Deep Learning Methods for Human Behavior Recognition

Jia Lu, Minh Nguyen, Wei Qi Yan
Auckland University of Technology
Auckland, 1010 New Zealand

Abstract— **In this paper, we investigate the problem of human behavior recognition by using the state-of-the-art deep learning methods. In order to achieve sufficient recognition accuracy, both spatial and temporal information was acquired to implement the recognition in this project. We propose a novel YOLOv4 + LSTM network, which yields promising results for real-time recognition. For the purpose of comparisons, we implement Selective Kernel Network (SKNet) with attention mechanism. The key contributions of this paper are: (1) YOLOv4 + LSTM network is implemented to achieve 97.87% accuracy based on our own dataset by using spatiotemporal information from pre-recorded video footages. (2) The SKNet with attention model that earns the best accuracy of human behaviour recognition at the rate up to 98.7% based on multiple public datasets.**

Keywords*— *Deep learning (DL), Convolutional neural network (CNN), Long short-term memory (LSTM), You only look once (YOLO), Selective kernel network (SKNet), Attention mechanism

I. INTRODUCTION

With decreasing costs of digital monitoring equipment such as cameras and microphones, video surveillance has been broadly applied to public places such as banks, museums, shopping malls, etc. which allows to monitor abnormal events. However, at present most of our video surveillance systems are still being run in traditional manner, which captures anomalies and associated evidences only through offline videos. It is challenging to make real-time alarms, pop up notifications, and monitor the incident scenes uninterruptedly. Hence, it is necessary to develop the technology for real-time human behavior recognition so as to curtail the security staff's workload and ameliorate the work efficiency.

In this paper, we design and implement deep learning methods which are time efficiency and outperform in training and testing. Moreover, deep learning models have manifestly exhibited excellent capabilities in human behavior recognition, how to make the model more stable and robust for human behavior recognition aptly has become the new challenge. In a nutshell, this research project aims to develop the cutting-edge deep learning methods so as to conduct human behavior recognition without too many manual operations. As the outcome of this research project, we anticipate reaching an overall up to 90% accuracy for the recognition in real time.

In intelligent surveillance [44, 56], human detection [1-5, 52], motion capture [6-9], human re-identification [10-12], gait recognition [13, 14, 15, 46-49, 51, 53], and human behavior analysis [16, 17, 54, 55] are employed to identify individuals in various scenes. The investigation of human behaviors in video footages is a redhot topic in the field of video surveillance. In these projects, human behavior analysis is not only for individuals (e.g., running, fainting, walking, etc.), but also for human-related events (talking, fighting, shoplifting, etc.) [18]. In this paper, our focus is on recognizing human behaviors by using both spatial and temporal information. We propose a novel YOLOv4 + LSTM net, which yields promising results. For the purpose of comparisons, we implement Selective Kernel Network (SKNet) with attention mechanism.

In section II of this paper, we will present our related work. Our methods will be explicit in Section III, our experimental results will be demonstrated in Section IV. The conclusion and future work of this paper will be delineated and envisioned in Section V.

II. RELATED WORK

With the increased capacity of computing devices, deep neural networks (DNNs) have gained massive attention to detect visual objects, which lead to a new era of computer vision [16, 19, 20, 57-66]. DNNs [21, 22, 45] encapsulate multiple hidden layers, the pretraining method is adopted to resolve the problem of an optimal local solution, the number of hidden layers is run up with "depth" in the neural network. Moreover, deep learning was implemented in both supervised and unsupervised models [23]. Apparently, the work has unmistakably unveiled the differences between deep neural networks and shallow neural networks in various aspects [24].

In human behavior recognition, the bounding box was taken into consideration for deep neural networks to resample the proposed pixels. Convolution neural networks (CNNs) contain multiple convolutional layers and subsampling operations, the outputs of convolutional layers are extracted as the feature maps which are flattened and fed into a fully connected layer.

Recent work indicates that if a convolutional network has shorter connections between layers from the input layer to the output layer, then the results will be more accurate and efficient. Huang *et al*. proposed a dense convolutional network (DenseNet) connecting each layer together in a feedforward manner [25]. In order to ensure the maximum information flow between the network layers, DenseNet directly connects all layers together. Moreover, in order to maintain the feedforward characteristics, each layer gets extra input from all the previous layers and passes its own feature map to all the subsequent layers. Moreover, DenseNet allows a layer to access feature maps from all of its preceding layers.

