Action Recognition with Non-Uniform Key Frame Selector

Haohe Li
Faculty of Electrical Engineering and Computer Science, Ningbo University, China
lhh_1023@hotmail.com

Chong Wang*
Faculty of Electrical Engineering and Computer Science, Ningbo University, China; Loctek Ergonomic Technology Corporation, China
wangchong@nbu.edu.cn

Shenghao Yu
Faculty of Electrical Engineering and Computer Science, Ningbo University, China
ysh_nbu@163.com

Chenchen Tao
Faculty of Electrical Engineering and Computer Science, Ningbo University, China
17681262923@163.com

ABSTRACT

Current approaches for spatiotemporal action recognition have achieved impressive progress, especially in temporal information processing. Meanwhile, the power of spatial information may be underestimated. Thus, a non-uniform key frame selector is proposed to pick the most representative frames according to the relationship between frames along the temporal dimension. Specifically, the reweight high-level frame features are used to generate an importance score sequence, while the key frames, in each temporal section, are selected based on the above scores. Such selected frames have richer semantic information, which has positive impact on the network training. The proposed model is evaluated on two action recognition, namely datasets HMDB51 and UCF101, and promising accuracy improvement is achieved.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Computer vision.

KEYWORDS

Action recognition, Key frame selection, Attention mechanism, Video summarization

ACM Reference Format:
Haohe Li, Chong Wang, Shenghao Yu, and Chenchen Tao. 2023. Action Recognition with Non-Uniform Key Frame Selector. In *2023 5th International Conference on Image Processing and Machine Vision (IPMV) (IPMV 2023), January 13–15, 2023, Macau, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3582177.3582182

1 INTRODUCTION

As one of the fundamental problems in computer vision, action recognition in videos has been extensively used in various domains, such as intelligent video surveillance, human-computer interaction, video sequence understanding, etc. [1-3]. Compared to the spatial information contained in still images, videos contain additional temporal information. Therefore, it is a challenging task that improving accuracy of action recognition through modeling videos with appropriate spatiotemporal representation.

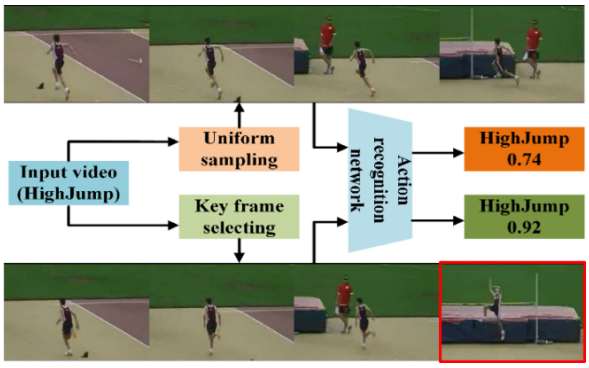

Figure 1: Comparison of uniform sampler and our key frame selector.

In the past decades, hand-crafted representation learning methods have received extensive research interest. Some work, such as Space-Time Interest Points (STIP) [4], 3D Histogram of Gradient [5] and Cuboids [6], extended the representations from isolated video frames to the temporal dimension with different feature encoding schemes. Improved Dense Trajectories (iDT) [7] further exploited the dense trajectory method [8] and achieved excellent results in video classification.

Since Convolutional Neural Networks (CNN) have achieved remarkable progress in image classification, researchers naturally extend this technology into video processing. Most current methods of action recognition are primarily based on two CNN frameworks. The first is two-stream convolutional networks [9, 10], which are devised to apply two separate recognition streams. One stream obtains spatial information from static video frames, while the other obtain motion information in the form of dense optical flow. On the other hand, 3D CNNs [11, 12] extract spatial and temporal information simultaneously from the video clip, which achieves wonderful action recognition results when trained on large-scale datasets. However, the capacity of existing 3D CNN architectures is

