Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **7**, Number **2**, May **2022**

Multi-object Crowd Real-time Tracking in Dynamic Environment Based on Neural Network

Fu-Quan Zhang

College of Computer and Control Engineering,
Minjiang University,
Fuzhou, China
Digital Media Art, Key Laboratory of Sichuan Province,
Sichuan Conservatory of Music,
Chengdu, China
8528750@qq.com

Lin-Juan Ma*

School of Computer Science and Technology,
Beijing Institute of Technology,
Beijing, China
*Corresponding Author: malinjuan@bit.edu.cn

Received January 2022 Revised March 2022

Abstract. *In this paper, it proposes a new lightweight neural network detection model and tracking algorithm that enables us to perform multi-object crowd tracking. The technique applied in our study is known as object detection and tracking which is based on neural network and can be calculated and reproduced intelligently according to the movement and behavior of the tracked crowd. Also, a number of experiments have been carried out to test the validity of the accuracy in the algorithm. And it is indicated that the results of the experiments have better performance and can contribute to multi-object crowd real-time tracking with high accuracy, which is similar to the state-of-the-art. All our preliminary results throw light on the nature of detection and tracking and result in a solution of multi-object crowd tracking, and it proves to be fruitful and encouraging.*
Keywords: Neural Network, Multi-object, Crowd tracking, Intelligent, Lightweight

1. **Introduction.** Recently, there have been studies highlighting multi-object crowd tracking based on neural networks, which has been a hot topic in computer vision. Due to the movement and behavior of the tracked crowd in the dynamic environment can be calculated and reproduced intelligently in specific application scenarios, it has received much attention from researchers and engineers and can be applied in many research fields, such as emergency evacuation, city planning, safety simulation, military exercise, and disaster scene simulation and command, etc.

Previously, most of the multi-object crowd tracking studies manly attaches much importance on traditional tracking algorithms. There are several traditional typical object tracking algorithms, including optical flow, Kalman Filter, mean-shift and so on. Optical flow [1] was first proposed by Gibson et al. in 1950, which is the movement of a target caused by the movement of the target scene or camera between two consecutive frames of images. Kalman et al. [2] in 1973 considered that Kalman filter is to predict the coordinates and velocities of the object position from a finite, noise-containing sequence of observations of its position. And mean-shift was presented by Fukunaga et al. [3] in

1975, where the basic idea is to use gradient climbing of probability density to find local optimum. Furthermore, these traditional typical tracking algorithms have been applied in person tracking and crowd tracking. Tangelder et al. [4] in 2014, proposed Fusion of optical flow based motion pattern analysis and silhouette classification for person tracking and detection. Then in the next year, Sahi et al. [5] proposed person tracking using mean shift with gray level grouping. And Sagun et al. [6] in 2017 used Kalman filter to track the trajectories of blobs for people counting and tracking from crowd video. In 2020, Wang et al.[7] proposed a new algorithm: linear-prefer coevolutionary algorithm to solve the problem in a multiobjective neural architecture search process.

While traditional tracking algorithms conducted on multi-object crowd tracking has made great progress, there are still some disadvantages in these methods. In real practice, they are lack of required template updating. Besides, the object tracking will be lost when moving object is occluded for a long time. What's more, spatial information was lacked of in histogram feature in the description of object color feature.

In order to solve the aboved problems, the tracking method based on neural networks is currently playing an important role in multi-object crowd tracking [8-16]. In literature, Chaudhari et al. [8] in 2019, presented to detect the crowd from the street, evaluate the density level and count individual from the crowd using Convolutional Neural Network. Meanwhile,Yang et al. [9] set forth to use pedestrian tracking algorithm for dense crowd based on Deep Learning. At the same time, Deng et al. [10] used neural network to tracking students' identity and motion in classroom. And in 2020, Franchi et al. [11] proposed to track pedestrians using a model composed of a Particle Filter and three Deep Convolutional Neural Networks. Also, Zhang et al. [12] proposed quantum genetic algorithm - learning vector quantization (QGA-LVQ) neural network to forecast the changes of traffic flow and can contribute to the pedestrians tracking. Then in 2021, Bornia et al. [13] proposed to use TensorFlow and Deep Learning to track people's motion and animated entities in the video sequences. In the meantime, Yang et al. [14] set forth the detection-tracking-correction strategy based on improved SSD and Cascade R-CNN to carry out real-time high-precision pedestrian tracking. Also, Wang et al. [15] proposed a transfer reinforcement learning-based object detection method to recognize the specific target including the crowd in the road environment. Then in 2022, Elsaid et al. [16] presented a hybrid attention-based Siamese network to enhance visual object tracking. Great improvements has been received in multi-object crowd tracking based on neural networks, however, in fact, there are still some challenges as follows:

