OpenPose-Based Yoga Pose Classification Using Convolutional Neural Network

Yuchen Liu*
WUSIE, South China University of Technology, Guangdong, China, 510006
*Corresponding author: 201830390426@mail.scut.edu.cn

Abstract. Human action recognition has various implementation, such surveillance system, elders care and construction alert, which arouse lots of interest of research in classification of still image. This paper mainly focusses on detecting the pose of Yoga. Comparing with traditional method using convolutional neural network, which is using original image as input to train the VGG network, extracting the skeleton images and feed them into Mobile net can impressively increase the accuracy. Dataset is collected from Kaggle website which contains five categories of labeled Yoga image. Openpose is an open-source API that can extract the human skeleton structure form the Yoga image based on the pose. With these skeleton image as input, the convolutional neural network will perceive everything important such as pose and angle of joints, rather than irrelevant features such as color and environment. Using Mobile net instead of common method to do classification with VGG, calculation time has been remarkably reduced and size of model is lighter which is able to be apply on single chip device. The result of model is impressive, showing high accuracy in both training data set and testing data set, which means no overfitting problem occurred in the experiment. Model size and demanding of hardware are also acceptable for a common personal computer.

Keywords: MobileNet, Openpose, Yoga pose classification.

1. Introduction

Human Action Recognition (HAR) is a newly developed technique that can comprehend action of human and label them with given classification name. Due to its various applications, HAR has aroused lots of passion around computer vision field. Representing human action with different data modalities such as, skeleton, RGB, infrared and point cloud, where massive useful information is encoded. Also, the type of data can vary with different scenarios, which shows the highly flexibility of application. Identifying the specific human movement or pose of body is the mission of HAR. It is already used extensively with video, but there is less relative research on doing so for still images. Dividing video data into fewer frames demonstrates better performance on action representation with less calculation needed and reduce redundant features without compromising the accuracy. This method will play a crucial role in varitires of situation such as judging althletes' movement by analysizing photos taken by high-speed camera, sending alert to adaults whose famaly elders live under house-care survallence systems if they get fallen down, calling supervisors when find some workers working without security in construction yard.

In the early field of human action recognition, the previous methods were mainly focus on employing Convolutional Neural Network (CNN) to process the whole picture [1-4]. The main obstacles for CNN to gain a high accuracy are background clutter, high intra-class variance and low inter-class variance among some action classes, changes in lighting source and variation in body pose [5]. In traditional CNN, the model will perceive every pixel in the image, which leads to an unpleased fact: some irrelevant factors, such as background, are perceived by the model. In some blurred image, features are harder to be extracted. In past decades, lots of types of CNN are applied to this task. The most famous one is VGG [6]. In this network, another crucial point of CNN structure design – depth, is addressed in this article. To this end, the research maintains the number of rest parameters in the architecture and add more convolutional layers to increase the depth, which is practicable since the small(3×3) convolutional filters are applied to every layer. This network is hardware-demanding since it has massive calculation of parameters. In this paper, another CNN model, Mobile Net, was applied, which is has less parameters.

The key point is to extract the pose of human. In this regard, Openpose can cope with extracting the skeleton shape from the images. Openpose proposed the method of partial affinity field (PAF) [7]. PAF is a representation composed of a set of vectors, which encode the unstructured pairing relationship (hand and elbow, elbow and shoulder) between a variable number of human body parts. Compared with previous methods, openpose effectively obtains paired scores from PAF without additional training steps. These scores are sufficient for greedy parsing to obtain high-quality results with real-time performance.

In this paper, an openpose-based CNN method was proposed for pose classification. This paper mainly concentrates on detecting Yoga pose specifically. Dataset is collected online which contains 5 types of poses of Yoga. The experimental results proved the proposed method can recognize Yoga pose effectively.

2. Mehodology

2.1.Dataset description and preprocessing

Yoga, a very well-known practice, by which people are able to relieve pressure and anxiety, are more and more popular now. Many poses

Of Yoga are known by public but the very well-known ones are the plank pose, goddess pose, tree pose, downward dog pose, and the warrior pose, which are considered in this study.

The dataset is download from Kaggle website and only one person shows up in the image. Every image is labeled manually with five names of different categories mentioned before. There are 1,081 images in training set and 470 images in testing set. In addition, the size of each image is not identical. Figure 1 presents a part of data in the collected dataset.

Since the model needs the same size of image as input, some measures are done as preprocessing. To be more specific, resizing was carried out to unify their size in 128×128. Furthermore, normalization is compulsory for the model which only takes images with 0–1-pixel value as input No augmentation is applied in this study.

