MOTION PREDICTION USING DEPTH INFORMATION OF HUMAN ARM BASED ON ALEXNET

JingYuan Zhu[1], ShuoJin Li[1], RuoNan Ma[2], Jing Cheng[1]

[1]School of Aerospace of Engineering, Tsinghua University, Beijing, China
[2]School of Economics and Management, Tsinghua University, Beijing, China

ABSTRACT

The development of convolutional neural networks(CNN) has provided a new tool to make classification and prediction of human's body motion. This project tends to predict the drop point of a ball thrown out by experimenters by classifying the motion of their body in the process of throwing. Kinect sensor v2 is used to record depth maps and the drop points are recorded by a square infrared induction module. Firstly, convolutional neural networks are made use of to put the data obtained from depth maps in and get the prediction of drop point according to experimenters' motion. Secondly, huge amount of data is used to train the networks of different structure, and a network structure that could provide high enough accuracy for drop point prediction is established. The network model and parameters are modified to improve the accuracy of the prediction algorithm. Finally, the experimental data is divided into a training group and a test group. The prediction results of test group reflect that the prediction algorithm effectively improves the accuracy of human motion perception.

KEYWORDS

Human Motion, Prediction, Convolutional Neural Network, Depth Information

1. INTRODUCTION

In recent years, more and more researches on human motion classification have been carried out. Some of them are mainly based on 3D skeleton while others are based on depth sequences. Identification, classification and prediction of human motion play significant roles in monitoring, human-machine interaction, body language application. The development of computer vision and image processing technology provides great room for its improvement as well. Besides, some artificial intelligence methods have also been put into use for this field.

Development and success during the past decade have already proved the effectiveness and availability in the field of computer vision. Sherif Rashad et al. [1] implemented a mobile application with machine learning method to improve security of mobile devices with predicting users' behaviour. Dilana et al. [2] applied machine learning methods including k-Nearest-Neighbors and Linear Discriminant Analysis to emotion classification. Abdelkarim Mars [3] proposed an Arabic online recognition system with neural network composed of Time Delay Neural Networks (TDNN) and multi-player perceptron (MLP). Alayna Kennedy et al. [4] made use of artificial neural networks to analyze information contained in electromyogram signals to classify human walking speed. In our work convolutional neural network is applied to make classification of human motion. This paper proposed a human motion classification model based on AlexNet with considerable correctness.

DOI: 10.5121/ijaia.2019.10402

2. RELATED WORK

Johansson's [5] motion-sensing experiment which was conducted in 1970s was considered as one of the pioneers in this field. In their experiment, several bright spots distributed on the human body are used to study human motion perception. Their research proves that human vision can not merely detect motion direction, but also detect different types of limb movement patterns, including recognition activities and different motion patterns. Many researches dealt with topics such as surface treatment, curves and contours, pattern recognition and feature extraction [6]. Arridhana Ciptadi et al. [7] presented a novel action representation based on encoding the global temporal movement of a human action. They found that temporal dynamics of an action are robust making it useful for action recognition and retrieval. Mosabbeb E et al. [8] proposed a distributed activity classification framework, in which they make use of several camera sensors to observing the scene. Their method could also be applied to human activity classification. Nicola Bellotto et al. [9] applied multi-sensor data fusion techniques and used the onboard laser range finder (LRF) for laser-based leg detection. They performed human tracking under complex environments.

Due to the progress of depth cameras, depth images are more and more used in motion prediction and recognition. Shotton J et al. [10] estimated the joint position of human based on the depth images. According to this method, the posture of human body could also be defined by the position of joint position. Ellis et al. [11] used a latency-aware learning formulation to train a logistic regression-based classifier that automatically determines distinctive canonical poses from data and uses these to robustly recognize actions in the presence of ambiguous poses such as balance and kick poses. Lu Xia [12] proposed a model-based approach detecting humans using a 2-D head contour model and a 3-D head surface model with data collected by Kinect sensor.

