The Advisable Technology of Key-Point Detection and Expression Recognition for an Intelligent Class System

Yusheng Zhao¹, a, Haotian Yan¹, b, Zhaoqing Wang², *
¹International School, Beijing University of Posts and Telecommunications Beijing, China
²School of Information & Communication Engineering, Beijing Information Science & Technology University, Beijing, China
*Corresponding author email: mj741561@163.com
zhaoyusheng@bupt.edu.cn; yanhaotian@bupt.edu.cn

Abstract The intelligent classroom system has a very wide range of application scenarios in modern education. With the continuous breakthrough of artificial intelligence technology, a real-time intelligent system that can judge students’ performance in class has a large demand market. In this paper, the real-time detection technology needed by the smart classroom system is studied from the aspects of behavior detection and expression recognition. In terms of behavior detection, this paper combines the two key point detection technologies CPM (Convolutional Pose Machines) and CMU (Carnegie Mellon University) OPENPOSE [1] [2]. In terms of expression recognition, in order to detect the position of the face in real time, this paper studies several popular target detection algorithms and improves the traditional CNN network, and proposes a real-time image classification network architecture [8].

1. Introduction
In the key point detection part, this paper analyzes the popular CPM (Convolutional Pose Machines) and CMU (Carnegie Mellon University) OPENPOSE algorithms, and describes and compares the main structures of the two algorithms. Finally, combined with the key point detection for the significance of this application scenario, we chose the CMU OPENSE algorithm for further exploration. In the expression detection section, based on the idea of the target detection algorithm RCNN [10]. In this paper, we use the deep learning method Single Shot MultiBox Detector (SSD) algorithm to select the bounding box, and then use our modified CNN to classify the selected faces [14]. Figure 1 shows the recognition results of combining the two algorithms. This article will elaborate on these two parts.
In Figure 1, the left column is the original image, and the right column is the result of the recognition. The output includes the structure of each human key point and its expression. The upper and lower lines correspond to the single-person detection and multi-person detection of the algorithm.

2. Literature review

In the key detection part, there are two main directions, one is top-down and the other is bottom-up. The top-up human bone key point localization algorithm mainly consists of two parts, human detection and the human body key point detection. That is, firstly, each person is detected by the target detection algorithm, and then the human skeleton key point detection is performed for a single person based on the detection frame, wherein the representative algorithms are RMPE, Mask R-CNN, etc. [5][6]. The bottom-up approach also consists of two parts, key point detection and key point clustering. That is, first all the key points in the picture need to be detected, and then all the key points are determined by the relevant strategy clustering into different individuals. The representative algorithms for modeling the relationship between key points is Associative Embedding [7]. The algorithms CPM and CMU OPENSE used in this paper are all bottom-up methods.

In 2016, Google proposed a new CNN, Xception, which greatly reduced the amount of computation of the CNN convolutional layer [9]. In 2014, RGB proposed the target detection algorithm RCNN. Its idea is to extract the region proposal first, and then use CNN to classify the regional proposal. Based on this idea, a faster target detection algorithm such as Fast RCNN and Faster RCNN is generated [11][12]. In 2016, RGB proposed YOLO with faster response, but its monitoring accuracy for small objects is not high [13]. In 2017, Weiliu et al. proposed a new target detection algorithm based on regression theory, Single shot detector, which has very fast response and high monitoring accuracy for small objects [14].
3. Key-point detection

3.1 Convolutional Pose Machines

CPM takes advantage of both the deep convolutional network and the spatial modeling of the Pose Machine framework [3]. The basic structure of the Pose Machine is shown in Figure 2. It was proposed by Shih-En Wei in 2016[1].

CPM uses Pose Machine framework to replace the original classifier with a full convolution layer, as shown in Figure 3[4].

When stage \( t = 1 \), CPM predicts joint points according to local image evidence. Using local information, local means that the acceptance field of the network is constrained to the local part of the output pixel value. Picture block. The input image passes through the full convolution network and outputs the predicted results of the joint points.

When stage \( t = 2 \), the input to the classifier includes: the original picture feature, the previous stage’s beliefs for each joint point, and the Gaussian center constraint of the generated center (used to bring the response back to the image center, handling multiple characters in the image. The situation, to achieve multi-objective attitude estimation).

When stage \( t > 2 \), the input of the classifier no longer includes the original picture feature, but is replaced by the convolution result of the previous layer. The other inputs are the same as \( t = 2 \). It is also three inputs.

Finally, through multiple iterations, the estimation of the relevant nodes is realized. However, the connection between the joint points cannot be achieved.

3.2 CMU OPENPOSE

The main idea of CMU OPENPOSE is to convolve separately from two branches, as shown in Figure 4.

