Scholar Publishing Group
International Journal of Neural Network
https://doi.org/10.38007/NN.2022.030202
ISSN 2790-2692 Vol. 3, Issue 2: 9-18

Gymnastic Movement Recognition Based on Depth Convolution Neural Network

Lu Xi*

Philippine Christian University, Philippine

xixiaolu2022@163.com

**corresponding author*

Keywords: Deep Learning, Convolutional Neural Networks, Gymnastics Exercise, Action Recognition

Abstract: With the development of deep learning technology and mobile Internet, more and more image-based artificial intelligence applications appear in people's lives. In the image information, the understanding of the characters in the picture has always been the focus of research and application, and it is also the basis of human-computer interaction. The human body key point detection technology can detect the joint point position of the target person in the image, so as to provide basic information for subsequent human-computer interaction applications and functions. The main purpose of this paper is to conduct research on gymnastics AR based on deep convolutional neural networks (CNN). This paper first expounds the mechanism contained in each component of the CNN. Analyzing the characteristics of CNN, compared with the multi-layer BP neural network, the transmission between the CNN neurons is combined with the local receptive field through weight sharing, which reduces the weight parameters while maintaining the network depth. The gradient disappearance problem can be avoided in the training process, and its network structure has good generalization ability. Through performance comparison experiments, it is found that whether using single-stream RGB data, optical flow data, or dual-stream fusion results, the recognition accuracy of deep neural networks is better than that of conventional networks. Illustrating the effectiveness of deep CNNs in gymnastics action recognition (AR).

1. Introduction

The difficulty of human body key point detection is that the changes in the posture of the characters in the image, the occlusion of the body shape of the characters, and the different sizes of people in the picture make the detection of the key points of the human body in the image

complicated and difficult. With the deepening of research, the current research work has been able to achieve better results in detection accuracy. However, with the development of the mobile Internet, it has become a trend to deploy deep learning models on mobile devices. The human key point detection model needs to consume a lot of computing and storage resources, so it is difficult to deploy on mobile devices with limited computing power and storage [1] -2].

In related research, Dessouky et al. proposed a cancelable speaker recognition scheme based on spectral tile selection [3]. Simulation results show that the proposed method is practical and meets the required criteria for reproducibility, safety, and performance. The proposed cancelable speaker recognition scheme achieves 98.75% accuracy using a CNN consisting of three layers. Martin et al. proposed a new dual spatiotemporal CNN (TSTCNN) [4]. When applied to table tennis, 20 table tennis strokes can be detected and identified. The proposed Twin architecture is a two-stream network, both consisting of 3 spatiotemporal convolutional layers followed by a fully connected layer where the data is fused, achieving 91.4% accuracy.

In this paper, based on the deep CNN, the research on gymnastics AR is carried out. This paper first analyzes the basic structure of CNN, and then expounds the characteristics of CNN; proposes a framework of human pose estimation algorithm based on depthwise separable convolution, and analyzes the evaluation indicators; this paper analyzes the structure of attention mechanism Expand the design, the structure of the attention mechanism constructed under the condition that the prominent spatiotemporal features are satisfied. Finally, the corresponding experiments were carried out.

2. Design Research

2.1. Basic Structure of CNN

CNN is one of the important structures of DNN, which imitates the visual perception mechanism of biology and can handle the task of supervised learning or unsupervised learning [5-6]. Each hidden layer includes convolution operations, pooling, and activation processing, which are processed in many different ways to enhance the ability to express things. More importantly, the CNN can process large-scale high-definition resolution images. [7-8]. Therefore, the CNN reduces the amount of computation by means of sparse connections, as shown in Figure 1, which represents the difference between sparse connections and full connections.

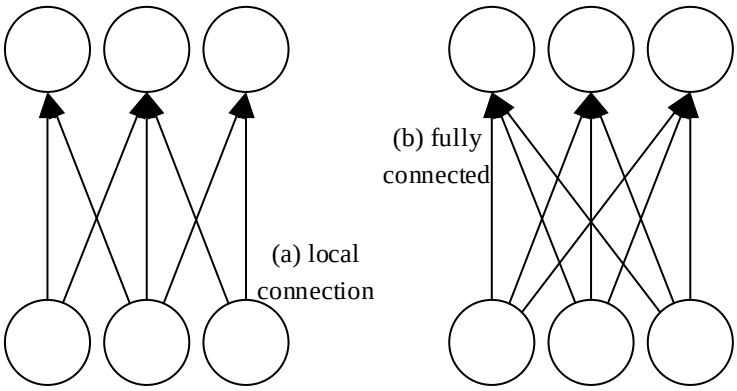

Figure 1. Schematic diagram of full connection and partial connection

2.2. Features of CNNs

CNN is a branch of deep learning, and its structure includes convolutional layers, pooling layers

and fully connected layers [9-10].

