TARGET TRACKING IN COMPLEX SCENES BASED ON COMPUTER VISION

RASTREAMENTO DE ALVOS EM CENAS COMPLEXAS COM BASE NA VISÃO COMPUTADORIZADA

SEGUIMIENTO DE BLANCOS EN ESCENAS COMPLEJAS CON BASE EN LA VISIÓN COMPUTADORIZADA

Original Article
Artigo Original
Artículo Original

Huanan Shang[1]
(Public health)

1. Huang S&T College, Henan Zhengzhou,China.

Correspondence:
Huanan Shang
Henan Zhengzhou, China, 450006.
usfxt41506@163.com

ABSTRACT

Objective: Use the deep learning network model to identify key content in videos. Methodology: After reviewing the literature on computer vision, the feature extraction of the target video from the network using deep learning with the time-series data enhancement method was performed. The preprocessing method for data augmentation and Spatio-temporal feature extraction on the video based on LI3D network was explained. Accuracy rate, precision, and recall were used as indices. Results: The three indicators increased from 0.85, 0.88, and 0.84 to 0.89, 0.90, and 0.88, respectively. This shows that the LI3D network model maintains a high recall rate accompanied by high accuracy after data augmentation. The accuracy and loss function curves of the training phase show that the accuracy of the network is greatly improved compared to I3D. Conclusion: The experiment proves that the LI3D model is more stable and has faster convergence. By comparing the accuracy curve and loss function curve during LI3D, LI3D-LSTM, and LI3D-BiLSTM training, it is found that the LI3D-BiLSTM model converges faster. ***Level of evidence II; Therapeutic studies - investigation of treatment results***.

Keywords: Computers; Computer Vision Systems; Public Health.

RESUMO

Objetivo: Usar o modelo de rede de aprendizagem profunda para identificar o conteúdo-chave em vídeos. Metodologia: Após revisão da literatura sobre a visão computadorizada, efetuou-se a extração da característica do vídeo alvo da rede utilizando o aprendizado profundo com o método de melhoramento de dados em séries temporais. Foi explanado o método de pré-processamento para aumento de dados e extração da característica espaço-temporal no vídeo baseado na rede LI3D. Foram utilizados como índices a taxa de precisão, precisão e recall. Resultados: Os três indicadores aumentaram de 0,85, 0,88, e 0,84 para 0,89, 0,90, e 0,88, respectivamente. Isso mostra que após o aumento dos dados, o modelo de rede LI3D mantém uma alta taxa de recuperação acompanhada de uma alta precisão. As curvas de precisão e função de perda da fase de treinamento demonstram que a precisão da rede é muito melhorada em comparação com a I3D. Conclusão: O experimento prova que o modelo LI3D é mais estável e que a convergência é mais rápida. Ao comparar a curva de precisão e a curva de função de perda durante o treinamento LI3D, LI3D-LSTM e LI3D-BiLSTM, verifica-se que o modelo LI3D-BiLSTM converge mais rapidamente. ***Nível de evidência II; Estudos terapêuticos – investigação de resultados de tratamento***.

Descritores: *Computadores; Sistemas de Visão Computacional; Saúde Pública.*

RESUMEN

Objetivo: Utilizar el modelo de red de aprendizaje profundo para identificar el contenido clave en los vídeos. Metodología: Después de revisar la literatura sobre visión por ordenador, se realizó la extracción de características del vídeo objetivo de la red utilizando el aprendizaje profundo con el método de aumento de datos de series temporales. Se explicó el método de preprocesamiento para el aumento de datos y la extracción de características espacio-temporales en el vídeo basado en la red LI3D. Se utilizaron como índices la tasa de exactitud, la precisión y recall. Resultados: Los tres indicadores aumentaron de 0,85, 0,88 y 0,84 a 0,89, 0,90 y 0,88, respectivamente. Esto demuestra que el modelo de red LI3D mantiene un alto índice de recuperación acompañado de una alta precisión tras el aumento de datos. Las curvas de precisión y de función de pérdida de la fase de entrenamiento muestran que la precisión de la red mejora mucho en comparación con la I3D. Conclusión: El experimento demuestra que el modelo LI3D es más estable y tiene una convergencia más rápida. Al comparar la curva de precisión y la curva de función de pérdida durante el entrenamiento de LI3D, LI3D-LSTM y LI3D-BiLSTM, se observa que el modelo LI3D-BiLSTM converge más rápidamente. ***Nivel de evidencia II; Estudios terapéuticos – investigación de resultados de tratamiento***.