978-1-7281-8579-8/20/$31.00 ©2020 IEEE

It can naturally scale to hundreds of layers without any optimization difficulties. All layers spread the weights within the same block, which shows that the features extracted from very early layers are supplied directly to deep layers through the same dense blocks [25]. Eventually, DenseNet produces a consistency of accuracy without any degradation or overfitting as the number of layers and associated parameters are raised.

Moreover, in order to dissolve the problem of overfitting and trim down the parameters of deep neural networks, Xie *et al.* propounded a highly modularized deep learning network (ResNeXt) for image classification [26]. It raises up the accuracy without ramping up the complexity of the deep learning method, meanwhile it effectively cuts off the number of hyperparameters. ResNeXt was motivated from the idea of VGG stacking blocks of the same shape and the split-transform-merge idea of Inception models [27, 42], which holds the robust scalability and is able to meliorate the accuracy without substantially altering the complexity of the model.

On the one hand, CNNs comprise of a series of convolution layers and pooling layers. We use the CNNs to apprehend the characteristics of the input image from the global receptive field so as to precisely savvy the image. On the other hand, convolutional networks carry out convolutional operations in 2D space, which is regarded as the networks that model spatial and channel-wise information within relevant receptive field. A good deal of work embarks on convolutional networks for digital image processing from the spatial domain. For instance, embedding multiscale information in Inception model [27] was offered to aggregate visual features of a variety of sizes of receptive fields so as to snatch a better performance; attention mechanism was brought in the spatial domain which gains positive results.

A simple and effective attention module for feedforward convolutional neural networks was set forth in 2018 which is called Convolutional Block Attention Module (CBAM) [28]. In the work, the goal of attention mechanism is on essential features and suppress unnecessary features [28, 50]. The input feature map passes through global max pooling and global average pooling based on width and height respectively, and then goes through the shared multilayer perceptron (MLP) network. The output features are based on elementwise and activated by using sigmoid function to generate the final channel attention. The attention is computed as eq.(1)

$$\begin{aligned} M_c(F) = {} & \sigma\Big(MLP\big(AvgPool(F)\big) + MLP\big(MaxPool(F)\big)\Big) \\ & = \sigma\Big(\mathrm{W}_1\big(\mathrm{W}_0(\mathrm{F}^c_{avg})\big) + \mathrm{W}_1\big(\mathrm{W}_0(\mathrm{F}^c_{max})\big)\Big) \end{aligned} \tag{1}$$

where F^c_{avg} and F^c_{max} denote the average pooling and max pooling in the attention module, W_1 and W_0 are the weights which share both input and ReLU activation function by using W_0.

The convolution kernels in SKNet [29] are presented to implement various receptive fields, which sustains three operators: Split, fuse, select. In neuroscience, the size of receptive field is constructed by using a stimulation mechanism. The method can make CNNs adjusting the size of their receptive field adaptively and competently for the input information. The attention models produce a diversity of receptive fields, the multiple SK units are then stacked into the SKNet.

YOLOv4 [30] was the recent deep learning model, which optimizes the algorithms such as SAM and PAN, etc. The model was trained based on a single GPU and owns the advantage of time efficiency. YOLOv4 takes the place of the spatial-wise attention by using point-wise attention, moreover, the concatenation was substituted by using the original shortcut connection. Afterwards, a Bag of Freebies (BoF) was utilized to be associated with a Bag of Specials (BoS) so as to uplift the overall performance, which encapsulates enlarging the receptive field by using attention mechanism.

From artificial neural networks to convolution neural networks, only one input is tackled, the previous input and the next input are stark irrelevant. To resolve this problem, the recurrent neural networks (RNN) for processing sequence data were suggested [31,32,33]. As a powerful multilayer neural network model, RNN is the long-term dependencies of the model based on time. Because of the gradient exploding and gradient vanishing, this limitation leads to the unstable variation of the errors in the model training. Thus, in order to utterly solve the problem of gradient vanishing, long short-term memory (LSTM) [34] was recommended. LSTM architecture embraces input gate, forget gate, and output gate. Gers *et al.* remedied the LSTM with forget gate, the network can purge unnecessary information and establish peephole connections [35,36].