When the BP neural network is used to train and recognize images, the recognition accuracy will decrease due to the increase of network parameters, and it will also lead to the appearance of overfitting. Compared with the multi-layer BP neural network, the transmission between the CNN neurons is combined with the local receptive field through weight sharing, which reduces the weight parameters while maintaining the network depth, and can avoid gradient disappearance during the training process. The emergence of the problem, its network structure has good generalization ability [11-12].

Assuming that the pixel size of the input image in the neural network is 1000×1000, in the BP neural network, the neurons are fully connected, as shown in Figure 1(b), the neurons between each hidden layer must 1000×1000=106 parameters are used for training. With the increase of the number of hidden layers, there will be more parameters between neurons, so it is obviously not feasible to use the full connection method for network training. In the CNN, each neuron only needs to be connected to the local pixels of the feature map of the previous layer as shown in Figure 1(a). If the local receptive field in the neural network is 10 × 10, the parameters to be trained are also there are only 100. The weight sharing of the network model of the CNN means that all neurons on the same feature map are trained with the same weight parameters, as shown in Figure 2 [13-14]. If the image contains 100 feature maps, the total number of convolution kernel parameters that need to be trained is 100×100=104, which is 108 times lower than the 1012 parameters of the fully connected network, and the local connection and weight sharing are extremely large It reduces the computational complexity of parameters, making it possible to train and recognize images in deep networks [15-16].

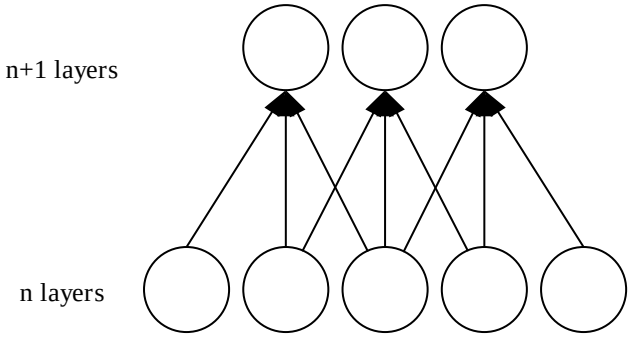

Figure 2. Schematic diagram of weight sharing

2.3. Algorithm Research

(1) Human pose estimation algorithm framework based on depthwise separable convolution

The main network architecture selected in this paper is the CPM model proposed by Carnegie Mellon University, which has high accuracy and good robustness. C represents the convolution layer, and P represents the pooling layer. In the first stage, The input image is first passed through three 9*9 convolution-pooling groups and a 5*5 convolutional layer to obtain the middle layer feature map x, and then they are input into the 9*9 convolutional layer and two 1*1 volumes Laminate [17-18]. In the first stage, the confidence of joint points is only calculated from the local image information, and then from the second stage onwards, the joint point confidence of the previous stage is used to learn the long-distance relationship between each joint point.

In each stage including the second stage, the output of the previous layer is combined with the intermediate layer x and then input to three 11*11 convolutional layers and two 1*1 convolutional layers for calculation. In each stage, the confidence of each joint point is output, and the loss

function in the calculation is:

$$f_t = \sum_{p=1}^{P+1} \left\| b_t^p - b_*^p \right\|_2^2 \tag{1}$$

Among them, t represents the t-th stage; p represents the number of the joint point; b*p represents the confidence of the p-th joint point in the ideal state.

(2) Evaluation indicators

With the development of human estimation, researchers have also put forward recognized evaluation indicators in evaluation methods. The evaluation indicators in this paper are the two most popular evaluation methods at present.

