varying density crowds through attention guided detection and density estimation”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018. p. 5197–5206. DOI: https://doi.org/10.1109/CVPR.2018.00545.

24. Sam, D. B., Surya, S. & Babu, R. V. “Switching convolutional neural network for crowd counting”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017. p. 4031–4039. DOI: https://doi.org/10.1109/CVPR.2017.429.

25. Jiang, Xiaohong, et al. “Crowd counting and density estimation by trellis encoder-decoder networks”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019. p. 6126–6135. DOI: https://doi.org/10.1109/CVPR.2019.00629.

26. Li, Y., Zhang, X. & Chen, D. “CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2018. p. 1091–1100. DOI: https://doi.org/10.1109/CVPR.2018.00120.

27. Zhang, Yingying, et al. “Single-Image crowd counting via multi-column convolutional neural network”. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2016. p. 589–597DOI: https://doi.org/10.1109/CVPR.2016.70.

28. Dong, Li, et al. “Crowd counting by using multi-level density-based spatial information: A Multi-scale CNN framework”. *Information Sciences.* 2020; Vol. 528: 79–91. DOI: https://doi.org/10.1016/j.ins.2020.04.001.

29. Zhao, Y., et al. “Leveraging heterogeneous auxiliary tasks to assist crowd counting”. *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2019. p. 12728–12737. DOI: https://doi.org/10.1109/CVPR.2019.01302.

Conflicts of Interest: the authors declare no conflict of interest

Received 08.08.2024
Received after revision 16.09.2024
Accepted 21.09.2024

DOI: https://doi.org/10.15276/hait.07.2024.17
УДК 004.932.72

Точний підрахунок натовпу для інтелектуальних систем відеоспостереження

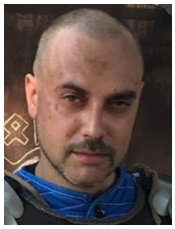
Добришев Руслан Євгенович[1)]
ORCID: https://orcid.org/0009-0007-8639-3157, rdobrishev@gmail.com
Максимов Максим Віталійович[1)]
ORCID: https://orcid.org/0000-0002-3292-3112, prof.maksimov@gmail.com. Scopus Author ID: 7005088554
[1)] Національний університет «Одеська Політехніка», пр. Шевченка, 1, м. Одеса, Україна, 65044

АНОТАЦІЯ

У статті представлено новий підхід на основі глибокого навчання для підрахунку натовпу в інтелектуальних системах відеоспостереження, що вирішує зростаючу потребу в точному моніторингу громадських місць у міських середовищах. Попит на точну оцінку кількості людей виникає через проблеми, пов'язані з безпекою, громадським порядком і ефективністю в міських зонах, особливо під час великих публічних заходів. Існуючі методи підрахунку натовпу, включаючи виявлення об'єктів на основі ознак і методи регресії, мають обмеження в умовах високої щільності через перекриття об'єктів, варіації освітлення та різноманітність людських фігур. Щоб подолати ці виклики, автори пропонують нову архітектуру енкодера-декодера на основі VGG16, яка включає ієрархічне вилучення ознак із використанням просторової та канальної уваги. Ця архітектура покращує здатність моделі керувати варіаціями щільності натовпу, використовуючи адаптивне підсумовування та дилатовані згортки для вилучення значущих ознак із щільних натовпів. Декодер моделі додатково вдосконалюється для обробки розріджених і густих сцен через окремі карти щільності, що підвищує її адаптивність і точність. Оцінка запропонованої моделі на еталонних наборах даних, включаючи Shanghai Tech і UCF CC 50, демонструє кращі результати порівняно з сучасними методами, з помітними покращеннями за метриками середньої абсолютної помилки та середньоквадратичної помилки. У статті підкреслюється важливість врахування змін у середовищі та різниці в масштабах у густонаселених середовищах, і показано, що запропонована модель ефективна як в умовах розрідженого, так і щільного натовпу. Це дослідження сприяє розвитку інтелектуальних систем відеоспостереження, пропонуючи більш точний і ефективний метод підрахунку натовпу з можливими застосуваннями у сфері громадської безпеки, управління транспортом і міського планування.

Ключові слова: підрахунок натовпу; інтелектуальні системи відеоспостереження; глибоке навчання; архітектура енкодера-декодера; оцінка карти щільності; ієрархічне вилучення ознак; згорткові нейронні мережі; моніторинг громадської безпеки

ABOUT THE AUTHORS



Ruslan Y. Dobryshev - PhD Student, Artificial Intelligence and Data Analysis Department, Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044
ORCID: https://orcid.org/0009-0007-8639-3157, rdobrishev@gmail.com
***Research field*:** Deep Learning, Crowd Analysis, Video processing, Motion tracking, Intelligent Surveillance, Computer Vision

Добришев Руслан Євгенович - аспірант кафедри Штучного інтелекту та аналізу даних. Національний університет «Одеська Політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

Maksym V. Maksymov - - Doctor of Engineering Sciences, Professor, Head of Software and Computer-oriented Technologies Department. Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: http://orcid.org/0000-0002-3292-3112. prof.maksimov@gmail.com. Scopus Author ID: 7005088554
***Research field*:** Development of methods and models for solving control problems in concentrated and distributed formulation

Максимов Максим Віталійович - д-р техн. наук, проф., зав. каф. Програмних і комп'ютерно-орієнтованих технологій. Національний університет «Одеська Політехніка», пр. Шевченка, 1. Одеса, 65044, Україна.