Video Human Action Recognition by the Fusion of Significant Motion and Coarse-fine Temporal Granularity Features

Hanhua Cao[1]

[1]Guangzhou Xinhua University

April 25, 2024

Abstract

Video human action recognition is an important academic issue, and due to the challenges involved in the problem, such as complex scenes, large changes in spatial scale, and irregular deformation of the identified targets, the algorithm generally has some shortcomings such as large parameter quantities and high computational costs. This paper proposes a new computing framework based on lightweight deep learning models, using time-domain Fourier transform to generate motion salience maps to highlight human motion areas in feature extraction in which video intrasegment and inter-segment Feature differences at different time scales are used to obtain action information at different time granularity, to conduct more accurate action model of the human action with varying spatio-temporal scales. Additionally, an action excitation method based on the deformable convolution is proposed to solve problems of irregular deformation, spatial multi-scale changes, and the loss of underlying information as the network depth increases. Data experiments are proposed to verify the effectiveness of the proposed algorithm in terms of computational efficiency and accuracy.

Video Human Action Recognition by the Fusion of Significant Motion and Coarse-fine Temporal Granularity Features

Kaishi Xu[1*], Hanhua Cao[2], Yang YI[2]

[1]*Huawei HiSilicon, Shanghai, P.R.China*

[2]*Guangzhou Xinhua University, Guangzhou, P.R.China*

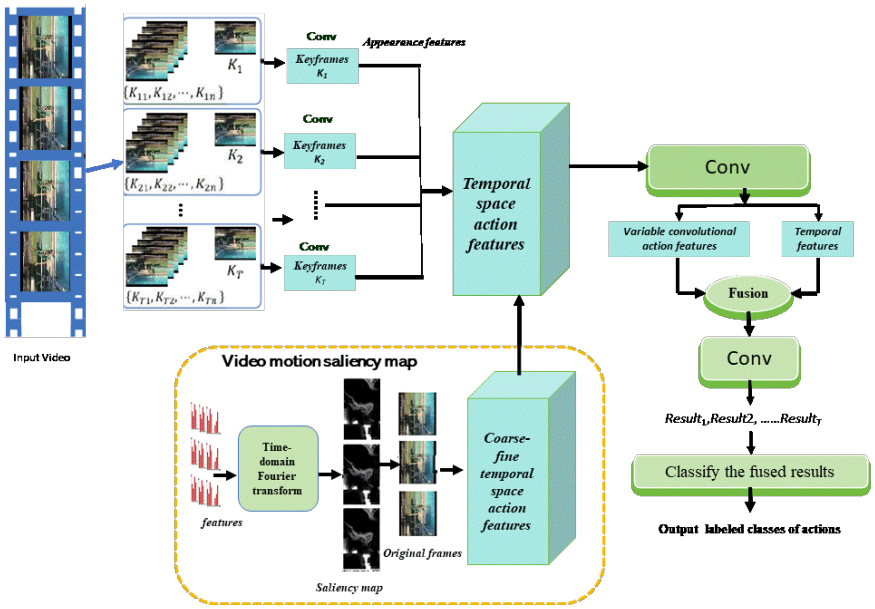
[*]E-mail:kaishixu@hotmail.com

Video human action recognition is an important academic issue, and due to the challenges involved in the problem, such as complex scenes, large changes in spatial scale, and irregular deformation of the identified targets, the algorithm generally has some shortcomings such as large parameter quantities and high computational costs. This paper proposes a new computing framework based on lightweight deep learning models, using time-domain Fourier transform to generate motion salience maps to highlight human motion areas in feature extraction in which video intrasegment and inter-segment Feature differences at different time scales are used to obtain action information at different time granularity, to conduct more accurate action model of the human action with varying spatio-temporal scales. Additionally, an action excitation method based on the deformable convolution is proposed to solve problems of irregular deformation, spatial multi-scale changes, and the loss of underlying information as the network depth increases. Data experiments are proposed to verify the effectiveness of the proposed algorithm in terms of computational efficiency and accuracy.