Mahasseni *et al.* proposed a 3D skeleton sequence based on video frames to regularize LSTM network [37]. Attention-enhanced graph convolutional LSTM (AGC-LSTM) with skeleton data was implemented for human behavior recognition, which adopts skeleton information as the input of LSTM, then spatiotemporal feature maps are extracted by using AGC-LSTM, while the LSTM has a strong ability to acquire temporal features. Combining the LSTM with graph structure together, the model effectively utilizes temporal and spatial information of input images [38].

Sharma *et al.* took use of attention mechanism to human action recognition, which mixes soft attention model with the LSTM to cope with long sequence data and learn the key point of the movement, thus achieved 81.44% accuracy based on UCF-11 dataset [39]. An end-to-end spatiotemporal attention model was accommodated for human behavior recognition [40]. The end-to-end recurrent pose-attention network (RPAN) was expounded by using CNN to generate the feature cube, the post attention mechanism shares the attention parameters through semantically-related human joints so as to attain high quality of human behavior recognition, which indicates 97.4% accuracy based on PennAction dataset [41].

Our contribution of this paper is to present the end-to-end deep learning methods, we firstly present a SKNet with attention mechanism for human behavior recognition, then YOLOv4 + LSTM method aided by using a GPU is employed to accelerate the deep learning processing so as to lessen the time costs.

III. OUR METHODS

In this paper, we spell out a spatial attention-based model SKNet which explicates more positive results than previous deep learning models in human behavior recognition. Pertaining to the spatial attention module, its focus is on "where" an informative part is. Figure 1 shows a block of SKNet with the spatial attention module, Eq. (2) presents how to compute the spatial attention

$$M_s(F) = \sigma\left(f^{7\times 7}([AvgPool(F); MaxPool(F)])\right) \qquad (2)$$

where $\sigma(\bullet)$ is the sigmoid function and $f^{7\times 7}$ is a convolution operation with the filter size of 7-by-7. The spatial attention module applies average-pooling and max-pooling operations along the channel axis and concatenates them to generate an efficient feature descriptor. Consequently, a convolution operation with a 7×7 filter is applied to produce the feature maps, a sigmoid function for normalization is offered to yield the final feature maps.

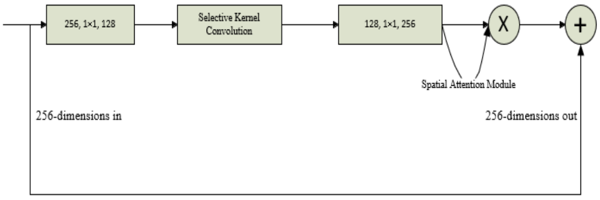


Fig. 1. The structure of the spatial attention module

In human behavior recognition, most of traditional machine learning methods are based on feature extraction techniques; furthermore, most of feature extraction techniques are based on spatial information, which may be affected by external settings. As most effective machine learning methods, iDT [43] has accomplished great progress in the field of human behavior recognition. In deep learning, most research work explicates that both spatial and temporal information is vital to motion features. Thus, LSTM is taken into account in our research project in order to extract the temporal information from each video frame. Fig. 2 shows the basic LSTM architecture of this paper.

In Fig. 1 and Fig. 2, we convert a video to the sequence of feature vectors so as to accurately present the features from each video frame. Consequently, LSTM network is applied to predict human behaviors. After combined the CNN and LSTM, the network achieves extremely high accuracy of human behavior recognition. The reason is that our convolutional operator is able to deal with each video frame independently. The model has the capability to restore the structure of each sequence and reshape the output to a vector sequence. CNNs compass the feature extractor, the output feature maps are generated from activation functions and relevant pooling layers, the feature maps exported from the CNNs will be imported as the input of the LSTM network.

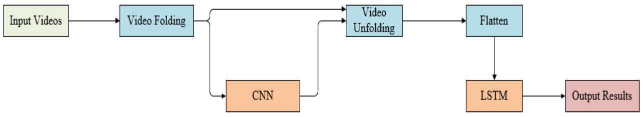


Fig. 2. The LSTM architecture for human behavior recognition

IV. EXPERIMENTAL RESULTS

In this paper, the open Weizmann dataset was selected, we also create our own dataset for the purpose of model testing and validation in this project. The Weizmann dataset contains ten classes, each of them contains nine videos which were taken by using a static camera with human behavior analysis, the dataset has nine participants in total. The resolution of Weizmann dataset is 180×144. In our experiments, we chose five classes which cover walking, skipping, running, jacking, and jumping. Fig. 3 shows the samples of the Weizmann dataset.

