25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Workshop on Computational Aspects of Deep Learning
Milan,Italy 10 | 15 January 2021

Compressed Video Action Recognition using Motion Vector Representation

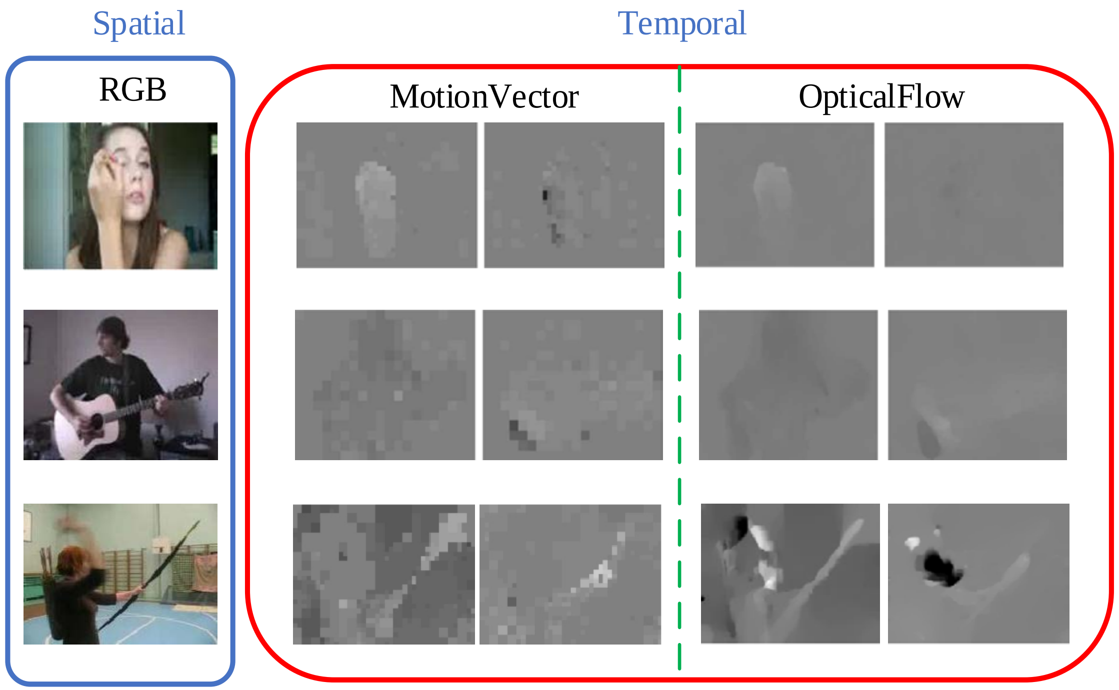
ChengHui Zhou[1], Xiaolei Chen[2], Pei Sun[2] , Guanwen Zhang[1*], Wei Zhou[1]
1 School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
2 CNPC logging Co.,Ltd, China

Reporter: Zhou ChengHui

January 10， 2021

CONTENTS

Introduction



We propose to train deep networks directly on compressed video representation, which is able to achieve competitive recognition performance.

