Towards Practical Compressed Video Action Recognition: A Temporal Enhanced Multi-Stream Network

Bing Li[1], Longteng Kong[1], Dongming Zhang[2], Xiuguo Bao[2], Di Huang[1],* and Yunhong Wang[1]
[1]IRIP Lab, School of Computer Science and Engineering, Beihang University, Beijing 100191, China.
[2]National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China.

Motivation

Recently, many studies attempt to recognize actions in compressed videos rather than regular ones, aiming to avoid the resource overhead of decoding. In the inference stage, they generally assume that all the observations of samples are available. However, in practical transmission, the compressed video packets are usually disorderly received and lost due to network jitters or congestions, as Fig. 1 shows. These facts limit the availability of existing methods.

In this work, we concentrate on practical compressed video action recognition, and consider to complement the missing video packets with partial received ones.

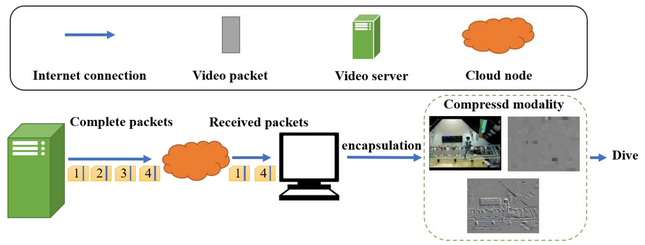

Fig. 1. Compressed video action recognition in practical scenarios. Number 2, 3 packets are lost during video transmission.

Contribution

- **We proposed a Temporal Enhanced Multi-Stream Network (TEMSN) for practical compressed video action recognition, as shown in Fig. 2.**
- **We use three modalities in compressed domain as complementary cues to capture richer information from compressed video packets.**
- **We design a temporal enhanced module based on Encoder-Decoder structure to generate more complete action dynamics.**
- **The proposed approach is evaluated on the HMDB-51 and UCF-101 datasets and state-of-the-art results are reached.**

Methodology

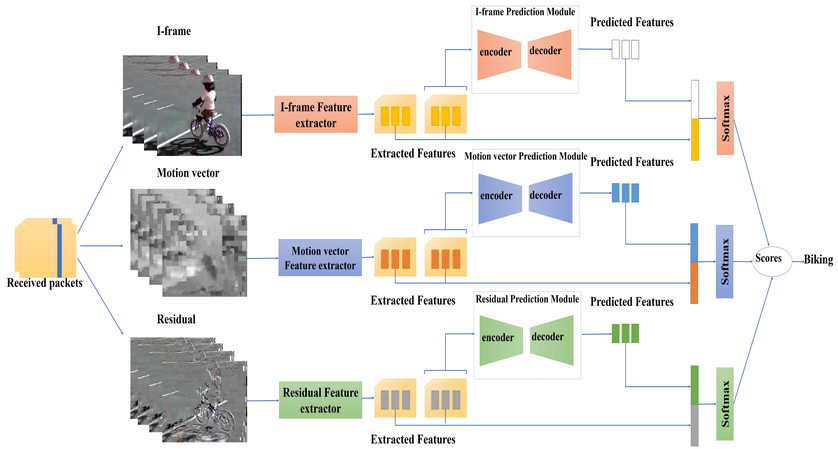

Fig. 2. Framework of the proposed Temporal Enhanced Multi-Stream Network (TEMSN).

Given limited compressed video packets, TEMSN takes three phases to make action recognition.

- **Multi-modal Encapsulation: We exploit the relation between the I-frame and P-frame to decouple the input, as Eq. 1, resulting residual, motion vector, and intra-frame.**

$$I_i^{(t)} = I_{i-D_i^{(t)}}^{(0)} + R_i^{(t)} \qquad (1)$$

- **Feature Extraction: We then transform the modalities into the feature spaces by the Multi-Stream Network which consists of three independent CNNs.**

- **Temporal Enhancement: The temporal enhanced module takes the packet features as input, and predicts the contiguous packets as Eq. 2. The original and synthesized features are concatenated to form global representation for final recognition.**

$$\begin{aligned} f_{pre} &= \varphi(s_t) \\ f_{RGB}^{(p_t)} &= \emptyset_{RGB}(f_{RGB}^{(p_{t-1})}) \\ f_{M\_V}^{(p_t)} &= \emptyset_{M\_V}(f_{M\_V}^{(p_{t-1})}) \\ f_{Res}^{(p_t)} &= \emptyset_{Res}(f_{Res}^{(p_{t-1})}) \end{aligned} \qquad (2)$$

Results

TABLE I
ACCURACY AVERAGED OVER THREE SPLITS ON HMDB-51 AND UCF-101 FOR STATE-OF-THE-ART COMPRESSED VIDEO-BASED METHODS.

Methods	HMDB-51	UCF-101	Ratio
EMV-CNN [10]	51.2	86.4	100%
DTMV-CNN [11]	55.3	87.5	100%
TTP [16]	58.2	87.2	100%
CoViAR [28]	59.1	90.4	100%
CoViAR‡	59.4	90.7	100%
CoViAR‡	57.3	88.5	50%
TEMSN (ours)	**61.1**	**91.8**	**100%**
TEMSN (ours)	**59.1**	**90.3**	**50%**

TABLE III
RESULTS OF THREE VIDEO PACKET PREDICTION SCHEMES ON HMDB-51.

Baselines	Random	Uniform	Normal
w/o Pre.	57.3	57.3	57.3
+Pre. (20%)	57.8	58.2	58.0
+Pre. (50%)	58.7	59.1	58.9
+Pre. (100%)	61.1	61.1	61.1

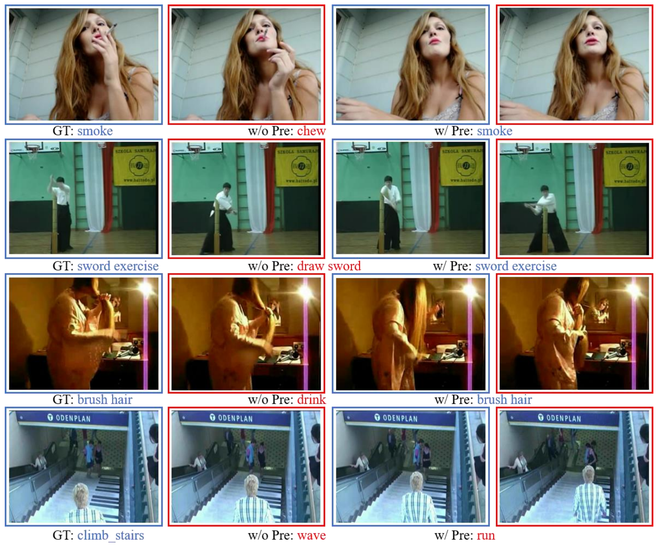

Fig. 3. Qualitative results of the proposed TEMSN on HMDB51, where each row belongs to an action. Frames in the blue box indicate the received ones, and in the red box indicate the lost ones.

Email: dhuang@buaa.edu.cn
libingsy@buaa.edu.cn
Postbox #406, No.37 Xueyuan Road,
Haidian District, Beijing, China, 100191

IRIP laboratory