*Corresponding author

IPMV 2023, January 13–15, 2023, Macau, China

ACM ISBN 978-1-4503-9792-6/23/01…$15.00
https://doi.org/10.1145/3582177.3582182

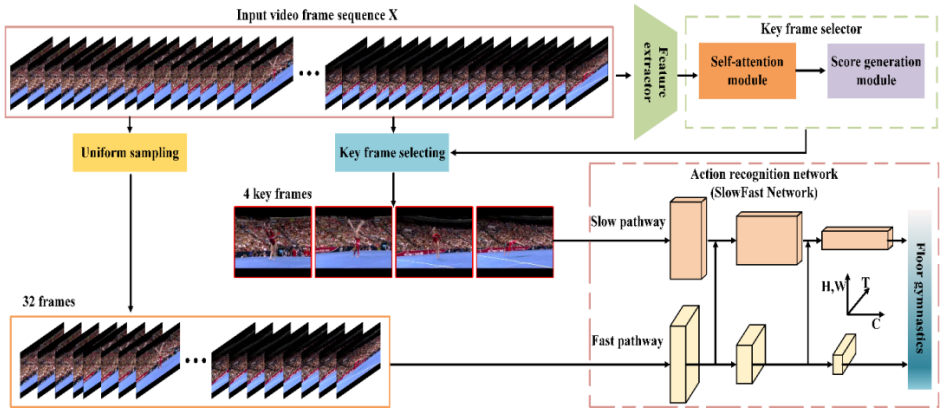

Figure 2: Overall framework of the proposed algorithm.

limited by the expensive computational cost and memory requirement. In addition, previous 3D CNNs treat temporal dimension as an extension of the spatial dimension. Although it is reasonable to handle the two spatial dimensions symmetrically in image domain [13], it is not appropriate to deal with the three dimensions of spatiotemporal information equally [14].

Combining the pros and cons of two types of networks, SlowFast Network [15] was proposed to employ different temporal rates to obtain branch inputs with different temporal lengths from a video clip. The first branch with several sparse frames as input is named as Slow pathway, which is designed to capture spatial semantic information. Relatively, the other branch with high temporal resolution input, namely Fast pathway, is utilized to capture motion information.

However, considering the nature of videos, limited sparse samples may not contain useful spatial semantic information. To address that, a non-uniform sampling scheme using key frames selector is introduced in this paper, which is inspired by video summarization [16-18]. For each isolated video frame, conventional 2D CNN networks are employed to extract its high-level features. Then, a self-attention network is used to reweight the frame features, followed by a score generation module that produces the importance scores. Those frames with higher scores are then picked as the key frames for action recognition. As shown in Figure 1, the proposed selector is capable to find more distinguished frames to achieve higher recognition accuracy.

2 METHODOLOGY

The overall framework of the proposed algorithm is shown in Figure 2. Firstly, GoogLeNet [19] is used to extract frame-level visual features from videos. Then a key frame selector, including the self-attention module and score generation module, is applied to calculate the importance score for each frame. The frames with the highest score in each temporal section are selected to provide more representative spatial information for the action recognition network, i.e. SlowFast Network [15] in this work.

2.1 Feature extractor

Given a frame sequence X:

$$X = \{x_i\}_{i=1}^{T}, \tag{1}$$

where $x_i \in R^{H\times W\times 3}$ is the i-th frame in the input frame sequence. H, W and 3 represent the height, width and the channel dimension of x_i , respectively. T is the frame number in X.

Considering the remarkable performance in the area of image classification, the GoogLeNet [19] that pre-trained on the ImageNet dataset, is utilized to extract high level features of each frame. Since only the features are what we need, its last fully connected layer is removed. Therefore, each frame x_i is fed into the pre-trained GoogLeNet, with a fixed size of 224×224. Then the output is a feature vector $f_i \in R^{1024}$.

$$f_i = g\left(x_i;\theta\right), \tag{2}$$

where $g(;\theta)$ is the feature extractor with the parameters θ. For the whole sequence, the corresponding feature vector can then be denoted as $F \in R^{T\times 1024}$.

