Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/32/20230178

Review of object tracking algorithms in computer vision based on deep learning

Xiao Luo

College of Computer and Network Security (Oxford Brookes College), Chengdu University of Technology Chengdu Sichuan Province 610059 China

luo.xiao@student.zy.cdut.edu.cn

Abstract. This paper is a survey of object tracking algorithms in computer vision based on deep learning. The author first introduces the importance and application of computer vision in the field of artificial intelligence, and describes the research background and definition of computer vision, and Outlines its broad role in fields such as autonomous driving. It then discusses various supporting techniques for computer vision, including correcting linear unit nonlinearities, overlap pooling, image recognition based on semi-naive Bayesian classification, human action recognition and tracking based on S-D model, and object tracking algorithms based on convolutional neural networks and particle filters. It also addresses computer vision challenges such as building deeper convolutional neural networks and handling large datasets. We discuss solutions to these challenges, including the use of activation functions, regularization, and data preprocessing, among others. Finally, we discuss the future directions of computer vision, such as deep learning, reinforcement learning, 3D vision and scene understanding. Overall, this paper highlights the importance of computer vision in artificial intelligence and its potential applications in various fields.

Keyword: Computer Vision, Target Tracking Algorithm, Convolutional Neural Networ.

1. Introduction

Computer vision is an important research direction in artificial intelligence field. Its goal is to enable computers to perceive and understand visual information, enabling automated analysis and understanding of highlight and video data. Through computer vision technology, computer can extract meaningful information from images and videos and perform advanced visual tasks, such as target detection, image segmentation, and pose estimation. Computer vision has gone through multiple stages of research and development, form edge detection and object recognition algorithms to deep learning and convolutional neural networks, and its capabilities have been significantly improved, enabling computer to automatically process image and video data, reducing labor costs and improving work efficiency. Compute vision is widely used in many fields. For example, in the field of autonomous driving, computer vision can identify and track vehicles and pedestrians on the road to assist the realization of intelligent driving systems. It plays an important role in face recognition and medical imaging fields.

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/32/20230178

2. Supporting technology

2.1. Rectified Linear Unit Nonlinearity

In machine learning and neural networks, an activation function is a function that maps an input to an output to introduce nonlinear properties. Modified linear unit (ReLU) is a commonly used activation function. ReLU:$f(x) = \max(0, x)$. Where x is the input and f(x) are the output, the function has the following characteristics: when x is greater than or equal to zero, the output is equal to the input f(x)=x, that is, the function remains linear. When x is less than zero, the output is zero. Moreover, ReLU function has three characteristics: fast training, avoiding gradient disappearance problem and sparse activation [1]. In deep convolutional networks, ReLU functions are usually used as nonlinear activation functions after convolutional layers, which can help networks better learn features and improve network performance. ReLU can also be used in combination with other technologies, such as dropout, to further improve the performance of deep neural networks [1].

2.2. Overlapping pooling

Overlapping pooling is a pooling operation used to reduce the spatial dimension in the convolutional neural network [1]. Different from common pooling operations, overlapping pooling allows pooling Windows to be partially overlapped with fixed steps, thus increasing sampling density. This allows the input to be sampled at a finer granularity, capturing more characteristic information.

2.3. Image recognition based on semi-naive Bayes classification

Here, the One-Dependence Estimator is usually adopted. As shown in the formula:

$$\mathrm{P}(\mathrm{Y}|\mathrm{X}) \propto \mathrm{P}(\mathrm{Y} = y_k) \prod_{j=1}^{n} P\ (X^{(j)} = x^{(j)}|\mathrm{Y} = c_k, \mathrm{p}x^{(j)}) \quad [2]$$

Consider random variables in the input space X and the output space Y. Where y is some value in the class tag collection y = $\{ y_1, y_2, \ldots . y_k \}$. $x^{(j)}$ is j^{th} first j a characteristic of samples. P (X|Y) represents the probability distribution of x given y.

2.4. Image human behavior recognition and tracking based on S-D model

Aiming at the problem that SNBC is not accurate enough to recognize human motion in moving images, an S-D algorithm combining DT and SNBC is proposed, and the corresponding S-D model is established. The basic structure of the S-D model is shown in the figure below.

