Zhao *et al. EURASIP Journal on Image and Video Processing* (2020) 2020:24
https://doi.org/10.1186/s13640-020-00501-x

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Improved two-stream model for human action recognition

Yuxuan Zhao[1], Ka Lok Man[1,2,3*], Jeremy Smith[4], Kamran Siddique[5] and Sheng-Uei Guan[1]

Abstract

This paper addresses the recognitions of human actions in videos. Human action recognition can be seen as the automatic labeling of a video according to the actions occurring in it. It has become one of the most challenging and attractive problems in the pattern recognition and video classification fields. The problem itself is difficult to solve by traditional video processing methods because of several challenges such as the background noise, sizes of subjects in different videos, and the speed of actions. Derived from the progress of deep learning methods, several directions are developed to recognize a human action from a video, such as the long-short-term memory (LSTM)-based model, two-stream convolutional neural network (CNN) model, and the convolutional 3D model.

In this paper, we focus on the two-stream structure. The traditional two-stream CNN network solves the problem that CNNs do not have satisfactory performance on temporal features. By training a temporal stream, which uses the optical flow as the input, a CNN can have the ability to extract temporal features. However, the optical flow only contains limited temporal information because it only records the movements of pixels on the *x*-axis and the *y*-axis. Therefore, we attempt to design and implement a new two-stream model by using an LSTM-based model in its spatial stream to extract both spatial and temporal features in RGB frames. In addition, we implement a DenseNet in the temporal stream to improve the recognition accuracy. This is in-contrast to traditional approaches which typically utilize the spatial stream for extracting only spatial features. The quantitative evaluation and experiments are conducted on the UCF-101 dataset, which is a well-developed public video dataset. For the temporal stream, we choose the optical flow of UCF-101. Images in the optical flow are provided by the Graz University of Technology. The experimental result shows that the proposed method outperforms the traditional two-stream CNN method with an accuracy of at least 3%. For both spatial and temporal streams, the proposed model also achieves higher recognition accuracies. In addition, compared with the state of the art methods, the new model can still have the best recognition performance.

Keywords: Action recognition, Two-stream CNN model, Spatial stream, LSTM-based model

1 Introduction

Action recognition aims to recognize the motions and actions of objects. In the human action recognition field, vision-based action recognition is one of the most popular and essential problems [1]. It requires approaches to track and distinguish the behavior of the subject through videos. Human action recognition is used in some surveillance systems and video processing tools [2]. In addition, the model of solving this kind of problem can also be used for other video classification tasks by transfer learning. Therefore, it is necessary to develop a new model to improve recognition accuracy.

Based on the rapid development of computer vision and neural networks, vast improvements have been achieved in the action recognition field [3, 4]. By using CNNs, spatial features from RGB video frames can be easily extracted, which is similar to its functions in image recognition [5, 6]. However, the critical challenge of video human action recognition is how to obtain and handle temporal features effectively. Compared with still

*Correspondence: ka.man@xjtlu.edu.cn
[1] Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Ren'ai Road, Suzhou, China
[2] imec-DistriNet, KU Leuven, Leuven, Belgium
Full list of author information is available at the end of the article

images, video contains valuable temporal information that can enhance the accuracy of the action recognition [7, 8]. How to get and use the temporal features has become an important task in the video classification problem.

Current solutions can be divided into two approaches. One direction is to use models that can extract the temporal features, such as the LSTM, in the final model [9]. LSTM uses three gates to decide which cell can be passed to the next layer or forgotten. Thus, it can keep the temporal information in the video. However, the input size of an LSTM would be much bigger than a CNN. Therefore, the training speed of such methods can be slow if they just rely on the LSTM. And according to our experiments, a single LSTM-based model still has space for improvement according to our experimental results shown in Section 3.2.

Another approach is to add an extra input stream, which can be the extracted temporal features using a CNN [9]. Optical flow is a popular input in this approach. It is a set of images, which presents the relative motion between the object and background in the video. Thus, the optical flow contains the features in time sequence. Since there are two inputs for the CNN, the traditional idea is to train two independence CNNs, one handling the RGB frames, and another one managing the optical flow. Then, it combines the results of both training streams and gets a final recognition result. This structure is known as the two-stream CNN model. However, the two-stream CNN model still has an apparent defect. The model does not contain the original temporarily information in the RGB video frames. Though the optical flow contains the temporal features, it only records the movements of pixels on the x-axis and the y-axis (Figs. 1 and 2).

