International Journal of Engineering

Journal Homepage: www.ije.ir

A Hybrid Deep Learning Architecture Using 3D CNNs and GRUs for Human Action Recognition

M. Savadi Hosseini, F. Ghaderi*

Human-Computer Interaction lab., Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran

P A P E R I N F O

Paper history:
Received 09 February 2020
Received in revised form 06 April 2020
Accepted 07 April 2020

Keywords:
Inflated 3D Convolutional Neural Networks
Recurrent Gate Unit
Human Action Recognition
Two-stream Architecture

A B S T R A C T

Video contents have variations in temporal and spatial dimensions, and recognizing human actions requires considering the changes in both directions. To this end, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) and their combinations have been used to tackle the video dynamics. However, a hybrid architecture usually results in a more complex model and hence a greater number of parameters to be optimized. In this study, we propose to use a stack of gated recurrent unit (GRU) layers on top of a two-stream inflated convolutional neural network. Raw frames and optical flow of the video are processed in the first and second streams, respectively. We first segment the video frames in order to be able to track the video contents in more details and by using 3D CNNs extract spatial-temporal features, called local features. We then import the sequence of local features to the GRU network, and use a weighted averaging operator to aggregate the outcome of the two processing flows, called global features. The evaluations confirm acceptable results for the two HMDB51 and UCF101 datasets. The proposed method resulted in a 1.6% improvement in the classification accuracy of the HMDB51 challenging dataset compared to the best reported results.

***doi**: 10.5829/ije.2020.33.05b.29*

1. INTRODUCTION

Over 500 million single video contents are uploaded to YouTube every minute [1]. This is only an example indicating the huge amount of video content being created and used continuously. Therefore, it is necessary to develop intelligent systems capable of video perception. Human action recognition in video content is one of the most important parts of such systems, and has attracted increasing attention in recent years. This research field has many applications including surveillance, security, and building management systems, self-driving cars, robot vision, smart cities, and etc.

Despite the efforts for improving the performance of the existing human action recognition algorithms, selecting appropriate features from the raw video frames is still a challenging task. In order to be able to percept the video content, it is essential to consider both temporal and spatial dynamics of the videos. In general, there are two main approaches used in literature: a) approaches that are based on hand-crafted features, and b) approaches that use deep learning algorithms in order to learn the appropriate representation.

While approaches based on hand-crafted features were growing, deep learning had great success in areas such as image, sound, and speech processing. In recent years, various architectures for action recognition were proposed that outperform the classic solutions to human action recognition problem. This is, in general, because deep learning algorithms do not rely on feature engineering by humans, and instead learn the appropriate representation of the available data [2]. To this end, variety of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) and their combinations have been used so far by researchers. However, they are mostly suffering from complexity of the networks and high number of parameters. In this paper we propose a hybrid model using inflated 3D CNNs and gated recurrent networks (GRUs), that

*Corresponding Author Institutional Email: *fghaderi@modares.ac.ir* (F. Ghaderi)

Please cite this article as: M. Savadi Hosseini, F. Ghaderi, A Hybrid Deep Learning Architecture Using 3D CNNs and GRUs for Human Action Recognition, International Journal of Engineering (IJE), IJE TRANSACTIONS B: Applications Vol. 33, No. 5, (May 2020) 959-965

outperforms the state-of-the-art results on the challenging HMDB51 dataset. Our contribution can be summarized as follows:

- We used a combination of 3D CNNs and GRUs for human action recognition task. The 3D CNNs are used to extract spatial-temporal features at a video segment level, while GRUs extract the features at a global level from the whole video.
- We used non-overlapping segments of video frames as input to the 3D CNNs. This helps to track the spatial-temporal dynamics of the video content in a segment-by-segment level.
- We used the weighted average operator to combine the outcome of optical flow and raw image streams. This operator improves the performance of the model.