Introduction: The computational frameworks of video human action recognition based on deep learning mainly include CNN, Recurrent Neural Network (RNN) and Transformer. Simonyan et al. [1] proposed a dual-stream CNN for video human action recognition, and Feichtenhofer et al. [2] used a residual network

(ResNet) to implement a dual-stream structure, though required additional optical flow images to obtain action features. Recurrent Neural Network (RNN) can model time series data [3], but it cannot fully utilize spatio-temporal information. Tran et al. [4] proposed the C3D network, Tran et al. [5] extended ResNet and proposed the Res3D network, and Diba et al. [6] designed the Temporal 3D Convolutional Network (T3D).

In addition, Girdhar et al. [7] combined Fast-RCNN and Transformer and proposed a Video action transformer network, while Bertasiu et al. [8] proposed a TimeSformer network based on a distributed attention



mechanism. Although Transformer-based video algorithms has achieved advanced accuracy results, its memory needed and computational overhead are large, thus the subsequent research are mainly focusing and trying to reduce memory costs and computational complexity.

Methods: A new framework for video human action recognition is proposed which does not require additional extraction of optical flow information, neither uses 3D convolution kernels and numerous attention modules. In addition, the method of using three dimensions of time, space, and channels to obtain more discriminative capabilities is developed. The time and spatial characteristics and action features of human behavior are stimulated and integrated in the three dimensions of time, space and channel to form more discriminative spatio-temporal action features, so that the computational network model can simultaneously process time and spatial features and action features, thereby improving the calculation effectiveness.

The main computational process is demonstrated in Figure 1, which is based upon deep learning model by coarse and fine time-granularity combining motion salience and multi-dimensional excitation.

1. Based on the dual-stream time segmentation deep learning network, the video is divided into *?* segments with equal time intervals and no overlap, ? = {?1, ?2, ?3, [?] , ?? }, and sample one frame as key-frame, called ?????, in each segment, ????? = {?1, ?2, ?3, [?] , ?? } and sample ? frames non-key-frame, as ??????.
2. By using time-domain Fourier transform, the pixel changes in video frames caused by action changes in the time dimension are utilized to obtain motion salient regions and generate motion salient maps. And by this graph, the original key-frames ????? and non-key-frames ?????? are achieved, to highlight the significant motion regions where human behavior is located and weaken irrelevant background information.

3. The feature extraction method of spatio-temporal differential based on coarse-fine time granularity is used to model the action explicitly. The features of the results from time and spatial differentiation on the coarse time granularity is taken to represent the long-term action changes between video segments. More refined action changes are represented by the results from spatio-temporal differentiation on the fine time granularity. Finally, integrating the long-term and the short-term action changes.
4. 2D CNN is used to extract video spatial information and generate appearance features, and then the appearance features are integrated with the action features, thus the new spatio-temporal characteristics of videos with stronger expressive abilities are constructed jointly.
5. In order to reduce computational complexity while ensuring the recognition accuracy of the algorithm, a deformable convolution based action feature excitation method (DCME) is used to excite the motion information in time and spatial action features. It utilizes two deformable convolutions of different scales to perform feature differentiation on adjacent key-frames in the complete video, obtaining excitation parameters that can enhance action features, thereby generating video level spatio-temporal action features.
6. Due to the varying importance of different channel features, a spatio-temporal channel excitation method based on time correlation is used to obtain temporal and spatial information of the channel dimension, and the two are fused to generate the time and spatial features of the channel dimension. Combine it with the video level action features of the time and space generated by action motivation methods to jointly construct stronger ability to describe action features with richer action and information.