Therefore, in this paper, we aim to solve the limitations of previous solutions. The idea is to make a combination of these two approaches. The new model keeps the temporal stream so that a CNN can still process temporal features from the optical flow. For the spatial stream, we use an LSTM-based model to replace the traditional CNN in order to extract more temporal features from the RGB frames.

2 Methodology

In this work, the proposed model can be mainly decomposed into three modules. They are a spatial stream with the LSTM, a temporal stream with a DenseNet, and a fusion layer with support vector machine (SVM) [10].

The overall structure of the model and the general training process are shown in Fig. 3. Firstly, the training data are RGB video frames and optical flow. The training process can be divided into three parts. For each video, a sequence of sampled RGB video frames is processed by the spatial stream. The LSTM-based model in this stream trains the data and gets a recognition result by marking grades for different labels. According to the input sequence of frames, the corresponding optical flow is input and processed in the temporal network. DenseNet is used in this stream for training. Because the whole training process is by supervised learning, every optical flow is also labeled. DenseNet also provides its recognition results. So far, there are two results from the above streams. Finally, the fusion layer will fuse the results of the two streams and get the final recognition result.

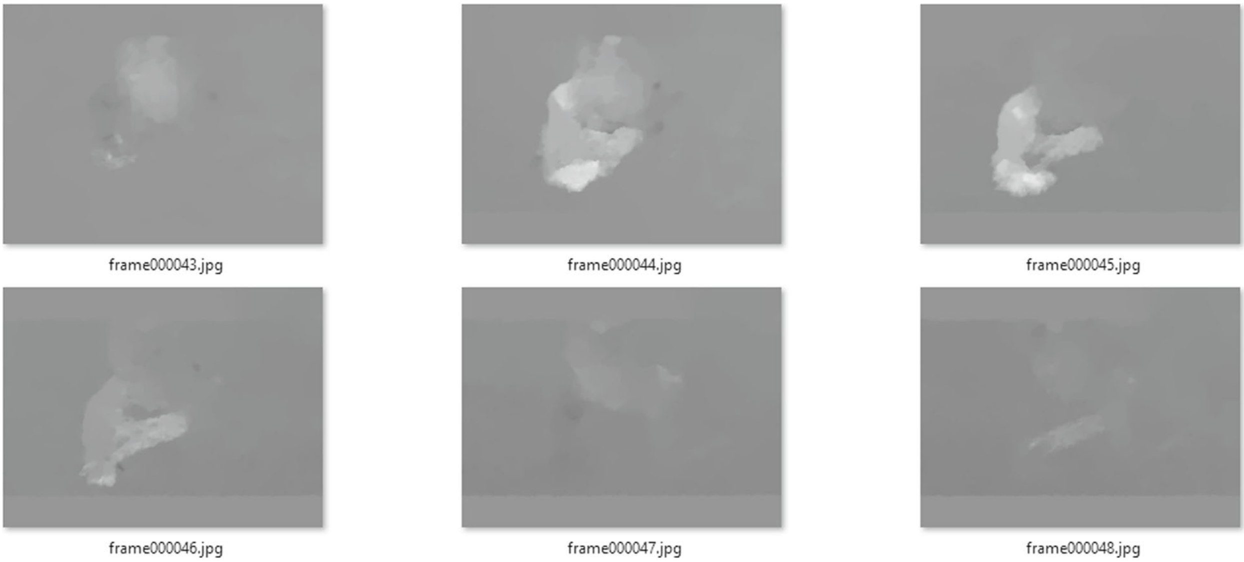

Fig. 1 Part of an optical flow of a video in UCF-101. It is a set of grayscale images and pays no attention to the RGB information. Conversely, it concerns the motion of the subject on the video

