A review of image feature extraction for object recognition

Steven Huang
Student ID: 19050155
Tutor: Dave Parry
Auckland University of Technology

Abstract

The computer visual system that imitates the human brain to realize the recognition and understanding of digital images is an important component of artificial intelligence. People have been working hard to research algorithms or models for image recognition and the key to this task is object feature extraction. This article will review the development of the image representation and feature extraction algorithms and will focus on the advantages and issues of the convolutional neural network (CNN) model.

1 Introduction

Attempts to understand intelligence and build artificial intelligence (AI) entities that can think and act reasonably as humans have never stopped. This effort is accompanied by the development of computers. Computer capabilities that have the ability to handle the following tasks: natural language processing (NLP), knowledge representation, machine learning, automatic reasoning, computer vision (CV), robotics, etc. can be considered AI [1] . The recognition and understanding of vision, speech, and natural language is probably the most important component of AI. Among them, the visual ability is a very important element of artificial intelligence, especially in the field of robotics or autonomous driving, etc. Nowadays, many intelligent CV applications have gradually entered various industries, such as healthcare, intelligence surveillance, transportation, agriculture, etc and have profoundly started to affect all aspects of people's lives. For image or video capture, today's advanced optical equipment can capture ultra-high resolution and even exceeds that of human eyes. However, it is difficult for computers to recognize and perceive the content in these digital images or videos like humans. At the same time, people's pursuit of the realization of automatic and intelligent computer vision has never stopped.

Early digital image recognition was based on knowledge of pixels in different regions or objects. In order to achieve segmentation of objects or regions in an image, most of the early algorithms are based on pixel-level cognition. That is, similar pixels may be belongs to the same region or object, and pixels with larger differences may be the boundary between the different object. The classic two segmentation method, Otsu's threshold algorithm [2] is to obtain the optimal threshold by exhaustively counting the inner and intro distance between the two classes of foreground and background under different thresholds. This pixel clustering algorithm has better performance when the foreground and background scenes are simple and there is a big difference between the foreground and the background. Other algorithms based on edge detection take advantage of the feature that pixels tend to have a sharp adjustment at object boundaries. The Canny edge detector [3] based on change calculation and intensity gradient has a good effect on digital image object edge detection. However, the versatility and robustness of object detection based on edge detection is poor because it requires a clearer edge and connectivity of the object to have a good detection effect. Other pixel clustering methods, Region-growing or Split-and-merge methods mainly rely on the assumption that adjacent pixels in one region or belongs to one object have similar values. In brief, these algorithms based on pixel-level cognition are straightforward and simple and perform well in recognition of special situation images.

Subsequent algorithms try to detect basic shape features in digital images for object recognition by selecting appropriate filters. This selection of filters based on object' basic shapes is obviously more refined than just based on pixel-level clustering. Basic edge detection operators such as Prewitt, Sobel, Roberts cross [4], etc. use differential to calculate the image gradient at common angles for edge recognition of objects. These artificially defined difference operators or filters can effectively detect pixel gradients in different directions, including horizontal, vertical, and specific diagonal directions, and this provides us with ideas for detecting the basic shape of objects. Harris and Stephens [5] proposed an algorithm that based on auto-correlation function can effectively detect cone and edge, which is helpful to understand the unconstrained real-world image.

In addition, blob detection algorithms based on differential or local extrema such as Laplacian of Gaussian (LoG), Difference of Gaussians (DoG) are often used in digital image object detection to enhance features. Crowley and Parker [6] proposed a DoG-based method to detect peaks and ridges for obtaining a multiple-scale shape representation of an image, adapting to different sizes, directions, or positions. Lowe [7] pointed out that local shape features in the digital images are invariant to translation, scaling, distortion, rotation, and are partially steady to illumination changes, affine, or 3D projection, this similar to the object recognition function of human visual cortex neurons. And based on the above theory, Lowe developed the Scale Invariant Feature Transform (SIFT) method for image feature extraction to effectively locate and match objects. In terms of face recognition applications, the model proposed by Papageorgiou et al. [8] uses three basic horizontal vertical and diagonal Haar wavelet filters to extract features before inputting the classifier, which has achieved success. Viola and Jones [9] based on Haar features, using the AdaBoost algorithm and an improved cascade classifier framework, making face detection achieve real-time performance. Overall, the filter-based shape feature extraction is significantly more calculation efficient than only relying on pixel classification or clustering method, and it also conforms to the characteristics of human vision, and the robustness and accuracy are improved to a certain extent.