Apart from using the data of skeleton, depth maps have also been applied to this field for a period of time. They are widely used especially in scene or object reconstruction and robots. Depth maps record the distance from the camera at a corresponding point in the scene in each pixel. Foix, S. et al. [13] introduced a 3D time-of-flight (ToF) camera which illuminates the scene with a modulated light source and observes the reflected light. The Kinect sensor (version 2) belongs to this kind of camera.

Many approaches applied to the field of gesture recognition are based on computer vision which is different from approaches detecting 3D skeleton coordinates or recognizing the whole body action stated above. In 1978, Stokoe [14] used four aspects including hand shape, position, orientation and movement to represent gesture. Many application systems related to gesture recognition have been put into use. Imagawa, K. et al. [15] proposed a system which could automatically analyze and annotate video sequences of technical talks. The system would track and recognize gestures to provide annotation. Triesch, J, et al. [16] presented a person-independent gesture interface implemented on a real robot allowing users to give simple commands to realize human-robot interaction. Nolker et al. [17] detected 2D location of fingertips by Local Linear Mapping neural network and mapped them to 3D position with which method they could recognize hand pose under different views.

A lot of related works on full-body movements have also been carried out. Bobick [14] presented several approaches to the machine perception of motion and discusses the role and levels of knowledge in each. Yang Wang et al. [18] considered the problem of describing the action being

performed by human figures of still images. Ji, Shuiwang et al. [19] developed a novel 3D CNN model for action recognition.

AlexNet is a convolutional neural network model designed by ImageNet competition winner Hinton and his student Alex Krizhevsky [20]. Due to the superiority of AlexNet, it has been applied to many fields of classification. For example, Yuan, Z. W. et al. [22] had applied AlexNetto improve the result of image retrieval with its fusion feature. Shenshen Gu et al. [23] applied this model to test whether there is a tennis ball in the picture. Jing Sun et al. [24] improved the accuracy of the scene image classification with their model based on AlexNet. The efficiency of AlexNet has also been tested with some datasets. Pal, Raju et al. [25] applied AlexNet to Animal Diagnostics Lab (ADL) dataset to extract features. Y Dang et al. [26] did their study on the evaluation of land cover classification using remote sensing images based on AlexNet.

It should be noted that most of the classification researches are based on large amount of accurate calculation based on kinematics data which requires massive amount of calculation. Besides, it would be difficult to make predictions of human motion. Taking these disadvantages into consideration, we tend to make use of convolutional neural networks(CNN) to analyze human kinematics data obtained from depth maps shot by Kinect sensor(version 2) and make prediction with an appropriate algorithm. The prediction model is mainly based on AlexNet trained with depth information.

3. PROJECT AND ALGORITHM

3.1. Project Structure

Ball throwing is chosen as the process to make analysis and prediction of in this project. The objective of the algorithm is to make analysis of the process of throwing ball and predict its drop point. Ball is thrown into the area which is divided into 4 parts of the same area size.

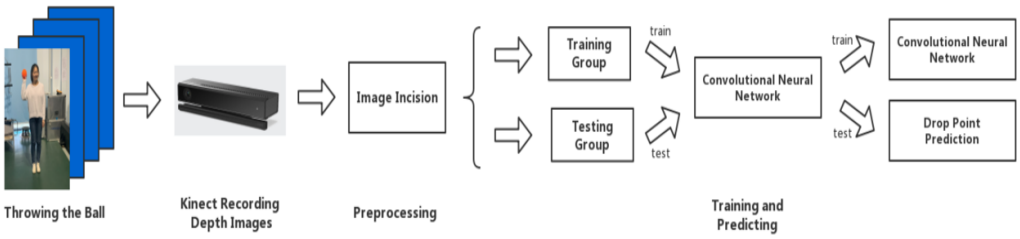

Figure 1. Project Flowchart

The human motion data is recorded in 30 frames depth maps with Kinect sensor v2. To increase effectiveness and shorten the time cost in training and predicting, those data would be preprocessed to keep the most important part and ignore other parts which have few impacts on training and prediction. Then they would be divided into two groups including training data and testing data for network model modified from AlexNet. The structure and parameters of the network would be adjusted to improve accuracy of drop point prediction. With these methods we could downsize the input data size and increase the efficiency and accuracy of the algorithm.