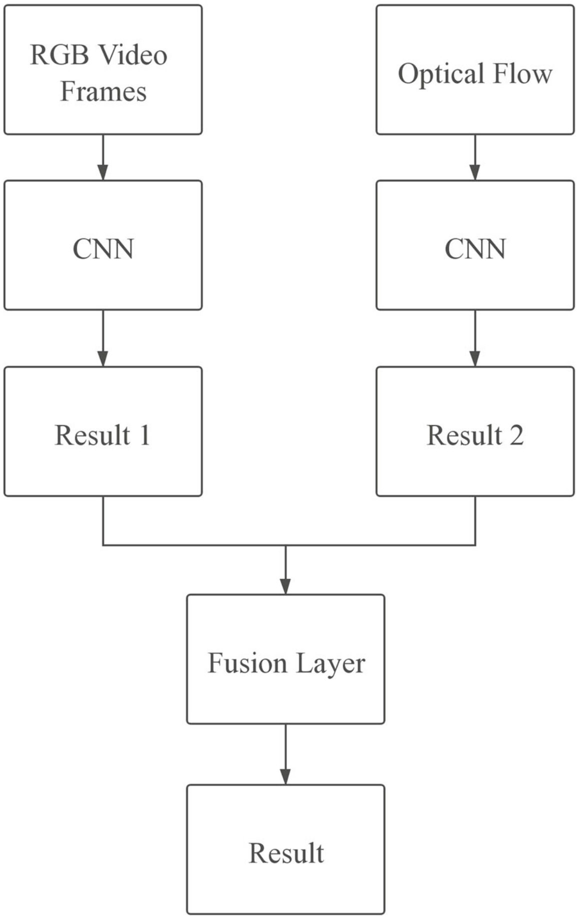

Fig. 2 The structure of the two-stream CNN model

2.1 Spatial stream

There is an LSTM-based model in the spatial stream, which uses a convolutional neural network for feature extraction and an LSTM network to do further classification (Fig. 4).

2.1.1 Convolutional neural network

The Visual Geometry Group (VGG16) is modified slightly in this model for the spatial feature extraction work [11]. VGG16 is a CNN provided by the Oxford University and has been widely used in the image classification

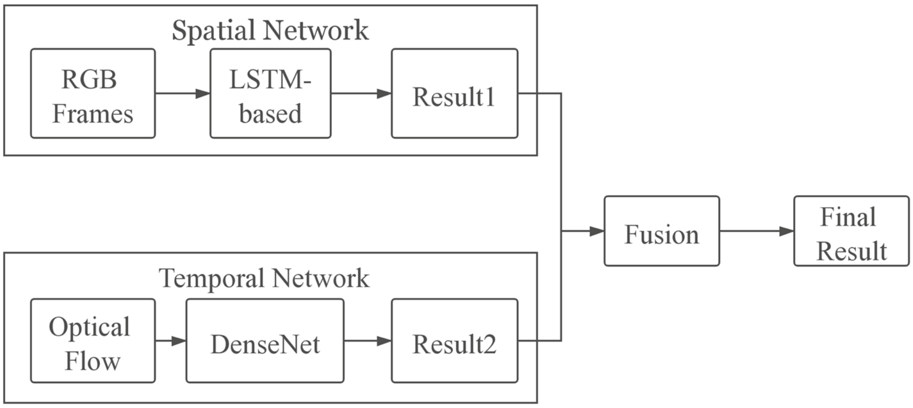

Fig. 3 The overall structure of the proposed two-stream model

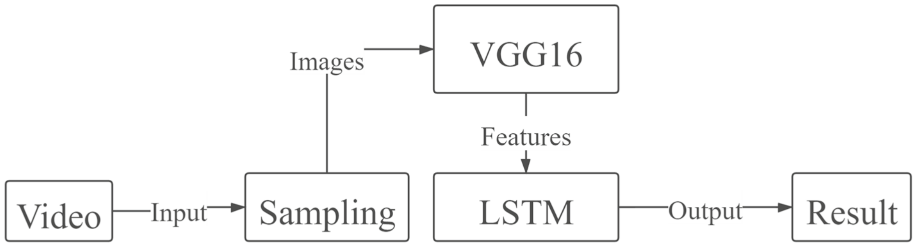

Fig. 4 The proposed structure in the spatial model

field. It is pre-trained on the ImageNet dataset [12], and the weights and layer configuration are available on the official website. In VGG16, there are sixteen hidden layers, includes thirteen convolutional 2D layers and three fully connected layers. The input RGB video frames are resized to 224 × 224 to fit the default input size of this model. There is no pooling applied to the output of the last convolutional layer, which means the output will be the 4D tensor output of the last convolutional layer. Finally, this model outputs the sequence of features of the input frames and pass these features to the next step (Figs. 4 and 5).

2.1.2 *Long short-term memory (LSTM)*

In consideration of the fact that CNN is mainly powerful for extracting the spatial features, it is necessary to utilize temporal features from these RGB video frames. Since the input of the whole model is a sequence of images associated in time, an LSTM is built in the current model [13]. The LSTM network in this model is set to be a single-directional structure. It contains one LSTM layer and two fully connected layers. Figure 6 shows the structure of the kernel of the LSTM layer, where σ and tanh represent the activation functions, c and h represent the cell state and the hidden state separately, and x is the input signal.