Average Precision (mAP): When evaluating the detection network, mAP is often an important indicator for analyzing and evaluating models [19-20]. When doing human body pose estimation, the bottom-up estimation method is to first predict the key points of the human body, and then connect the key points. Its calculation formula is:

$$OKS = \frac{\sum_i \exp(-d_i^2 / 2s^2 k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{1}$$

Among them, i is the number of the joint point; vi is the visible situation of the i-th key point, if it is greater than 0, it is visible; δ(v>0) means that only the visible situation is calculated into the evaluation system; s is the character scale factor, Represents the square root of the area of the entire image occupied by the human body; d2 is the Euclidean distance between the predicted joint point and the real joint point; ki is the set normalization constant.

3. Experimental Study

3.1. Structural Design of Attention Mechanism

Most of the current research results of attention mechanisms related to video AR are coupled with long-term and short-term memory networks, that is, ignoring the spatial features of videos. Therefore, the structure of the attention mechanism constructed in this paper needs to satisfy the condition of prominent spatiotemporal features.

The structure of the attention mechanism of the fully convolutional network proposed in this paper is shown in Figure 3, which consists of the following parts:

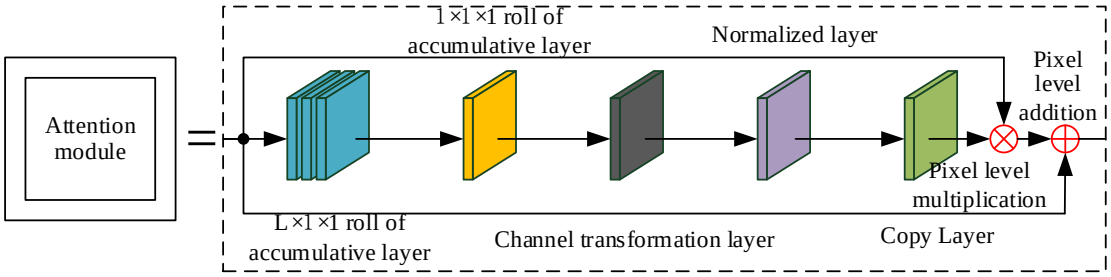

Figure 3. Structure of attention mechanism

(1) Convolutional layer

The weight parameters of the convolutional layers are used to learn to highlight salient regions in RGB images and optical flow images. There are two convolutional layers (blue and yellow) in

Figure 3. The size of the convolution kernels are all 1×1, and the number of convolution kernels is L and 1, respectively.

The reason for designing a multi-channel L×1×1 convolution structure (the blue part in Figure 3) instead of letting the input of the attention module directly pass through the single-channel 1×1×1 convolution structure is to To reduce the loss of input information, the second is to increase the amount of parameters to facilitate the learning of the attention module.

(2) Channel transformation layer (Reshape)

The main role of the channel transformation layer is to change the dimension of the output feature map.

(3) Normalization layer (Softmax)

The Softmax function is used to normalize the feature map, and the feature map after the normalization layer becomes a probability map.

(4) Duplication layer (Tile)

The generated probability map is copied multiple times to ensure that the number of channels of the probability map is equal to the number of channels of the input feature map of the attention module, so that the probability map and the input feature map can be multiplied at the pixel level.

(5) Short-circuit structure

The re-weighted feature map output after the above four steps will be input to the classification layer, but if the probability map generated by the attention mechanism is not accurate enough, the feature map participating in the classification will have noise, which is detrimental to the recognition effect influences. In order to avoid this situation, before the feature map output, the attention module designs a short-circuit structure, the purpose is to connect the input feature map and the weighted output feature map by pixel-level addition to suppress noise as much as possible data. Short-circuit structures are also used in related research to suppress data noise and retain effective information in feature maps.

3.2. Introduction to the Experimental Environment and Process Architecture

This deep learning-based video behavior algorithm experiment is mainly completed on the laboratory server. The configuration of the experiment in terms of hardware environment is shown in Table 1, and the configuration in terms of software environment is shown in Table 2.