The manuscript is structured as follows. In the next section, we briefly review the literature. The proposed method will be described in Section 3. We devote the fourth section to the evaluation, which begins with the introduction to the dataset and concludes by comparing the results with other studies. Finally, discussion and conclusion sections are presented in Sections 5 and 6, respectively.

2. RELATED WORKS

The classic approach to human action recognition problem is based on using hand-crafted features and using them for classification. Bobick et al. [3] used motion energy images and motion history images to identify actions. Hu et al. [4] extracted foreground gradient-oriented histograms to identify actions by separating the foreground images of motion history. Roh et al. [5] expanded the motion history images from 2D to 3D. Dense trajectories [6] were a turning point in methods based on hand-crafted features. These trajectories are extracted using dense sampling and recording of locomotor characteristics. Wang et al. [7] provided improved directions by estimating camera movements. Dense trajectories require heavy computation. Eleonora et al. [8] have eliminated the associated paths by identifying stationary points. The most popular approach based on hand-crafted features is the bag of visual words [9, 10] or the alternatives derived from it [11, 12]. The bag of visual words utilizes valuable features such as dense trajectories as well as powerful coding techniques such as Fisher vectors. Peng et al. [13] used a stack of Fisher vectors instead of one layer of it. Fernando et al. [14] presented a functional approach to extract the temporal evolution of motions. This approach was expanded hierarchically [15]. Simonyan et al. [16] presented a two-stream architecture that used a video frame as space stream and a stack of optical flow for temporal representation. Space stream is trained using ImageNet dataset [17]. Splitting networks [18, 19] expanded the architecture by increasing the depth of its two-stream architecture. Although this architecture uses a larger number of frames (3 or 7), they are still insufficient to identify complex operations and similar classes. A linear time encoding approach [20] was proposed to solve this problem. Zhu et al. [21] used more time frames to derive the time features to obtain the time pyramid integration layer and hence obtained better results.

On the other hand, considering video as 3D cubes, 3D CNNs can well capture the spatial-temporal features. Ji et al. [22] developed a 3D-based convolutional grid architecture that resulted in high computational cost due to large filters. The authors of [23], reduced the computational cost by using $3 \times 3 \times 3$ filters in all layers. The success of deep networks in classifying 2D images encouraged researchers to transfer successful 2D architectures to their 3D models. Carreira and Zisserman [24] achieved promising results by modifying the Inception-V1 [25] architecture to a 3D model. In a comparative study, Kurmanji et al. compared the effectiveness of 2D and 3D CNNs in hand gesture recognition task [26].

Although 3D CNNs incorporate spatio-temporal features, they are not time-sensitive in chronological order. Actions in different videos usually follow a certain order, regardless of the different speeds. That's why RNNs can be more useful. One of the most widely used types of these networks is long short-term memory (LSTM) [27]. Yong et al. [28] proposed a hierarchical approach based on RNNs. Although LSTMs have performed well in other areas such as voice and text processing [29-33], they have not been successful in human action recognition. Joe et al. [34] used an LSTM layer after a CNN network. They also used an aggregation layer in a similar network.

3. PROPOSED METHOD

In this method, local spatial-temporal features are first extracted using inflated 3D convolutional neural networks (I3D) [24] from the single shot video streams. The local features are extracted from a small portion of the video. The output sequences of the I3D network, called local spatial-temporal features, are then fed to a multilayer grid of GRU cells. Finally, the outcome of the two spatial and temporal flows, called global features, are pooled together in the weighted average layer. The details of the proposed architecture, illustrated in Figure 1, are presented in the sequel.

3. 1. Video Segmentation As depicted in Figure 1, the video sequence of length N is divided into T equal-sized segments (probably except the last segment) of length L. Each of these segments, called spatial-temporal

subsequences, are used in two separate processing flows to generate spatial and temporal features. To be able to extract temporal features, optical flow is calculated for consecutive frames.