1) They are subject to the large amount of calculation and long time consuming, which is not applicable in the case of real-time performance.

2) Most algorithms models are too complex to deploy and use, which can influence the accuracy of the tracking.

3) Little effort has been paid to the category and the location of the multi-object crowd in dynamic environment.

Therefore, to fill the gaps, the main goal of this study is to perform multi-object crowd intelligent tracking in dynamic environment using the novel lightweight neural network LK-Yolo. To improve the real-time tracking performance, the KCF tracker instead of traditional tracker is added in the network, which is fast, effective and has strong robustness. And the novel lightweight neural network is presented in our study to overcome the complexity of the deployment and utilization and realize the trade-off between the speed and accuracy. Finally, the problem of the category and the location of the multi-object crowd in dynamic environment can be resolved by combing neural network with the tracker. The main contributions of our work are as follows:

The novel lightweight neural network LK-Yolo is presented to improve the performance of the multi-object crowd intelligent tracking.

The KCF tracker is merged with neural network to realize the category, location and tracking.

In our work, there are 4 sections as follows: in section 2, it introduces the materials and methods of the proposed neural network. And in section 3, it shows and states the results of the experiments of our work. Finally, it summarizes the paper in section 4

2. **Materials and Methods.** The architecture of the proposed LK-Yolo neural network can be seen in Figure 1, which can perform multi-task crowd real-time tracking. There are mainly three parts of the LK-Yolo neural network, including backbone part, head part and prediction part. When the image or frame input, the information can be acquired by the model. The feature can be extracted in the presented backbone network Resnet50-vd and high-level semantic feature map from various scale images should be constructed through Path Aggregation Network (PAN). To improve the performance of the LK-Yolo, SE module is added near the tail of the model. Also KCF tracker is fused in the model to complete the task of multi-task crowd tracking in real-time.