2.2 Key frame selector

Inspired by the concept of video summary [16-18], a key frame selector is designed to pick the important frames in the video. As shown in Figure 2, the network is composed of a self-attention module and a score generation module, while the detailed diagram is presented in Figure 3.

2.2.1 Self-attention module. Normally, an attention function can be described as mapping a query vector and a set of key and value pairs to an output, where the output is a weighted sum of the value vectors [20]. Following the above description, the first step of our self-attention module is to produce the query and key-value pairs:

$$Q = m_q\left(F;\theta_q\right), \tag{3}$$

$$K = m_k\left(F;\theta_k\right), \tag{4}$$

$$V = m_v\left(F;\theta_v\right), \tag{5}$$

where m_q, m_k and m_v present the linear transformation with parameters θ_q, θ_k and θ_v. Then Q, K and V are the desired query, key and value vectors, which have the same dimension as the sequence feature vector F.

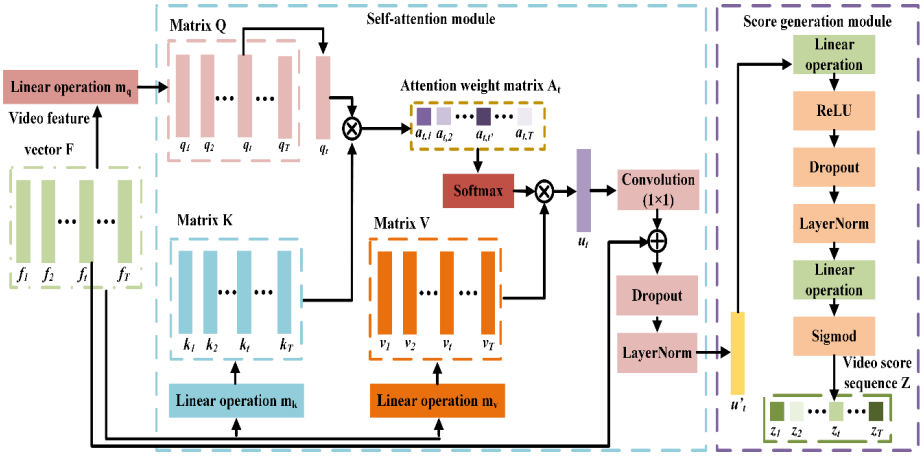

Figure 3: Diagram of the proposed key frame selector.

The dot-product attention is adopted here and the matrix of the attention weight is computed as:

$$A = \frac{QK^*}{\sqrt{d}}, \tag{6}$$

where A is the dot-product attention weight matrix, d is the scaling factor and K^* is the transpose of K. After that, the self-attention function can be defined as:

$$U = softmax\,(A)\,V \tag{7}$$

Furthermore, a 1×1 convolution and a residual operation are adopted to get the final output U':

$$U' = LayerNorm\,(DropOut\,(F + WU)), \tag{8}$$

where W is the weight of the 1×1 convolution operation.

2.2.2 Score generate module. The goal of the score generation module is to produce a frame-level score for each frame in the input sequence, which indicates its importance and contribution for later action recognition. A two-layer network is designed to perform frame score regression. The first layer has a ReLU activation followed by dropout and layer normalization [21], while the second layer has a single hidden unit with a sigmoid activation,

$$U'' = LayerNorm\left(DropOut\left(ReLU\left(m\left(U'\right)\right)\right)\right), \tag{9}$$

$$Z = Sigmod\left(m\left(U''\right)\right), \tag{10}$$

where $Z \in R^T$ is the frame-level score sequence and m denotes the linear mapping operation. A higher score means richer spatial semantic information in the certain frame, which may be more useful for the following action recognition network.