Figure 1. Sample data in the collected dataset.

2.2.Openpose

It is believed that when CNN tries to learn the feature of the image, every kind of feature will be perceived, including different faces, the color of clothes and the environment which are all irrelevant from pose detection. Openpose can extract skeleton information from pictures in form of a skeleton man on black background.

Having the image passed through the first ten layers of vgg-19 network to generate a group of characteristic images input is the first stage.

The main part of the network is divided into two large modules. The first is the PAF module, which is used to train PAF. Here you can see intuitively that it is a group of vectors. The second is the confidence map module, which is used to train the key parts of each person's body in the image. Confidence map is the probability that each pixel on the map may be a part of the body. Each body part corresponds to a confidence map.

2.3.CNN model

Convolutional Neural Network is famous algorithm in image classification field [8-10]. It takes am image with fixed size as input and process it with a convolutional layer to extract the features. The output through a dense layer is an array with a length which fits the number of categories. After applying the softmax process to get the probability, the index corresponding to the maximum value in the array is the prediction given by the model

Rather than traditional Visual Geometry Group Network (VGG), which is widely used in image classification competition and has remarkable performance, the model used in this article is Mobile Net. Comparing with VGG network, which uses convolutional kernel whose channel number fits the channel of the input image, Mobile Net uses a special design kernel method called Depthwise Separable Convolution. Decomposing the convolutional kernel into depthwise convolution where each convolution kernel is applied to each channel and pointwise convolution, which is used to combine the output of channel convolution, can significantly reduce the number of parameters and maintain a high level of accuracy. Meanwhile, the under sampling is directly completed by using 2 as stride during convolution, which saves the time of under sampling again with pooling and can improve the operation speed. the model parameters are only 1/32 of VGG.

In the last dense layer, the number of neuron nodes is 5 due to there are 5 categories of images. The whole program is running in TensorFlow environment which enables the computer to do matrix calculation on GPU: Nvidia 3080. The optimizer selected here is Adam that take the advantages of gradient descent algorithm with adaptive learning rate and momentum gradient descent algorithm, it can not only adapt to sparse gradient but also alleviate the problem of gradient oscillation. Crossentropy is applied in this research as loss function and accuracy is evaluation metrics. The final result is accessed after running for 10 epochs.

3. Result and discussion

3.1.Performance of the proposed model

Table 1. Performance of two different methods

Model	Training loss	Training accuracy	Testing loss	Testing accuracy
Original CNN	0.0023	0.9991	0.6681	0.8191
CNN based on openpose	0.1213	0.9543	0.1778	0.9555

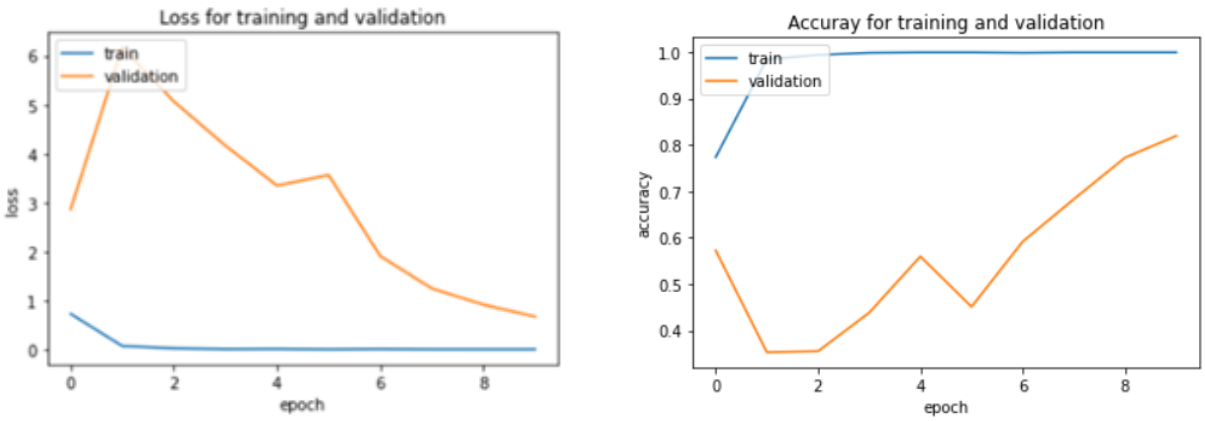

Figure 2. Accuracy and loss without openpose..