Descriptores: *Computadoras; Sistemas de Visión Computacional; Salud Pública.*

DOI: http://dx.doi.org/10.1590/1517-8692202228052021_0532
Article received on 12/11/2021 accepted on 12/22/2021

INTRODUCTION

Motion pattern recognition based on video content, a hot trend in the field of computer vision, has important research significance and application value in intelligent transportation, network security, medical care, nursing treatment and human-computer interaction.[1] By detecting and identifying video content, people can quickly access key information in video content and use this key information to help people solve many of the problems they face today.[2] The main research work is to use the deep learning network model to identify the key content in the video by identifying the content in the video. It mainly introduces from the aspects of video data preprocessing, data augmentation, feature extraction, feature aggregation and multimodal feature fusion.

Realizing human-computer intelligent interaction can promote the development of virtual reality; by extracting the motion information of the target object in video surveillance, the elevator can be dispatched reasonably, the elevator utilization rate can be improved, and the working time and work quota of traffic police or security personnel can be optimized.[3] With the rapid development of artificial intelligence, motion pattern recognition based on deep learning has become a research hotspot. The recognition rate of motion patterns in complex scenes has been greatly improved, which has laid the necessary foundation for the application of relevant research in practice Lei J et al.2017.[4]

When extracting video features, we must take into account the spatial and temporal dimensions, and if only the video frame level features are extracted and modeled, the natural association between the upper and lower frames of the video will be lost.[5] This isolated feature cannot completely represent the content of a video. By fine-tuning the I3D network structure, a lightweight video feature extraction network structure LI3D is proposed, and the video features are extracted by means of migration learning. The model can effectively extract spatial and temporal information in the video and is innovative.[6]

The study is divided into three parts: the first part is a literature review. The second part is based on the video target feature extraction of computer vision deep learning network, expounding the data augmentation preprocessing method and the video spatiotemporal feature extraction method based on LI3D network. The third part is the verification of the proposed method, and the experiment confirms the superiority of the method.

Video target feature extraction based on computer vision deep learning network

Data augmentation pretreatment

Data augmentation, one of the commonly used techniques in deep learning, is mainly used to increase the training data set and make the data set as diverse as possible, so that the trained model has stronger generalization ability, and at the same time, the data can contain as much feature information as possible, so that the trained model can have better robustness and accuracy. In the process of preprocessing the video data, different data augmentation strategies are processed for the spatiotemporal characteristics of the video data. In the time series, multi-time scale frame extraction processing is performed, which can effectively extract key frames for video content with different motion characteristics, so that the information of the input network can better represent the content of the video; In the spatial aspect, the video clipping processing of multi-space scale is carried out, and the specific content in the key frame image is extracted by three different clipping methods, which further expands the feature diversity of the data set, and more feature information can be obtained at the same time. (Figure 1 and 2)

The data enhancement method in time series makes the training data show diversity in timing, enabling the network to extract richer timing features. Similarly, in terms of spatial scale, the multi-space scale clipping process can make each frame of the input network have more spatial information, thereby extracting more spatial information and achieving data enhancement in space. The multi-spatial video clipping strategy used is shown in Figure 3, which are mainly three ways: center clipping, random clipping, and key frame scaling.

The specific processing is as follows: key frame scaling is the size required to directly scale the original image to the network, and scaling the image will enhance its smoothness and sharpness; center clipping is the size of the image required to cut out the network from the center of each key frame, and these key frames can get more important information about the center of the image; random cropping is a random cropping on key frame images to get an image that fits the network input size. Through the above three ways of data augmentation, the convergence speed of the network can be effectively improved, and the generalization ability of the network can be enhanced.