Table 1. Experimental hardware environment configuration

Equipment name	Model or size
CPU	Intel i7-6850K
GPU	Nvidia GTX 1080Ti ×4
Memory	64 GB
A main board	ASUS X99-E
Hard disk	1TB

Table 2. Experimental software environment configuration

Operating system	Ubuntu 16.04
Programing language	Python 3.6
Deep learning framework	PyTorch 1.1
Mainly dependent libraries	OpenCV 4.2 FFmpeg 4.0 NumPy 1.18 Matplotlib 3.1.3

In terms of hardware, four-way Nvidia 1080Ti GPU and Intel i7 CPU play an important role due

to the high computational complexity and speed requirements of deep learning and 3D convolutional networks, while ensuring the iterative efficiency of various comparative experiments. In terms of software, since a large number of video clipping and sampling operations need to use FFmpeg tools, a large number of video frame image preprocessing is completed by OpenCV, deep learning model building and training are mainly completed using PyTorch framework, and finally NumPy and Matplotlib are used to complete result processing and analysis Work.

The overall experiment content process is shown in Figure 4, which is mainly divided into the data processing of video and video frame sequences and the construction of input data pipelines, the establishment of network and parameter indicators, the training part, and the result analysis part.

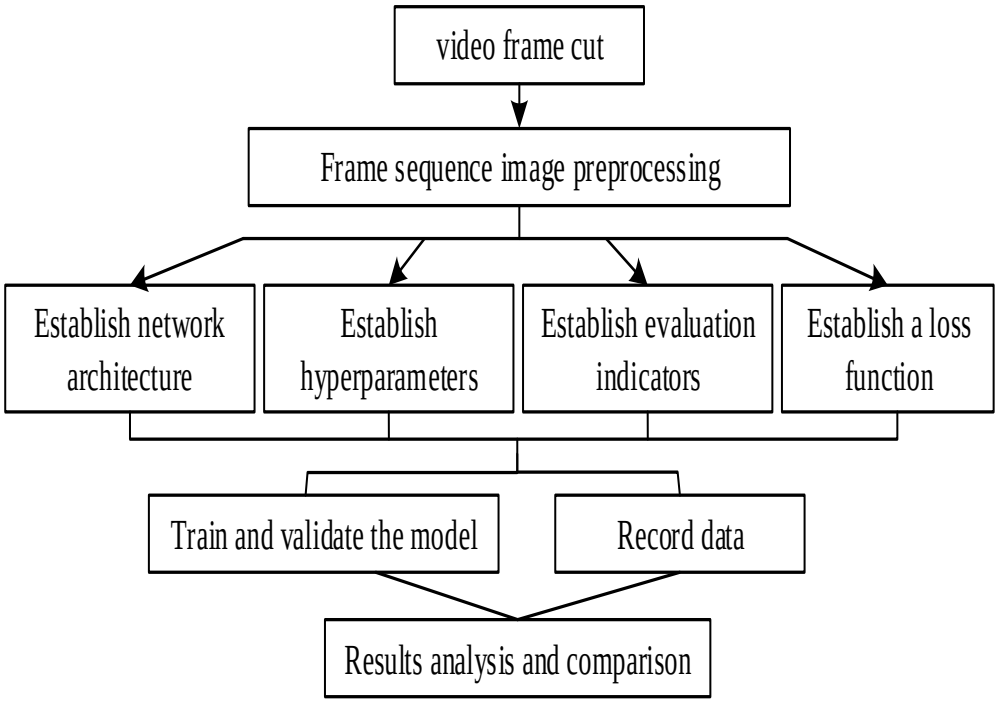

Figure 4. The overall process of the experimental content

4. Experiment Analysis

4.1. Performance Comparison

In this subsection, deep neural network based methods for human AR will be evaluated. Experiments were carried out with deep neural network, VGGNet-16 network, ResNet101 network and BN-Inception network respectively to test the AR accuracy, optical flow and two-stream composite RGB optical flow using RGB data and optical flow data in these four networks data. The experimental results are shown in Table 3.