Related equations [13]:

$$f_t = \sigma(W_f \cdot [\, C_{t-1}, h_{t-1}, x_t \,]\,) + b_i \tag{1}$$

$$i_t = \sigma(W_i \cdot [\, C_{t-1}, h_{t-1}, x_t \,]\,) + b_f \tag{2}$$

$$\widetilde{c}_t = \tan(W_c \cdot [\, C_{t-1}, h_{t-1}, x_t \,]\,) + b_c \tag{3}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{c}_t \tag{4}$$

$$o_t = \sigma(W_0 \cdot [\, C_t, h_{t-1}, x_t \,]\,) + b_0 \tag{5}$$

$$h_t = o_t \cdot \tan\,(C_t) \tag{6}$$

The LSTM uses three gates to determine which information is useful. The above equations describe how the three gates work in the LSTM. W in these equations represents the matrix of parameters. The forget gate chooses the information which will not be used in the current cell. The input gate determines the input of the next cell. The output gate determines the hidden state outputted from the current cell. In particular, the LSTM layer in this model has 512 hidden units in every cell. The output shape of this block equals the number of classes in the dataset. To reduce the possibility of overfitting, a dropout layer is added between full-connected layers.

Since the training speed of an LSTM is much slower than the CNN, if the input data of this stream is the standard sampled files of UCF-101, the training time will be too long. According to tests, if the input data of the spatial stream is the image set provided by the Graz University of Technology, the training time of each epoch will be more than 2 h under the processing of an Nvidia RTX2080ti GPU card. Therefore, a sampled script is added before the whole model to extract 25 frames from each video.

2.2 Temporal stream

This section describes the convolutional neural network we used in the temporal stream. The difference of this stream is that it uses a stack of optical flow images. As shown in Fig. 7 [9], there are five variations of the optical flow-based input. In this work, we choose (d) and (e) as the input data. A stack of optical images which contains ten x-channel and ten y-channel images is considered as an input. Therefore, the input shape is (20,224,224). Graz University of Technology provides the file of the optical flow.

We choose the DenseNet in the temporal stream [14]. DenseNet uses several dense blocks in its structure, and every dense block contains several convolutional layers. Unlike the VGG net, layers in the same block are related to each other. Therefore, every layer contains the output features of all the previous layers in the same block. The relation among different layers is enhanced a lot in the DenseNet. Figure 8 [14] are the basic structure of this network. The advantage of this model is that it needs fewer