Branch 1 (Part Confidence Maps): Find all joint points, there are two parts. The first part is stage \( t = 1 \),
which accepts the input as the original image, and then outputs confidence maps to each joint point according to the classifier. The second part is stage \( t \geq 2 \), the input it accepts is all the confidence maps and the original image obtained by the previous stage, and its output is still confidence maps. Looping until convergence.

Branch2 (Part Affinity Fields): Its structure is basically the same as the first branch. The only difference is that each of the confidence maps of the output contains a certain type of connection (which can be simply understood as a bone).

The core of this algorithm is to successfully represent these connections as joints of joints, thus realizing the basic structure of the human body. As shown in Figure 5.

![Figure 5](image.png)

Figure 5. The output of CMU OPENPOSE [2]

4. A real-time expression detection system with a few parameters

4.1. A deep learning method based on regression for face detection

RCNN, an object detection method, has achieved unprecedented accuracy, but because it does not meet the real-time characteristics requirements, we use the SSD method to detect human faces [10][14]. The SSD is applied to a neural network and generates a series of bounding boxes for each feature map cell on different levels of feature maps, and then uses the IoU algorithm to find the bound closest to ground truth. The response of SSD is very fast, which can achieve the mAP of 73.1% with a processing speed of 58 frames per second [14]. The efficiency of SSD is ample for real-time detection.

4.2. Modified CNN for real-time classification

The traditional CNN network generates many parameters during training. Due to the increasing in computation, CNN has a serious delay in classifying expressions. Reducing the parameters is one of the main ideas of accelerating the network, aiming to recognize facial expressions in real time on devices without high performance. Because the fully-connected layer produces many parameters, the fully-connected layer of the CNN should be considered for removal. This paper uses a Depth-wise Separable Convolution (DSC) combined with the residue parts of CNN (Figure 6). The DSC is divided into two phases. Suppose the convolution kernel’s size is \( Dn \cdot Dn \) and the input is \( M \) feature maps. In the depth-wise convolution phase, \( M \) convolution kernels perform one-to-one convolution with \( M \) feature maps to generate \( M \) results. In the point-wise convolution phase, convolution operations are performed with \( N \) convolution kernels whose size is \( 1 \cdot 1 \) and \( M \) results from the last phrase, so \( N \) results are generated, namely, the input of \( M \) channels are converted into the outputs of \( N \) channels [9]. Fig6 shows the whole network structure, and through the comparison we found that DSC greatly reduces parameters and computation.

The improved network (Figure 7) consists of four layers of DSCs, with a batch normalization and a ReLU function sequenced after each convolutional layer. Additionally, an average pooling layer and a softmax function are placed at the end of the modified network. Finally, there are almost 60,000 parameters included in the model, which is 80 times smaller than the original CNN.
Figure 6. The number of convolution operations of CNN is $DK \cdot DK \cdot M \cdot N \cdot DF \cdot DF$, and DSC is $DK \cdot DK \cdot M \cdot DF \cdot DF + M \cdot N \cdot DF \cdot DF$. The CNN convolution kernel parameter is $DK \cdot DK \cdot N \cdot M$, and the convolution kernel parameter is $DK \cdot DK \cdot M+N \cdot M$.

Figure 7. The improved network

Figure 8. This matrix shows the accuracy of the classification

5. Conclusion

This paper studies two important techniques applied to the smart class system, key point detection and expression recognition. We compared two key point detection algorithms, CPM and CMU OPENSE, because CMU OPENSE is more suitable for multi-person scenes, and can detect key points of hands. CMU OPENSE is more suitable for building the behavior detection module of smart class system. For the traditional CNN, based on the idea of reducing the parameters and the computation time, we rebuild a network that can recognize the expression in real time by deleting the full connection layer, using depth-seperable convolution and detecting the face areas with SSD.

Reference

[1] Wei, S. E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4724-4732).

[2] Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2016). Realtime multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1611.08050.

[3] Ramakrishna, V., Munoz, D., Hebert, M., Bagnell, J. A., & Sheikh, Y. (2014, September). Pose machines: Articulated pose estimation via inference machines. In European Conference on
Computer Vision (pp. 33-47). Springer, Cham.

[4] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[5] Fang, H., Xie, S., Tai, Y. W., & Lu, C. (2017, October). Rmpe: Regional multi-person pose estimation. In The IEEE International Conference on Computer Vision (ICCV) (Vol. 2).

[6] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017, October). Mask r-cnn. In Computer Vision (ICCV), 2017 IEEE International Conference on (pp. 2980-2988). IEEE.

[7] Newell, A., Huang, Z., & Deng, J. (2017). Associative embedding: End-to-end learning for joint detection and grouping. In Advances in Neural Information Processing Systems (pp. 2277-2287).

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[9] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. arXiv preprint, 1610-02357.

[10] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[11] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[12] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[13] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[14] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.