The subsequences are formed in non-overlapping manner, i.e., frames t to $t+L-1$ are used in the current subsequence, and frames $t+L$ to $t+2L-1$ are used in the following subsequence, where L is the number of frames passed to the next step simultaneously as a 3D tensor. Obviously, in this form there is no duplicate frame in consecutive subsequences. Converting a long video to a sequence of short segments can reduce the complexity of extracting spatial-temporal features and hence, enhance the learning process.

3. 2. 3D CovNet The spatial-temporal subsequences derived in the previous step can be used to extract relevant features. The set of features extracted from each subsequence is called local spatial-temporal features. In recent years, many successful architectures have been presented for feature extraction and image classification, many of which have been developed using trial and error approach. Following the proposed approach in [24], we simply convert successful image classification models from 2D to 3D instead of repeating the chore process of evaluating different spatial-temporal models. This is done by repeating two-dimensional filters in the direction of time axis in order to convert the square filters into cubic ones. In this study, we used the I3D architecture presented by Carreira and Zisserman [24]. The inception module used in this architecture is shown in Figure 2.

3. 3. Recurrent Layers After extracting the local features by I3D, a sequence of feature vectors is obtained that include the temporal order of local motions. We pass this sequence to a network of gated recurrent unit (GRU) cells. It should be noted that it is possible to configure and implement GRUs that perform similar to LSTMs, while they have less complexity and fewer number of parameters. We call the output of the GRU layers global features, because they contain information about the whole video stream including all subsequences.

3. 4. Aggregation Layer We use two spatial and temporal streams in the network, as mentioned previously. Usually, the outcome of the two streams are aggregated using average operator. In such cases, the impact of space and time flows on the human action recognition task is considered to be equal. However, different conditions e.g., the action types, length of movies, illuminations, etc. may affect the contribution of each flow in the final prediction. Here, we use the weighted average instead of simple average so that the two streams are not forced to have identical contributions. To this end, the weighted average is defined as follows:

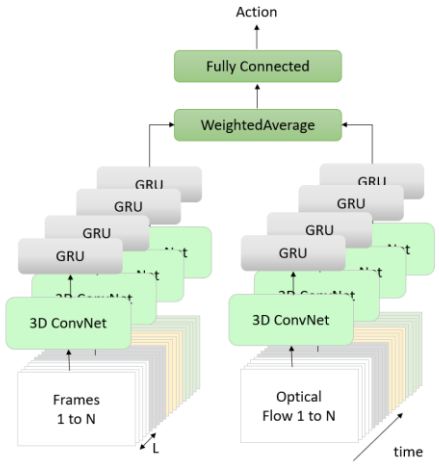

Figure 1. Overall architecture of the proposed method. First, optical flow of the video frames is calculated, to obtain two separate temporal and spatial processing flows. Next, we divide each stream into subsequences of length L. The subsequences are fed to 3D ConvNets independently, providing two sequences of local features. We then feed these feature sequences to a GRU network for each stream. Finally, using a weighted average layer, we merge the output of the two streams and give the output to a softmax layer for action recognition

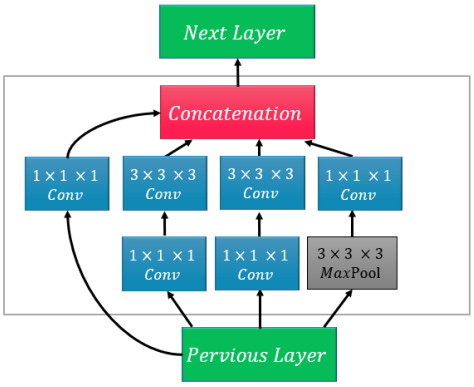

Figure 2. Inflated Inception module. Illustrated is the main module used in I3D architecture for our proposed method

$$weighted\ Average = \alpha \times A + (1-\alpha) \times B \quad (1)$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are the outputs of the spatial and temporal flows. The value of α can be adjusted manually, however, in sake of appropriate value for this parameter we assigned the task to the deep network to learn the parameter.