In summary, the difference between our presented algorithm and the existing 2D CNN methods that integrate spatio-temporal and motion features lies in: (1) the introduction of motion salience which making this type of 2D CNN method more robust for recognizing human action in complex scene videos; (2) By combining the interaction information between frames within video segments and between frames within video segments, the time and spacial features and action features were built and fused at coarse and fine time granularity, respectively. A more accurate model of spatio-temporal action features at the video segment level was carried out in the time dimension; (3) Adopting two feasible variable convolutions with different scales to stimulate the actions features with diverse spatial scale changes and irregular deformations, and meanwhile, the features ae combined with the characteristics in the channel dimension, so the spatio-temporal action characteristics are stimulated in both spatial and channel dimensions.

Data Set: Four popular benchmarks in video recognition are used in the data experiments of this paper, UCF101 [9] (101 categories of actions and 13320 video clips), HMDB51 [10] (51 action categories and 6766 video clips), Kindetics-400 [11] (400 action categories, each containing at least 400 video clips), and Something Something [12] (with two versions, SSV1 and SSV2, both containing 174 action categories). All the data are divided into training and testing sets using the official designated division rules. The scene related datasets are UCF101, HMDB51, and Kinetics-400, while the time related datasets are Something-Something.

Computational Process and Prameters: Using Ubuntu 18.04, Intel Xeon E5-2620-2.10GHz CUP, 64GRAM, equipped with 4 GeForce GTX 1080 Ti GPUs, using Python, using a deep learning framework of Pytorch 1.8+CUDA 11.0.

Training phase. Firstly, the video was first divided into equal interval ? segments, and then randomly sample 5 frames in each segment, selecting the 4th frame as the key-frame. The short edge size of these frames is fixed at 256, and then they are adjusted to 224 x 224 using random cropping and core cropping as inputs to the network. Therefore, the input of the network is [?, ?, 3, 224, 224], where ? is the batch size and ? is the number of segments or key-frames for each video. In this experiment, ? is set to 8 or 16.

ResNet50 is taken as the backbone network and initialize the computation with a pre-trained model on ImageNet. The learning rate starts from 0.01, and a small batch stochastic gradient descent algorithm (mini batch SGD) with a momentum of 0.9 is used, with a weight decay rate of 0.0001. Adjustments are made at the 30th, 40th, and 50th epochs, resulting in a total of 70 epochs being trained.

Testing phase. Two testing strategies are used to balance recognition rate of accuracy and efficiency. One strategy is core drop/1 clip, which cuts the core of each sampled video frame to 224 x 224 as network input. However, due to the randomness and sparsity of sampling, this strategy results in low recognition accuracy. Another sampling strategy with high recognition accuracy is 3-????/10-?????, which means each video is sampled 10 times, and the resulting video frames are cropped 3 times as inputs to the network. All recognition results are averaged to obtain the finals. This strategy obtains more comprehensive information due to multiple sampling and cropping, resulting in higher recognition accuracy but lower recognition efficiency.

The evaluation indicators used in this paper are divided into efficiency and accuracy, among which the efficiency evaluation indicator refers to the number of floating-point calculation (FLOPs), which is the computational complexity of the model. Evaluate the recognition accuracy rate of the network using the average accuracy (A). According to the different discrimination rules for correctly classified positive samples, they are divided into average accuracy (??), Top-1 accuracy ($?_{???1}$) (the principle for discriminating positive samples is that the classification result with the highest probability in the prediction result is the correct classification result), and Top-5 accuracy ($?_{???5}$) (the principle for discriminating positive samples is that the first five classification results with the highest probability in the prediction result contain the correct classification result).

Ablation Experiment : The goal of the ablation experiment is to improve the effectiveness of the algorithm. The original video frames and the video frames with prominent motion regions were used as input for action feature extraction, and experimental verification was conducted on four experimental datasets, as shown in Table 2 (the recognition accuracy of UCF101 and HMDB51 was obtained when the video segment number was 16, while the other two datasets were obtained when the video segment number was 8).