3.2. Prediction Algorithm

Convolutional neural networks model is made use of to analyze data obtained from depth maps. The algorithm is modified from AlexNet which had been proved available in many fields. Compared with other classical machine learning classification algorithms, the AlexNet has many advantages which would help it perform better in classification. Firstly, AlexNet uses RELU as the activation function of CNN whose effect would be better for deeper networks. Secondly, to avoid the problem of over-fitting while training the model, AlexNet takes advantage of dropout layers to randomly ignore part of changes in neurons. Besides that, the usage of overlapping max-pooling layers which are different from the average-pooling layers that had been widely used in CNN before could avoid the fuzzification effect and improves the richness of features at the same time. In AlexNet model, the addition of LRN layer could create a competition mechanism for the activity of local neurons under which would make relatively larger values become larger and correspondingly the neurons with lower feedback would be suppressed. All of these improvements could help AlexNet model to make more accurate classification.

The original structure of AlexNet includes five convolution layers, two overlapping max-pooling layers distinctively following the first two convolution layers and two fully connected layers. The output of the last fully connected layer will then be put into the Softmax classifier to get the final result.

In the process of analysis, we tried to modify the network model through methods including adding dropout layers into our model and adjust loss function to obtain better classification results. The dropout layer could reduce the severity of over-fitting while different loss functions could lead to different results so an appropriate loss function would help us train the model with reasonable loss values. Apart from those, it would be significant to find a matched scale of network. Parameters such as learning rate and batch size also need adjustment.

4. EXPERIMENT

4.1. Laboratory Equipment

Experimenters are students picked out from our lab. The whole experiment was also carried out in our lab as well. To make sure that the effect of the model wouldn't be seriously limited by the amount of our data, we repeated the experiments 400 times to get depth image data needed for network training.

In the experiment, the equipment includes a Kinect sensor v2 with a tripod for fixation, a square infrared module whose area is 2.25 square meters, several computers to control the sensor and module. Apart from those, a ball is needed for experimenters to throw.

Kinect sensor v2 is used to record the depth maps in the process of experimenters' throwing balls. It is controlled by a C++ program to record the first 30 frames of depth maps since the experimenter's right arm is raised higher than his waist. The square infrared module is controlled by algorithm coded by C++ to record the first drop point of falling ball in every experiment. The module could detect not only the drop area but also the exact coordinate of the drop point according to which the area of drop point could be calculated.

In the experiment of data collection, Kinect sensor v2 is placed to record the process of experimenters' body motion, especially their right arms. Distance between experimenters and the square infrared module's center is 195 centimeters. Experimenters are asked to stand with their feet separated at a distance of 30 centimeters. Kinect sensor v2 is fixed at the height of 135 centimeters. Distance between Kinect sensor v2 and the module is 27 centimeters. Distance between experimenters and the module is 95 centimeters. With different parameters, the networks model would provide us with various results.

4.2. Experimental Methods

The experimenters are asked to throw the ball with their right arms into an area which is divided into four parts of the same size with similar postures. The main goal of the algorithm is to analyze the process of ball throwing and predict the drop point of the ball according to motion of human body, mainly human arms.

All of the data collected by Kinect sensor are recorded in depth maps. Kinect sensor is activated whenever the assistant's hand is higher than his waist and would begin to record the 30 frames. The size of every frame is 217088 while the length of one frame is 512 and the width is 424. The value of every point is normalized to the range from 0 to 255. Taking the great size of every frame and that even more than 50 per cent of a frame is actually empty without any information and that the drop point is mainly determined by the motion of experimenters' upper bodies into consideration, those frames are cropped to a relatively smaller size with the width of 110 and height of 129. After pre-processing, experimenters' bodies would take up more than 70 per cent space of a picture.

