HARVARD MEDICAL SCHOOL

MASSACHUSETTS GENERAL HOSPITAL

Gordon Center for Medical Imaging

Identity-aware Facial Expression Recognition in Compressed Video

Xiaofeng Liu, Linghao Jin, Xu Han, Jun Lu, Jane You, Lingsheng Kong

https://liu-xiaofeng.github.io/ liuxiaofengcmu@gmail.com

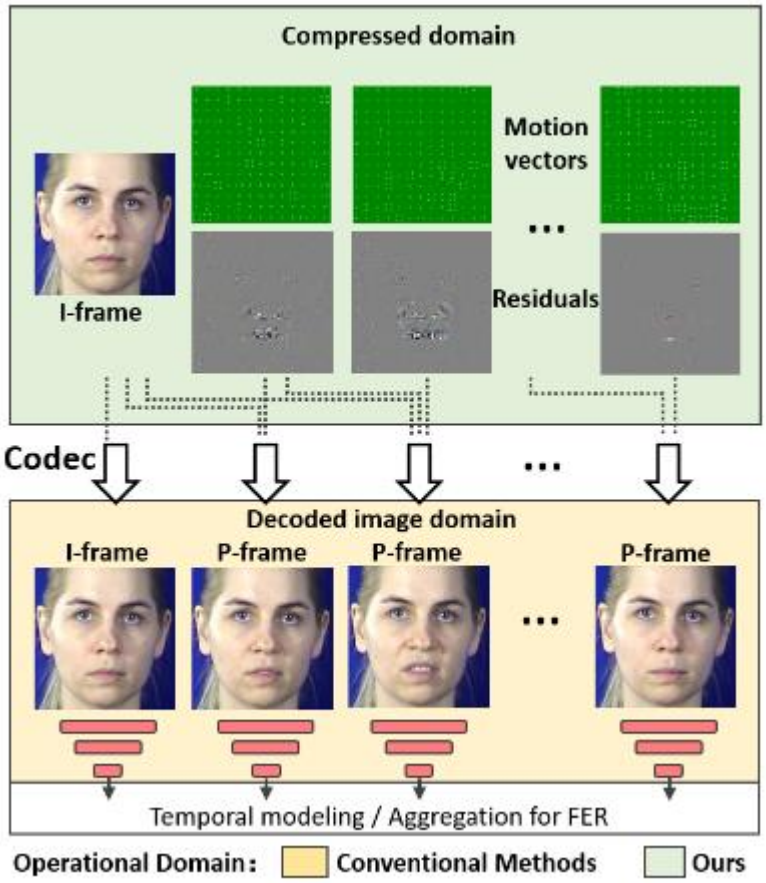
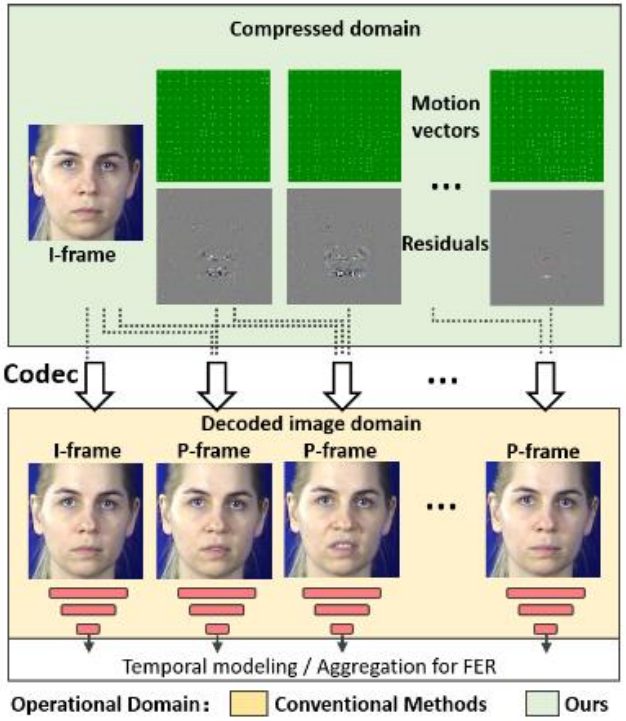


Illustration of the typical video compression and the scheme of conventional FER methods, which first decode the video and then feed it into a FER network.