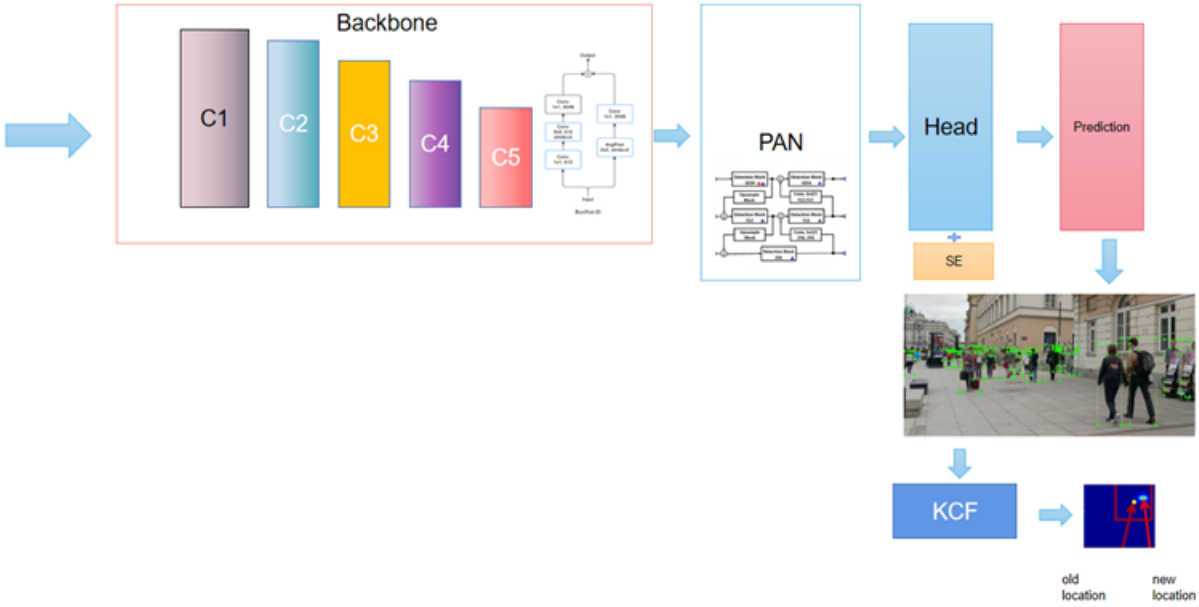

FIGURE 1. The architecture of LK-Yolo neural network

2.1. **Network Architecture.** The backbone network adopted in our work is Resnet50-vd, which is the variant of the Resnet50. It is known that the basic idea of ResNet is to add the directly connected channel, which is similar to Highway Network, allowing the original input information to be transmitted directly to the following layers. And two residual modules will be used in Resnet network structure, one is two 3×3 convolutional layers connected together as a residual module, the other is three convolutional layers connected together as a residual module, including 1 ×1 convolutional layer, 3 ×3 convolutional layer and 1 ×1 convolutional layer. Thus it can be seen that Resnet50 contains 49 convolution layers and a full connection layer. As shown in Table 1, it can be divided into seven parts in Resnet50 network structure. In the first part, it does not contain residual block and mainly focuses the convolution, regularization, activation function, maximum pooling

calculation on the input. And in the second, third, fourth and fifth parts of the structure all contain residual blocks. Through the convolution calculation of the first five parts, the pooling layer will transform it into a feature vector, and finally the classifier will calculate this feature vector and output the category probability. While ResNet50-vd regarded as

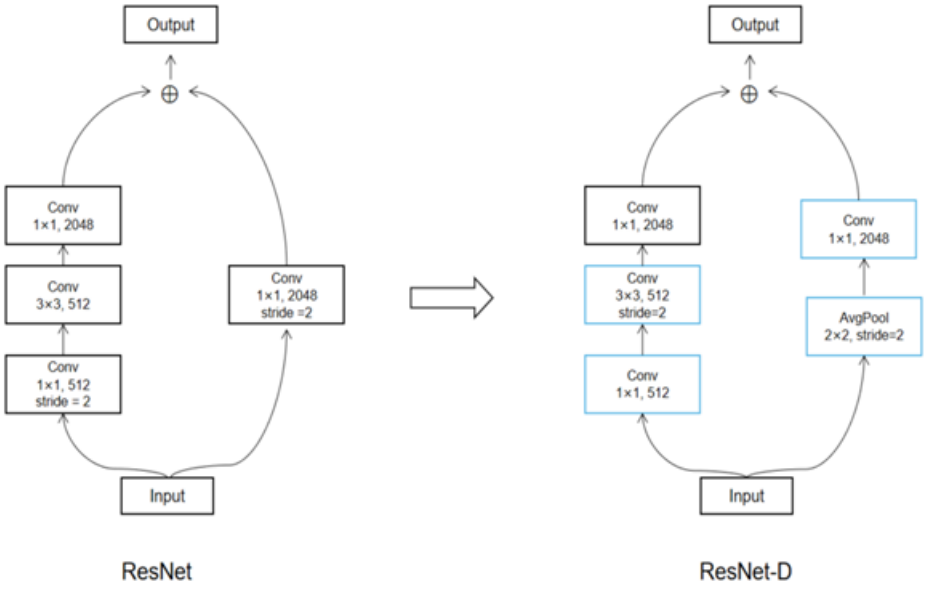

FIGURE 2. The comparison of ResNet and ResNet-D

our backbone network, which is the ResnetNet-D Network with 50 convolutional layers, the unit ResNet-D is shown in Figure 2. Compared to Resnet50, the conv7x7 was improved to three conv3x3 in the head part to mitigate the lost information in ResNet50-vd. And in the downsample part of each stage, it has been modified from $(1 \times 1\, s2) \rightarrow (3 \times 3) \rightarrow (1 \times 1)$ to $(1 \times 1) \rightarrow (3 \times 3\, s2) \rightarrow (1 \times 1)$, also in the shortcut part, from $(1 \times 1\, s2)$ to $avgpool(2) + (1 \times 1)$, to alleviate the information loss. Finally, the accuracy can be improved without increasing the amount of calculation in our network.