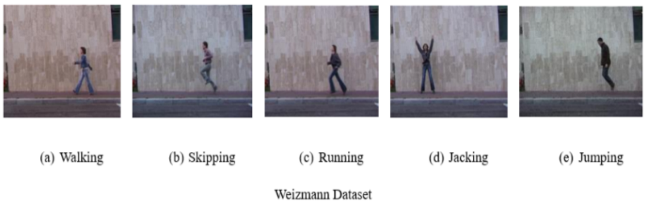


Fig. 3. The example of the Weizmann dataset

Our own dataset comprises the samples from a total of 20 videos with five classes of human behavours, which were shot by using a static camera. The resolution of this dataset is 1280×720. Our own dataset comprises of 3,200 frames and 2,000 frames were chosen for model training, 1,200 frames were selected for model testing, Fig. 3 indicates the samples of our own dataset.

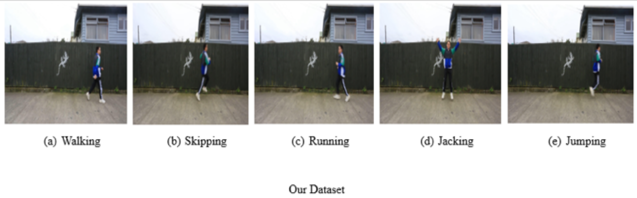
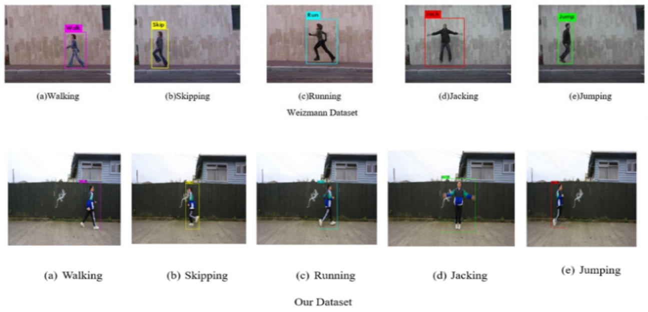


Fig. 4. The example of our own dataset

Our focus of this paper is chiefly on the proposed deep learning methods and how they affect our outcomes. We adopted LSTM + YOLOv4 model with class score fusion to fulfil the human behavior recognition. Moreover, an attention mechanism based on SKNet net was verified in this paper. In Fig.5, we demonstrate the video frames of our results by utilizing the two selected datasets.

Fig.6 exhibits the training and validation loss by using the Weizmann dataset with the SKNet net and attention model. From Fig. 6, we showcase that SKNet models are able to achieve 86.224% accuracy. Moreover, by combining the attention model with the SKNet net, the accuracy reaches to 97.194%.

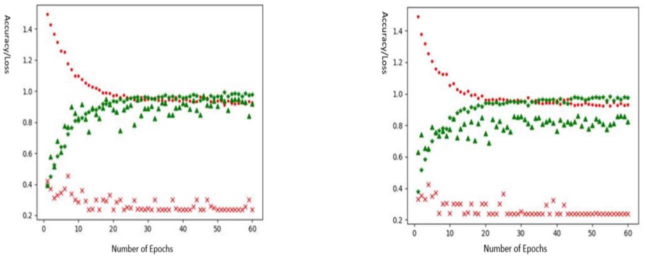

(a)Walking (b)Skipping (c)Running (d)Jacking (e)Jumping

Weizmann Dataset

(a) Walking (b) Skipping (c) Running (d) Jacking (e) Jumping

Our Dataset

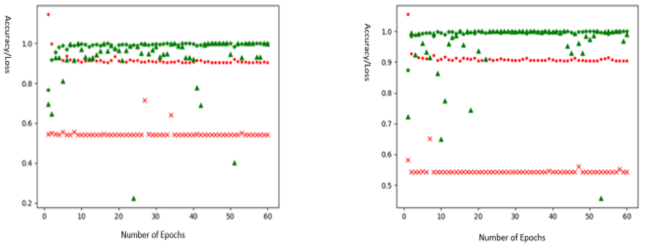
Fig. 5. The results based on two selected datasets