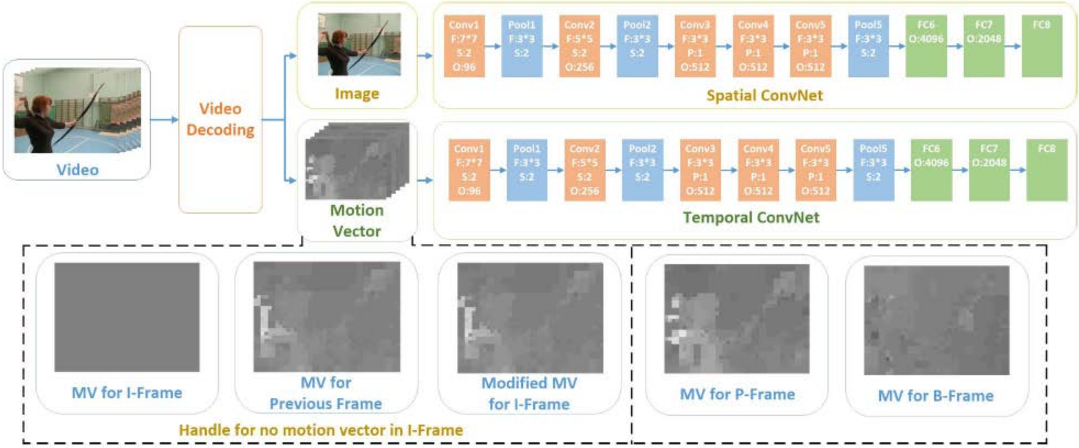
The reasons to choice motion vectors instead of optical flow:

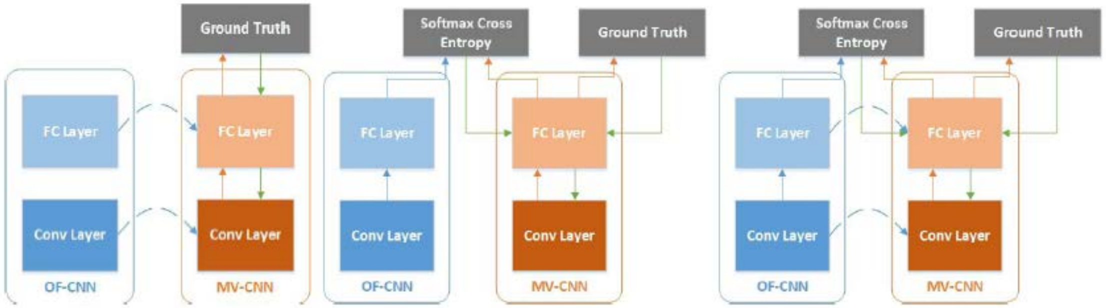
a) **The calculation of OF is too expensive to realize real-time application;**
b) **MV can be directly extracted from the compressed video;**
c) **Although motion vector contains some noise, it still represents motion information similar to optical flow**

Dataset	Spatial-Resolution	TV-L1 Flow(GPU)(FPS) (RTX 1080 Ti)	High Performance MV(CPU)(FPS)
UCF101	320*240	28.2	676.7
HMDB51	320*240	28.2	676.7

Introduction -- Related work

Real-Time Action Recognition With Deeply Transferred Motion Vector CNNs --- Transactions on Image Processing2018