2.2. **Path Aggregation Network.** The Path Aggregation Network (PAN) [14] is utilized in our proposed neural network, as shown in Figure 3. Compared to FPN, it has a novel bottom-up enhancement path to shorten the information path and improve the feature pyramid by using precise location signals stored in low-level features. Also, adaptive feature pooling is presented to restore corrupted information paths between each candidate region and all feature levels, aggregate each candidate region at each feature level to avoid arbitrary assigned results.

2.3. **KCF Tracker.** . In our work, KCF tracker is added in our neural network. It is considered that KCF is a discriminant tracking method, which generally trains an object detector in the tracking process, uses the object detector to detect whether the predicted position in the next frame is the object, and then updates the training set with the new detection results to update the target detector. And in the training of the object detector, the object area is generally selected as the positive sample while the surrounding area of the object as the negative sample. The main contributions of the KCF is the positive and negative samples are collected by the cyclic matrix of the object area, and the object detector is trained by ridge regression. What's more, the operation of a cyclic matrix is

TABLE 1. The structure of Resnet50 network

Layer name	Output size	50-layer	
Conv1	112×112	7×7 , 64 , stride 2	
Conv2_x		3×3 max pool, stride 2	
	56×56	1×1, 64 3×3, 64 1×1, 256	×3
Conv3_x	28×28	1×1, 128 3×3, 128 1×1, 512	×4
Conv4_x	4×14	1×1, 256 3×3, 256 1×1, 1024	×6
Conv5_x	7×7	1×1, 512 3×3, 512 1×1, 2048	×3
	1×1	average pool, 1000-d fc, softmax	

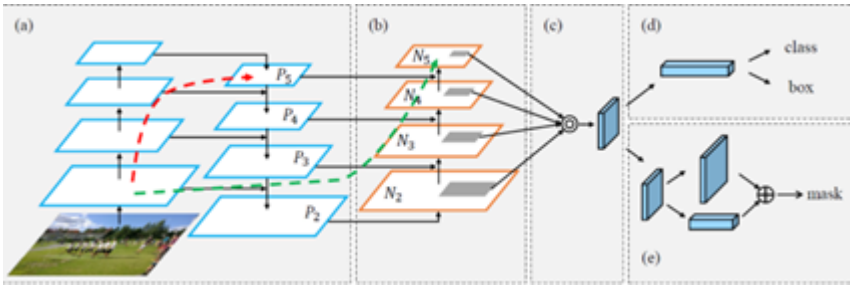

FIGURE 3. The illustration of PAN [14]

transformed into the Hadamad product of vectors by the diagonalizable property of the cyclic matrix in Fourier space, which greatly reduce the amount of computation, improve the speed of calculation, so that the algorithm can meet the real-time requirements. And the main steps of the KCF algorithms can be shown in Figure 4. Firstly, in the It frame, sample near the current position pt and train a regressor which can calculate the response of a small window sample. Secondly, in the It+1 frame, samples are taken near the position pt of the previous frame, and the response of each sample is judged by the previous regressor. Finally, the sample with the strongest response is taken as the frame position pt+1.

2.4. **Activation Function.** In our work, Mish activation function has been chosen as the activation function in our presented neural network, which is shown in Figure 5. It is the smooth non-monotone activation functions which can be defined as Formula 1 and Formula 2. There are several advantages of Mish activation function. Firstly, it is the nonmonotone function to keep negative values small and thus stabilize the network gradient flow. Secondly, due to smooth function, good generalization ability and effective optimization ability contained, which can improve the quality of results. Finally, good performance has been obtained in it for deep neural networks compared with ReLU.