HARVARD MEDICAL SCHOOL | MASSACHUSETTS GENERAL HOSPITAL | Gordon Center for Medical Imaging

Usually transfer compressed video——should we decode it for FER?

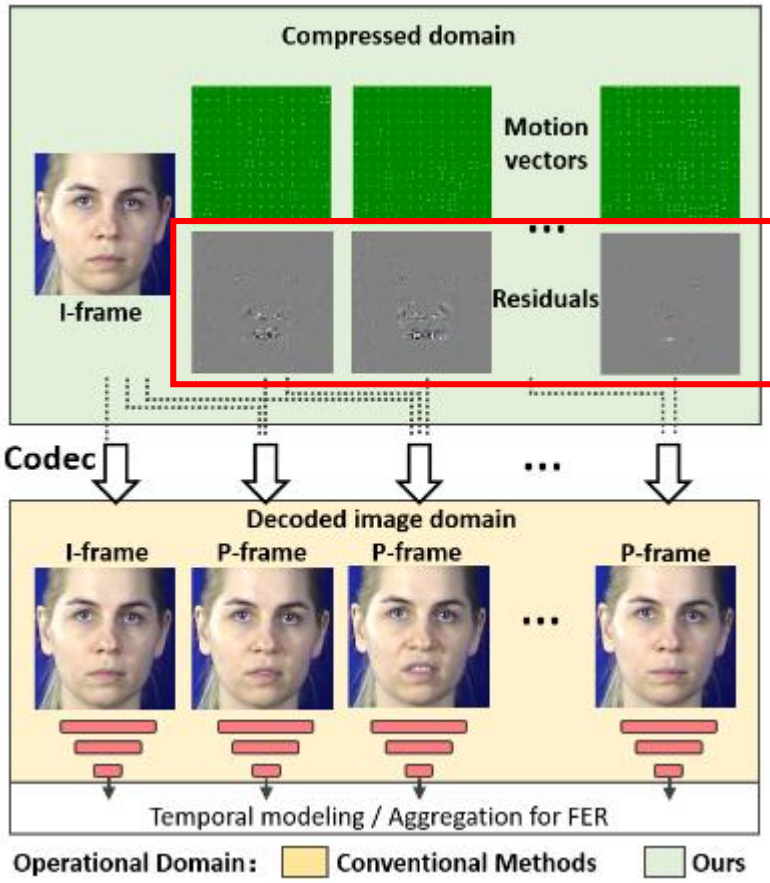
The residuals (P-frames) have expression info!



To illustrate the format of input video, we choose the MPEG-4 as an example [43]. The compressed domain has two typical frames, i.e., I frame and P frames. Specifically, the I frame $I \in \mathbb{R}^{h\times w\times 3}$ is a complete RGB image. We use h and w to denote its height and width respectively. Besides, the P frame at time t $P^t \in \mathbb{R}^{h\times w\times 3}$ can be reconstructed with the stored offsets, called residual errors $\Delta^t \in \mathbb{R}^{h\times w\times 3}$ and motion vectors $\mathcal{T}^t \in \mathbb{R}^{h\times w\times 2}$.