4. EVALUATION

In this section, we evaluate the proposed method. First, the datasets are introduced. Then, the impact of the

proposed innovations is examined. Finally, the effectiveness of the proposed method in comparison with other methods is discussed.

4. 1. Dataset We use HMDB51 [35] and UCF101 [36] datasets to evaluate the proposed method. The HMDB51 dataset was created from YouTube videos and contains 6766 videos in 51 classes. The UCF101 dataset consists of 13320 videos in 101 different classes. For each of these datasets, three separate splits are presented by the data providers for train/test purpose. We used the same splits in our experiments. The results are reported in terms of classification accuracy to be comparable with those of the other studies.

4. 2. Implementation Details To implement the proposed method, we used the Keras library on the Tensorflow open-source platform. Keras is an open source library that enables the deep learning models to be implemented quickly. We also used Numpy library for data processing and OpenCV for low-level video processing. An NVIDIA GeForce GTX 1080 Ti GPU was utilized in our experiments. To derive the optical flow, we used the TV-L1 algorithm [37] which uses all pixels. Moreover, we used RMSProp algorithm as optimizer. The learning rate of this algorithm was set to 0.001 and the weight decay is 0.01. Two layers of GRU with a width of 256 neurons were used for initial evaluation. We have also used the I3D in [24] that was pre-trained on the Kinetics dataset [38]. We have also used the ReLU activation function in each layer. The final classification was done by a softmax layer, which follows the fully connected layer.

4. 3. Investigating the Effect of Innovations In this section, we examine the effect of innovations in performance of the model. First, the role of the weighted average layer on the performance of the model is evaluated. Afterwards, the effect of using spatial-temporal segmentation and GRU are investigated.

Aggregation layer The use of averages operator for aggregation is a common practice. However, when we average, we implicitly assume the effect of the two spatial and temporal streams are identical, whereas based on the application and the data, this may not be the case. In a data-deriven approach, we assigned the task of learning the contribution of each stream to the network. The results of using the weighted average are shown in Table 1.

Video segmentation The results of some typical L values on the performance of the method are presented in Tables 2 and 3. It can be seen that except for split 1 of HMDB51 dataset, L=20 provides the best result. Therefore, this value of L is used in the rest of the experiments.

Using GRUs Our proposed method uses GRUs instead of LSTMs because of their lower number of parameters. As shown in Table 4, two layers of GRU outperform LSTM.

4. 4. Comparison with Other Methods Here, we compare our results with those of other studies. First, we report the best obtained results from our evaluations (Table 5). The mean for the best results is calculated on three standard splits for each dataset. The selected studies are those that have recently achieved good results on the datasets. Many of these methods have used deep learning algorithms. In Tables 6 and 7, the results of the proposed method are compared with those of the other studies on the HMDB51 and UCF101 datasets, respectively. As can be seen, the result of our proposed method is 1.6% better than that of the best available method.

TABLE 1. Evaluation of the effect of weighted average layer on classification accuracy for split-1 of datasets

DataSets	Temporal	Spatial	Average	W. Average
HMDB51	78.36	80.13	83.32	84.59
UCF101	89.91	90.36	92.36	95.65

TABLE 2. Evaluation of the effect of spatial-temporal sub-sequence lengths on the classification accuracy of the HMDB51 dataset

Splits	L=10	L=20	L=30
Split1	84.59	83.89	76.07
Split2	82.91	83.59	77.03
Split3	83.01	84.08	78

TABLE 3. Evaluation of the effect of spatial-temporal subsequence length on the classification accuracy of the UCF101 dataset

Splits	L=10	L=20	L=30
Split1	94.95	95.65	91.02
Split2	96.48	96.87	92
Split3	94.49	95.76	91.26

TABLE 4. Comparing different recurrent layers on split 1 of HMDB51 dataset

Architecture	Parameter	Accuracy
2 GRU layers with 256 units	2.768.692	84.59
2 LSTM layers with 256 units	3.687.220	78.91
1 LSTM layers with 512 units	6.321.716	82.32

TABLE 5. The best obtained results on the two HMDB51 and UCF101 datasets using our method.