$$f(x) = x.\tanh\left(\zeta(x)\right) \tag{1}$$

$$\zeta(x) = \ln\left(1 + e^{x}\right) \tag{2}$$

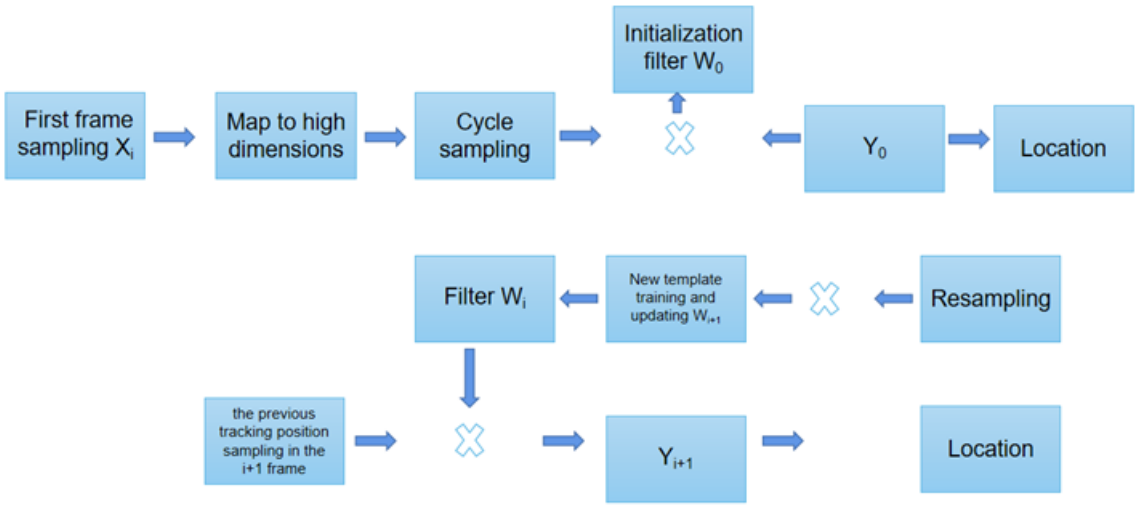

FIGURE 4. Main steps of the KCF algorithms

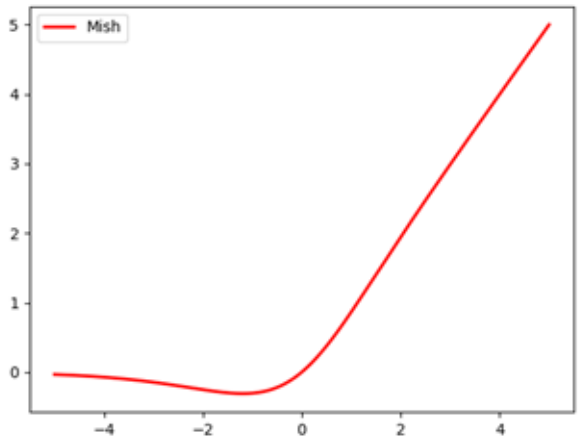

FIGURE 5. Mish activation function

2.5. **Dataset.** As shown in Figure 6, the dataset trained in our network is CrowdHuman [17] dataset, which is large, annotated and highly diverse can be used for person detection in a crowd. The dataset has 15,000 images for training, 4,370 images for validation, and 5,000 images for testing, for a total of 470K human instances from the training and validation subset. And each image in the dataset contained 23 people, with a variety of occlusion and ach human instance is annotated with a head bounding box, a human-visible area bounding box, and a full-body bounding box.

2.6. **Loss Function.** The IOU is regarded as the positioning confidence, which can be shown in Formula 3. As the most commonly used indicator in object detection, it is not only used to determine positive samples and negative samples in the anchor-based method, but also used to evaluate the distance between the predict box and the ground-truth. In our work, the objectness can be calculated as shown in Formula 4, where p is the predicted value of IoU between the detection box and the ground truth, α is the balance parameters.

