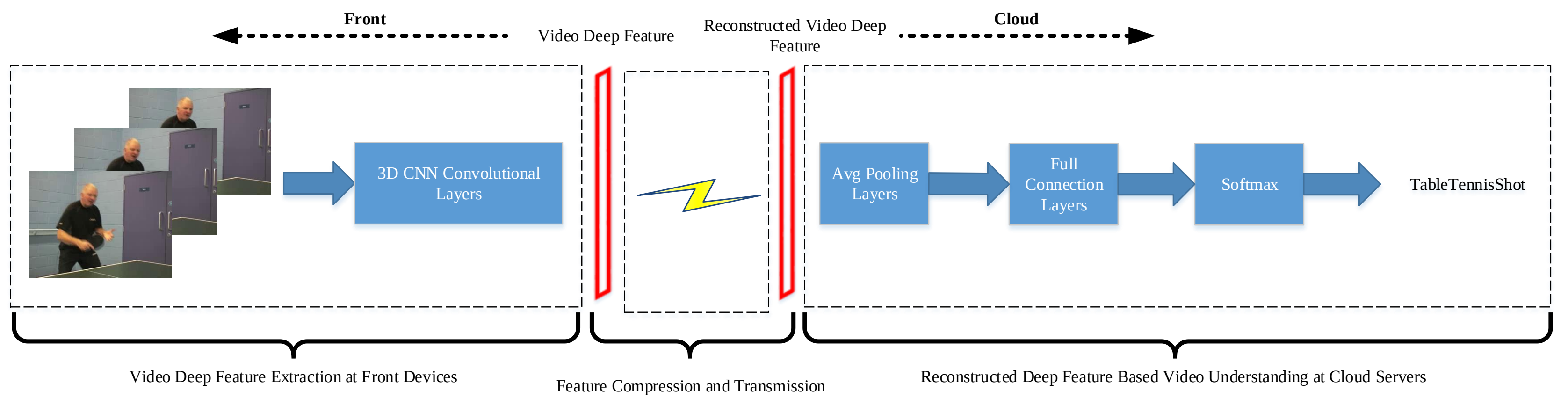
Binary Representation and Compression of 3D CNN Features for Action Recognition

Peiyin Xing, Peixi Peng, Yongsheng Liang, Tiejun Huang, Yonghong Tian
{pyxing,pxpeng,tjhuang,yhtian}@pku.edu.cn, liangys@hit.edu.cn

1.Motivation

A common framework of the action recognition is to collect the videos from different cameras into a cloud center firstly, and then perform the 3D CNN on the cloud server. Although directly, this framework will bring a huge burden to the cloud server and video transmission. To handle this challenge, the "front-cloud" collaborative processing architecture, as shown in the following figure, can be used.


Front
Video Deep Feature
Reconstructed Video Deep Feature
Cloud
3D CNN Convolutional Layers
Avg Pooling Layers
Full Connection Layers
Softmax
TableTennisShot
Video Deep Feature Extraction at Front Devices
Feature Compression and Transmission
Reconstructed Deep Feature Based Video Understanding at Cloud Servers

To handle this challenge, there are two main issues need to be addressed:

a) The first issue is which type of feature should be extracted from the 3D CNN by the front-end devices.
b) The second issue is how to reduce the transmitted data amount in the front-end devices effectively without decreasing the recognition accuracy in the cloud server. To reduce the data size of transmitted features, quantization for limited bit representation and compression for data redundancy removal are often used.