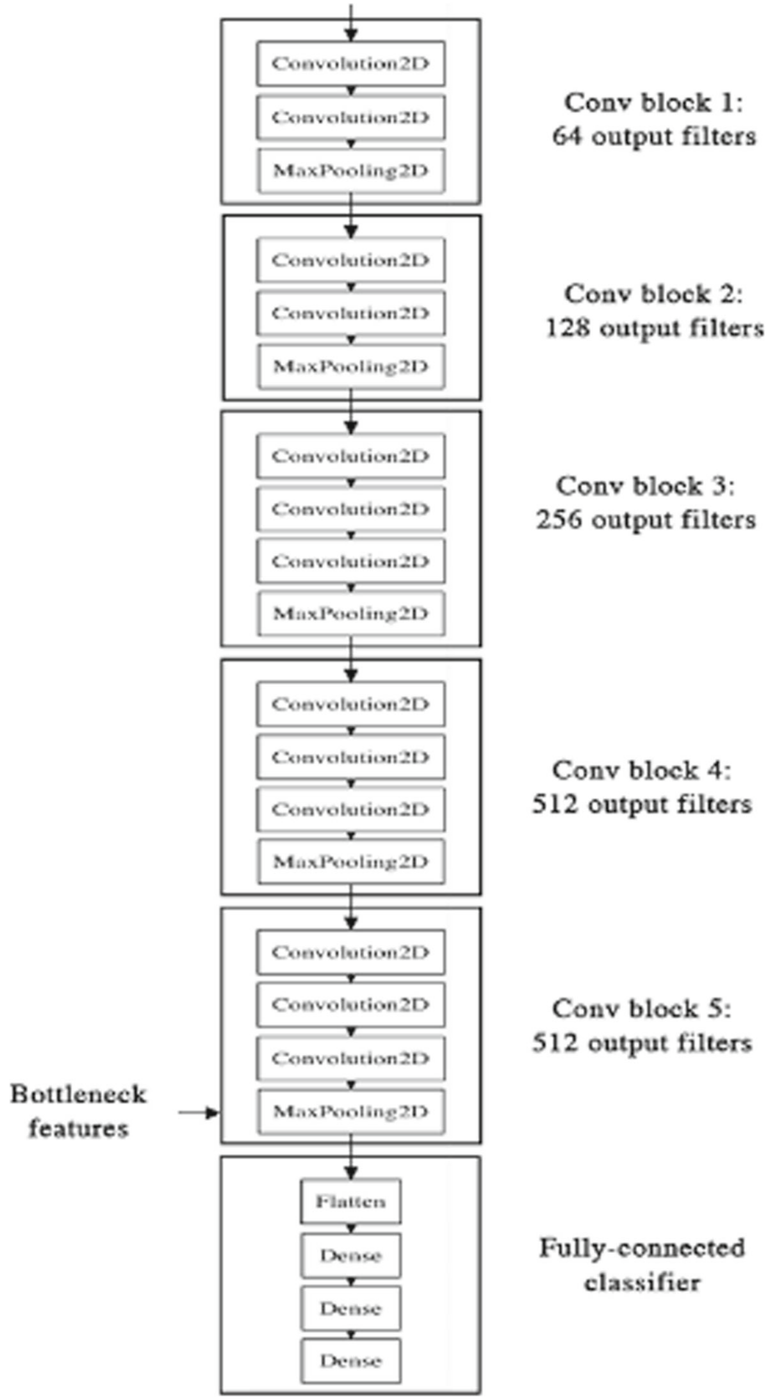

Fig. 5 The structure of VGG16

feature images than other models. Due to the enhanced relation among layers, more information can be collected in a single-dense block. Therefore, we do not need a lot of parameters and feature images to ensure the stability of the whole training process. Furthermore, the vanishing gradient problem will be solved because of the dense connection.

In this work, we use a basic DenseNet-121 as the proposed model. It contains four dense blocks and 58 convolutional layers in total. Since the optical input of UCF-101

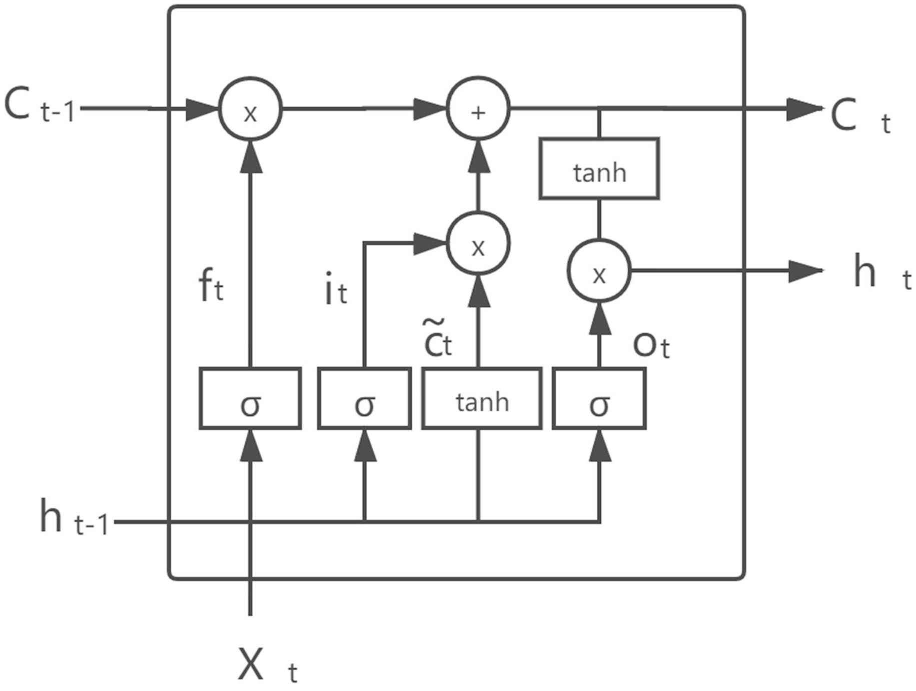

Fig. 6 The structure of the LSTM

is large enough, it is easy to avoid the over-fitting problem. Therefore, in the final classifier of this stream, we do not need to create several softmax layers for different datasets. As a replacement, we keep the original softmax layer for the UCF-101 classification.

2.3 Fusion layer

Although the proposed model modifies both the spatial stream and the temporal stream, the outputs of these two streams are not changed. Each stream outputs its classification results separately. The choices of the method in the fusion layer can be similar to the traditional two-stream CNN model. According to previous experiments of the two-stream CNN model [9], the SVM has a better performance than the average method. Therefore, we choose the SVM in our proposed model.

3 Experiments and results

3.1 Dataset and implementation

The proposed model in this paper is evaluated on the UCF-101 human action recognition dataset [15]. It contains 13320 labeled videos belong to 101 human action categories, such as punching, boxing, and walking. All these 101 categories can be divided into five types: Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments, and Sports. All videos in this dataset are realistic and collected from YouTube. The UCF-101 dataset does not have a pre-divided training set and testing set. It gives the three official guides of training and testing splits for both action recognition and action detection.