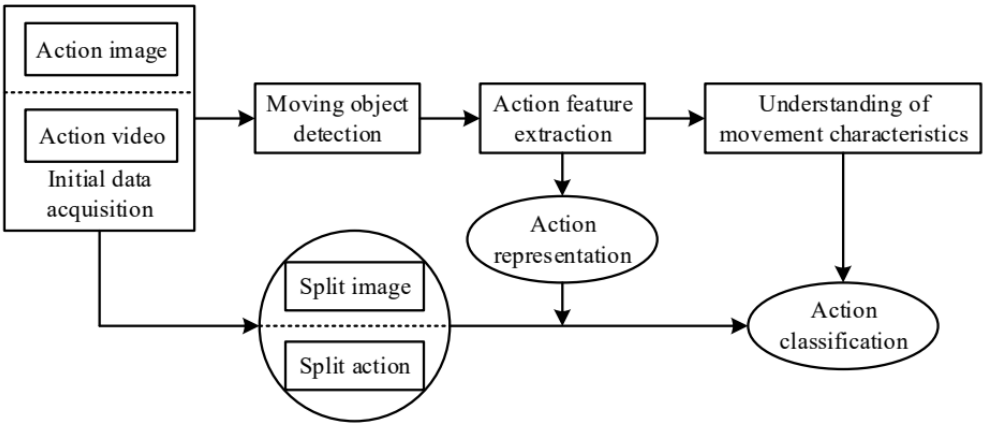

Figure 1. S-D model structure [2].

The S-D algorithm is used to calculate the optical flow information in the moving image and extract the movement trajectory of the athlete. Through feature extraction and processing of trajectory, SNBC is used for training and classification to realize recognition and tracking of human motion. For each

DOI: 10.54254/2755-2721/32/20230178

point in the image sequence, sampling points with smaller eigenvalues are deleted according to the eigenvalue threshold. The motion coordinates of the next frame are calculated by the median filter and the optical flow portion. The following formula can be used to obtain the human movement trajectory, so as to realize the movement recognition and tracking.

$$T = 0.001 * \max_{i \in I} min(\lambda_i^1, \lambda_i^2) \quad [2]$$

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega_t)|x_t, y_t \quad [2]$$

$$\omega_t = (u_t, v_t) \quad [2]$$

2.5. Object tracking algorithm based on convolutional neural network

Convolutional neural network is a layered model that can learn features directly from image original pixels [3]. Firstly, in the first stage, the 32*32 pre-processed black and white image is input into the convolution layer composed of 6 5*5 filters, and the feature map size is 28*28. Then, the nonlinear transformation is carried out by ReLU function. Next, use the maximum under sampling layer to select the maximum value in each 2*2 field. This network structure makes the network robust to micro translation. In the second stage, convolution and ReLU transformation are similarly carried out, and the advanced features are transformed into one-dimensional vectors by maximum subsampling feature mapping. Through these two stages of operation, the network is able to extract higher-level features.

2.6. Object tracking algorithm based on particle filter and convolutional neural network

Particle filter is a recursive Bayesian filtering algorithm, which uses sequential Monte Carlo important sampling method to represent the posterior probability. The core idea is to approximate a posterior probability distribution using a series of random particles. A particle filter has two main components: a state transition model that generates candidate samples based on previous particle samples. Observe the model and calculate the similarity between the candidate samples and the model of this objective view. A given observation sequence: $y_{1:t} = [y_1, \ldots\ldots, y_t]$, target tracking system's goal is to estimate the posterior probability density function of the target at time t $p(x_t|y_{1:t})$. According to Bayesian theory, the posterior probability density can be expressed as:

$$p(x_t|y_{1:t}) \propto p(y_t|x_t) \int p\,(x_t|x_{t-1}) p(x_{t-1}|y_{1:t-1}) dx_{t-1} \quad [3]$$

In the above formula, $p(x_t x_{t-1}), p(y_t x_t)$ are the dynamic model and the observation model, respectively. ssThe optimal target state x at the last time t can be obtained from the maximum posterior probability:

$$x_t^* = arg \max_{x_t} p\,(x_t y_{1:t}) = x_t^i = arg \max_{x_t^i} w_t^i \quad [3]$$