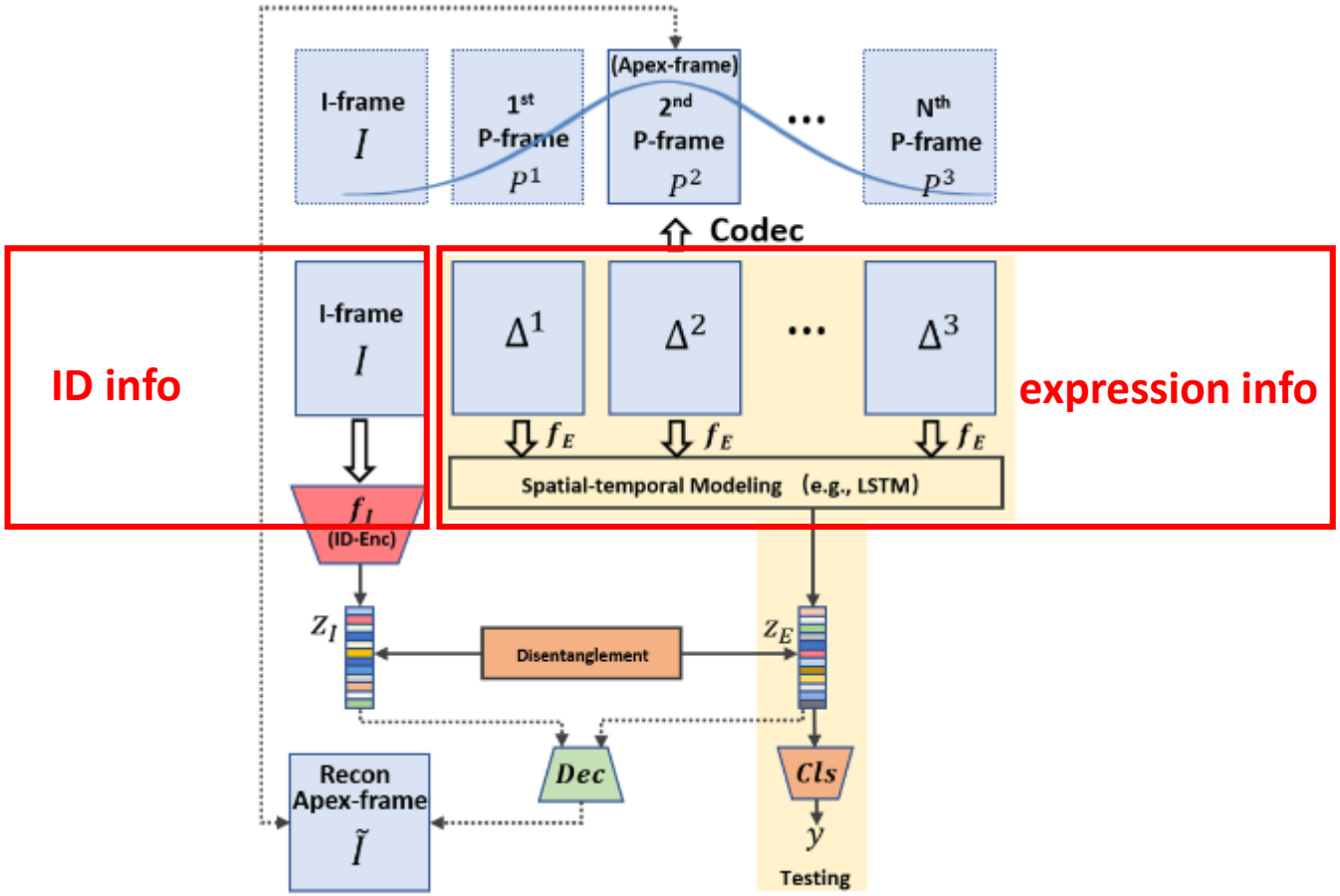
Noticing that the motion vectors $\mathcal{T}^t$ has much lower resolution, since its values within the same macroblock are identical. Considering the micro movements of facial expression in each frame, the coarse $\mathcal{T}^t$ usually not helpful for the FER. For P frame reconstruction $P_i^t = P_{i-\mathcal{T}_i^t}^{t-1} + \Delta_i^t$, where index all the pixels and $P^0 = I$. Then, $\mathcal{T}^t$ and Δ^t are processed by discrete cosine transform and entropy-encoded.