Machine learning methods based on statistical theory were widely used in the object recognition task. Unlike the previous method, which usually only relies on a single image, statistical models are based on more samples. Although this will improve accuracy, the feature extraction of image objects still relies on manually selected filters. The advantage of the statistical model is that after learning from a large number of samples and multiple artificially selected features, the importance of different features will be obtained. The different importance means that the effectiveness of different filters for certain object recognition and this mixing will reduce accidental errors than using only a single fixed filter for the image. Lin et al. [10] demonstrated the association with local binary patterns (LBP) and histograms of oriented gradients(HOG) as a feature extractor to train an SVM classifier that can support 1000 classes, and the accuracy reaches 52.9 percent on the large-scale ImageNet dataset [11][12][13]. Grauman and Darrell [14] proposed a linear pyramid match kernels

method is suitable for object recognition and classification relay on pyramid multi-resolution image histograms. Lazebnik et al. [15] showed a model of pyramid matching kernel and SIFT descriptor combined with SVM for feature learning and screening, showing a significant improvement in performance on challenging scene classification tasks. Encouraged by the experiment of Lazebnik et al. [15], combined with HOG feature extractor and sliding window technology, object classification, and localization are realized [16]. In short, the statistical machine learning model can automotive evaluate and select key feature filter extractors from multiple features, which increases robustness and shows a certain degree of intelligence.

Most of the current outstanding and popular object recognition models are mainly based on CNN architecture. The convolution kernel used for convolution operations is not much different from the filter used for feature extraction in the previous model. Convolution parameters are completely learned without pre-processing or manual selection of some feature extraction filters. The key to CNN's superior ability in feature extraction lies in the adjustment of convolution parameters according to the supervision direction and the automation and optimization of this adjustment. In addition, the supervised CNN model relies on artificially annotated object labels (such as categories, bounding box positions, segmentation masks) to learn and adjust parameters, rather than propose appropriate filters based on the artificial summary of the object. This greatly increases the robustness for identifying objects with special characteristics (e.g. zebra, dairy cows, color texture animal or objects, etc.) compare to traditional algorithms. Inspired by the visual nervous system revealed by Hubel and Wiesel, Fukushima and Miyake [17] proposed a hierarchical cascaded neuron network architecture, which became the original CNN model. On the other hand, the neural network models with backpropagation (BP) algorithm enables feature extraction (or the convolution kernel parameters) to be automatically adjusted, so there is no need to input the pre-obtained feature vectors but directly the original image in the object recognition applications. LeCun et al. [18] showed a neural network model applied the BP algorithm, resulting in an error rate of only one percent, and this real-world application prove the power of the BP algorithm. The ability to automatically adjust the parameters brought by the BP algorithm allows us to see that through large numbers of sample training, the neural network model has the learning ability. Besides, Russakovsky et al. [19] pointed out that the error rate of image classification from the ImageNet Large-scale Visual Recognition Challenge (ILSVRC)10 to ILSVRC14 is decreasing year by year, the most significant decline occurred in the year 2012, with the best error rate result dropping from 25.8 percent in the previous year to 16.4 percent. In ILSVRC11, the winner used traditional SIFT descriptors as the image feature extractor plus linear SVM classifier framework for image classification [20]. However, Krizhevsky et al. [21] demonstrated a deep neural network model with the BP algorithm reduces the error to 16.4 percent, even far ahead of the second place by 26.2 percent in image classification tasks. This phenomenon illustrates the great potential of the CNN model in image object recognition, and it starts to attract more attention in the computer vision and deep learning field. Since then, the CNN model has become the prevailing model in various computer vision basic tasks. In general, the automatic and powerful feature extraction capabilities of the CNN model come from the BP algorithm and the guarantee of the supervised optimization direction for minimizing the loss function.