Splits	HMDB51	UCF101
Split1	84.59	95.65
Split2	83.59	96.87
Split3	84.08	95.76
AVERAGE	84.08	96.09

TABLE 6. Comparison of the results on HMDB51 dataset.

Method	Description	year	Accuracy
Proposed method	I3D + GRU	2020	84.08
Wang et al. [39]	HAF+BOW/FV	2019	82.48
Zhu et al. [21]	DTPP	2018	82.01
Wang et al. [40]	SVMP + I3D	2018	81.3
Roy et al. [41]	-	2018	81.1
Carreira and Zeisserman [24]	Two-Stream I3D + Kinetics pre-training	2017	80.7
Fiza et al. [42]	iDT+DA-VLAD	2018	80.1
Tran Do et al. [43]	-	2018	78.7
Zhu Yi et al. [44]	-	2018	78.7
Shuang et al. [45]	-	2018	74.2

TABLE 7. Comparison of the results on UCF101 dataset

Method	Description	year	Accuracy
suggested method	-	2019	96.09
Choutas et al. [46]	I3D + PoTion	2018	98.2
Carreira and Zeisserman [24]	Two-Stream I3D + Kinetics pre-training	2017	98
Zhu et al. [21]	DTPP	2018	98
Ali Diba et al. [47]	HATNet (32 frames)	2019	97.7
Tran Do et al. [43]	-	2018	97.3
Zhu Yi et al. [44]	Hidden Two-Stream (I3D)	2018	97.1
Yang et al. [48]	Full IF-TNN	2019	96.2
Xuyang et al. [45]	-	2018	96
Zheng et al. [49]	S-TPNet + iDT	2019	96

5. DISCUSSIONS

Convolutional neural networks have been used widely for different image processing applications. By learning a 2D convolution kernel, these types of networks are able to learn appropriate representation of the input images. Although, the CNNs have shown promising results in image classification, they are not good for video processing in general. This is because the convolution operator can only extract spatial features, and misses the temporal dynamics of the video content. Three-dimensional CNNs are used to overcome this problem, in which a 3D kernel is used to extract both spatial and temporal features.

In this research, in an attempt to use both spatial and temporal features, we used subsequences of raw images and their derived optical flow (a widely used technique for extracting temporal features) in two separate processing flows. The frames of the subsequence are passed to the 3D CNNs and the output of the CNNs are used as the input to the RNNs. Finally, the RNN outputs of the spatial and temporal flows are combined using a weighted average mechanism to generate the input to the softmax layer.

One of the major problems with LSTM networks is their high number of parameters and hence difficulty of training. In case of small number of training examples, this may yield an overfitted model. To reduce the number of network parameters, we used GRUs instead of LSTMs, which in most cases have almost similar performance as LSTMs and at the same time have fewer parameters. The outcome of this approach is presented in Table 4. As it can be seen the classification accuracy is increased using the GRU layers, while the number of parameters is reduced about 56% compared with the case of using LSTM.

As shown in Tables 6, our method outperforms the state-of-the-art reported results. However, as presented in Table 7, our approach may not perform the best for the UCF101 dataset. The reason might be from the fact that the I3D network used in our approach is not fine-tuned. This could have been the case for the HMDB51 dataset as well, but because the UCF101 dataset has more classes and some of them are very similar, the whole network needs to be trained end-to-end. This problem can be solved by fine-tuning the last few layers of the I3D to achieve better performance.

Contribution of temporal and spatial extracted features is applications-specific and depends on different conditions of the dataset. Therefore, instead of using the average pooling technique, we use a weighted average and let the model to learn the share of each stream based on the characteristics of the existing dataset. Table 1 shown that although using a simple average provides better results compared with the single temporal or spatial features, a weighted average pooling outperforms the simple average for both datasets.