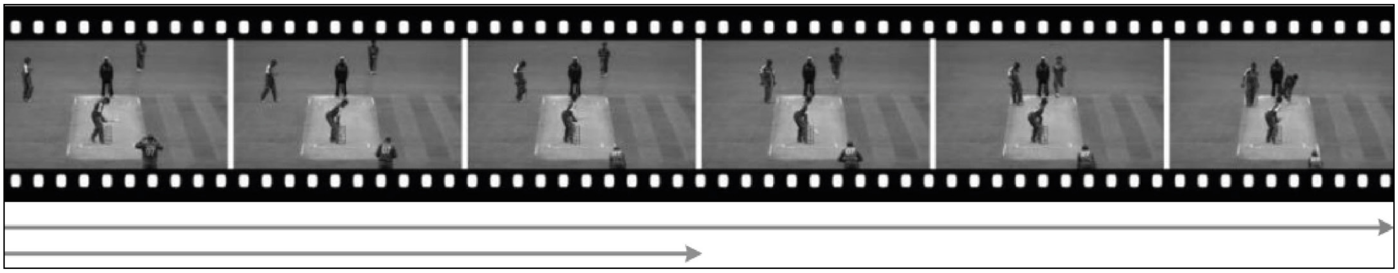

Figure 1. Shorter Sequence Frame Extraction Strategy Diagram.

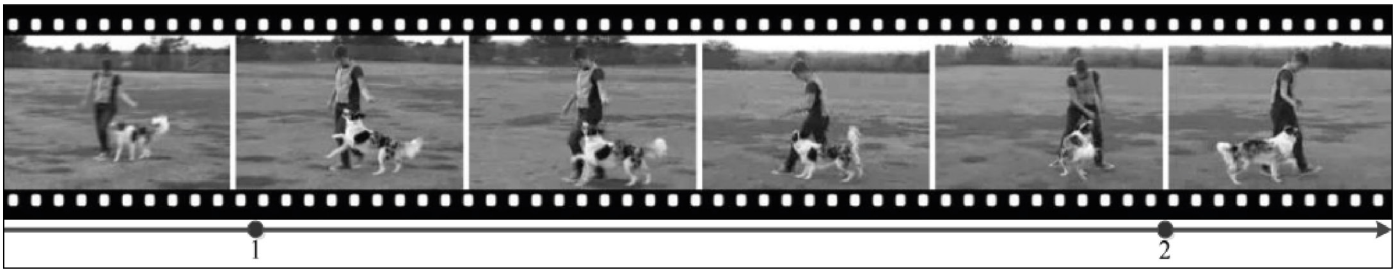

Figure 2. A schematic diagram of a long sequence frame extraction strategy.