2 CNN achievements

With further development, the CNN model also has many improvements and extensions. The emergence and combination of various functional layers, such as convolutional layers, pooling layers, fully connected layers, and different activation layers, etc, make CNN efficient and more accurate in feature compression extraction. Besides, the layers of CNN model is getting deeper and deeper, which effectively improves the detailed feature extraction ability of high-resolution images. AlexNet [21], VGGNet [22], GoogLeNet [23] were all winners of the year in ILSVRC image classification and object location detection. These three more and more deep layer CNN networks with excellent object detection capabilities have become the backbone architecture of many applications. Although the deep CNN model has advantages in solving high-resolution images and extracting detailed features, it also brings gradients vanishing problems. He et al. [24] revealed this problem and proposed a residual learning network called ResNets, which effectively solved the problem of deep network degradation. Subsequently, because of the excellent feature extraction capabilities of the deep CNN model, good results have been obtained in a variety of basic CV tasks, such as image classification, object detection, image segmentation, motion analysis, scene reconstruction, etc.

Despite the performance of the above-mentioned deep CNN model has been greatly improved in basic image classification or target detection, these models are only a rough classification of the entire single image or the detection of a single object. How to effectively detect multiple objects of different categories and locate them in an image is still needs a lot of effort. Girshick [25] showed that the R-CNN model that first extracts region proposals and then sent them to the CNN model for classification can simultaneously detect and locate multiple objects in one image. Girshick [26] proposed Fast R-CNN, by using CNN to realize the region proposal in the R-CNN model, thereby improving the calculation speed. Ren et al. [27] present a regional proposal network (RPN), which can estimate the category and location of the object at the same time through the classifier and regressor, which reaches an outstanding performance. Although the R-CNN series models have good detection performance, the multi-step architecture makes the computational complexity of the model very high. Unlike previous models, the You Only Look Once (YOLO) model uses a one-stage unified CNN to directly predict the bounding box and class probability, which is the real-time object detection [28]. Another one-stage detection model, Single Shot MultiBox Detector (SSD) utilizes multiple different scale convolution feature maps to jointly predict the position and object class, trade-off part of the calculation and increasing the accuracy [29]. The RefineDet which uses the anchor optimization module, which greatly reduces the bounding-box search space and improves the accuracy, so as to have the efficiency of the one-stage model and the accuracy of the two-stage model [30]. The one-stage model may be the trend of efficient multi-object recognition and location and reducing the amount of calculation will be the improvement direction.

On the other hand, the CNN model also has a good performance in image segmentation and recognition. Compared with object bounding box detection, image segmentation needs to detect different objects and their more precise boundaries. The goal of image segmentation is to allot a label to each pixel to simplify the image representation to easily obtain or analyse meaningful content. This is also another effective way for object detection or understanding of image content, and image segmentation can be regarded as a pixel classification or clustering problem. Image segmentation is generally divided into semantic segmentation and instance segmentation. Long et al. [31] developed an end-to-end fully convolutional deep CNN model, which can effectively predict pixel-wise classification and semantic segmentation images. Another Encoder-Decoder based segmentation models are to use a CNN to extract key information and then connect to a deconvolution network to predict the segmentation image. Noh et al. [32] show a model that associates with a VGG back-end encoder network and a deconvolution decoder network that can directly output the segmentation mask image, and this network achieved the highest accuracy (72.5 percent) in the PASCAL VOC 2012 dataset segmen-

tation challenge. Ronneberger et al. [33] proposed another architecture encoder-decoder image segmentation model, U-Net, which is very suitable for medical image segmentation. The U-Net model saves the information on multi-scale feature maps, reduces information loss, and reduces the training dependence on lots of samples. Hence, it is very suitable for medical image segmentation where it is always not easy to obtain enough samples. In terms of instance segmentation, there are also many excellent CNN architectures emerging. He et al. [34] present a multi-instance object segmentation model Mask-RCNN, which was inspired by the Faster R-CNN model for multi-object bounding box detection and segmentation mask prediction. This model only adds a branch based on the R-CNN model, which can effectively detect objects' bounding-box and simultaneously predict a segmentation mask for each object. The Mask R-CNN model performance far exceeds other models in the COCO 2016 object detection challenge, it has achieved promising results at the object' instance-level recognition. Subsequently, inspired by Mask-RCNN, many new instance segmentation frameworks were proposed, such as R-FCN, SharpMask, DeepMask, PolarMask, etc [35]. These CNN-based deep learning frameworks have an impressive performance on instance segmentation tasks, and the precise object segmentation mask makes the understanding of the image further than bounding-box detection.

3 CNN disadvantages

Although the excellent performance in various applications reveals the unlimited potential of intelligent computer vision capabilities, the CNN model also has some problems worthy of attention. The main problems of CNN are high computational complexity, the large samples dependence, and the correct spatial relationship constraint of each part of the object.