2.2.3 Key Frame Selection. Given the score sequence Z, N key frames can be selected from the input frame sequence X. To be specific, X and Z are divided into N subsequences.

$$X = \left\{\hat{X}_i\right\}_{i=1}^{N}, \tag{11}$$

$$Z = \left\{\hat{Z}_i\right\}_{i=1}^{N}, \tag{12}$$

And the i-th key frame is selected from the $\hat{X}_i$ with the highest score in $\hat{Z}_i$.

$$h_i = argmax\left(\hat{Z}_i\right), \tag{13}$$

where h_i is the index of the highest score in $\hat{Z}_i$. And the selected key frames $\tilde{X}$ are expressed as follows:

$$\tilde{X} = \left\{\hat{X}_{i,h_i}\right\}_{i=1}^{N} \tag{14}$$

2.3 Action Recognition Network

The SlowFast network [15] is selected as the action recognition network in this paper, which has a two-stream network architecture with different frame rates. Specifically, the stream with a larger frame rate is named as the slow pathway, while the other stream is the fast pathway. The ResNet-50 is used as the backbone network for each pathway.

Naturally, the non-uniform key frames $\tilde{X}$ selected in Section 2 are suitable for the slow pathway, while the uniformly sampled frames are fed into the fast path. The final loss function L_c for the action recognition module can then be defined in the form of cross-entropy loss:

$$L_c = -\frac{1}{C}\sum_{i=0}^{C-1} y_i \log\left(p_i\right), \tag{15}$$

where p_i is the recognition probability of the i-th category and C is the number of action categories. And $y_i = 1$ if the input video X belongs to the i-th action categories.

3 EXPERIMENTS

In this section, the evaluation datasets and experiment implement details are presented. Then the experiment results of our method are rendered, which are also compared with the state-of-the-art (SOTA) methods.

Dataset: two public datasets, namely UCF101 [22] and HMDB51 [23] are used to evaluate our method. UCF101 is a fundamental action dataset with 13,320 realistic videos distributed in 101 classes. Each video lasts 3-10 seconds and consists of 100-300 frames on average. HMDB51 contains 51 action categories, with total 6,766

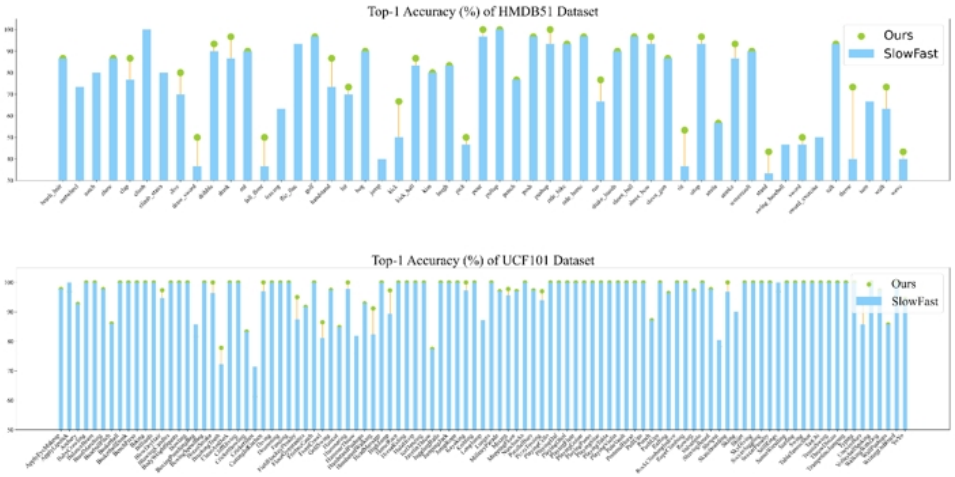

Figure 4: Top-1 accuracy of our method and SlowFast Network for each category on HMDB51 dataset (a) and UCF101 dataset (b).