In order to improve the computational efficiency, the algorithm to choose only track the target position and size, $x_t = (p_t^x, p_t^y, w_t, h_t)$, in order to target the abscissa and ordinate, width and length. It is assumed that the dynamic model of two consecutive frames follows Gaussian distribution.

$$p(x_t x_{t-1}) = M(x_t; x_{t-1},) \quad [3]$$

2.7. Residual learning framework

The residual learning framework solves the problem of gradient disappearance and gradient explosion in deep neural network training [5]. It introduces a residual term so that the output of each layer includes residuals, that is, the difference between the current layer output and the input. This design makes the current layer output, and improves the model performance and generalization ability. Residuals are the key components, including convolution layer, batch normalization layer and activation function layer. Residuals are generated by transformation and added to input, which deepens the learning ability of the network [4].

DOI: 10.54254/2755-2721/32/20230178

3. Challenges and solutions

3.1. Construct a deeper convolutional neural network

Increasing network depth can improve accuracy, but building deeper overpasses requires consideration of training time, computational resources, and overfitting challenges [5]. Gradient disappearance and gradient explosion: In deep networks, gradients may gradually decrease or increase, resulting in problems such as gradient disappearance or gradient explosion, making the network difficult to train and converge. Increased computing and storage resource requirements: Deeper networks require more computing resources and storage space to handle more parameters and intermediate computation results, increasing the pressure on computing and storage. Overfitting problem: Deeper networks have stronger representation capabilities, but also tend to overfit training data and are difficult to generalize to new data samples. Activation functions (ReLU) are used to mitigate the gradient disappearance problem [1], regularization and random deactivation are used to reduce overfitting problems, and normalization is used to accelerate training and improve the stability of the network. As well as the use of residual links and attention mechanisms to help gradient propagation and optimize network structure.

3.2. Processing of large data sets

Large image databases such as ImageNet provide rich data resources, but processing these data is still challenging. Efficient algorithms and processing methods are needed to improve speed and accuracy [5]. Labeling data is expensive and time-consuming, and the dataset may be affected by class imbalance and data bias. In addition, processing big data requires powerful computing resources and storage space. In order to cope with these problems, it is necessary to effectively process data, solve class imbalance and bias, and use distributed computing to meet the processing requirements. Data storage management can be used in distributed storage systems or cloud services to increase reliability and access speed. Data preprocessing and cleaning can improve quality and reduce noise effects. Parallel computing breaks down tasks and leverages distributed computing frameworks to accelerate them. Virtual datasets can simulate real datasets by generating models, which can be used for data analysis, model training and testing. Large image databases such as ImageNet provide rich data resources, but processing these data is still challenging. Efficient algorithms and processing methods are needed to improve speed and accuracy [5]. Labeling data is expensive and time-consuming, and the dataset may be affected by class imbalance and data bias. In addition, processing big data requires powerful computing resources and storage space. In order to cope with these problems, it is necessary to effectively process data, solve class imbalance and bias, and use distributed computing to meet the processing requirements. Data storage management can be used in distributed storage systems or cloud services to increase reliability and access speed. Data preprocessing and cleaning can improve quality and reduce noise effects. Parallel computing breaks down tasks and leverages distributed computing frameworks to accelerate them. Virtual datasets can simulate real datasets by generating models, which can be used for data analysis, model training and testing [6].

4. Future development direction

The future of computer vision will focus on several key directions. In the future, computer vision will focus on several key directions: deep learning and neural network evolution, designing more complex structures to improve expression and generalization; Reinforcement learning and autonomous decision making through interaction with the environment; Realize adaptive learning under limited annotated data to quickly adapt to new fields or tasks [7].