In our experiments, the final accuracy is the average accuracy of all three splits. The ImageNet is used for the

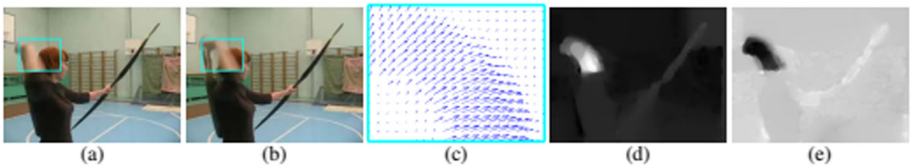

Fig. 7 Optical flow. **a**, **b** RGB video frames, which are similar to the input data of the spatial stream. However, there are some rectangles that highlight the moving area in the frame. **c** The dense optical flow in the outlined area. **d** *y*-channel images of the displacement vector. **e** *x*-channel images of the displacement vector

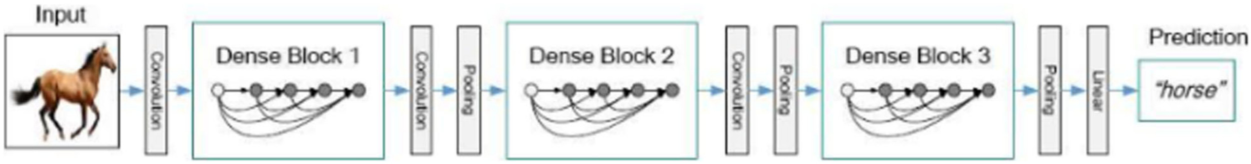

Fig. 8 The structure of the DenseNet

pre-trained part for both two streams. Finally, we evaluate the spatial stream, the temporal stream, and the whole model separately to check the individual performance of every component in the proposed model. We use Keras and an Nvidia RTX2080ti GPU for the experiments. The GPU is powered by the Turing GPU architecture, which is useful for the research of deep learning. Besides, it has an 11 GB GDDR6 frame buffer.

3.2 Result and discussion

In this section, we compare the experimental results between the proposed model and other state-of-the-art methods. We also discuss the implications and current limitations of our work.

3.2.1 Spatial stream

We compare the accuracy of different neural networks on the UCF-101. Five scenarios are considered: (a) the spatial stream ConvNet, which was used in Google's first two-stream CNN model [9]; (b) VGG16; (c) VGG16 and a bidirectional LSTM; and (d) VGG16 and a single-directional LSTM. In our experiments, all methods are pre-trained by ImageNet. In addition, the dropout is set to be 0.5. Besides the top 1 accuracies, the top 5 accuracies of LSTM-based models are also shown in Table 1.

According to the results in Table 1, the method which uses a simple VGG16 has a poor performance. The spatial stream ConvNet can improve the top 1 accuracy by more than 40%. Furthermore, two LSTM-based models, which are used in the proposed model, can improve by another 16% on that basis. Between these two LSTM-based models, the single-directional LSTM seems to have better accuracy than the bidirectional one; however, the difference is less than 2%.

Table 1 Different models accuracies on video frames of UCF-101

	Top 1 %	Top 5%
Spatial Stream ConvNet [9]	72.7	
VGG16	32.1	51.3
Inception V3 [16]	54.55	79.92
VGG16+LSTM (bidirectional)	88.1	96.72
VGG16+LSTM (single directional)	90.81	98.61

The performance includes both top 1 and top 5 accuracies

3.2.2 Temporal stream

Here, we compare the performances of different methods in the temporal stream. From Table 2, we can see the DenseNet can get the highest top 1 accuracy, which is 3% higher than the ResNet101. As in the previous experiment, VGG16 get the lowest top 1 accuracy.

3.2.3 Whole model

We compare the performance of the proposed two-stream model with state-of-the-art methods on UCF-101. The performance is measured by the average accuracy on all three splits of the UCF-101 dataset. In the experiment, we use the control variable method. If the model uses Kinetics for the pre-trained part, it will certainly get a higher accuracy of the recognition. To make the comparison fair, all models in these experiments are only pre-trained on ImageNet. However, we keep the CNN backbone of each method different so that the final accuracy can be authentic. In addition, only the top 1 accuracy is considered in this experiment because most of the methods in this table do not provide the top 5 accuracy for comparison.