SKNet + Attention mechanism (ACC=97.194%) SKNet (ACC=86.224%)

Fig. 6. The training and validation loss by using SKNet and attention mechanism based on Weizmann dataset

In this paper, all our experiments necessitate large amount of computations, we chose the batch size 8 and learning rate 0.001. Moreover, the number of the epoch is set to 60. In Fig. 6 and Fig. 7, the green dots represent the training and validation accuracy, the red dots stand for the training and validation loss.



SKNet + Attention mechanism (ACC=98.644%) SKNet (ACC=96.982%)

Fig. 7. SKNet and Attention mechanism training and validation loss during the training process by using our own dataset

Fig.7 shows the training and validation loss by using SKNet net and attention mechanism with our own dataset. The SKNet net individually achieved 96.982% accuracy with the assistance of our own dataset; the SKNet net after combined with attention mechanism is able to earn 98.644% accuracy. Compared with the two models, the accuracy grows 1.662%. For better comparing the models based on various datasets, we selected the Weizmann dataset and our own dataset. For both datasets, the number of classes is the same. The both datasets subsume the same static video frames. In our experiments, the number of epochs is 60, batch size is 8, and the learning rate is 0.001.

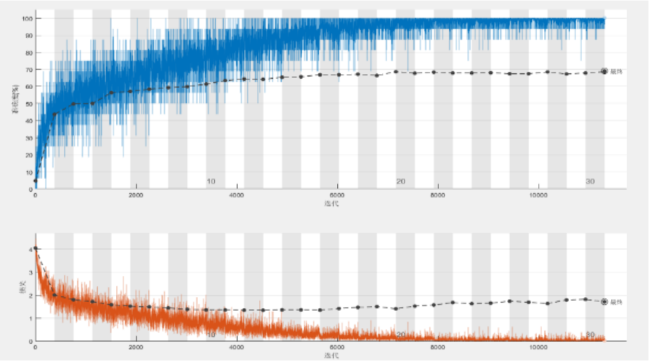


Fig. 8. CNN+LSTM training and validation loss during the training process by using our own dataset

Fig.8 shows the training and validation loss based on our own dataset by using the model CNN+LSTM. Regarding CNN+LSTM, it got 98.53% accuracy based on our own dataset. In this experiment, the proportion between training and validation sets was set to 90:10. The number of epochs is assigned as 30 with 11,310 iterations, the batch size is 16, the learning rate is 0.0001.

Throughout our experiments, we chose the public dataset and the deep learning methods to compare our experimental results. The deep learning models with attention mechanism are much stable and robust in human behavior recognition. Table I shows the comparison of our deep learning models for human behavior recognition by using multiple datasets.

In Table I, the SKNet net with attention mechanism shows the positive results for human behavior recognition. The network SKNet with attention mechanism is able to get 97.19% accuracy which has 0.90% growth of the total accuracy compared with YOLOv3 net. YOLOv4 net for human behavior recognition reaches the highest result based on Weizmann dataset which has 97.36% accuracy; based on our own dataset, it is up to 98.66% accuracy.

TABLE I. THE ACCURAACY (%) COMPARISONS OF DIFFERENT DEEP LEARNING METHODS IN HUMAN BEHAVIOR RECOGNITION

Methods		SKNet +Attention	SKNet	ResNet	DenseNet	YOLOv3	YOLOv4
Weizmann dataset	Accuracy	**97.19**	86.22	91.40	97.62	96.29	97.36
	Precision	93.00	85.00	88.80	95.30	94.60	96.20
Our dataset	Accuracy	**98.64**	96.98	-	-	96.37	98.66
	Precision	100	99.00	-	-	95.28	97.50

Moreover, we took use of LSTM net to extract the temporal information, YOLO nets were employed to extract the spatial information, finally we combine these two networks together by using score fusion and our own datasets. Table 2 illustrates the comparison of YOLO + LSTM model and YOLO nets based on our dataset. In this paper, by combining YOLOv3 net and LSTM net to extract both spatial and temporal information, we are able to achieve the accuracy 97.58%, which has 1.21% growth compared with only extracting spatial information by using YOLOv3. Moreover, YOLOv4 net gets the accuracy 97.36%. By combining YOLOv4 with LSTM net, the total accuracy is up to 97.87%.