The field of 3D vision and scene understanding will be further developed to enable accurate understanding and simulation of objects, people and actions in complex scenes. At the same time, more emphasis will be placed on multimodal vision, blending visual data from different sensors and information sources to achieve more comprehensive visual understanding and interaction. With the popularity of 3D vision systems, the design and optimization of deep learning models that process 3D data becomes important [8]. The article mentions the development of 3D CNNs (three-dimensional

DOI: 10.54254/2755-2721/32/20230178

Convolutional neural networks), and the application of geometric deep learning in computer graphics, robotics, video classification and other fields. Future research may focus on how to design more efficient 3D CNNs to apply deep learning to more complex 3D computer vision tasks. Robustness and privacy protection are becoming increasingly important, and research will focus on developing models and algorithms that can withstand adversarial attacks, as well as designing privacy-secure data processing and information transmission methods. However, differential privacy protection mechanisms can negatively affect model accuracy for unusual data or long-tail distributions [9]. Future research could delve into the impact of differential privacy on individual samples and propose solutions to improve the accuracy of the model on these data. Finally, the concept of lifelong learning will play an important role in the above direction. Models will continue to learn from new data and adapt to changing environments [10]. These common advances will advance the application of computer vision in areas such as autonomous driving, medical imaging, and intelligent safety. Therefore, computer vision will contribute strong visual perception and understanding capabilities to the development of artificial intelligence.

5. Conclusion

Computer vision is an important research direction in the field of artificial intelligence, which aims to enable computers to perceive and understand visual information. Through the introduction of technologies such as deep learning and neural networks, computer vision has made significant progress in image and video processing, and is widely used in multiple fields such as autonomous driving, face recognition, and medical imaging. In the future, the development of computer vision will focus on deep learning, reinforcement learning, three-dimensional vision and scene understanding. These developments will promote the application of computer vision technology in various fields, providing more powerful visual perception and understanding capabilities for the development of artificial intelligence.

References

[1] Krizhevsky, A, Sutskever, I and Hinton, G (2017) ImageNet Classification with Deep Convolutional Neural Networks Available at: ImageNet classification with deep convolutional neural networks | Communications of the ACM Access date: July 16, 2023

[2] Song, Y (2021) Research on Sports Image Recognition and Tracking Based on Computer Vision Technology Available at: Research on Sports Image Recognition and Tracking Based on Computer Vision Technology | IEEE Conference Publication | IEEE Xplore Access date: July 16, 2023

[3] Tian, Y and Cao, D (2022) Computer vision recognition and tracking algorithm based on convolutional neural network Available at: (PDF) Computer vision recognition and tracking algorithm based on convolutional neural network (researchgate.net) Access date: July 16, 2023

[4] He, K et.al (2015) Deep Residual Learning for Image Recognition Available at: [1512.03385] Deep Residual Learning for Image Recognition (arxiv.org) Access date: August 15, 2023

[5] Simonyan, K and Zisserman, A (2015) VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION Available at: [1409.1556] Very Deep Convolutional Networks for Large-Scale Image Recognition (arxiv.org) Access date: July 16, 2023

[6] Sun, S et.al (2020) The virtual training platform for computer vision Available at: The virtual training platform for computer vision | IEEE Conference Publication | IEEE Xplore Access date: August 16, 2023

[7] Zhang, Y , Wu, Y and Chen, H (2023) Research progress of visual simultaneous localization and mapping based on Deep learning Available at: Research progress of visual simultaneous localization and Mapping based on deep learning - CNKI (cdut.edu.cn) Access date: August 15, 2023

[8] Chen, X and Guo, H (2023) A Futures Quantitative Trading Strategy Based on a Deep Reinforcement Learning Algorithm Available at: A Futures Quantitative Trading Strategy

DOI: 10.54254/2755-2721/32/20230178

Based on a Deep Reinforcement Learning Algorithm | IEEE Conference Publication | IEEE Xplore Access date: August 15, 2023

[9] Shaqwi, F et. Al (2021) A Concise Review of Deep Learning Deployment in 3D Computer Vision Systems Available at: A Concise Review of Deep Learning Deployment in 3D Computer Vision Systems | IEEE Conference Publication | IEEE Xplore Access date: August 15, 2023

[10] Golatkar, A (2022) Mixed Differential Privacy in Computer Vision Available at: Mixed Differential Privacy in Computer Vision | IEEE Conference Publication | IEEE Xplore Access date: August 15, 2023