In the process of throwing, experimenters first lift their right arms from the front of their bodies and throw the ball out when the right hands are at almost the same height as their overhead. Experimenters are asked to try their best to keep other parts of their body steady enough to guarantee that the fall of the ball is only influenced by the motion of their arms.

Figure 2. Photo of Experiments

Figure 3. Kinect Sensor and Infrared Module

5. EXPERIMENTAL RESULTS AND DISCUSSION

The 30 frames of depth images recorded in one experiment also contain experimenters' body motion information after the ball is thrown away. With observation and algorithm testing, the first 10 frames recorded since the experimenters' right hands are raised higher than their waists were finally decided to be the data put into the network model.

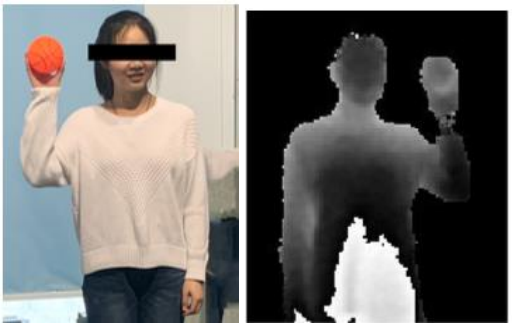

Figure 4. Photo and Depth Image of Throwing Ball

The convolutional neural networks model based on AlexNet is firstly trained and tested with data collected from our experiments. Result of prediction accuracy obtained through cross-validation is shown in Figure 5. The accuracy of prediction is slightly over 50 percent which can't be put into use in reality. Therefore, modifications of parameters and structure of networks are needed to increase effectiveness and availability of the action classification algorithm.

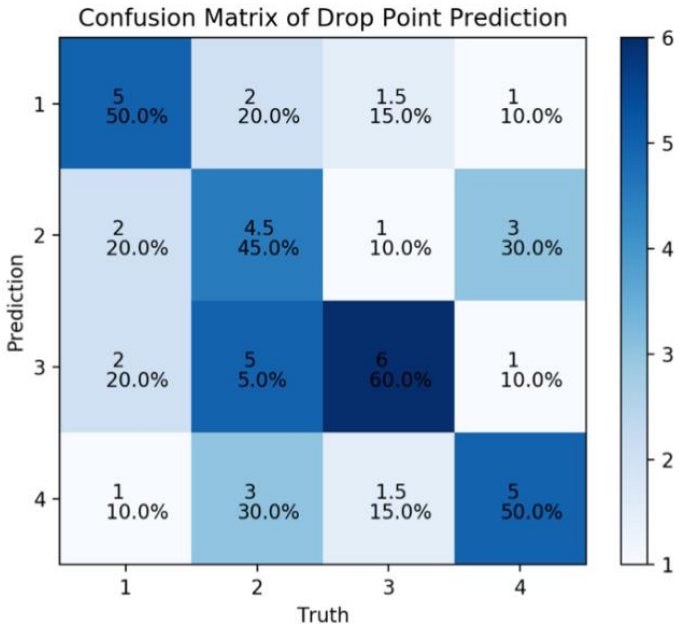

Figure 5. Prediction Result 1

As for the convolutional neural networks, it is important to decide the scale of the net which is closely linked to effectiveness and accuracy of training and predicting. Since the depth images have much in common, the neural networks model has great possibility to meet the problem of over-fitting which would badly effect the model's output.

To avoid the problem of over-fitting, we downsize the network model to a relatively proper one. After enough adjustment and experimenting, we finally confirmed the number of neurons in the net which is appropriate for input data.