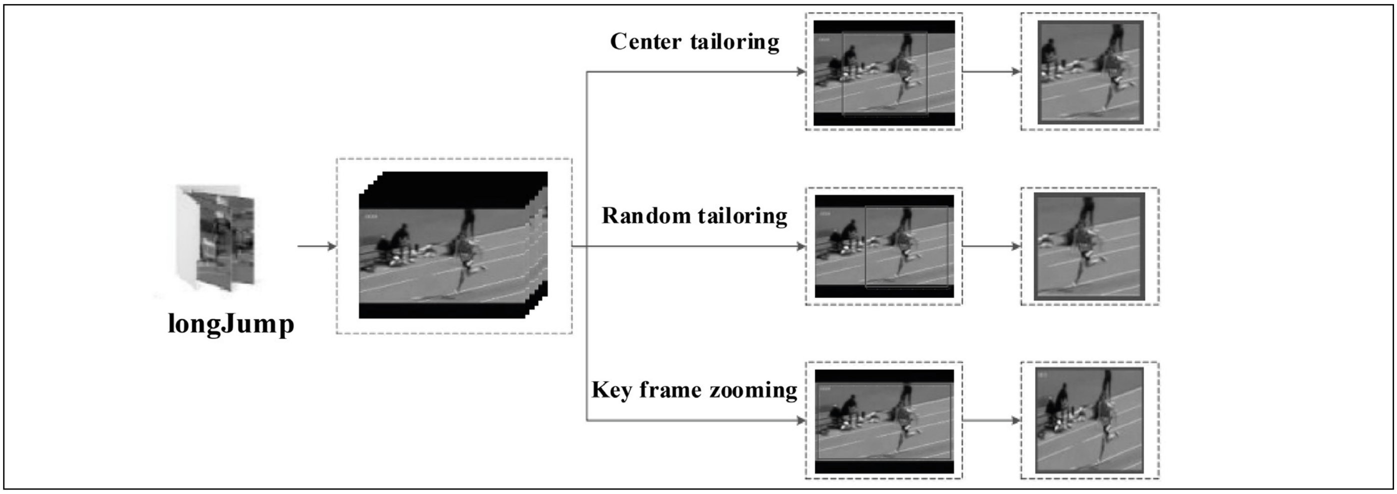

Figure 3. Three ways of data augmentation.

Space-time feature extraction based on LI3D network

In the process of image content recognition and classification, the acquisition of features is the key to determine the quality of algorithm recognition. Around this problem, many traditional characterization methods have been developed, such as SIFT, HOG and so on. In recent years, with the continuous development of deep learning, feature descriptors based on convolutional neural networks (CNN) are increasingly used in the field of image content recognition and classification. Compared to traditional feature descriptors, the process of CNN extracting feature descriptors is equivalent to training one filter (convolution kernel), and these filters are equivalent to the detection operators in the traditional feature extraction methods, which differ from the traditional feature extraction methods in that: the detection operators of traditional feature extraction methods such as SIFT and HOG are generally designed by humans and are summarized by a large amount of prior knowledge, while these filters are obtained through data-driven autonomous learning in the process of neural network training. In the process of video feature extraction, the 3D convolutional network is used to simultaneously learn the information of time and space, and obtain descriptors that can simultaneously represent temporal and spatial features, which is obviously impossible to achieve by traditional feature extraction methods. However, due to the introduction of convolution operations in the time dimension, the traditional 3D convolutional neural network C3D is extremely computationally intensive during training. Moreover, due to the limited amount of video data, it is not possible to provide better pre-training weights. Therefore, the efficiency of C3D to extract video timing features is not high, and the recognition accuracy of the network is not good enough. I3D solves these two problems very well. Firstly, I3D utilizes the network architecture of InCeption-v2. Compared with the traditional C3D network, the I3D network has deeper features. Moreover, I3D can use the pre-training weight of Inception-v2 on the image large-scale dataset ImageNet, which greatly reduces the computational load of model training and improves the robustness and generalization ability of the network.

Experimental Design and Analysis

Experiment 1: Comparing the performance of the proposed network structure LI3D and the traditional I3D network structure, in the case of the same iteration 50 generations, it is compared from multiple dimensions such as network parameter quantity, test time consumption, test accuracy rate, accuracy rate and recall rate. The accuracy rate, accuracy and recall rate are calculated as follows: accuracy = all samples with the correct prediction / total sample; accuracy = predict positive class as positive class / all forecast as positive class; recall rate = predict positive class as positive class / all positive class. The results of Experiment 1 are shown in Table 1. From the results in the table, it is easy to see that the LI3D network structure can effectively reduce the amount of network parameters, and the network parameter amount has dropped from 12.28M to 8.16M, which is a total reduction of 33.6%. Fewer network parameters enable the network to have better data fitting capabilities and to minimize network overfitting. The accuracy rate on the UCF101 test set increased from 0.82 to 0.84, the accuracy increased from 0.84 to 0.88, and the recall rate increased from 0.81 to 0.83. Moreover, since the LI3D network uses the superposition of the 3*1*1 convolution kernel and the 1*3*3 convolution kernel instead of the 3*3*3 convolution kernel in the BD network, the calculation amount of the network can be effectively reduced, and the test time of the network was reduced from the original 2.08s to 1.62s, a decrease of 22.12%. Network test time is closer to real-time effects.

On this basis, the data augmentation operation was added to compare the changes in accuracy rate, accuracy and recall rate of I3D and LI3D after the data was augmented in the same iteration of 50 generations. The comparison results are shown in Table 2.