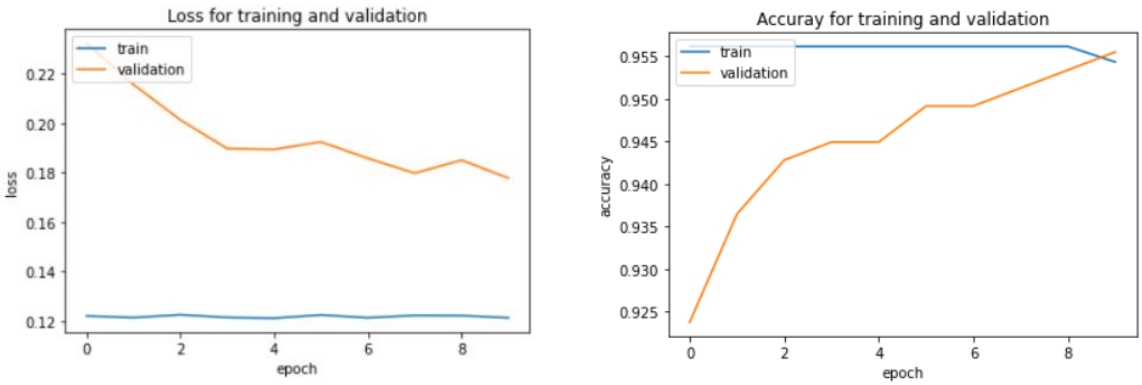

Figure 3. Accuracy and loss with openpose.

In terms of the experimental results, it can be observed that the proposed OpenPose-based method achieved a better accuracy in classification of Yoga images shown in Table 1, Figure 2 and Figure 3. In traditional CNN, the training accuracy is 0.9991 and training loss is 0.0023 while the testing accuracy is 0.8191 and testing loss is 0.6681. In the modified version, the CNN based on openpose, the training accuracy is 0.9543 and training loss is 0.1213, while the testing accuracy is 0.9555 and testing loss is 0.1778

The fact that the original CNN performs better on trainset with 0.9991 accuracy while it has only 0.8191 accuracy on testing set shows that the overfitting problem occurs in this method. Comparing with original CNN, the CNN based on openpose has two close and equally high accuracy on both training set and testing set around 0.955, only 0.05 lower than CNN on training set while 0.14 higher than CNN on testing set, which proves that no overfitting happened, and a satisfying result is achieved. The openpose-based methods unified the characteristics of poses and make the CNN model focus only on the postures.

4. Conclusion

High accuracy shows that CNN is really efficient in Yoga image classification. Also, the belief that CNN perceived everything, relevant or not, in the image, is verified since the skeleton images have better accuracy in testing. This result demonstrates that, extracting the major bone structure, the skeleton image, removing the background and colors, is a better choice of input dataset. As a light version of classification model, Mobile net shows remarkable performance using only 10 epochs training while the model size is small. Reducing the calculation and size of the model, Mobile net can be applied on small single chip microcomputer equipment that can help people who is doing Yoga adjust their pose. However, the cost on extracting the skeleton image by using openpose is still relative massive considering the low consummation of device of mobile net. In further research, some other method of extracting skeleton graph should be tested.

References

[1] Nasiri, A., et al. "Pose estimation-based lameness recognition in broiler using CNN-LSTM network." Computers and Electronics in Agriculture 197 (2022): 106931.

[2] Cassinis, L., et al. "On-ground validation of a CNN-based monocular pose estimation system for uncooperative spacecraft: Bridging domain shift in rendezvous scenarios." Acta Astronautica 196 (2022): 123-138.

[3] Garg, S., "Yoga pose classification: a CNN and MediaPipe inspired deep learning approach for real-world application." Journal of Ambient Intelligence and Humanized Computing (2022): 1-12.

[4] Qiu, Y., et al. "Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training." Biomedical Signal Processing and Control 72 (2022): 103323.

[5] Girish, D., et al. "Understanding action recognition in still images." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020

[6] Simonyan, K. et al. "Very deep convolutional networks for large-scale image recognition.", arXiv preprint arXiv:1409.1556 (2014).

[7] Osokin, D., "Real-time 2d multi-person pose estimation on cpu: Lightweight openpose." arXiv preprint arXiv:1811.12004 (2018).

[8] O'Shea, K., "An introduction to convolutional neural networks." arXiv preprint arXiv:1511.08458 (2015).

[9] Albawi, S., "Understanding of a convolutional neural network." 2017 international conference on engineering and technology (ICET). Ieee, 2017.

[10] Gu, J., et al. "Recent advances in convolutional neural networks." Pattern recognition 77 (2018): 354-377.