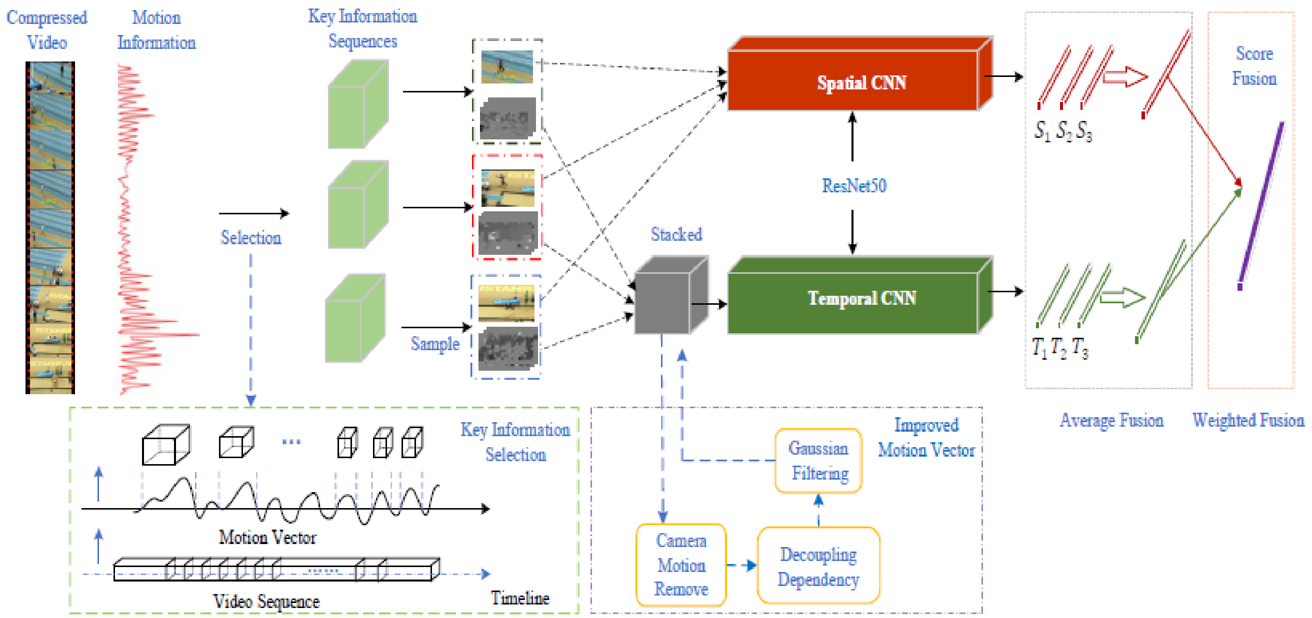


Zhang propose four training strategies which leverage the knowledge learned from OF CNN to enhance the accuracy of MV CNN.

Problems:

a) It still densely samples the video frames in a short time range;
b) It has risks of missing key information in video;
c) It still requires optical flow as an additional supervision.