Loss function plays an important role in CNN training. Neural networks model's training process is to minimize the value of loss function. At first the loss function only contains MSE between truth and prediction. To further avoid the model from effect of over-fitting, modification of the loss function by adding regular terms was carried out. The regular term chosen is L^2regular term which is shown as below. This modification proves useful for alleviating over-fitting and improving the model's accuracy.

$$\Omega(F(\vec{x};\vec{w})) = \ell_2 \frac{\|\vec{w}\|_2^2}{2n} (\ell_2 > 0)$$

With regular terms added into the loss function, the results of model's prediction accuracy were increased further as shown in Figure 6.

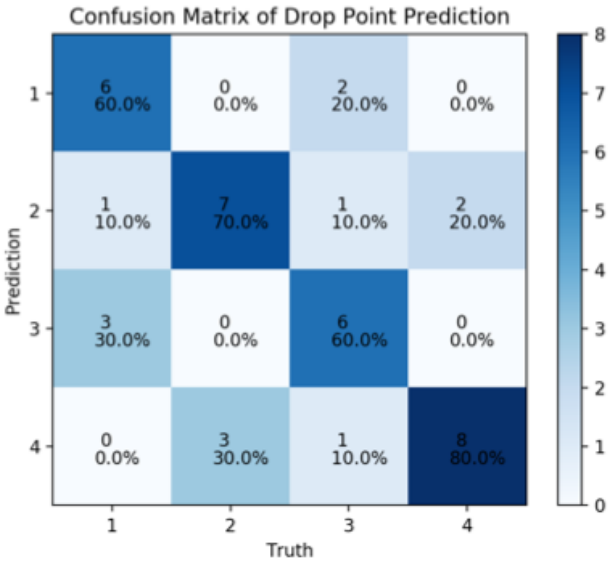

Figure 6. Prediction Result 2

The normalization layers are often added to networks model in order to control the cost of the training process. However, under some circumstances, normalization layers would actually decrease the accuracy of neural network models. Normalization layers sometimes effects the calculation of weight in training processes since it will decrease some parameters' influence on the network output. Frames put into the model has the width of 110 and height of 129, value of every point varies from 0 to 255 reflecting depth information. Values of loss function are finite in training. After analyzing and testing, normalization layers are aborted to obtain higher accuracy.

Adding dropout layers is a common way of preventing over-fitting. Dropout layer's existence makes some of the implicit layer nodes in the networks stop working randomly during the training process. Dropout layers could reduce the interaction between feature detectors which could improve the effectiveness and accuracy of the model.

Dropout layers are added into the model with various sequences to test the accuracy of training and accuracy. After lots of attempts, it is found that adding the dropout layers after the first max-pooling layer and the second max-pooling layer could give out the best result which is shown in Figure 7.

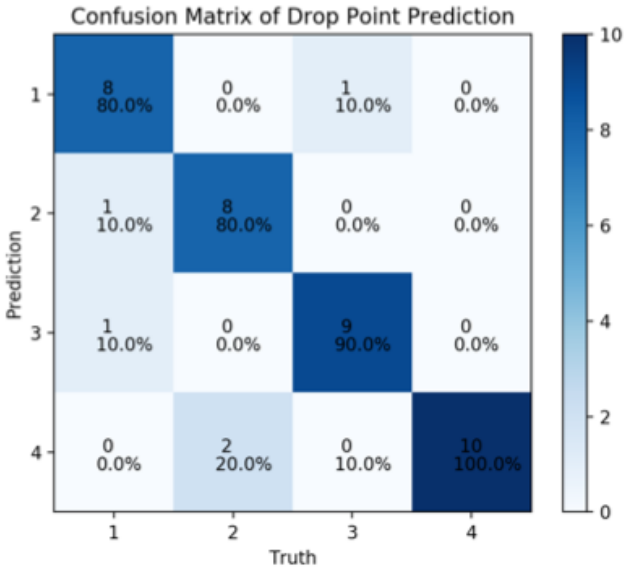

Figure 7. Prediction Result 3

After modification made to the convolutional neural network model, the most appropriate structure of convolutional neural networks model to make prediction of drop points is acquired. The networks structure is shown in Figure 8.

