Research on Video Behaviour Recognition Methods Based on Deep Learning

Yan Liu[1], Haimin Zhang[2]

1. Yanjing Institute of Technology, Yanjing 224051, China.

2. Anhui Institute of Information Technology, Hefei 241000, China.

Abstract: Currently, in the field of computer vision, video behaviour recognition has become a hot content and has been applied in many emerging industries. This paper explores the application of deep learning method in video behaviour recognition technology by taking the multi-temporal information fusion recognition method as an example, and investigates the practical application value of this method by constructing a model of artificial neural network and taking an experimental approach, in order to have positive significance to the development of intelligent human-computer interaction technology in China.

Keywords: Video Behaviour; Deep Learning; Human Gesture

Introduction

Compared to static images, video behaviour recognition is mainly about recognising changes in consecutive frames of an image for a moving subject, by acquiring a sequence of images and thus using a computer to analyse and process them to recognise the behaviour of a moving subject in a video. Behaviour recognition technology is used in many fields and has significant research value. This paper explores more accurate video behaviour recognition methods based on deep learning algorithms.

1. The concept of video behaviour recognition

Human behaviour is complex, with single limb movements as well as movements of multiple body joints; it includes single person behaviour as well as interaction between two or more people, and also between people and objects [1]. The semantic information presented by different human behaviours varies greatly, and when developing intelligent HCI technologies, the different human behaviours need to be recognised in order to provide reliable support for the relevant technologies. Video Behaviour Recognition (VBR) is the process of learning from video captured by sensors through a model of computer vision algorithms to analyse the characteristics of behavioural activities and facilitate computer understanding and recognition of complex human behaviour. Video behaviour recognition mainly includes capturing data information and pre-processing, extracting and expressing behavioural features, and recognising and classifying human behaviour in video [2].

2. Deep learning based video behaviour recognition methods

2D convolution can extract the spatial information in video data extraction, but video behaviour recognition needs to extract the temporal dimension information in it based on spatial information, and 2D convolution cannot meet this need, so 3D convolution needs to be developed. When applying 3D convolution for video behaviour recognition, the extraction of video data features can be carried out in both temporal and spatial dimensions, and the behavioural features can be used effectively, thus improving the accuracy of video behaviour recognition [3].

2.1 Artificial Neural Networks

Artificial neural network (ANN) is a mathematical model developed on the basis of the neural network of observing organisms, which has strong learning ability and is widely used in machine learning.ANN is applied in the field of machine learning mainly by constructing network models through artificial neurons and using connection strength as a learning parameter, so as to achieve video behaviour recognition. Biological neural networks include structures such as neurons, contacts and cells, which enable organisms to perceive and judge things in the outside world. The specific workflow is: acquiring external signals, transmitting signals to neurons, activating signals, processing signals, and sending signals to the next level of neurons [4]. In the body of a living creature, the neurons are connected in a complex way, like a huge network structure, and the signals output from the neural network structure can help the creature to establish autonomous consciousness and make judgments about the signals from the outside world. The ANN is then used to integrate the input data with the built-in parameters, and to activate the data using the excitation function. In the specific calculation, the input data can be used as the input data and the weighted parameters, and the value after the weighting of the two can be expressed as Z. The activation function is, and the output signal at the moment t+1 is expressed as y(t+1). In the calculation, the artificial neural model is given by:

$$y(t+1) = f\left(Z = \sum_{i=1}^{n} \omega x_i(t) \right)$$

The artificial neural network is a topology composed of neurons at different levels, including an input layer, an output layer and a hidden layer, where the neurons in the input layer pass information about the received data into the neurons in the hidden layer, where the data are processed, where a single hidden layer is able to change the dimensional information of the data, where multiple hidden layers are able to extract the features of the data and transform the data features into semantic information, and where the output layer is able to output the data [5].

For an artificial neural network to be able to solve complex problems, an activation function needs to be applied to introduce non-linearities into the model data, which can be done through a Sigmoid activation function with the function equation:

$$g(z) = \frac{1}{1+e^{-z}}$$

When applying this activation function, the range of values can be transformed to (0,1), maintaining the original input data magnitude and reducing the difference, which is suitable for situations where the data difference is large.