Table 3 gives a quantitative comparison of the experimental results. According to this table, our proposed two-stream model get the highest top 1 recognition accuracy among all methods, which is 92.5%. Compared with the state-of-the-art two-stream CNN method, the proposed model outperforms it by more than 3%. The traditional LSTM-based model achieves 69.1% accuracy, which is 23% less than our proposed model. Compared with another traditional direction of video classification, the C3D, the accuracy of our method is 10% higher than its accuracy. In addition, we also compare other state-of-the-art methods such as RGB-I3D and TSN. Benefitting from the advanced temporal stream, the proposed model can also have higher recognition accuracy than these methods. Besides the top 1 accuracy, in the spatial stream, our method only uses 25 sampled frames as the input, while all

Table 2 Performances on the optical flow of different CNNs

	Top 1%	Top 5%
ResNet101 [17]	76.1	
VGG16	30.1	46.5
DenseNet121 (proposed method)	79.63	80.12

The performance includes both top 1 and top 5 accuracies

Table 3 State-of-the-art performance comparison on UCF101

	Pre-trained	CNN backbone	UCF-101%
Two stream CNN [9]	ImageNet	VGG16	88.7
Conv + LSTM [8]	ImageNet	AlexNet	69.1
C3D [18]	ImageNet	VGG11	82.3
RGB-I3D [19]	ImageNet	Inception v1	84.5
TSN [20]	ImageNet	Inception v2	86.4
3D Hybrid Model [21]	2D CNN	C3D	89.4
Two-stream model (proposed model)	ImageNet	DenseNet	92.5

The accuracy is the average accuracy for all three splits of the dataset

state-of-the-art methods use the standard frame dataset of UCF-101, which has much more frames of each video. Though the input volume is smaller, the proposed method still has a better performance.

In summary, the proposed model achieves higher recognition accuracies in both the spatial stream and the temporal stream than the traditional two-stream CNN model. In addition, compared with other state-of-the-art approaches, the proposed model can get the highest overall top 1 accuracy.

3.2.4 Discussion

The study presented an innovative two-stream model for video human action recognition. The model enhances the function of the spatial stream. According to the results of the experiments, each stream of the proposed model can have a good recognition performance. Finally, the whole model can achieve higher top 1 accuracy than previous deep learning models.

Here are some potential impacts of this study:

It provides a new solution for the temporal features extraction problem. It shows that even if the LSTM-based model in the spatial stream is a combination of two basic networks, the two-stream model can still have a high recognition accuracy. In the future, this structure can have further improvements.

The study shows that the optical flow can still be improved if we use advanced CNN.

The proposed model can be applied in video description tasks by the help of natural language description methods [22].

The proposed model can be used for smart city surveillance such as the unforseeable event detection and traffic control [23].

Though the proposed method outperforms the state-of-the-art methods, it still has limitations that needed to be solved improved in the future.

The proposed model increases the complexity compared with either the LSTM method and the two-stream CNN. The whole model needs more parameters in both streams. Compared with traditional methods, both the LSTM-based model and the DenseNet requires more training time. This drawback limits the size of the input, especially for the spatial stream. Currently, input data for each video are 25 RGB frames. If the training speed is improved, more frames can be added for the recognition. As a result, the model can achieve higher accuracy in experiments.

The fusion layer is not well developed. We still use the SVM, which is a traditional method in the fusion layer. This part still has room for improvement.

4 Conclusion

We propose an innovative deep learning model, which is used for human action recognition. The basic structure of this model is the two-stream structure. However, unlike the traditional two-stream CNN model, the proposed method aims to extract both spatial and temporal features from the RGB video frames in the spatial stream. In order to achieve this objective, we use the LSTM-based model to replace the traditional convolutional neural network in its spatial stream. Furthermore, we implement a DenseNet to improve the performance of the temporal stream. According to the experimental results, with respect to the traditional two-stream model and other neural networks, the key achievement of our proposed method is to obtain the highest top 1 accuracy among the human action recognition tasks of UCF-101 dataset.

Acknowledgements

The research presented in this paper is supported by the Research Management Center, Xiamen University Malaysia under XMUM Research Program Cycle 3 (Grant No: XMUMRF/2019-C3/IECE/0006). In addition, it is also supported by by Xi'an Jiaotong-Liverpool University, Suzhou, China, and the Research Development Fund (RDF-15-01-01). Ka Lok Man wishes to thank the AI University Research Centre (AI-URC), Xi'an Jiaotong-Liverpool University, Suzhou, China, for supporting his related research contributions to this article through the XJTLU Key Programme Special Fund (KSF-P-02).