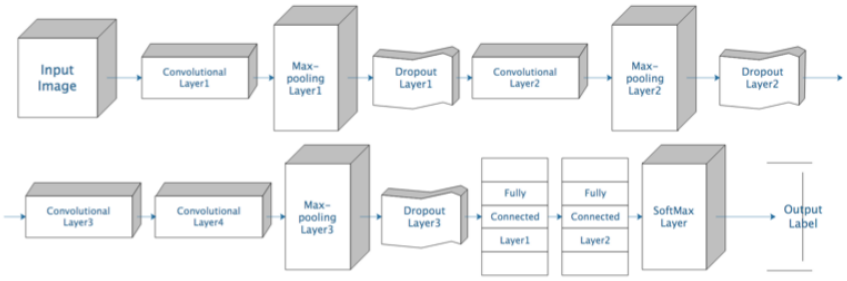

Figure 8. Model Structure

Other parameters such as learning rate and batch size are also adjusted. With this modified model, the accuracy of prediction would be higher than 80 per cent. The high level of accuracy for prediction proves that the model is able to be applied to the field of predicting human's arm motion.

6. CONCLUSION AND FUTURE WORK

This paper proposed a model to make classification of human motion based on AlexNet. To make the model designed by us reliable enough, the convolutional neural network model is trained with depth images data of human motion obtained from about 400 times of experiments. We made several modifications to the convolutional neural network model to make it appropriate for the target we want to reach. The classification results showed that our convolutional neural network model could be applied to make classification of human motion and make prediction of ball throwing. With accuracy our model has reached, it could be used to catch things like balls thrown out by people practically. Apart from that, our model could also be applied to other fields including human-machine interaction, video monitoring, body action language and medical care.

In future work we would try to make prediction from different angles by adding more images recorded from different angles. Besides, limited by size of data, drop area is only divided into four parts so far. It is possible for the network model to make prediction of drop area which is divided by finer meshes with larger amount of data. All depth images recorded since experimenters raised their right hands higher than their waists are used as training data without distinguishing. They would take up the same weight in training and testing. Surely it may be more reliable and reasonable to distribute those frames of suitable weights before training the model. In conclusion, enrichment of dataset and improvement of algorithm will help us gain better results.

ACKNOWLEDGEMENTS

In the whole process of this project, we got adequate support from Tsinghua University. We want to express our thanks for Dr. Ou Ma for his help and guidance.

REFERENCES

[1] Sherif, R. and Jonathan, B., 2018. "Behavior-based security for mobile devices using machine learning techniques". *International Journal of Artificial and Applications(IJAIA),* 9(4).

[2] Dilanam, H., Lin, Z., and Harald C. 2018. "Performance evaluation of various emotion classification approaches from physiological signals". *International Journal of Artificial and Applications(IJAIA),* 9(4).

[3] Abdelkarim Mars and Georges Antoniadis. 2016. "Arabic online handwriting recognition using neural network". *International Journal of Artificial and Applications(IJAIA),* 7(5).

[4] Alayna Kennedy and Rory Lewis. 2016. "Optimization of neural network architecture for biomechanicclassdfication tasks with electromyogram inputs". *International Journal of Artificial and Applications(IJAIA),* 7(5).

[5] Johansson, G., 1973. "Visual perception of biological motion and a model for its analysis". *Perception Psychophysics,* 14(2), pp. 201–211.

[6] IEEE, 2002. "Proceedings of ieee conference on computer vision and pattern recognition". *IEEE Computer Society Conference on Computer Vision Pattern Recognition, Cvpr*.

[7] CiptadiA.,Goodwin, M.S.,andRehg,J.M.,2014."Movement pattern histogram for action recognition and retrieval". *European Conference on Computer Vision.*

[8] Adeli, M. E., Raahemifar, K., and Fathy, M., 2013. "Multi-view human activity recognition in distributed camera sensor networks". *Sensors,* 13(7), pp. 8750–8770.