6. CONCLUSION

In this research, we addressed the challenges of using recurrent neural networks in recognizing human actions in video content by developing a new architecture that

utilizes basic deep image classification algorithms. We also succeeded in extracting global features using a hybrid architecture of three-dimensional convolutional neural networks, derived from successful two-dimensional counterparts, and a stack of recurrent neural networks. Low-level inflated 3D neural networks, transform the sequence of video frames into shorter sequences by extracting locally useful spatial-temporal features. This enables recurrent neural networks to better capture global features.

When structuring the spatial-temporal subsequences, we used non-overlapping windows. The main reason behind this decision was to reduce the computational costs. One can investigate about the effects of overlapping windows on the overall performance of the proposed method. Moreover, only few values of L have been investigated (Tables 2 and 3). Another direction of future research would be to find the optimal value of L and to analyze the dependence between the datasets and the optimal value of L.

Our proposed approach improved the classification accuracy 1.6% on the challenging HMDB51 dataset compared with the best available method. For the UCF101 dataset, one can obtain better results by training the whole networks, rather than using pre-trained components.

7. REFERENCES

1. Hale, J., "More than 500 hours of content are now being uploaded to youtube every minute", ***Santa Monica, CA: Tubefilter***, (2019) https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/..
2. Goodfellow, I., Bengio, Y. and Courville, A., "Deep learning, MIT press, (2016).
3. Bobick, A.F. and Davis, J.W., "The recognition of human movement using temporal templates", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 23, No. 3, (2001), 257-267.
4. Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y. and Huang, T.S., "Action detection in complex scenes with spatial and temporal ambiguities", in 2009 IEEE 12th International Conference on Computer Vision, IEEE. (2009), 128-135.
5. Roh, M.-C., Shin, H.-K. and Lee, S.-W., "View-independent human action recognition with volume motion template on single stereo camera", ***Pattern Recognition Letters***, Vol. 31, No. 7, (2010), 639-647.
6. Wang, H., Kläser, A., Schmid, C. and Liu, C.-L., "Action recognition by dense trajectories", in CVPR 2011, IEEE. (2011), 3169-3176.
7. Wang, H. and Schmid, C., "Action recognition with improved trajectories", in Proceedings of the IEEE international conference on computer vision. (2013), 3551-3558.
8. Vig, E., Dorr, M. and Cox, D., "Space-variant descriptor sampling for action recognition based on saliency and eye movements", in European conference on computer vision, Springer. (2012), 84-97.
9. Peng, X., Wang, L., Wang, X. and Qiao, Y., "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice", ***Computer Vision and Image Understanding***, Vol. 150, (2016), 109-125.
10. Sivic, J. and Zisserman, A., "Video google: A text retrieval approach to object matching in videos", in null, IEEE. (2003), 1470.
11. Liu, L., Wang, L. and Liu, X., "In defense of soft-assignment coding", in 2011 International Conference on Computer Vision, IEEE. (2011), 2486-2493.
12. Perronnin, F., Sánchez, J. and Mensink, T., "Improving the fisher kernel for large-scale image classification", in European conference on computer vision, Springer. (2010), 143-156.
13. Peng, X., Zou, C., Qiao, Y. and Peng, Q., "Action recognition with stacked fisher vectors", in European Conference on Computer Vision, Springer. (2014), 581-595.
14. Fernando, B., Gavves, E., Oramas, J., Ghodrati, A. and Tuytelaars, T., "Rank pooling for action recognition", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 39, No. 4, (2016), 773-787.
15. Fernando, B., Anderson, P., Hutter, M. and Gould, S., "Discriminative hierarchical rank pooling for activity recognition", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016), 1924-1932.
16. Simonyan, K. and Zisserman, A., "Two-stream convolutional networks for action recognition in videos", in Advances in neural information processing systems. (2014), 568-576.
17. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database", in 2009 IEEE conference on computer vision and pattern recognition, IEEE. (2009), 248-255.
18. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X. and Van Gool, L., "Temporal segment networks for action recognition in videos", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 41, No. 11, (2018), 2740-2755.
19. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X. and Van Gool, L., "Temporal segment networks: Towards good practices for deep action recognition", in European conference on computer vision, Springer. (2016), 20-36.
20. Diba, A., Sharma, V. and Van Gool, L., "Deep temporal linear encoding networks", in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. (2017), 2329-2338.
21. Zhu, J., Zhu, Z. and Zou, W., "End-to-end video-level representation learning for action recognition", in 2018 24th International Conference on Pattern Recognition (ICPR), IEEE. (2018), 645-650.
22. Ji, S., Xu, W., Yang, M. and Yu, K., "3d convolutional neural networks for human action recognition", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 35, No. 1, (2012), 221-231.
23. Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M., "Learning spatiotemporal features with 3d convolutional networks", in Proceedings of the IEEE international conference on computer vision. (2015), 4489-4497.
24. Carreira, J. and Zisserman, A., "Quo vadis, action recognition? A new model and the kinetics dataset", in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017), 6299-6308.
25. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., "Going deeper with convolutions", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2015), 1-9.
26. Kurmanji, M. and Ghaderi, F., "Hand gesture recognition from rgb-d data using 2d and 3d convolutional neural networks: A comparative study", ***Journal of AI and Data Mining***, Vol. 8, No. 2, (2020), 177-188.