Authors' contributions

Yuxuan Zhao designed and implemented the models and experiments. Ka Lok Man proposed the topic and corrected the design. Seng-Uei Guan and Jeremy S. Smith helped to improve the experiment setting. Kamran Siddique helped the preparation of datasets and gave useful guides for the model design. The author(s) read and approved the final manuscript.

Availability of data and materials

Please contact the corresponding author for data requests.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Ren'ai Road, Suzhou, China. [2] imec-DistriNet, KU Leuven, Leuven, Belgium. [3] Swinburne University of Technology, Sarawak, Malaysia. [4] Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, UK. [5] Department of Information and Communication Technology, Xiamen University Malaysia, Sepang, Malaysia.

Received: 8 November 2019 Accepted: 9 March 2020
Published online: 17 June 2020

References

1. S. Hongeng, R. Nevatia, F. Bremond, Video-based event recognition: activity representation and probabilistic recognition methods. Comput. Vis. Image Underst. **96**(2), 129–162 (2004)
2. H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, D.-S. Chen, A comprehensive survey of vision-based human action recognition methods. Sensors. **19**(5), 1005 (2019)
3. H. Jhuang, T. Serre, L. Wolf, T. Poggio, in *2007 IEEE 11th International Conference on Computer Vision*. A biologically inspired system for action recognition (IEEE, 2007), pp. 1–8. https://doi.org/10.1109/iccv.2007.4408988
4. H. Wang, C. Schmid, in *Proceedings of the IEEE International Conference on Computer Vision*. Action recognition with improved trajectories, (2013), pp. 3551–3558. https://doi.org/10.1109/iccv.2013.441
5. S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2012)
6. A. Krizhevsky, I. Sutskever, G. E. Hinton, in *Advances in Neural Information Processing Systems*. ImageNet classification with deep convolutional neural networks, (2012), pp. 1097–1105. https://doi.org/10.1145/3065386
7. Z. Zhang, D. Tao, Slow feature analysis for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **3**, 436–450 (2012). https://doi.org/10.1109/tpami.2011.157
8. J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Long-term recurrent convolutional networks for visual recognition and description, (2015), pp. 2625–2634. https://doi.org/10.21236/ada623249
9. K. Simonyan, A. Zisserman, in *Advances in Neural Information Processing Systems*. Two-stream convolutional networks for action recognition in videos, (2014), pp. 568–576
10. C. Gold, P. Sollich, Model selection for support vector machine classification. Neurocomputing. **55**(1-2), 221–249 (2003)
11. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint (2014). arXiv:1409.1556
12. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. ImageNet: a large-scale hierarchical image database (IEEE, 2009), pp. 248–255. https://doi.org/10.1109/cvpr.2009.5206848
13. S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
14. G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Densely connected convolutional networks, (2017), pp. 4700–4708. https://doi.org/10.1109/cvpr.2017.243
15. K. Soomro, A. R. Zamir, M. Shah, Ucf101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint (2012). arXiv:1212.0402
16. X. Xia, C. Xu, B. Nan, in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. Inception-v3 for flower classification (IEEE, 2017), pp. 783–787. https://doi.org/10.1109/icivc.2017.7984661
17. C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, in *Thirty-First AAAI Conference on Artificial Intelligence*. Inception-v4, Inception-ResNet and the impact of residual connections on learning, (2017)
18. D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, in *Proceedings of the IEEE International Conference on Computer Vision*. Learning spatiotemporal features with 3D convolutional networks, (2015), pp. 4489–4497. https://doi.org/10.1109/iccv.2015.510
19. J. Carreira, A. Zisserman, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Quo vadis, action recognition? A new model and the kinetics dataset, (2017), pp. 6299–6308. https://doi.org/10.1109/cvpr.2017.502
20. L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, in *European Conference on Computer Vision*. Temporal segment networks: towards good practices for deep action recognition (Springer, 2016), pp. 20–36. https://doi.org/10.1007/978-3-319-46484-8_2
21. Z. Hu, E.-J. Lee, in *2019 IEEE International Conference on Computation, Communication and Engineering (ICCCE)*. Human motion recognition based on improved 3-dimensional convolutional neural network (IEEE, 2019), pp. 154–156
22. A. Dilawari, M. U. G. Khan, A. Farooq, Z.-U. Rehman, S. Rho, I. Mehmood, Natural language description of video streams using task-specific feature encoding. IEEE Access. **6**, 16639–16645 (2018)
23. S. Kang, W. Ji, S. Rho, V. A. Padigala, Y. Chen, Cooperative mobile video transmission for traffic surveillance in smart cities. Comput. Electr. Eng. **54**, 16–25 (2016)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.