Table 3. Accuracy of different networks on RGB and optical flow

Convolutional Network	RGB	Optical Flow	Fusion
VGGNet-16	78.9%	86.2%	91.4%
BN-Inception	84.2%	86.9%	92.2%
ResNet101	83.9%	85.6%	92.1%
Deep Convolutional Neural	84.6%	87.4%	93.5%

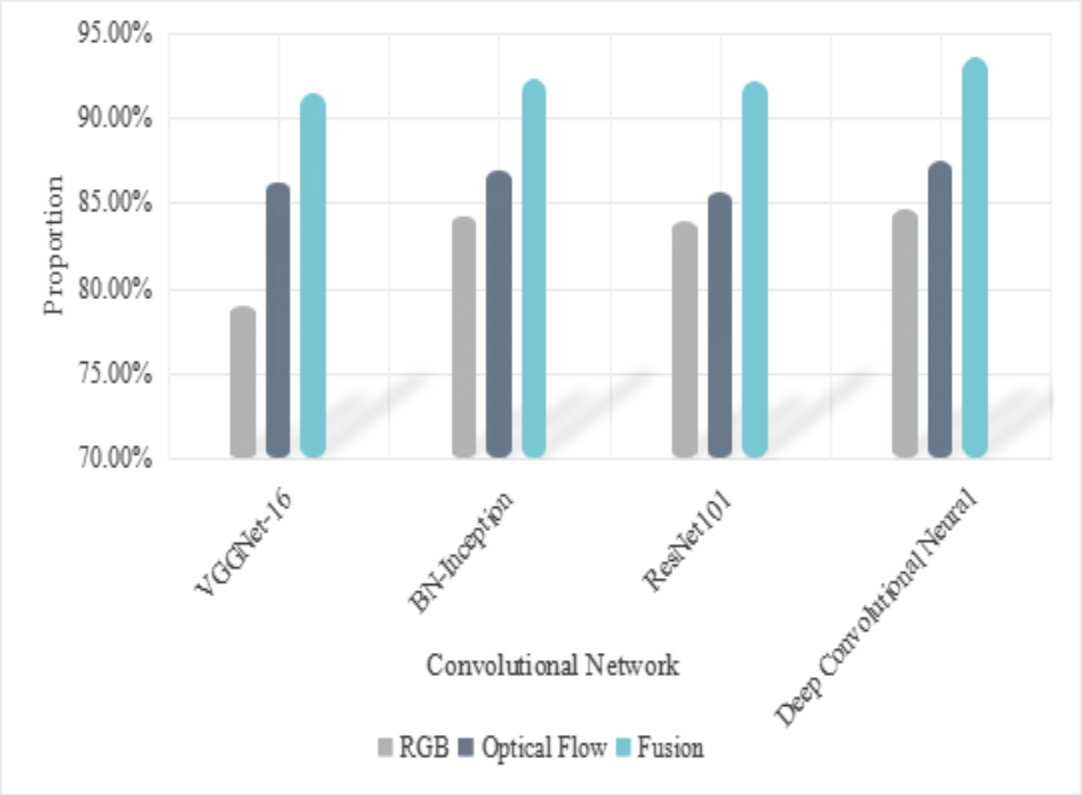

Figure 5. Accuracy analysis of different networks on RGB and optical flow

As can be seen from the results in the figure, whether using single-stream RGB data, optical flow data, or dual-stream fusion results, the recognition accuracy of using deep neural networks is better than that of conventional networks. Illustrating the effectiveness of deep CNNs in gymnastics AR.

4.2. Influence of Different Video Time Segments K

The number K of video clips distributed in a short time has a great influence on the classification results, so it is very important to control the value of K to achieve a good classification effect. When the K value is 1, it is normal not to use the video time division technique, and increasing the K value is expected to improve the recognition performance of the model. In the experiment, set the value of K from 1 to 9 and use the same method to calculate the function. The experimental results are summarized in Table 4.

Table 4. Classification accuracy of deep CNNs under different K values

K	RGB	Optical Flow	Fusion
1	83.9%	85.8%	92.6%
3	85.8%	87.9%	93.8%
5	87.7%	88.5%	94.9%
7	86.4%	88.6%	94.7%
9	87.9%	88.8%	95.2%