TABLE II. THE ACCURACY (%) COMPARISONS OF YOLO + LSTM NETWORK AND YOLO METHODS BASED ON OUR OWN DATASET

Our Dataset	Walk	Skip	Run	Jack	Jump	Accuracy
YOLOv3+LSTM	97.28	96.41	98.46	100.00	95.76	97.58
YOLOv3	96.55	92.15	97.82	100.00	95.33	96.37
YOLOv4+LSTM	98.13	97.04	98.12	100.00	96.06	97.87
YOLOv4	97.53	95.69	97.79	100.00	95.79	97.36

V. CONCLUSION AND FUTURE WORK

In this paper, we have presented multiple deep learning models to accomplish human behaviour recognition. Throughout our experiments, we see that deep learning models are well implemented. The combination of SKNet with attention mechanism shows positive results for human behavior recognition. Moreover, by adopting YOLOv4, LSTM net with class score fusion to acquire both spatial and temporal information also exhibits positive results, which has 1.50% growth of accuracy compared with only using the YOLOv3 net.

From the experimental outcomes, we see that most of deep learning models could be extended through either depth of network layers or width of the layers so as to improve the accuracy of the methods. However, convolution-based deep learning uplifts the efficiency without beefing up the complexity. Meanwhile, YOLOv4 + LSTM network is also outperformed in this paper.

In our future work, we will add the attention module and spectrum information into the proposed models in order to achieve better accuracy in human behavior recognition, we will embark on multi-person behavior recognition. Additionally, more complex human behaviors with interactions such as talking, fighting, robbery etc. will be probed in the near future.

REFERENCES

[1] Chu, X., Zheng, A., Zhang, X., & Sun, J. Detection in crowded scenes: One proposal, multiple predictions. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12214-12223, 2020.

[2] Huang, J., Zhu, Z., Guo, F., & Huang, G. The devil is in the details: Delving into unbiased data processing for human pose estimation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5700-5709, 2020

[3] Luo, Y., Zhang, C., Zhao, M., Zhou, H., & Sun, J. Where, what, whether: Multi-modal learning meets pedestrian detection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14065-14073, 2020

[4] Wu, J., Zhou, C., Yang, M., Zhang, Q., Li, Y., & Yuan, J. Temporal-context enhanced detection of heavily occluded pedestrians. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13430-13439, 2020

[5] Zhang, Z., Gao, J., Mao, J., Liu, Y., Anguelov, D., & Li, C. STINet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11346-11355, 2020

[6] Kiciroglu, S., Rhodin, H., Sinha, S. N., Salzmann, M., & Fua, P. ActiveMoCap: Optimized viewpoint selection for active human motion capture. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 103-112, 2020

[7] Zhang, Y., An, L., Yu, T., Li, X., Li, K., & Liu, Y. 4D association graph for realtime multi-person motion capture using multiple video cameras. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1324-1333, 2020

[8] Xu, L., Xu, W., Golyanik, V., Habermann, M., Fang, L., & Theobalt, C. EventCap: Monocular 3D capture of high-speed human motions using an event camera. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4968-4978, 2020

[9] Habermann, M., Xu, W., Zollhofer, M., Pons-Moll, G., & Theobalt, C., DeepCap: Monocular human performance capture using weak supervision. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5052-5063, 2020

[10] Ahmed, S. M., Lejbolle, A. R., Panda, R., & Roy-Chowdhury, A. K. Camera on-boarding for person re-identification using hypothesis transfer learning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12144-12153, 2020

[11] Huang, Y., Zha, Z. J., Fu, X., Hong, R., & Li, L. Real-world person re-identification via degradation invariance learning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14084-14094, 2020

[12] Zeng, K., Ning, M., Wang, Y., & Guo, Y. Hierarchical clustering with hard-batch triplet loss for person re-identification. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13657-13665, 2020

[13] Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., ... & He, Z., GaitPart: Temporal part-based model for gait recognition. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14225-14233, 2020

[14] Li, X., Makihara, Y., Xu, C., Yagi, Y., & Ren, M., Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13309-13319, 2020