First, the architecture of the CNN model is complex and has numerous parameters. Therefore, training a model requires a lot of calculations, and usually requires hardware with powerful parallel computing capabilities (e.g. GPU). Second, the CNN model needs plenty samples to drive to obtain more accurate results. Due to various limitations, it is usually impossible to obtain adequate samples, which may lead to uncertain prediction results. Not only that, the sample needs to be broad enough to cover as many scenarios as possible to avoid biased results.

In addition, the CNN model may rely more on the small local features of the object, rather than the global object shape. After the image has undergone multiple deep convolution operations, the detailed features are retained well, but the spatiotemporal relationship between local part has been broken up, and the overall shape or other global representation information would lose. Geirhos et al. [36] stated that the CNN model combines various small detailed shape (low-level features) extracted by the convolutional layers to predict a category. Therefore, it is easy for the CNN model to notice minute details and ignore global features, which may lead to recognition errors. For example, the spatial relationship of various parts of a human face image will be disrupted after convolution operations. If a face image with transposition of facial organs is deliberately sent to the CNN model, it will still be incorrectly recognized as a face. These features of the CNN model also make the recognition ability of overlapping objects very poor. Jaderberg et al. [37] introduced the CNN model with the Spatial Transformer, which can learn the invariance to translation, scaling, distortion, rotation, and general deformations, and obtaining the most advanced performance on multiple benchmarks. Hinton et al. [38] also proposed the Capsule Networks, which encapsulates the neural representing the different attributes of the same object into a logical group for learning, which greatly reduces the error rate compared to other models in 3D object recognition task on smallNORB [39] dataset. Although the aforementioned two models have tried to solve these shortcomings, these problems may require further research to overcome in the future.

4 Conclusion

In summary, with the support of the BP algorithm, the CNN model can obtain excellent feature extraction capabilities in digital images through amounts of samples. The CNN model is good at image object recognition, location, or segmentation tasks due to its powerful feature extraction capabilities. This automatic acquisition and learning ability of the feature extractor for images show computer vision intelligence like a human learn from what he/she sees but does not know. On the other hand, the CNN model is far less intelligent than humans, it cannot have a comprehensive evaluation for context variant images (e.g. spatially variant, illumination change, different shooting angles, deformation, and texture, etc) due to its own limitations.

References

[1] S. Russell and P. Norvig, "Artificial intelligence: a modern approach," 2002.

[2] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[3] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[4] R. C. Gonzales and R. E. Woods, "Digital image processing," 2002.

[5] C. G. Harris, M. Stephens, *et al.*, "A combined corner and edge detector.," in *Alvey vision conference*, vol. 15, pp. 10–5244, Citeseer, 1988.

[6] J. L. Crowley and A. C. Parker, "A representation for shape based on peaks and ridges in the difference of low-pass transform," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 156–170, 1984.

[7] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, pp. 1150–1157, Ieee, 1999.

[8] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pp. 555–562, IEEE, 1998.

[9] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[10] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang, "Large-scale image classification: Fast feature extraction and svm training," in *CVPR 2011*, pp. 1689–1696, IEEE, 2011.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 886–893, IEEE, 2005.

[12] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *2009 IEEE 12th international conference on computer vision*, pp. 32–39, IEEE, 2009.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[14] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2, pp. 1458–1465, IEEE, 2005.

[15] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 2169–2178, IEEE, 2006.

[16] H. Harzallah, F. Jurie, and C. Schmid, "Combining efficient object localization and image classification," in *2009 IEEE 12th international conference on computer vision*, pp. 237–244, IEEE, 2009.

[17] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and cooperation in neural nets*, pp. 267–285, Springer, 1982.

[18] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," *Advances in neural information processing systems*, vol. 2, pp. 396–404, 1989.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge,"

[20] J. Sánchez and F. Perronnin, "High-dimensional signature compression for large-scale image classification," in *CVPR 2011*, pp. 1665–1672, IEEE, 2011.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "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, 2014.

[26] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[30] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4203–4212, 2018.

[31] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[32] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, pp. 1520–1528, 2015.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[34] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[35] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *arXiv preprint arXiv:2001.05566*, 2020.

[36] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," *arXiv preprint arXiv:1811.12231*, 2018.

[37] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, pp. 2017–2025, 2015.

[38] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with em routing," in *International conference on learning representations*, 2018.

[39] Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2, pp. II–104, IEEE, 2004.