27. Hochreiter, S. and Schmidhuber, J., "Long short-term memory", ***Neural Computation***, Vol. 9, No. 8, (1997), 1735-1780.
28. Du, Y., Wang, W. and Wang, L., "Hierarchical recurrent neural network for skeleton based action recognition", in Proceedings of the IEEE conference on computer vision and pattern recognition., (2015), 1110-1118.
29. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H. and Courville, A., "Describing videos by exploiting temporal structure", in Proceedings of the IEEE international conference on computer vision. (2015), 4507-4515.
30. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y., "Show, attend and tell: Neural image caption generation with visual attention", in International conference on machine learning. (2015), 2048-2057.
31. Vinyals, O., Toshev, A., Bengio, S. and Erhan, D., "Show and tell: A neural image caption generator", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2015), 3156-3164.
32. Sutskever, I., Vinyals, O. and Le, Q.V., "Sequence to sequence learning with neural networks", in Advances in neural information processing systems. (2014), 3104-3112.
33. Graves, A., Jaitly, N. and Mohamed, A.-r., "Hybrid speech recognition with deep bidirectional lstm", in 2013 IEEE workshop on automatic speech recognition and understanding, IEEE. (2013), 273-278.
34. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R. and Toderici, G., "Beyond short snippets: Deep networks for video classification", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2015), 4694-4702.
35. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. and Serre, T., "Hmdb: A large video database for human motion recognition", in 2011 International Conference on Computer Vision, IEEE. (2011), 2556-2563.
36. Soomro, K., Zamir, A.R. and Shah, M., "Ucf101: A dataset of 101 human actions classes from videos in the wild", ***arXiv preprint arXiv:1212.0402***, (2012).
37. Pérez, J.S., Meinhardt-Llopis, E. and Facciolo, G., "Tv-l1 optical flow estimation", ***Image Processing On Line***, Vol. 2013, (2013), 137-150.
38. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T. and Natsev, P., "The kinetics human action video dataset", ***arXiv preprint arXiv:1705.06950***, (2017).
39. Wang, L., Koniusz, P. and Huynh, D.Q., "Hallucinating bag-of-words and fisher vector idt terms for cnn-based action recognition", ***arXiv preprint arXiv:1906.05910***, (2019).
40. Wang, J., Cherian, A., Porikli, F. and Gould, S., "Video representation learning using discriminative pooling", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018), 1149-1158.
41. Roy, D., Murty, K.S.R. and Mohan, C.K., "Unsupervised universal attribute modeling for action recognition", ***IEEE Transactions on Multimedia***, Vol. 21, No. 7, (2018), 1672-1680.
42. Murtaza, F., HaroonYousaf, M. and Velastin, S.A., "Da-vlad: Discriminative action vector of locally aggregated descriptors for action recognition", in 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE. (2018), 3993-3997.
43. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. and Paluri, M., "A closer look at spatiotemporal convolutions for action recognition", in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. (2018), 6450-6459.
44. Zhu, Y., Lan, Z., Newsam, S. and Hauptmann, A., "Hidden two-stream convolutional networks for action recognition", in Asian Conference on Computer Vision, Springer. (2018), 363-378.
45. Sun, S., Kuang, Z., Sheng, L., Ouyang, W. and Zhang, W., "Optical flow guided feature: A fast and robust motion representation for video action recognition", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2018), 1390-1399.
46. Choutas, V., Weinzaepfel, P., Revaud, J. and Schmid, C., "Potion: Pose motion representation for action recognition", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018), 7024-7033.
47. Diba, A., Fayyaz, M., Sharma, V., Paluri, M., Gall, J., Stiefelhagen, R. and Van Gool, L., "Holistic large scale video understanding", ***arXiv preprint arXiv:1904.11451***, (2019).
48. Yang, K., Fu, J., Guo, X., Lu, Y., Qiao, P., Li, D. and Dou, Y., "If-ttn: Information fused temporal transformation network for video action recognition", ***arXiv preprint arXiv:1902.09928***, (2019).
49. Zheng, Z., An, G., Wu, D. and Ruan, Q., "Spatial-temporal pyramid based convolutional neural network for action recognition", ***Neurocomputing***, Vol. 358, (2019), 446-455.