It is not difficult to see from Table 2 that after the data is augmented, the recognition performance of I3D and LI3D networks has been greatly improved, and the accuracy rate, accuracy and recall rate of LI3D are more obvious than I3D. The three indicators increased from 0.85, 0.88, and 0.84 to 0.89, 0.90, and 0.88, respectively. This shows that after data augmentation, the LI3D network model maintains a high recall rate while maintaining a high accuracy for the positive rate. Figures 4 and 2.19 show the accuracy and loss function curves of the training phase before and after the data augmentation of the I3D and LI3D networks before and after the data augmentation.

The blue curve and the yellow curve in figure 4 are the accuracy curves of the traditional I3D model before and after the data augmentation.

Table 1. Performance comparison of I3D and LI3D networks (before data augmentation).

Model index	Network parameters	Testing time-consuming	accuracy rate	Accuracy	recall
I3D	12.28M	2.80s	0.82	0.84	0.81
LI3D	8.16M	1.62s	0.84	0.88	0.83

Table 2. I3D and LI3D Network Performance Comparisons (After Data Enlargement).

Model index	accuracy rate	Accuracy	recall
I3D	0.85	0.88	0.84
LI3D	0.89	0.90	0.88

It is not difficult to see from the figure that after the data augmentation operation, the accuracy of the network is greatly improved compared with before. However, the accuracy rate after data augmentation is only reached before the increase of LI3D data. The LI3D model has been able to achieve better accuracy after data augmentation in the same iteration of 50 generations. It can be seen that the data augmentation operation can effectively improve the recognition performance of the network. Under the premise of the network model with the fine-tuning, LI3D, the accuracy is more obvious than the traditional I3D network. Figure 5 is a comparison of the loss function curves of I3D and LI3D iterative 50 generations in the case of data augmentation and no data augmentation. It can be seen from Fig. 5 that after the data is augmented, the convergence speed of the network is faster, the network is more stable, and the robustness is better. Among them, the network convergence rate of the combination of LI3D and data augmentation strategy is the fastest.

Experiment 2: After experiment 1, it can be known that the optimal performance configuration method is data augmentation strategy and LI3D model. Based on the experiment 1, Experiment 2 compares the performance of the LI3D, LI3D-LSTM and LI3D-BiLSTM models under the data augmentation strategy and the same iteration 50 generations. The effect of timing feature analysis on the LI3D model is analyzed. Table 3 shows the accuracy rate, accuracy, and recall rate of LI3D, LI3D-LSTM, and LI3D-BiLSTM under the same iteration of 50 generations.

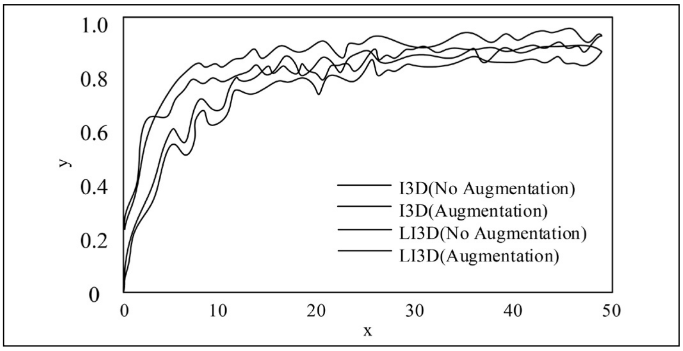

Figure 4. Accuracy curves of I3D and LI3D data before and after augmentation.

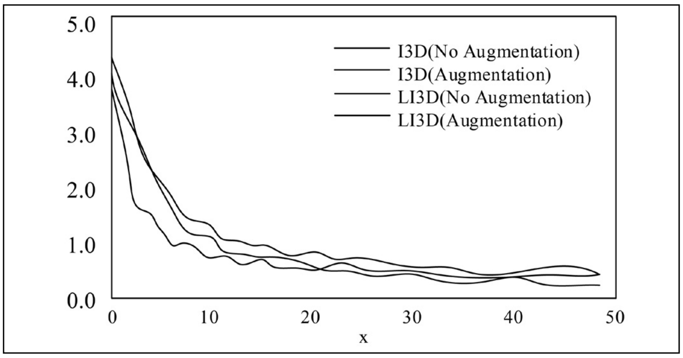

Figure 5. Loss function curve before and after light-I3D data augmentation.