The typical compression algorithms are only developed to compress the file size, and the encoded format can be very different with the RGB images w.r.t. the statistical and structural properties. Therefore, a tailored processing network is necessary to accommodate the compressed format. Considering the structure of residual images Δ^t are much simpler than the decoded images, it is possible to utilize simpler and faster CNNs $f_E : \mathbb{R}^{h\times w\times 3} \rightarrow \mathbb{R}^{512}$ to extract the feature of each frame [3], [44]. Practically, we follow the CNN in the typical CNN-LSTM FER structure [44], [3], but with fewer layers to explore the information in Δ^t. Noticing that f_E is shared for all frames, and only needs to store one f_E in processing.

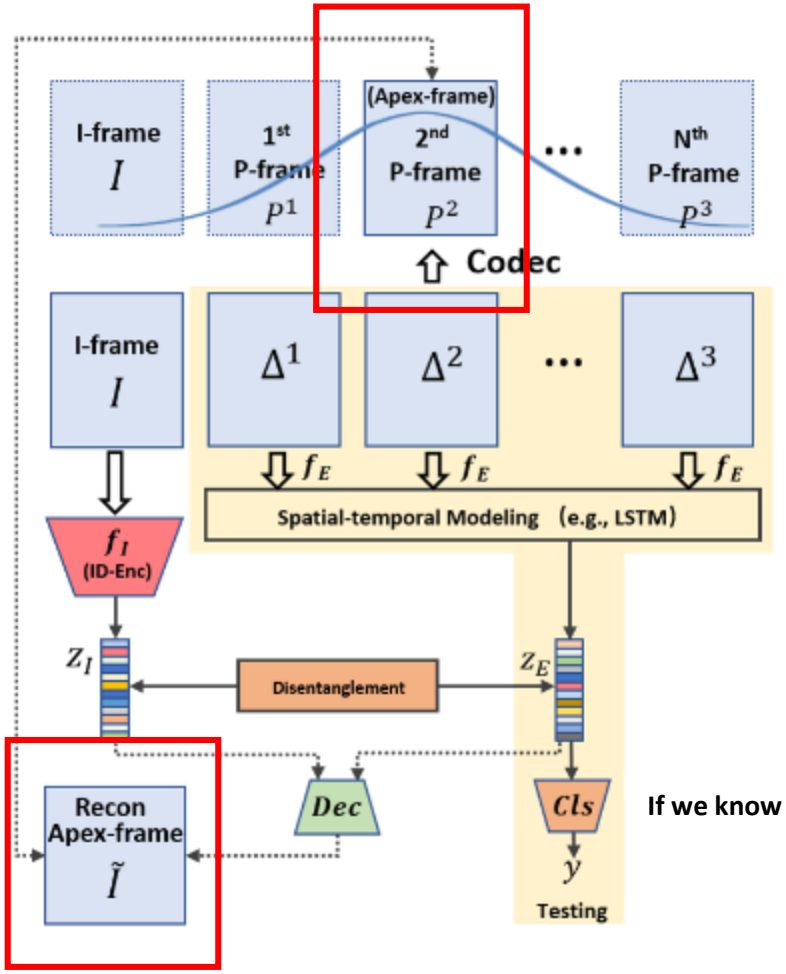


The residuals (P-frames) have expression info!

Illustration of the typical video compression and the scheme of conventional FER methods, which first decode the video and then feed it into a FER network.



Liu, Xiaofeng, B. V. K. Vijaya Kumar, Jane You, and Ping Jia. "Adaptive deep metric learning for identity-aware facial expression recognition." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 20-29. 2017.
Liu, Xiaofeng, Site Li, Lingsheng Kong, Wanqing Xie, Ping Jia, Jane You, and B. V. K. Kumar. "Feature-level frankenstein: Eliminating variations for discriminative recognition." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 637-646. 2019.



We extract expression in P-frames and identity from the I-frame to achieve id-exp disentanglement.

If we know the apex frame, the performance can be boost.

$$\mathcal{L}_1 = ||I_{Apex} - \hat{I}_{Apex}||_2^2$$

The illustration of IFERCV framework for identity-aware facial expression recognition in compressed video domain.