2.2 Convolutional neural networks

Convolutional neural networks (CNNs) are feedforward neural networks with convolutional computation and depth structure, and are deep learning algorithms that can be used to directly input the original data information and automatically extract the data features, which is an obvious advantage in the application of video processing. After random initialisation, the convolutional kernel can be continuously adjusted to obtain the weight parameters. During the calculation, the convolutional kernel slides over the image to obtain the response of the image and enhance some of its features. When performing the calculation of the 3D convolution, the following equation can be used:

$$V_{ij}^{xyz} = f\left(\sum_{m} \sum_{l=0}^{Li-1} \sum_{w=0}^{Wi-1} \sum_{h=0}^{Hi-1} W^{lwh}{}_{i,j,m} \cdot V_{i-1,m}^{(x+l)(y+W)(z+h)} + b_{i,j} \right)$$

In the above equation, where f represents the activation function, m represents the feature map, Li and Wi represent the length and width of the convolution kernel, Hi represents the convolution kernel time length, W is represents the connection weights, and bi,j represents the feature map bias.

The pooling layer is mainly located between the convolutional layers, sampling the feature map, shrinking the input image and reducing the fit. In specific applications, mean pooling and maximum pooling can mainly be performed.

2.3 Deep learning framework

Deep learning frameworks are mainly used as tools and libraries in applications where the construction of deep learning models can be carried out. In general, deep learning frameworks such as Caffe, Tensorflow and Pytorch can be used for video behaviour recognition, Caffe can be written in C++ for modularity, speed and expression design, Tensorflow can be used for a large number of mathematical operations and can be deployed in a server for the purpose of video behaviour recognition. Pytorch framework can be used to build dynamic neural networks based on GPU acceleration, and has strong advantages in terms of simple code and easy debugging.

2.4 Video behaviour recognition method with multilayer spatio-temporal information fusion

When applying convolutional neural for data feature extraction, the input image will be reduced in resolution in the pooling operation. The multilayer spatio-temporal information fusion recognition method (MLSIFN) includes three modules: base skeleton, feature fusion and classification. In specific applications, the base skeleton module can generate spatio-temporal features, and after embedding the semantic features, the semantic information can be introduced from the high level to the low level spatio-temporal features, and then through feature fusion, the spatio-temporal information can be aggregated to improve the representational capability of the spatio-temporal features. The underlying skeleton network is constructed by setting the size of the 3D convolution and pooling layers, and the features are connected by 2 fully connected layer networks for dimensionality reduction. When performing semantic embedding, this can be computed by the following fusion algorithm:

$$H_l = Upsample(M_{l+1}) + M_l$$

Where Hl is the L-layer spatio-temporal features after embedding the semantics, and Ml+1 and Ml represent the spatio-temporal features with layers l+1 and l, respectively. The classification module can perform maximum pooling of spatio-temporal features, using the human body as the foreground, transforming the feature vectors through the fully connected layer, and then using the corresponding functions to recognise the data.

In the feature fusion experiments, the MLSIFN can be used to explore the high and low layer fusion strategies and to analyse the strategies for characterisation capability enhancement. In the process of conducting behavioural recognition feature extraction, the low-level features have more spatial detail information, the high-level features have more semantic information, and the mid-level features have weaker and more interfering semantic information. Therefore, when fusing, the staff adopt the strategy of fusing the low-level features with the high-level features.

Conclusion

In conclusion, this paper presents the construction and training methods of artificial neural network models, and in the process of video behaviour recognition, the multilayer spatio-temporal fusion behaviour recognition method is introduced. Data features are extracted through 3D convolutional neural networks, and then spatio-temporal features are aggregated using feature fusion and then embedded with semantics to achieve enhanced semantic representation by incorporating high-level semantics in the low-level feature information. Finally, comparative experiments are conducted to explore the effectiveness of the application of the multilayer spatio-temporal information fusion structure.

References

[1] Liu SS. Deep learning based RGB video human behavior recognition [D]. Shandong University, 2021.

[2] Zhou YY. Research on human abnormal behavior recognition method based on deep learning for surveillance video [D]. Henan University of Technology, 2021.

[3] Pu ZX, Ge YX. A small-sample video behavior recognition algorithm based on multi-feature fusion[J]. Journal of Computer Science,2023,46(03):594-608.

[4] Huang M, Shang RX, Qian HM. A composite deep neural network for human behavior recognition in video[J]. Pattern Recognition and Artificial Intelligence, 2022, 35(06): 562-570.

[5] Li C, He M, Wang Y, Luo L, Han W. A review of video behavior recognition techniques based on deep learning[J]. Computer Application Research, 2022, 39(09): 2561-2569.

Fund projects:

Digital Teaching and Research Special Project *Research on abnormal behavior recognition methods for smart campus video surveillance* (project no: 2021YITSRFSZ008)

About Authors:

Liu Yan, 1981, female, Han nationality, master's degree,main research direction: deep learning and big data analysis

Zhang Haimin, 1983, male, Han nationality, master's degree,main research direction: image processing and artificial intelligence