2. The Main Contribution

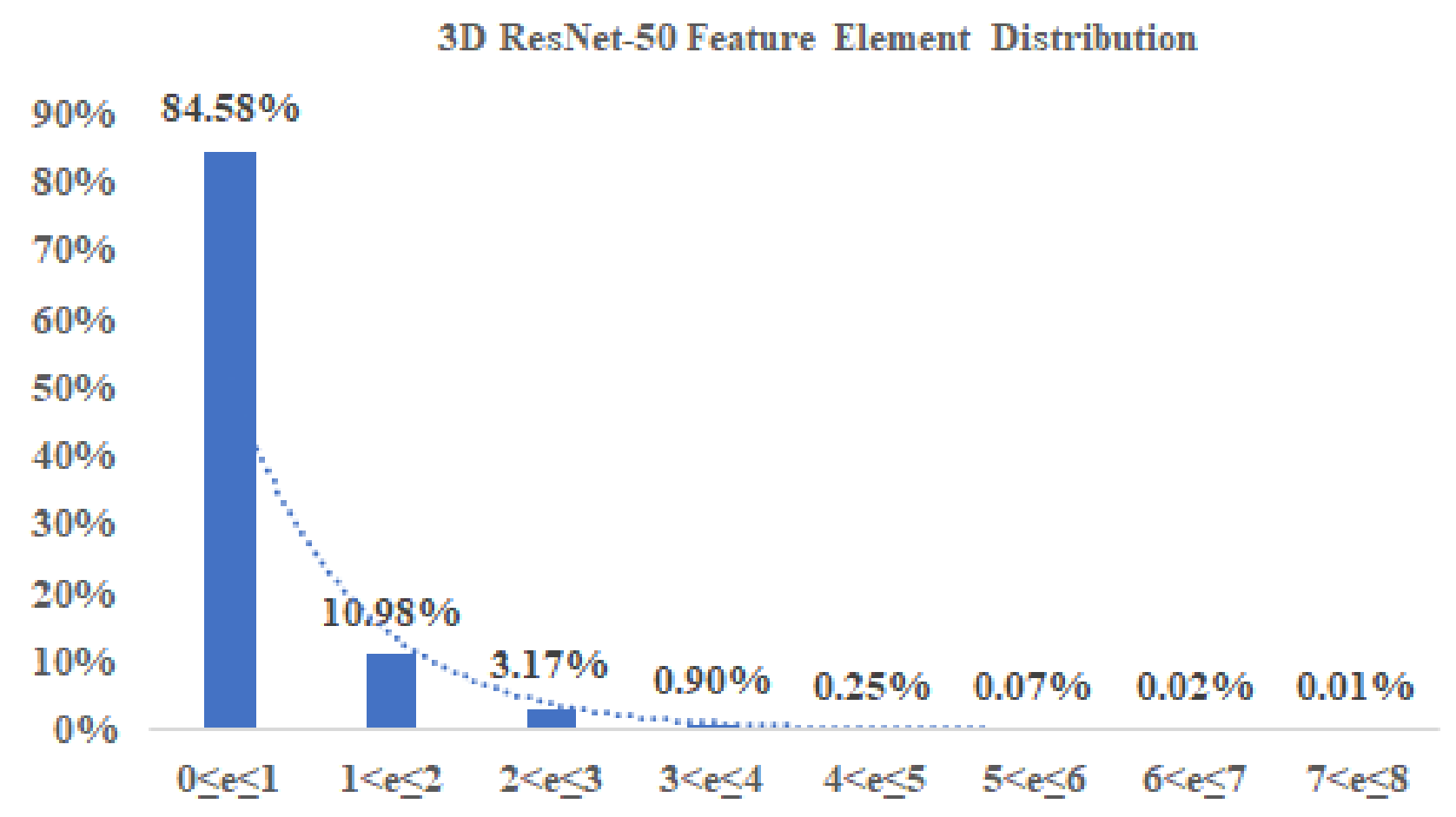
1) We propose a novel "front-cloud" collaborative processing framework is proposed for 3D CNN based action recognition. The data amount need to be transmitted is only 3% of the original encoded videos.
2) A logarithmic quantization with a fixed threshold is proposed to represent the features with limited bits. The features can be represented with only 1 bit with slight accuracy drop.
3) Inter-prediction coding is used to remove the temporal redundancy in the 3D CNN features by reshaping the feature as a video clip.