[15] Liu, J., Liu, Y., Wang, Y., Prinet, V., Xiang, S., & Pan, C., Decoupled representation learning for skeleton-based gesture recognition. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5751-5760, 2020

[16] Lu, J., Shen, J., Yan, W., & Bacic, B., "An empirical study for human behavior analysis," International Journal of Digital Crime and Forensics, 11-27, 2017

[17] Lu, J., Yan, W., & Nguyen, M., Human behaviour recognition using deep learning. In IEEE International Conference on Advanced Video and Signal Based Surveillance, 1-6, 2018

[18] Xiang, T., & Gong, S., Video behaviour profiling and abnormality detection without manual labelling. In IEEE Conference on Computer Vision (ICCV), 1238-1245, 2005

[19] Asadi-Aghbolaghi, M., Clapes, A., Bellantonio, M., Escalante, H. J., Ponce-López, V., Baró, X., ... & Escalera, S., A survey on deep learning based approaches for action and gesture recognition in image sequences. In IEEE International Conference on Automatic Face & Gesture Recognition, 476-483, 2017

[20] Herath, S., Harandi, M., & Porikli, F., "Going deeper into action recognition: A survey," Image and Vision Computing, 60, 4-21, 2017

[21] Hinton, G. E., & Salakhutdinov, R. R., "Reducing the dimensionality of data with neural networks," Science, 504-507, 2006

[22] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A., Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI Conference on Artificial Intelligence, 2017

[23] Ji, S., Xu, W., Yang, M., & Yu, K., 3D convolutional neural networks for human action recognition. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 221 – 231, 2013

[24] Liu, J., Shahroudy, A., Xu, D., & Wang, G., Spatio-temporal LSTM with trust gates for 3D human action recognition. In European Conference on Computer Vision, pp. 816-833, 2016

[25] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q., Densely connected convolutional networks. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700-4708, 2017

[26] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K., Aggregated residual transformations for deep neural networks. In IEEE Conference on Computer Vision and Pattern Recognition, 1492-1500, 2017

[27] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A., Going deeper with convolutions. In IEEE Conference on Computer Vision and Pattern Recognition, 1-9, 2015

[28] Woo, S., Park, J., Lee, J. Y., & So Kweon, I., CBAM: Convolutional block attention module. In European Conference on Computer Vision (ECCV), pp. 3-19, 2018

[29] Li, X., Wang, W., Hu, X., & Yang, J., Selective kernel networks. In IEEE Conference on Computer Vision and Pattern Recognition, 510-519, 2019

[30] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. YOLOv4: Optimal speed and accuracy of object detection, 2020. *arXiv:2004.10934*.

[31] Sutskever, I., Martens, J., & Hinton, G. E. Generating text with recurrent neural networks. In International Conference on Machine Learning, 1017-1024, 2011

[32] Sutskever, I., Vinyals, O., & Le, Q. V. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems, 3104-3112, 2014

[33] Graves, A., & Jaitly, N., Towards end-to-end speech recognition with recurrent neural networks. In International Conference on Machine Learning, 1764-1772, 2014

[34] Hochreiter, S., & Schmidhuber, J. "Long short-term memory," Neural Computation, 9(8), 1735-1780, 1997

[35] Gers, F. A., Schmidhuber, J., & Cummins, F., "Learning to forget: Continual prediction with LSTM," Neural Computation, 12(10), 2451-2471, 2000

[36] Gers, F. A., & Schmidhuber, E. "LSTM recurrent networks learn simple context-free and context-sensitive languages," In IEEE Transactions on Neural Networks, 12(6), 1333-1340, 2001

[37] Mahasseni, B., & Todorovic, S., Regularizing long short-term memory with 3D human-skeleton sequences for action recognition. In IEEE Conference on Computer Vision and Pattern Recognition, 3054-3062, 2016

[38] Si, C., Chen, W., Wang, W., Wang, L., & Tan, T., An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In IEEE Conference on Computer Vision and Pattern Recognition, 1227-1236, 2019

[39] Sharma, S., Kiros, R., & Salakhutdinov, R., Action recognition using visual attention, 2015. arXiv:1511.04119.