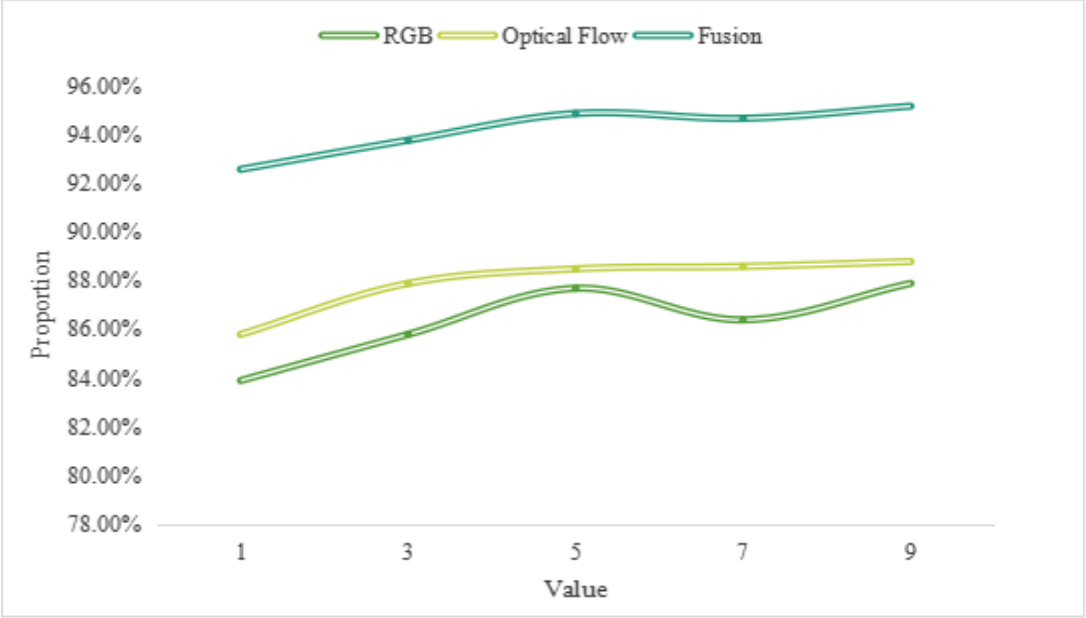

Figure 6. Classification accuracy analysis of deep CNNs under different K values

It can be seen from the experimental results that by increasing the K value of the number of short clips, the correct classification rate increases accordingly. For example, when the K value is 5, the binomial result is 2.4% better than when the K value is 1. This shows that using more time slices and creating long-term models helps to bring richer information and better temporal order models per video. When the value of K continues to increase, the recognition function tends to be saturated.

5. Conclusion

Nowadays, the application of CNN models has penetrated into various fields, such as character recognition, text recognition, face detection and analysis, etc. In the recognition process of CNN, there are two unique features, one is small-scale connection, also known as local connection; the other is sharing a weight, also known as weight sharing. Through these two features, the parameters of the CNN in the calculation process are greatly reduced, and the generalization ability of the network is also improved. In order to achieve high accuracy of the network framework, the existing human pose estimation models only focus on deepening the neural network. Reflecting on this approach, although the accuracy of the model is constantly increasing, it has the disadvantages of large computational load, high computational cost, and high hardware requirements. Based on the deep CNN, this paper studies gymnastics AR; proposes a human pose estimation algorithm framework based on depthwise separable convolution, and analyzes the evaluation indicators; this paper designs the attention mechanism structure to meet the requirements of prominent space and time. The structure of the attention mechanism constructed under the condition of features.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] *Malik Z, Shapiai M. Human action interpretation using CNN: a survey. Machine Vision and Applications, 2022, 33(3):1-23. https://doi.org/10.1007/s00138-022-01291-0*

[2] *Kahaki S, Nordin M J, Ahmad N S, et al. Deep CNN designed for age assessment based on orthopantomography data. Neural Computing and Applications, 2020, 32(13):9357-9368. https://doi.org/10.1007/s00521-019-04449-6*

[3] *El-Moneim S A, Nassar M A, Dessouky M I, et al. Cancellable template generation for speaker recognition based on spectrogram patch selection and deep CNNs. International Journal of Speech Technology, 2022, 25(3):689-696. https://doi.org/10.1007/s10772-020-09791-y*

[4] *Martin P E, Benois-Pineau J, R P áteri, et al. Fine grained sport AR with Twin spatio-temporal CNNs. Multimedia Tools and Applications, 2020, 79(27):20429-20447. https://doi.org/10.1007/s11042-020-08917-3*

[5] *Uppal H, Sepas-Moghaddam A, Greenspan M, et al. Depth as Attention for Face Representation Learning. IEEE Transactions on Information Forensics and Security, 2021, PP(99):1-1. https://doi.org/10.1109/TIFS.2021.3053458*