Table 1: Top-1 Accuracy performance comparison on HMDB51

Method	Pretraining	Accuracy (%)
IDT [7]	-	61.7
Two-stream [9]	ImageNet	59.4
TSN [25]	ImageNet	69.4
I3D [26]	ImageNet+Kientics-400	74.5
ResNext101 [27]	Kinetics-400	70.1
R(2+1)D [28]	Kinetics-400	74.5
S3D-G [29]	ImageNet+Kinetics-400	75.9
LGD-3D [30]	ImageNet+Kinetics-600	75.7
SlowFast [15]	Kinetics-400	75.2
Ours	Kinetics-400	76.4

video clips. Both datasets have a three split evaluation protocol provided by the organizers.

Training: The weights pre-trained on the SumMe dataset [24] is used for the key frame selector. For the action recognition network, the clips of 64 frames from the full-length video are randomly sampled and divided into 32 uniformly frames, while 4 key frames are generated by the proposed seletor. As the spatial domain, the shorter side of input clip is randomly sampled in 256×320 pixels while the other side changed according to the corresponding proportion. After that, the input clip, or its horizontal flip, is randomly cropped to 224×224 pixels. Adam is chosen as the optimizer and the initial learning rate is set to 0.001.

Inference: Following the protocol in [15], 10 clips from each video are uniformly sampled and each clip has 3 spatial crops. For each clip, a clip-level prediction score is obtained through the network. The video-level score is computed by averaging all the clip-level scores of a video.

The average Top-1 accuracy over three splits on both HMDB51 and UCF-101 are reported in Table 1 and Table 2, which are compared with other SOTA methods. As shown in Table 1, the accuracy of original SlowFast Network [15] on HMDB51 is 75.2%, higher than TSN [25] and I3D [26] by 5.8% and 0.7%, respectively. It is worth noting that our key frames selector can further improve the accuracy by 1.2%, which is benefited from valid semantic information. As for UCF101, the propose model obtains the 96.3% top-1 accuracy, which is constantly higher than the vanilla SlowFast Network [16] and comparable to the best SOTA method LGD-3D [29].

The accuracy of each category of both datasets is presented in Figure 4. It can be seen, our method has achieved significant boosts at categories “clap, dive, handstand, kick, sit, stand, throwing, etc.” on HMDB51 and “brushing teeth, handstandwalking, highjump, etc.” on UCF101.

4 CONCLUSION

In this paper, a simple yet effective key frame selector has been proposed for the action recognition task. To be specific, the frame features are reweighted by the self-attention module and a score sequence is generated. The key frames are selected with the highest scores in the score sequence. Experimental results have shown the effectiveness of the model. Although our study has presented an effective method to pick the key frames in video, future work will be devoted to integrate more spatiotemporal information into the process of key frames selection.

Table 2: Top-1 Accuracy performance comparison on UCF101

Method	Pretraining	Accuracy (%)
IDT [7]	-	86.4
Two-stream [9]	ImageNet	88.0
TSN [25]	ImageNet	94.2
I3D [26]	ImageNet+Kientics-400	95.4
ResNext101 [27]	Kinetics-400	94.5
R(2+1)D [28]	Kinetics-400	96.8
S3D-G [29]	ImageNet+Kinetics-400	96.8
LGD-3D [30]	ImageNet+Kinetics-600	97.0
SlowFast [15]	Kinetics-400	96.2
Ours	Kinetics-400	96.3

ACKNOWLEDGMENTS

This work was supported by Zhejiang Provincial Natural Science Foundation of China (LY20F030005) and National Natural Science Foundation of China (61603202).