[40] Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J., An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In AAAI Conference on Artificial Intelligence, 4263-4270, 2017

[41] Du, W., Wang, Y., & Qiao, Y., RPAN: An end-to-end recurrent pose-attention network for action recognition in videos. In IEEE International Conference on Computer Vision, 3725-3734, 2017

[42] Simonyan, K., Zisserman, A., Very deep convolutional networks for large-scale image recognition, ICLR, 2015

[43] Wang, H., Schmid, C., Action recognition with improved trajectories. In IEEE International Conference on Computer Vision (ICCV), 2013

[44] Yan, W. Introduction to Intelligent Surveillance, Springer, 2019.

[45] Liu, Z., Yan, W., Yang, B., Image denoising based on a CNN model IEEE ICCAR, 389-393, 2018

[46] Wang, X., Yan, W., "Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory," Int. J. Neural Syst. 30(1): 1950027:1-1950027:12, 2020

[47] Wang, X., Yan, X., "Cross-view gait recognition through ensemble learning," Neural Compution and Applications, 32(11), 7275-7287, 2020

[48] Wang, X., & Yan, W. Q. "Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory," International Journal of Neural Systems, 30(01), 1950027, 2020

[49] Wang, X., Zhang, J., Yan, W., "Gait recognition using multichannel convolution neural networks," Neural Comput. Appl. 32(18): 14275-14285, 2020

[50] Ji, H., Liu, Z., Yan, W., Klette, R., Early diagnosis of Alzheimer's disease based on selective kernel network with spatial attention, ACPR 2 (1), 503-515, 2019

[51] Wang, W., Yan, W., "Human gait recognition based on SAHMM," IEEE/ACM Transactions on Biology and Bioinformatics, 2019

[52] Zheng, K., Yan, W., Nand, P., "Video dynamics detection using deep neural networks," IEEE Trans. Emerg. Top. Comput. Intell. 2(3): 224-234, 2018

[53] Liu, C., Yan, W., "Gait recognition using deep learning," Handbook of Research on Multimedia Cyber Security, 214-226, 2020

[54] Lu, J., Nguyen, M., Yan, W., "Comparative evaluations of human behaviour recognition using deep learning," Handbook of Research on Multimedia Cyber Security, 176-189, 2020

[55] Yan, W. Computational Methods for Deep Learning, Springer, 2021

[56] Zhou, L., Yan, W. Q., Shu, Y., & Yu, J. "CVSS: A cloud-based visual surveillance system". International Journal of Digital Crime and Forensics (IJDCF), 10(1), 79-91, 2018

[57] Liu, X., Nguyen, M., Yan, W., Vehicle-related scene understanding using deep learning, ACPR Workshop, 2019

[58] Ji, H., Yan, W., Klette, R., Early diagnosis of Alzheimer's disease using deep learning, ICCCV 2019

[59] Zhang, Q., Yan, W. Currency recognition using deep learning, IEEE AVSS, 2018

[60] Shen, Y., Yan, W., Blind spot monitoring using deep learning, IEEE IVCNZ, 2018

[61] Zhang, Q., Yan, W., Kankanhalli, M. "Overview of currency recognition using deep learning," Journal of Banking and Financial Technology 3 (1), 59–69, 2019

[62] Gu, Q., Yang, J., Yan, W. Q., Li, Y., & Klette, R. Local Fast R-CNN flow for object-centric event recognition in complex traffic scenes. In *Pacific-Rim Symposium on Image and Video Technology*, 439-452, 2017

[63] Gu, Q., Yang, J., Yan, W. Q., & Klette, R. Integrated multi-scale event verification in an augmented foreground motion space. In Pacific-Rim Symposium on Image and Video Technology, 488-500, 2017

[64] Song, C., He, L., Yan, W. Q., & Nand, P. An improved selective facial extraction model for age estimation. In International Conference on Image and Vision Computing New Zealand (IVCNZ), 2019

[65] Al-Sarayreh, M., Reis, M. M., Yan, W. Q., & Klette, R. A sequential CNN approach for foreign object detection in hyperspectral images. In International Conference on Computer Analysis of Images and Patterns, 271-283, 2019

[66] Al-Sarayreh, M., Reis, M. M., Yan, W. Q., & Klette, R. Deep spectral-spatial features of snapshot hyperspectral images for red-meat classification. In International Conference on Image and Vision Computing New Zealand (IVCNZ), 2018