[9] Nicola, B. and Huosheng, H., 2009. "Multisensor-based human detection and tracking for mobile service robots". *IEEE Transactions on Systems Man Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man Cybernetics Society,* 39(1), pp. 167–81.

[10] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A., 2013. *Real-Time Human Pose Recognition in Parts from Single Depth Images*.

[11] Ellis, Chris, Masood, Zain, S., Tappen, Marshall, F., LaViola, Joseph, J., Jr, and Sukthankar, 2013. "Exploring the trade-off between accuracy and observational latency in;action recognition". *International Journal of Computer Vision,* 101(3), pp. 420–436.

[12] Lu, X. L. X., 2011. "Human detection using depth information by kinect". *IEEE Computer Society Conference on Computer Vision Pattern Recognition, Cvpr.*

[13] Foix, S., Alenya, G., and Torras, C., 2011. "Lock-in time-of-flight (tof) cameras: A survey". *IEEE Sensors Journal,* 11(9), pp. 1917–1926.

[14] Stokoe, W. C., 1980. "Sign language structure". *Annual Review of Anthropology,* 9(9), pp. 365–390.

[15] Imagawa, K., Lu, S., and Igi, S., 1998. "Color-based hands tracking system for sign language recognition" *IEEE International Conference on Automatic Face & Gesture Recognition*.

[16] Triesch, J., and Malsburg, C. V. D., 1999. "A gesture interface for human-robot-interaction". *IEEE International Conference on Automatic Face & Gesture Recognition*

[17] Nlker, C., Ritter, H., Fakultt, T., and Bielefeld, U., 1998. "Illumination independent recognition of deictic arm pos- tures". In Conference of the IEEE Industrial Electronics Society.

[18] Bobick,A.F.1997."Movement, activity and action:the role of knowledge in the perception of motion".Philosophical Transactions of the Royal Society of London, 352(1358), pp. 1257–1265.

[19] Wang, Y., Jiang, H., Drew, M. S., Li, Z. N., and Mori, G., 2006. "Unsupervised discovery of action classes". *IEEE Computer Society Conference on Computer Vision Pattern Recognition*.

[20] Shuiwang, J., Ming, Y., and Kai, Y., 2013. "3d convolutional neural networks for human action recognition". *IEEE Transactions on Pattern Analysis Machine Intelligence,* 35(1), pp. 221–231.

[21] Krizhevsky A, Sutskever I , Hinton G. 2012. "ImageNet Classification with Deep Convolutional NeuralNetworks". *Advances in neural information processing systems*, 25(2).

[22] Yuan, Z. W., and Zhang, J., 2016. "Feature extraction and image retrieval based on alexnet".*EighthInternational Conference on Digital Image Processing.*

[23] Gu, S., Lu, D., Yue, Y., and Chen, X., 2017. "A new deep learning method based on alexnet model and ssd model for tennis ball recognition". *IEEE International Workshop on Computational Intelligence Applications.*

[24] Jing, S., Cai, X., Sun, F., and Zhang, J., 2016. "Scene image classification method based on alex-net model". *International Conference on Informative Cybernetics for Computational Social Systems.*

[25] Pal, R., and Saraswat, M., 2018. [IEEE 2018 Eleventh International Conference on Contemporary Computing (IC3) - Noida, India (2018.8.2-2018.8.4)] 2018 *Eleventh International Conference on Contemporary Computing (IC3) - Enhanced Bag of Features Using Alexnet and Improved Biogeography-Based Optimization for Histopathological Image Analysis*. pp. 1–6.

[26] Dang, Y., Zhang, J., Deng, K., Zhao, Y., and Fan, Y. U., 2017. "Study on the evaluation of land cover classification using remote sensing images based on alexnet". *Journal of Geo-Information Science,* 19(11).