[6] *Vesala G T, Ghali V S, Subhani S, et al. Convolution Neural Networks Based Automatic Subsurface Anomaly Detection and Characterization in Quadratic Frequency Modulated Thermal Wave Imaging. SN Computer Science, 2022, 3(3):1-13. https://doi.org/10.1007/s42979-022-01055-7*

[7] *Memon F A, Khan U A, Shaikh A, et al. Predicting Actions in Videos and Action-Based Segmentation Using Deep Learning. IEEE Access, 2021, PP(99):1-1. https://doi.org/10.1109/ACCESS.2021.3101175*

[8] *Kumar N, Sukavanam N. A weakly supervised CNN model for spatial localization of human activities in unconstraint environment. Signal, Image and Video Processing, 2020, 14(5):1009-1016. https://doi.org/10.1007/s11760-019-01633-y*

[9] *Divya R, Peter J D. Smart healthcare system-a brain-like computing approach for analyzing the performance of detectron2 and PoseNet models for anomalous action detection in aged people with movement impairments. Complex & Intelligent Systems, 2022, 8(4):3021-3040. https://doi.org/10.1007/s40747-021-00319-8*

[10] *Bennett S M, Hindin J S, Mohatt J, et al. Proof of Concept Study of an Oral Orthotic in Reducing Tic Severity in Tourette Syndrome. Child Psychiatry & Human Development, 2021, 53(5):953-963. https://doi.org/10.1007/s10578-021-01178-7*

[11] *Chuah W, Tennakoon R, Hoseinnezhad R, et al. Deep Learning-Based Incorporation of Planar Constraints for Robust Stereo Depth Estimation in Autonomous Vehicle Applications. IEEE Transactions on Intelligent Transportation Systems, 2021, PP(99):1-12.*

[12] *Vigneshwaran B, Iruthayarajan M W, Maheswari R V. Recognition of shed damage on 11-kV polymer insulator using Bayesian optimized convolution neural network. Soft Computing, 2022, 26(14):6857-6869. https://doi.org/10.1007/s00500-021-06629-w*

[13] *Pabitha C, Vanathi B. Densemask RCNN: A Hybrid Model for Skin Burn Image Classification and Severity Grading. Neural Processing Letters, 2021, 53(9):1-19. https://doi.org/10.1007/s11063-020-10387-5*

[14] *Mohan H M, Rao P V, Kumara H, et al. Non-invasive technique for real-time myocardial infarction detection using faster R-CNN. Multimedia Tools and Applications, 2021, 80(17):26939-26967. https://doi.org/10.1007/s11042-021-10957-2*

[15] Sahay A, Amudha J. Integration of Prophet Model and Convolution Neural Network on Wikipedia Trend Data. Journal of Computational and Theoretical Nanoscience, 2020, 17(1):260-266. https://doi.org/10.1166/jctn.2020.8660

[16] Gopi C S, Sivareddy C, Prasad K M, et al. A Proficient Scheme for Detecting Breast Cancer by Classification Techniques. Journal of Computational and Theoretical Nanoscience, 2020, 17(8):3453-3457. https://doi.org/10.1166/jctn.2020.9209

[17] Uppal H, Sepas-Moghaddam A, Greenspan M, et al. Depth as Attention for Face Representation Learning. IEEE Transactions on Information Forensics and Security, 2021, PP(99):1-1. https://doi.org/10.1109/TIFS.2021.3053458

[18] Vesala G T, Ghali V S, Subhani S, et al. Convolution Neural Networks Based Automatic Subsurface Anomaly Detection and Characterization in Quadratic Frequency Modulated Thermal Wave Imaging. SN Computer Science, 2022, 3(3):1-13. https://doi.org/10.1007/s42979-022-01055-7

[19] Jaraut P, Abdelhafiz A, Chenini H, et al. Augmented CNN for Behavioral Modeling and Digital Predistortion of Concurrent Multiband Power Amplifiers. IEEE Transactions on Microwave Theory and Techniques, 2021, PP(99):1-1.

[20] Singh B, Sur A , Mitra P. Steganalysis of Digital Images Using Deep Fractal Network. IEEE Transactions on Computational Social Systems, 2021, PP(99):1-8. https://doi.org/10.1109/TCSS.2021.3052520