Table 3. Comparison of LI3D, LI3D-LSTM and LI3D-BiLSTM Test Indicators.

Model index	LI3D	LI3D-LSTM	LI3D-BiLSTM
accuracy rate	0.85	0.86	0.90
Accuracy	0.86	0.89	0.91
recall	0.84	0.87	0.90

It can be seen from the data in the table that LSTM and Bi-LSTM can improve the accuracy rate, accuracy and recall rate of the network by applying the timing characteristics of LI3D extraction. This shows that LSTM and Bi-LSTM can make good use of the timing relationship in the data and improve the characterizing ability of the feature to the video content. Moreover, Bi-LSTM can simultaneously utilize the context of the time series data, and the performance of Bi_LSTM is superior to that of LSTM. Figures 6 and 7 show the accuracy and loss function curves for LI3D, LI3D-LSTM, and LI3D-BiLSTM under the same iteration of 50 generations.

CONCLUSION

Based on computer vision technology, the target tracking in complex scenes is studied. The difference between 3D convolution and traditional 2D convolution is studied, and fine-tuning based on the structure of the I3D network, using a lightweight LI3D model, comparing the parameter quantity and test time consumption of the LI3D model and the traditional I3D model. And the difference in the accuracy rate, accuracy, and recall rate of the network test phase before and after the application data augmentation strategy. The superiority of the data augmentation strategy plus the LI3D model is verified. At the same time, the accuracy curve and loss function curve of the I3D model and the LI3D model in the training phase are compared under the same iteration 50 generations. The experiment proves that the LI3D model is more stable and the convergence is faster. Finally, based on the LI3D model, the BiLSTM feature analysis module is added, the accuracy, accuracy and recall rate of the LI3D, LI3D-LSTM and LI3D-BiLSTM test phases are compared under the condition of iterative 50 generations, and the effect of BiLSTM context feature association strategy on timing feature characterization is verified. At the same time, by comparing the accuracy curve and loss function curve during LI3D, LI3D-LSTM and LI3D-BiLSTM training, it is verified that the LI3D-BiLSTM model converges faster. In a word, the improved feature aggregation layer is fused, and a multi-feature coding soft association splicing feature aggregation module with adjustable feature clustering center is proposed, which has important practical significance. In the subsequent research, the process of feature extraction needs to be extracted in a targeted manner to improve the recognition ability of the network for key content in the video, thereby improving the recognition accuracy and computational efficiency of the network.

The author declare no potential conflict of interest related to this article

AUTHORS' CONTRIBUTIONS: Each author made significant individual contributions to this manuscript. Huanan Shang: writing and execution.

REFERENCES

1. Liu Y, Wang Q, Yan Z, Hu H. A Novel Trail Detection and Scene Understanding Framework for a Quadrotor UAV With Monocular Vision. IEEE Sensors Journal. 2017;17(20):6778-87.
2. Cai C, Fan B, Weng X, Zhu Q, Su L. A target tracking and location robot system based on omnistereo vision. Industrial Robot. 2017;44(6):741-53.
3. Harik EHC, Guérin F, Guinand F, Brethé JF, Pelvillain H, Parédé JY. Fuzzy logic controller for predictive vision-based target tracking with an unmanned aerial vehicle. Advanced Robotics. 2017;31(7):368-81.
4. Lei J, Wang L, He Y, Zhang Z. Image segmentation method for robot vision. Systems Engineering & Electronics. 2017;39(7):1653-9.
5. Madrigal F, Hayet JB. Motion priors based on goals hierarchies in pedestrian tracking applications. Machine Vision & Applications. 2017;28(11):1-19.
6. Choe KW, Kardan O, Kotabe HP, Henderson JM, Berman G. To search or to like: Mapping fixations to differentiate two forms of incidental scene memory. Journal of Vision. 2017;17(12):1-22.