Persian Abstract

چکیده

محتوای ویدئو در محورهای زمان و مکان تغییر می‌کند و برای شناسایی اعمال انسان در ویدئو باید تغییرات در راستای هر دو محور مورد بررسی قرار گیرد. بدین منظور شبکه‌های عصبی پیچشی (CNN) و شبکه های عصبی بازگشتی (RNN) و ترکیبات آن‌ها برای دنبال کردن دینامیک ویدئو مورد استفاده قرار می‌گیرد. معماری‌های ترکیبی معمولا شامل مدل‌های پیچیده‌تر هستند و بنابراین هزینه محاسباتی بیشتری در جریان آموزش پارامترهای زیاد آن‌ها تحمیل می‌شود. در این تحقیق، استفاده از یک معماری شامل لایه‌هایی از واحدهای بازگشتی دروازه‌ای (GRU) بر روی شبکه‌های عصبی پیچشی متورم دو جریانی پیشنهاد شده است. در جریان اول فریم‌های خام، و در جریان دوم شار نوری پردازش می‌شود. در ابتدا به منظور ایجاد قابلیت دنبال کردن تغییرات در ویدئو با جزئیات بیشتر، قاب‌های ویدئو قطعه بندی شده و با استفاده از CNN های سه بعدی ویژگی‌های زمانی–مکانی، که در اینجا ویژگی‌های محلی نامیده می‌شوند، از آن‌ها استخراج می‌شوند. سپس دنباله ایجاد شده از ویژگی‌های محلی به شبکه GRU داده شده و خروجی دو جریان پردازشی را، که ویژگی‌های سراسری می‌نامیم، با استفاده از عملگر میانگین وزن‌دار تجمیع می‌کنیم. ارزیابی‌ها نتایج قابل قبول روش پیشنهادی را بر روی مجموعه داده‌های HMDB51 و UCF101 تایید می‌کنند. این روش برای مجموعه داده پرچالش HMDB51 نسبت به بهترین جواب گزارش شده در سایر پژوهش‌ها حدود ۱٫۶٪ در صحت دسته‌بندی بهبود ایجاد نموده است.