Table 1: Validation of the validity of the motion-significant map was used.

method	UCF101 (??%)	HMDB51 (??%)	Kinetics-400($?_{???1}$%)	SSV1 ($?_{???1}$%)
MDT-Net1[a]	94.9	71.2	75.1	53.2
MDT-Net2[b]	96.8	73.5	77.2	53.5

[a] Enter the raw video frame

[b] Input to video frames enhanced using motion-significant maps

It can be noticed that the method with the processing of motion salience maps improves its accuracy rate significantly on scene related datasets, such as UCF101 increased from 94.9% to 96.8%, HMDB51 increased from 71.2% to 73.5%, and Kinetics-400 increased from 75.1% to 77.2%. However, the time related and simple background dataset SSV1 does not show significant improvement in recognition accuracy. The main purpose of using motion salience maps is to reduce the interference of irrelevant background information. Since scene related datasets have cluttered background, the process of highlighting motion salience areas could help the computational network to focus on human action, and so the obtained results are more accurate.

Comparison with State-of-the-art: Compare our presented method with three basic 2D CNN methods TSN [3], TSM [13], and TEA [14], using TSN as the benchmark (almost all the advanced and popular 2D CNN based methods are all originated from TSN).

The experimental results showed that our presented method had better recognition accuracy rate than TSN and TSM on all four datasets, however the improvement compared to TEA [14] was relatively small or almost unchanged. The reason is that TSN [3] does not include modules that can describe temporal information and lacks the information of motion. Therefore, for scene related datasets UCF101, HMDB51, and Kinetics-400, its recognition accuracy rate is only slightly lower than the other three methods, while for time related datasets SSV1, it lags far behind.

Table 2: Comparison Results between our presented method and Three Benchmark 2D CNN Methods (The recognition accuracy of UCF101 and HMDB51 in the table was obtained when the video segment number was 16, while the other two datasets were obtained when the video segment number was 8.)

data set	method	??/????1%a	?(A)%
UCF101	TSN (2016) [3]	94.0	-
	TSM (2019) [13]	94.5	+0.5
	TEA (2020) [14]	96.9	+2.9
	This paper	96.8	+2.8
HMDB51	TSN (2016) [3]	68.5	-
	TSM (2019) [13]	70.7	+2.2
	TEA (2020) [14]	73.3	+4.8
	This paper	73.5	+5.0
Kinetics-400	TSN (2016) [3]	72.5	-
	TSM (2019) [13]	74.7	+2.2
	TEA (2020) [14]	76.1	+3.6
	This paper	77.2	+4.7
SSV1	TSN (2016) [3]	19.7	-
	TSM (2019) [13]	44.8	+25.1
	TEA (2020) [14]	51.9	+32.2
	This paper	53.5	+33.8

In addition, this paper also investigates the trade-off between computational complexity (FLOPs) and accuracy. In the scenario of Kinetics, among the four methods, TSN [3] has the highest computational complexity and the lowest recognition accuracy. For the time related SSV1, the computational complexity of TSN is half that of the other three, and the corresponding recognition accuracy rate is also half that of the other three, which proves the importance of incorporating time into the features more specifically.

Conclusion: A lightweight and efficient algorithm framework to solve the problem of human action recognition is discussed in this paper, in which a coarse-fine time granularity algorithm based on motion salience and multi-dimensional excitation and a method of action feature extraction based on motion salience and spatio-temporall difference are developed. Meanwhile, a method of motivating action features by deformable convolution and spatio-temporal channel excitation according to time related information are also be employed. The data experiments on most popular benchmarks have verified the advantage of our presented algorithms on the rate of accuracy and computational efficiency, and our future work will focus on exploring more strategies to improve the robustness of the algorithm.

Author contributions: **Kaishi Xu:** Conceptualization; supervision; writing-reviewing; editing; data curation; software; investigation. **Hanhua Cao** : writing-reviewing; editing; software; Investigation and methodology. **Yang YI** : writing—original drift; software; acquisition; Funding acquisition; Investigation and methodology.

Acknowledgments: This paper is partly supported by Key Discipline Project of Guangzhou Xinhua University with No.2020XZD02, Guangdong Key Discipline Scientific Research Capability Improvement Project with No.2021ZDJS144 and No.2022ZDJS151.