Method -- Overview

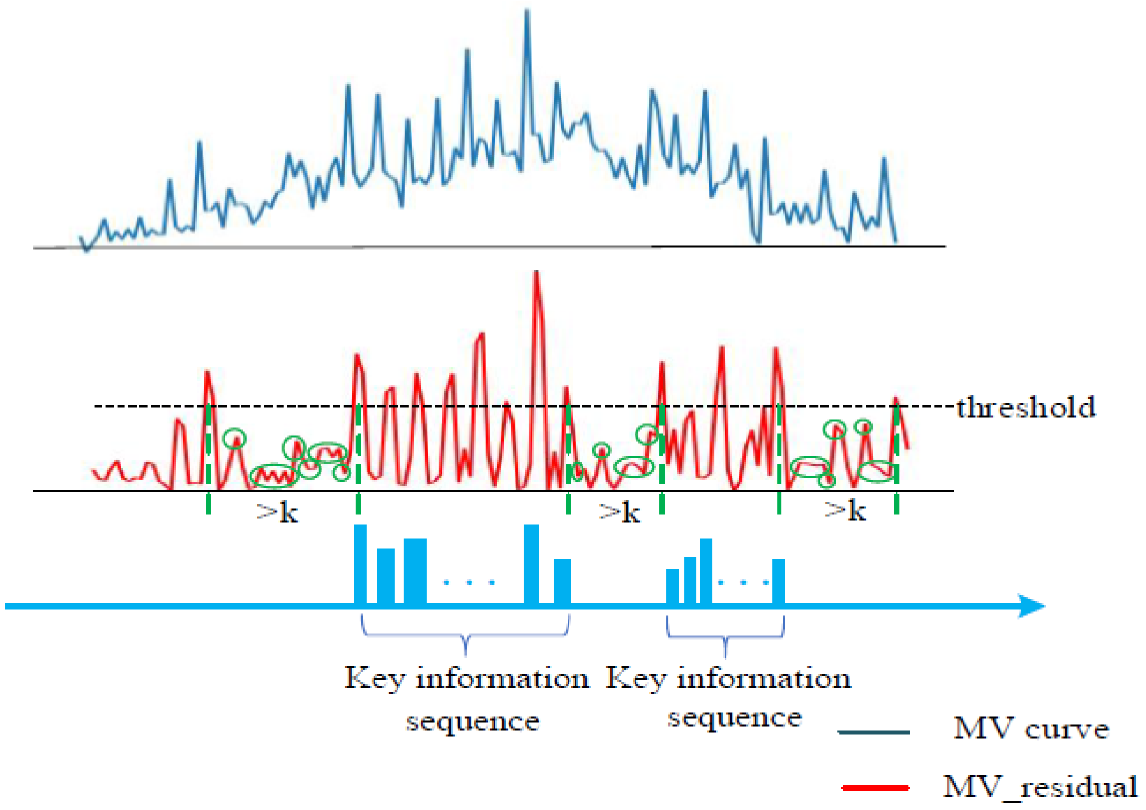


In this paper, we propose a novel approach for compressed video action recognition using motion vector representation.

Our model consists of three parts:

a) Key Information Sequence Selection;
b) Improved Motion vector;
c) Feature Fusion;

Method -- Key Information Sequence Selection (KIS)



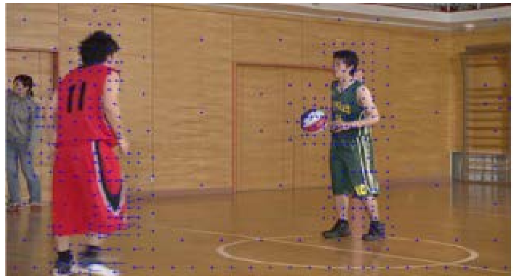
We exploit motion vector as an objective criteria to detect key information sequences in video. The core part of the key information selection algorithm is locating active parts of motion vector curve.

$$MV_{frame} = \sum_{t=1}^{T} MV_{block}(t) \qquad (1)$$

$$MV_i = \frac{MV_{frame(i)}}{\max(MV_{frame})} \qquad (2)$$

$$MV_{residual} = |MV_i - MV_{i+1}| \qquad (3)$$

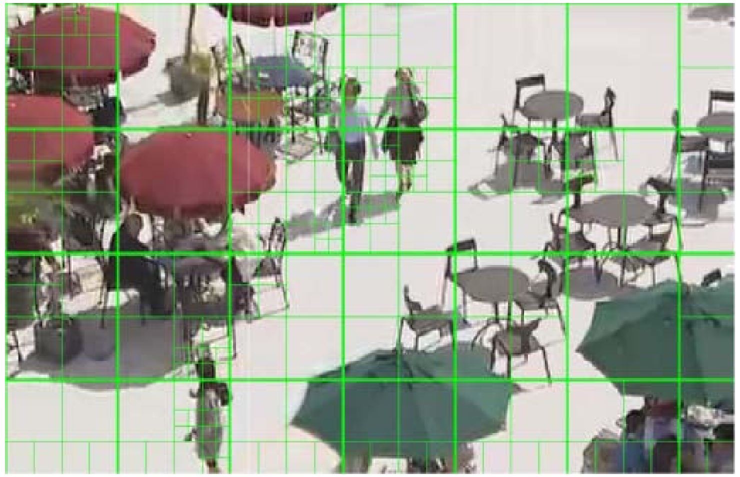
Method -- Improved Motion vector



Camera Stop

Camera Move

Motion vectors are mainly composed of two factors: the object motion and camera motion. Moving objects usually attract more visual attention than background.