TABLE I: Comparison of various methods on the CK+ dataset in terms of average recognition accuracy of seven expressions. Note that in order to make the comparison fair, we do not consider image-based and 3D geometry based experiment setting and models [15], [18], [1].

Method	Accuracy	Landmarks	Test speed
STM-ExpLet (2014) [51]	94.19	×	-
LOMo (2016) [52]	95.10	✓	-
DTAGN (2015) [53]	97.25	✓	-
PHRNN-MSCNN (2017) [54]	98.50	✓	-
(N+M)-tuplet (2018) [55]	93.90	✓	12fps
C3D-GRU (2019) [56]	97.25	×	-
CTSLSTM (2019) [57]	93.9	✓	-
SC (2019) [58]	97.60	✓	-
G2-VER (2019) [59]	97.40	×	-
LBVCNN (2019) [47]	97.38	×	-
Mode variational LSTM (2019) [44]	97.42	×	11fps
IFERCV	97.44	×	**35fps**
IFERCV+Adversarial disentanglement[23]	98.38	×	**35fps**
IFERCV-$\hat{I}$	97.16	×	**35fps**
IFERCV+$\mathcal{T}^{-t}$	97.85	×	29fps

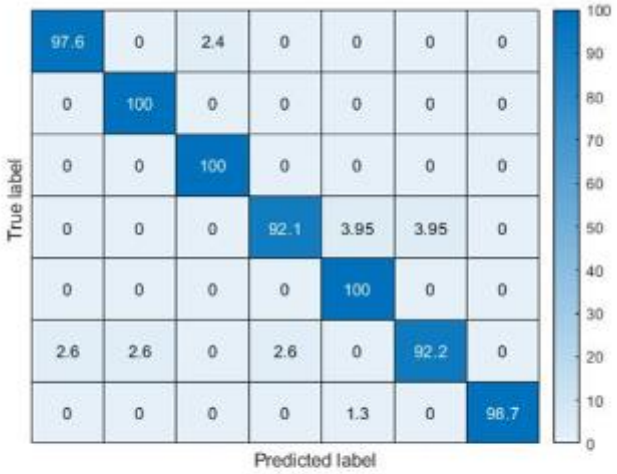


Fig. 3: Confusion matrix of IFERCV on CK+ datasets.

TABLE II: Comparison of various methods on the AFEW dataset in terms of average recognition accuracy of seven expressions. *optical flow is used.

Method	Model type	Accuracy	Test speed
CNN-RNN (2016) [62]	45.43	Dynamic	-
Undirectional LSTM (2017) [63]	48.60	Dynamic	-
HoloNet (2016) [64]	44.57	Static	-
DSN-HoloNet (2017) [65]	46.47	Static	-
DenseNet-161 (2018) [66]	51.44	Static	-
DSN-VGGFace (2018) [67]	48.04	Static	-
FAN (2019) [5]	51.18	Static	-
CTSLSTM (2019) [57]	51.2	Dynamic	-
C3D-GRU (2019) [56]	49.87	Dynamic	-
DSTA (2019)* [68]	42.98	Dynamic	-
E-ConvLSTM (2019)* [69]	45.29	Dynamic	4fps
Mode variational LSTM (2019) [44]	51.44	Dynamic	11fps
IFERCV	51.62	Dynamic	**34fps**
IFERCV+Adv Disentanglement[23]	52.01	Dynamic	**34fps**
IFERCV+$\mathcal{T}^t$	**51.86**	Dynamic	30fps

Fig. 4: Confusion matrix of IFERCV on AFEW datasets.

HARVARD MEDICAL SCHOOL

MASSACHUSETTS GENERAL HOSPITAL

Gordon Center for Medical Imaging

Identity-aware Facial Expression Recognition in Compressed Video

Xiaofeng Liu, Linghao Jin, Xu Han, Jun Lu, Jane You, Lingsheng Kong

https://liu-xiaofeng.github.io/ liuxiaofengcmu@gmail.com