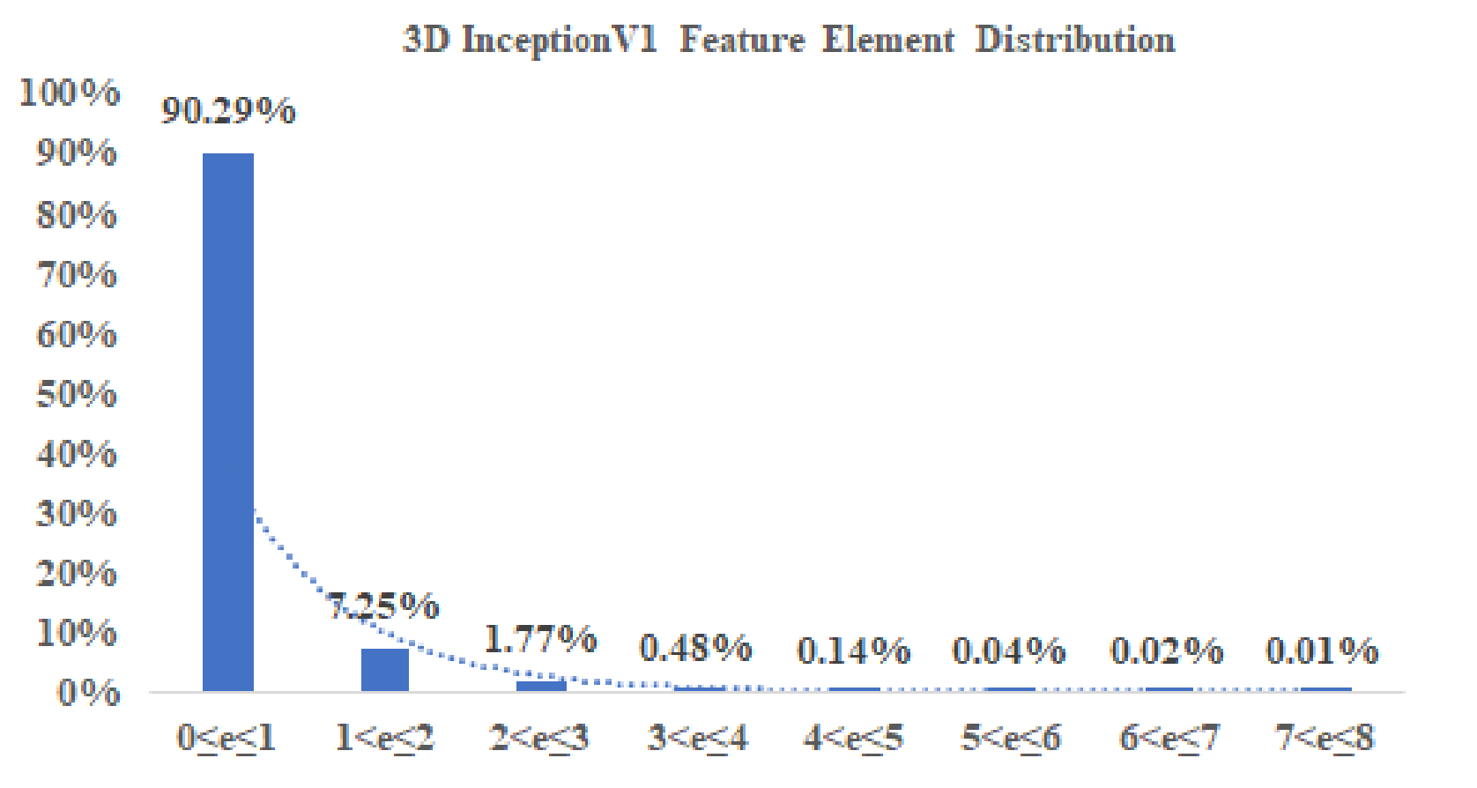
3. 3D CNN Feature Analysis

We choose the output of 3D CNN convolution layer as the feature of action recognition, and divide the network into two parts: feature extraction network and feature analysis network. The convolution part of 3D CNN is placed on the front-end devices, and the features (the output of the convolution part) is sent to the cloud to compute the rest network layers.

1) About half of feature elements are zero, and more than 80% elements are between 0 and 1, showing as an exponential distribution.
2) The 3D CNN features are produced by convolution of a video clip rather than irrelevant images. As similar to the video, there are high temporal redundancy among the 3D CNN feature elements.