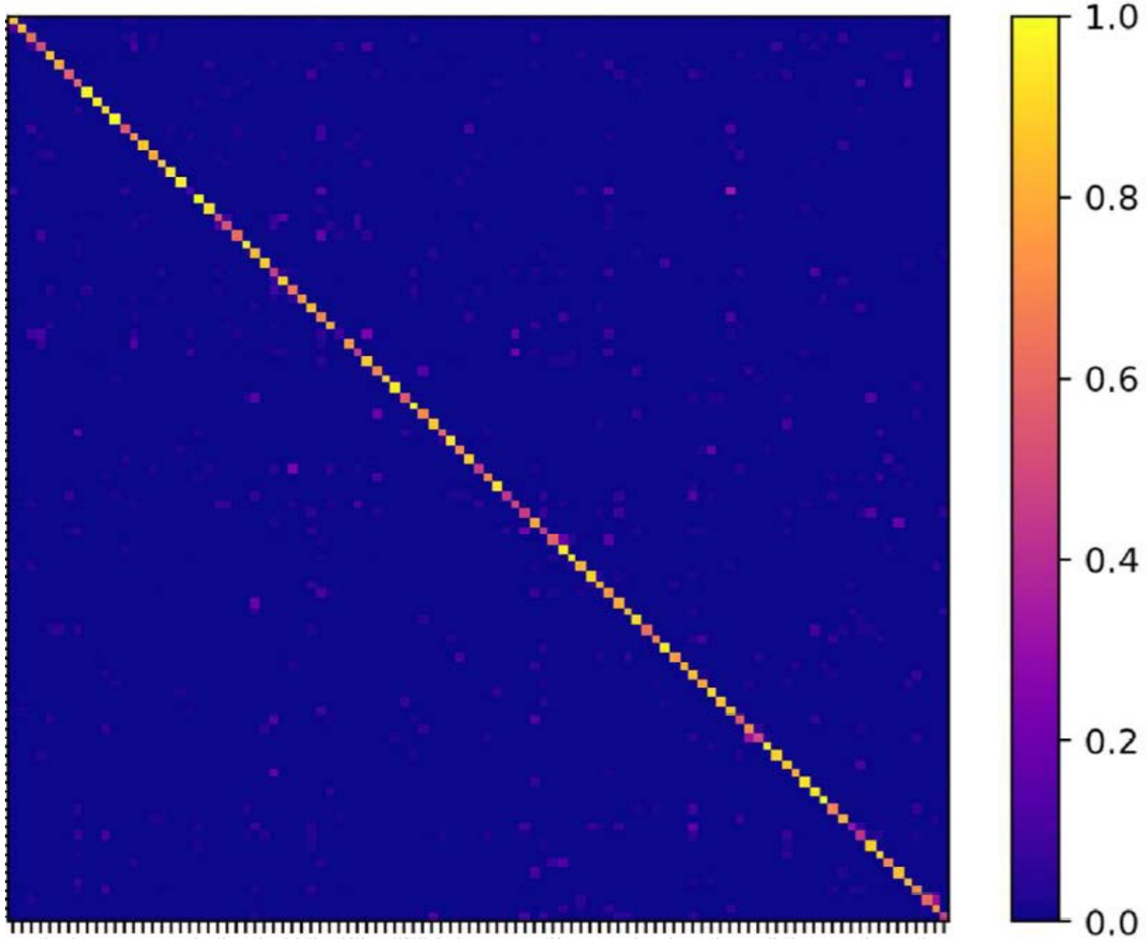


CU Size	Split depth
64 * 64	0
32 * 32	1
16 * 16	2
8 * 8	3

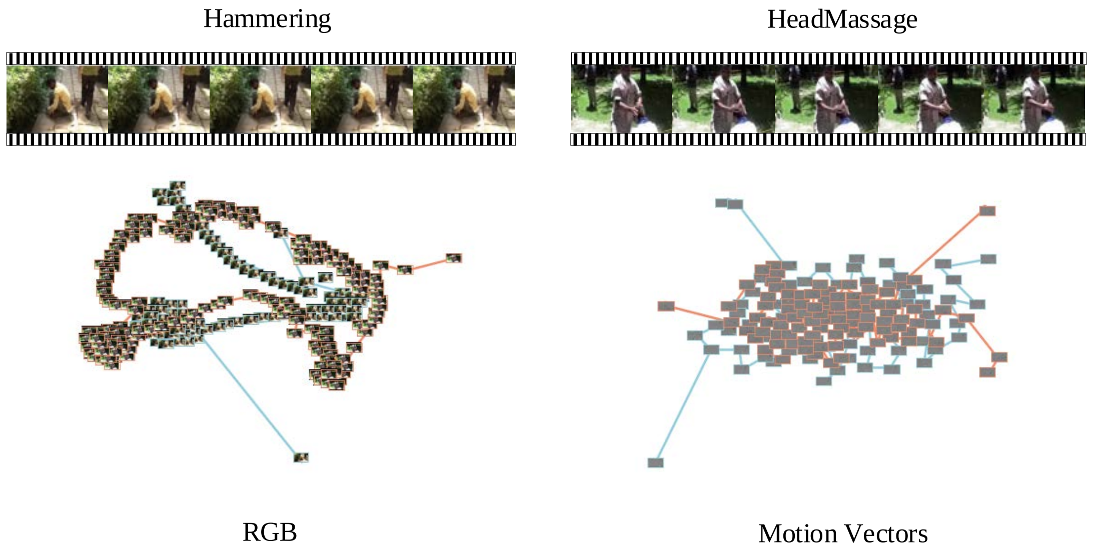
$$R_B^K = \left\{ (i',j') \,\middle|\, |D_{i'j'}^k < \frac{1}{|P^k|} \sum_{(i,j)\in P^k} D_{ij}^k \right\}$$

$$\max hist\left(\bigcup_{i,j\in R_B^K} A(M_{i,j}^k) \right)$$

Results



We show the confusion matrix for our algorithm on UCF101. It can be shown that algorithm performs well in most videos for Human action category like Billiards and CleanAndJerk. However, algorithm performs worse in class HeadMassage and Hammering. For HeadMassage, algorithm always misclassies into Hammering.



Results

RESULTS ON UCF101

Releted Algorithm	Accuracy	FPS
MDI + RGB	76.9%	<131
C3D(1 net)(GPU)	82.3%	313.9
DTMV + RGB-CNN	86.4%	390
Two-stream CNNs(GPU)	88.0%	14.3
Two-stream I3D (RGB + Flow)	93.4%	<14
TSN (RGB + Optical Flow) (GPU)	94.0%	14
TSN (RGB + RGBDiff) (GPU)	91.0%	340
Ours	92.1%	461.5

RESULTS ON HMDB-51

Releted Algorithm	Accuracy	FPS
MDI + RGB[59]	42.8%	<131
C3D(1 net)(GPU)[60]	50.3%	313.9
DTMV + RGB-CNN[62]	55.3%	390
Two-stream CNNs(GPU)[25]	46.4%	14.3
Two-stream I3D (RGB + Flow)[61]	66.4%	<14
TSN (RGB + Optical Flow) (GPU) [27]	69.4%	14
TSN (RGB + RGBDiff) (GPU) [27]	65.7%	340
Ours	60.3%	461.5

References

① He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Las Vegas: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.

② Zhang B, Wang L, Wang Z, et al. Real-time action recognition with deeply transferred motion vector cnns[J]. IEEE Transactions on Image Processing, 2018, 27(5): 2326-2339.

③ Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.

④ Wang L, Xiong Y, Wang Z, et al. Temporal segment networks for action recognition in videos[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(11): 2740-2755.

⑤ Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J].Advances in neural information processing systems, 2014: 568-576.

⑥ Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]. Santiago: IEEE International Conference on Computer Vision (ICCV)，2015: 4489-4497.

⑦ Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]. Honolulu: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 6299-6308.

⑧ Wang H, Schmid C. Action recognition with improved trajectories[C]. Sydney: IEEE International Conference on Computer Vision (ICCV), 2013: 3551-3558.

THANKS FOR YOUR ATTENTION