Conflict of interest statement: The authors declare no conflicts of interest.

Data availability statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Simonyan K.,Zisserman A. :Two-stream convolutional networks for action recognition in videos. *Proceedings of the 27th International Conference on Neural Information Processing Systems.* 568-576

(2014).https://dl.acm.org/doi/10.5555/2968826.2968890

2 Feichtenhofer C., Pinz A., Wildes R P.: Spatiotemporal multiplier networks for video action recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* . 4768-4777 (2017).https://doi.org/10.1109/CVPR.2017.787

3 Wang L., Xiong Y., Wang Z., Wang Z.,Qiao Y.,Lin D.,Tang X.,Gool L. Temporal segment networks: Towards good practices for deep action recognition. *Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016.* 20-36 (2016).https://doi.org/10.48550/arXiv.1608.00859

1. Tran D., Bourdev L., Fergus R., Torresani L.,Paluri M.: Learning spatiotemporal features with 3d convolutional networks.*2015 IEEE International Conference on Computer Vision (ICCV).* 4489-4497 (2015).https://doi.org/10.1109/ICCV.2015.510
2. Tran D., Ray J., Shou Z.,Chang S.,Paluri M.: Convnet architecture search for spatiotemporal feature learning [EB/OL]. *The IEEE International Conference on Computer Vision and Pattern Recognition* .1-12(2017).https://doi.org/10.48550/arXiv.1708. 05038
3. Diba A., Fayyaz M., Sharma V., Karami A.,Arzani M., Yousefzadeh R.,Gool L. Temporal 3d convnets: New architecture and transfer learning for video classification [EB/OL]. *The IEEE International Conference on Computer Vision and Pattern Recognition* .1-9(2017).https://doi.org/10.48550/ arXiv. 1711. 08200
4. Girdhar R., Carreira J., Doersch C., Zisserman A.: Video action transformer network. *The IEEE International Conference on Computer Vision and Pattern Recognition* . 244-253 (2019). https://doi.org/10.48550/arXiv.1812.02707
5. Bertasius G., Wang H., Torresani L.: Is space-time attention all you need for video understanding. *The IEEE International Conference on Computer Vision and Pattern Recognition* . 2(3)-4 (2021). https://doi.org/10.48550/arXiv.2102.05095
6. Soomro K., Zamir A R., Shah M. Ucf101: A dataset of 101 human actions classes from videos in the wild[EB/OL]. *The IEEE International Conference on Computer Vision and Pattern Recognition* . 1-7(2012).https://doi.org/10.48550/arXiv. 1212.0402
7. Kuehne H., Jhuang H., Garrote E.,Poggio T., Serre T. Hmdb: A large video database for human motion recognition. *2011 International Conference on Computer Vision* .2556-2563(2021). https://doi.org/10.1109/ICCV.2011.6126543
8. Carreira J., Zisserman A. : Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* .6299-6308 (2017).https://doi.org/10.1109/CVPR.2017.502
9. Goyal R., Ebrahimi Kahou S., Michalski V., Materzyńska J.,Westphal S.,Kim H.,Haenel V.,Fruend I.,Yianilos P.,Mueller-Freitag M.,Hoppe F., Thurau C.,Bax I.,Memisevic R.: The "something something" video database for learning and evaluating visual common sense.*2017 IEEE International Conference on Computer Vision (ICCV).* 5842-5850 (2017).https://doi.org/ 10.48550/arXiv.1706.04261
10. Lin J., Gan C., Han S. Tsm: Temporal shift module for efficient video understanding. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* .7083-7093(2019).https://10.1109/ICCV. 2019.00718
11. Li Y., Ji B., Shi X., Zhang J.,Kang B.,Wang L.: Tea: Temporal excitation and aggregation for action recognition.*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* .909-918(2019).https://10.1109/cvpr42600.2020.00099