REFERENCES

[1] S. Herath, M. Harandi and F. Porikli, "Going Deeper into Action Recognition: A Survey," in Image and Vision Computing, pp. 4-21, 2017.
[2] X. Wang, "Intelligent Multi-Camera Video Surveillance: A Review," in Pattern Recognition Letters, pp. 3-19, 2013.
[3] C. Camporesi, M, Kallmann and J. J. Han, "VR Solutions for Improving Physical Therapy," in IEEE Virtual Reality, pp. 77-78, 2013.
[4] I. Laptev, "On Space-Time Interest Points," in ICCV, pp. 107-123, 2003.
[5] A. Klaser, M. Marszałek, and C. Schmid, "A Spatio-Temporal Descriptor based on 3d-Gradients," in BMVC, pp. 1-10, 2008.
[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Features", in ICCV, pp. 65-72, 2005.
[7] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in ICCV, pp. 3551-3558, 2013.
[8] H. Wang, A. Kl¨aser, C. Schmid, and C.-L. Liu, "Dense Trajectories and Motion Boundary Descriptors For Action Recognition," in ICCV, pp. 60-79, 2013.
[9] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition In Videos," in NIPS, 2014.
[10] C. Feichtenhofer, A. Pinz and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," in CVPR, pp. 1933-1941, 2016.
[11] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," in IEEE Trans. PAMI, pp. 221-231, 2013.
[12] D. Tran, L. Bourdev, R. Fergus, "Learning Spatiotemporal Features with 3D Convolutional Networks," in ICCV, pp. 4489-4497, 2015.
[13] J. Huang and D. Mumford, "Statistics of Natural Images and Models," in CVPR, pp. 541-547, 1999.
[14] Y. Weiss, E. P. Simoncelli, and E. H. Adelson, "Motion Illusions as Optimal Percepts," in Nature Neuroscience, pp. 598–604, 2002.
[15] Feichtenhofer C, Fan H, Malik J, *et al.*, "SlowFast Networks for Video Recognition," in ICCV, pp. 6202-6211, 2019.
[16] Fajtl J, Sokeh H S, Argyriou V, *et al.*, "Summarizing videos with attention," in ACCV, pp. 39-54, 2018.
[17] Li P, Ye Q, Zhang L, *et al.*, "Exploring Global Diverse Attention Via Pairwise Temporal Relation for Video Summarization," in Pattern Recognition, 2021.
[18] Otani M, Nakashima Y, Rahtu E, *et al.*, "Rethinking the evaluation of video summaries," in CVPR, pp. 7596-7604, 2019.
[19] Szegedy C, Liu W, Jia Y, *et al.*, "Going Deeper with Convolutions," in CVPR, pp. 1-9, 2015.
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, *et al.*, "Attention Is All You Need," in NIPS, pp. 5998-6008, 2017.
[21] Ba J L, Kiros J R, Hinton G E, "Layer Normalization," in arXiv:1607.06450, 2016.
[22] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Action Classes from Videos in The Wild," in arXiv:1212.0402, 2012.
[23] H. Kuehne, H. Jhuang, and E. Garrote, *et al.*, "HMDB: A Large Video Database for Human Motion Recognition," in ICCV, pp. 2556-2563, 2011.
[24] Gygli M, Grabner H, Riemenschneider H, *et al.*, "Creating Summaries from User Videos," in ECCV, pp. 505-520, 2014.
[25] Wang L, Xiong Y, Wang Z, *et al.*, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," in ECCV, pp. 20-36, 2016.
[26] Carreira J, Zisserman A, "Quo Vadis, Action Recognition? A New Model and The Kinetics Dataset," in CVPR, pp. 6299-6308, 2017.
[27] Hara K, Kataoka H, Satoh Y, "Can Spatiotemporal 3D CNNS Retrace The History of 2D CNNS and ImageNet?" in CVPR, pp. 6546-6555, 2018.
[28] Tran D, Wang H, Torresani L, *et al.*, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," in CVPR, pp. 6450-6459, 2018.
[29] Xie S, Sun C, Huang J, *et al.*, "Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-Offs in Video Classification," in ECCV, pp. 318-335, 2018.
[30] Qiu Z, Yao T, Ngo C W, *et al.*, "Learning Spatio-Temporal Representation with Local and Global Diffusion," in CVPR, pp. 12056-12065, 2019.