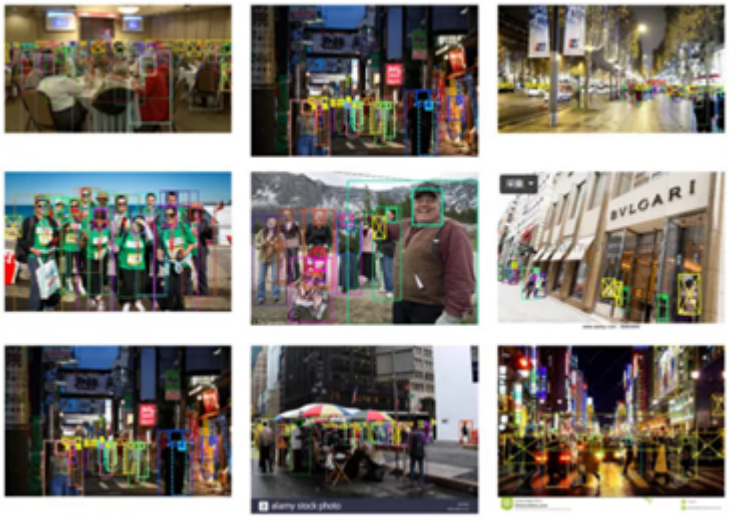

FIGURE 6. CrowdHuman dataset

And the p can be calculated and learned through the loss function for positive sample as shown in Formula 5, namely, binary cross entropy, where t represents the IoU between the detection box and the ground truth, $\sigma(.)$ represents sigmoid activation function.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

$$s = objectness^{1-\alpha} * p^{\alpha} \tag{4}$$

$$loss = -t * \log(\sigma(p)) - (1-t) * \log(1-\sigma(p)) \tag{5}$$

2.7. **Attention Mechanism.** To improve the performance of our neural network LK-Yolo, the SE module is adopted. And it is found that the model can be more accurate when the SE module is more near the tail of the network. The innovation of it lies in focusing on the relationship between channels, thus the model can automatically learn the importance of different channel feature. Firstly, the Squeeze operation is used for the convolutional feature map to get the channel level global feature. Secondly, the Excitation would be operated on the global feature to learn the relationship between channels, also the weights can be obtained. Finally, the final feature should be acquired when the weights multiplied by the original feature map.

3. **Results and Discussions.** Table 2 shows details of accuracy of different models on the CrowdHuman dataset. It is suggested that the proposed LK-Yolo neural network has better performance than other common neural network methods, although the accuracy is a little lower than Yolov4. What's more, it has made great progress than the traditional simulation method, and the parameters of traditional crowd tracking simulation is shown in Figure 7. When put into use in practice, our proposed model is more suitable for multi-object crowd real-time tracking due to its few parameters. In a word, the present results are consistent with the original assumptions.

4. **Conclusions.** It is concluded that the results of our proposed LK-Yolo neural network for multi-object crowd real-time tracking is fruitful and promising, which is highly corresponds to much of the work carried out on the experiments and theory. Despite great advantages mentioned above, there are still some limitations that is clear in our study. It is still a challenge to realize the trade-off between the speed and accuracy when

TABLE 2. The accuracy of different models

Method	Parameters	Accuracy
Yolov3	61.9M	89%
Yolov4	64.46M	93%
Ours	26M	92.1%

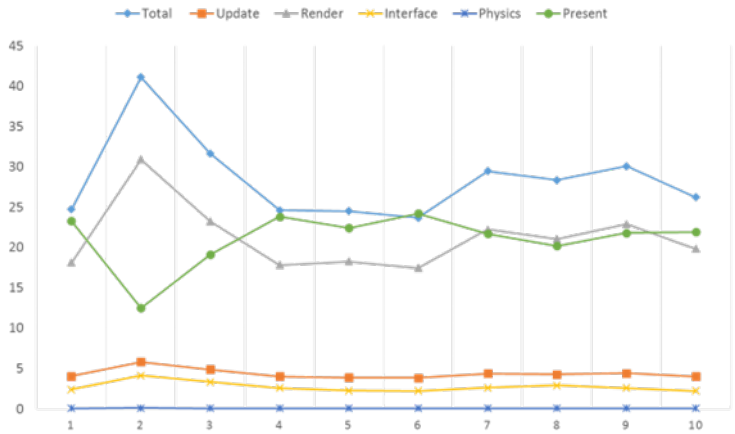

FIGURE 7. The parameters of traditional crowd tracking simulation

performing multi-object crowd real-time tracking, and the crowd dataset needs to be updated and improved for the sound model obtained. To sum up, the problem remains to be solved in the further work.