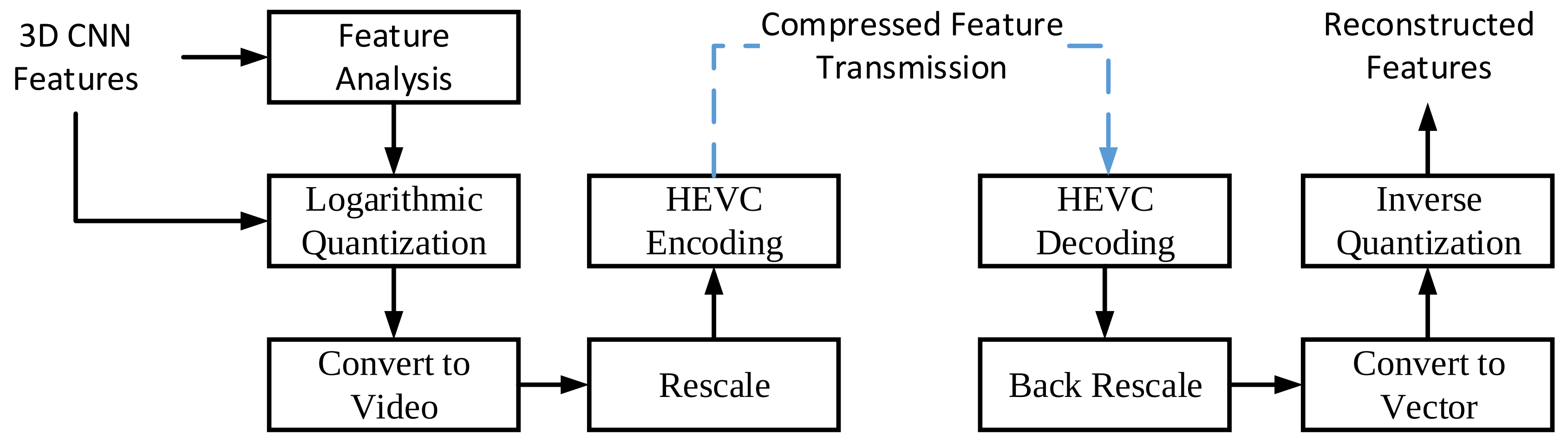
Value	Zeros	None-Zeros
3D ResNet-50	48.60%	51.40%
3D InceptionV1	48.83%	51.17%





4. The Proposed Compression Framework

a) The 3D CNN features will be quantized by the logarithmic quantization;
b) The maximum value determined by feature analysis is implicit in encoder and decoder;
c) Convert the features to a video clip and rescale the element values for HEVC inter coding.



Quantization

$$Q_k = round\left(\frac{log_2(V+1)}{log_2(\max(V)+1)} \times (2^k - 1)\right)$$

5. Experiments

QP	3D ResNet-50				3D InceptionV1			
	diff top1(%)	diff top5(%)	Fea. Com. Ratio	Comp. Fea./Video	diff top1(%)	diff top5(%)	Fea. Com. Ratio	Comp. Fea./Video
10	0.05	0.16	50.53	3.38	0.32	0.13	95.87	0.87
20	0.05	0.16	61.89	2.76	0.32	0.13	111.72	0.75
30	0.05	0.16	91.60	1.84	0.32	0.13	153.53	0.55
40	0.16	0.16	187.76	0.85	0.24	0.08	318.04	0.25
42	0.16	0.11	241.61	0.65	0.08	-0.02	390.59	0.20
44	0.02	0.16	296.82	0.53	0.16	0.00	509.15	0.16
46	-0.08	0.06	751.60	0.21	-0.19	-0.45	1220.34	0.06
48	-0.32	-0.34	2361.55	0.07	**-0.79**	**-1.24**	**3510.56**	**0.02**
50	**-0.98**	**-0.63**	**5029.12**	**0.03**	-1.67	-2.48	6629.42	0.01

The accuracy loss and compression performance are shown in the table. Can be seen in the table, the feature compression ratio are 5029.12 (3D ResNet-50) and 3510.56 (3D In-ceptionV1) with top1 accuracy loss 0.98% (3D ResNet-50) and 0.79% (3D InceptionV1). The ratio of compressed features to original videos are 0.03 (3D ResNet-50) and 0.02 (3D InceptionV1). The feature compression ratio of 3D InceptionV1 is only half of 3D ResNet-50, because the feature dimension of 3D InceptionV1 is half of the 3D ResNet-50. The dimension of 3D InceptionV1 is (30, 1024, 4, 8, 8) and the dimension of 3D ResNet-50 is (30, 2048, 4, 8, 8) in our experiment. Compared with transmitting the original videos, aggregating compressed features seems to be appropriate solution for video big data analysis.