Acknowledgment. This work is partially supported by Digital Media Art, Key Laboratory of Sichuan Province, Sichuan Conservatory of Music, Project No. :21DMAKL01; Supported by the first batch of industry-university cooperation collaborative education project funded by the Ministry of Education of the People's Republic of China, 2021, Project No. :202101071001; Supported by Minjiang College 2021 school-level scientific research project funding, Project No. :MYK21011.The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] J.J.Gibson, The perception of the visual world, *The American Journal of Psychology*, vol.64, no.3, pp. 440-444, 1951.

[2] R.E. Kalman, New results in linear filtering and prediction theory, *Journal of Basic Engineering*, vol.83, no.83, pp. 109, 1961.

[3] K. Fukunaga, L. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition., *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 32-40, 1975.

[4] H. Tangelder, E. Lebert, G.J. Burghouts, K.V. Zon, M.J.D. Uyl, Fusion of optical flow based motion pattern analysis and silhouette classification for person tracking and detection, *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 9253, 92530L, 2014.

[5] A. Sahi, K.Talele, Person tracking using mean shift with gray level grouping, 2015 International Conference on Communication, *Information & Computing Technology (ICCICT)*, 14933470, pp. 1-5, 2015.

[6] M. Sagun , B.Bolat . A novel approach for people counting and tracking from crowd video, *IEEE International Conference on Innovations in Intelligent Systems & Applications*, pp. 277-281, 2017.

[7] E.K. Wang, S.P. Xu, C.M. Chen, N. Kumar, Neural-Architecture-Search-Based multiobjective cognitive automation system, *IEEE Systems Journal*, vol. 15, no. 2, pp. 2918-2925, 2020.

[8] M.D. Chaudhari, A. Ghotkar, Human face detection in crowd and density analysis using neural network approach, *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI)*, pp. 1208-1215, 2019.

[9] G. Yang, Z.H. Chen, Pedestrian tracking algorithm for dense crowd based on deep learning, *2019 6th International Conference on Systems and Informatics (ICSAI)*, pp. 568-572, 2019.

[10] W.Y. Deng, Y. Wang, J. Men, Deep tracking students identity and motion in classroom, *2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 149-152, 2019.

[11] G. Franchi, E. Aldea, S. Dubuisson, I. Bloch, Tracking hundreds of people in densely crowded scenes with particle filtering supervising deep convolutional neural networks, *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 2071-2075, 2020.

[12] F.Q. Zhang, T.Y. Wu, Y.O. Wang, R. Xiong, G.Y. Ding, P. Mei, L.Y. Liu, Application of quantum genetic optimization of LVQ neural network in smart city traffic network prediction, *IEEE Access*, vol. 8, pp. 104555-104564, 2020.

[13] J.Bornia, A.Frihida, O.Debauche, S.A. Mahmoudi, P. Manneback, Deep learning and tensorflow for tracking people's movements in a video, *5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications*, pp. 1-6, 2021.

[14] S. Yang, Z. Chen, X. Ma, X. Zong, Z. Feng, Real-time high-precision pedestrian tracking: a detection–tracking–correction strategy based on improved SSD and Cascade R-CNN, *Journal of Real-Time Image Processing*, vol. 19, pp. 287-302, 2021.

[15] K. Wang, C.M. Chen, M.S. Hossain, G. Muhammad, S. Kumar, S. Kumari, Transfer reinforcement learning-based road object detection in next generation IoT domain, *Computer Network*s, vol. 193, 108078, 2021.

[16] A.O. Elsaid, M.M. Fouad, T.E. Ghoniemy, Att-SiamMask: Attention-based network for enhanced visual object tracking, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 13, pp. 2073-4212, 2022.

[17] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X.Y. Zhang, J. Sun, CrowdHuman: A benchmark for detecting human in a crowd, arXiv.1805.00123, 2018.