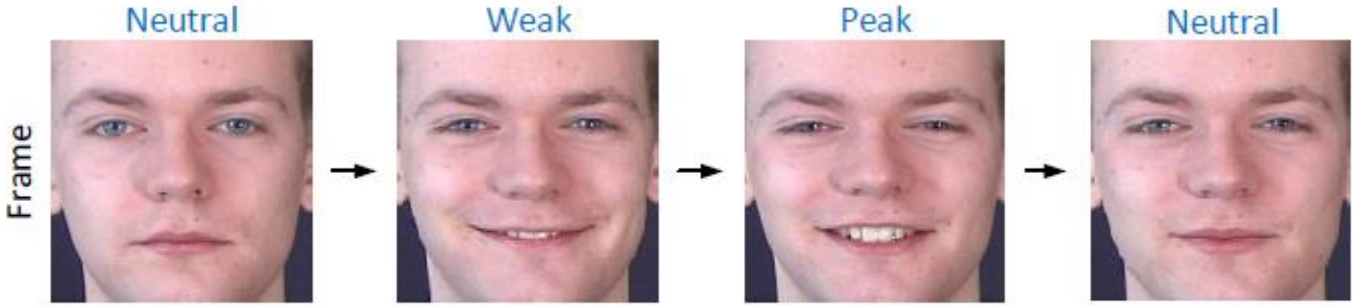
Video-based Facial Expression Recognition using Graph Convolutional Networks

Daizong Liu[1], Hongting Zhang[1], and Pan Zhou[1]

[1]Huazhong University of Science and Technology

Video-based Facial Expression Recognition

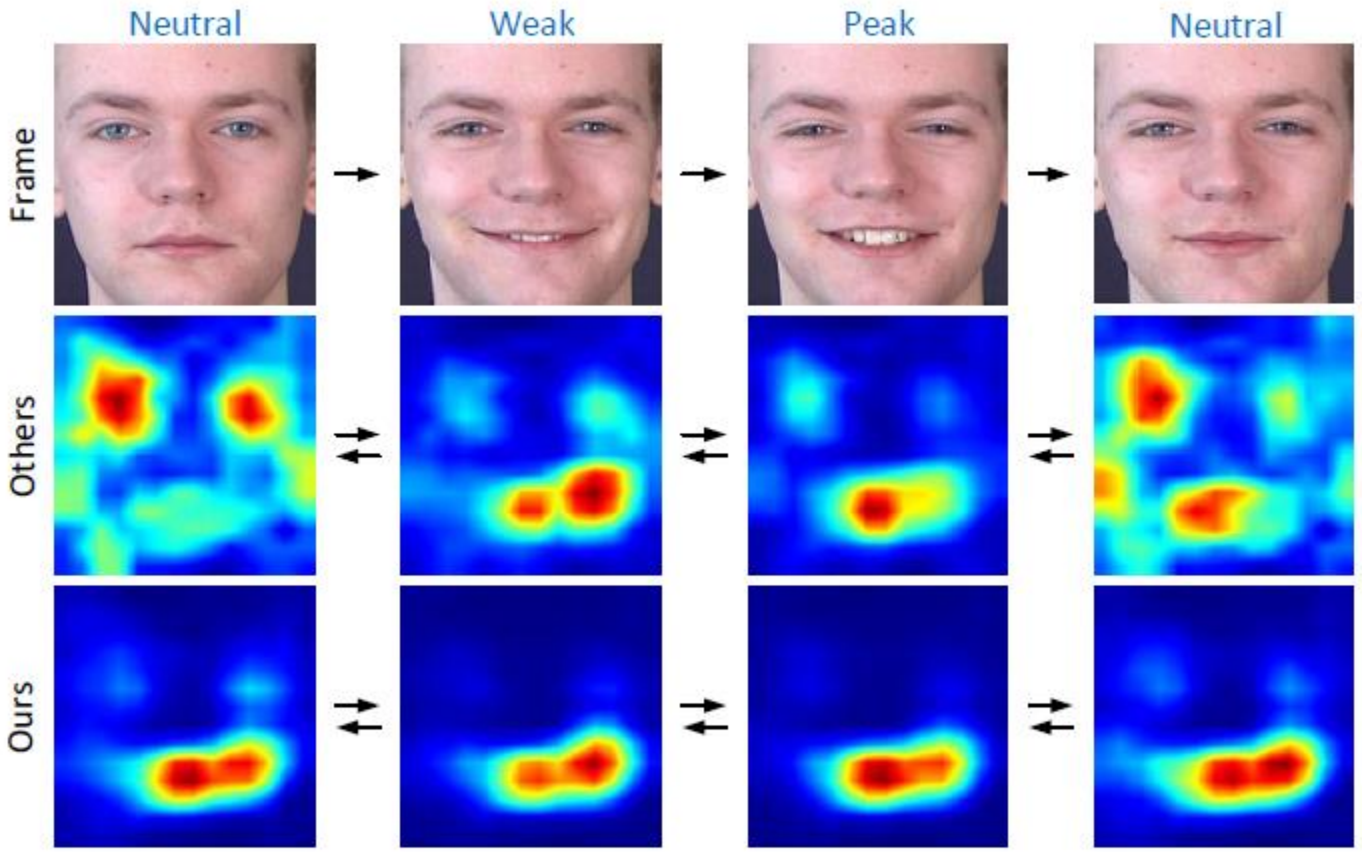
- Video based classification task for facial expression recognition
- Inputs: a facial video sequence
- Outputs: corresponding expression id



Expression: “Happy”

“Video-based Facial Expression Recognition using Graph Convolutional Networks”
D Liu, H Zhang, P Zhou, **ICPR 2020**

Challenges

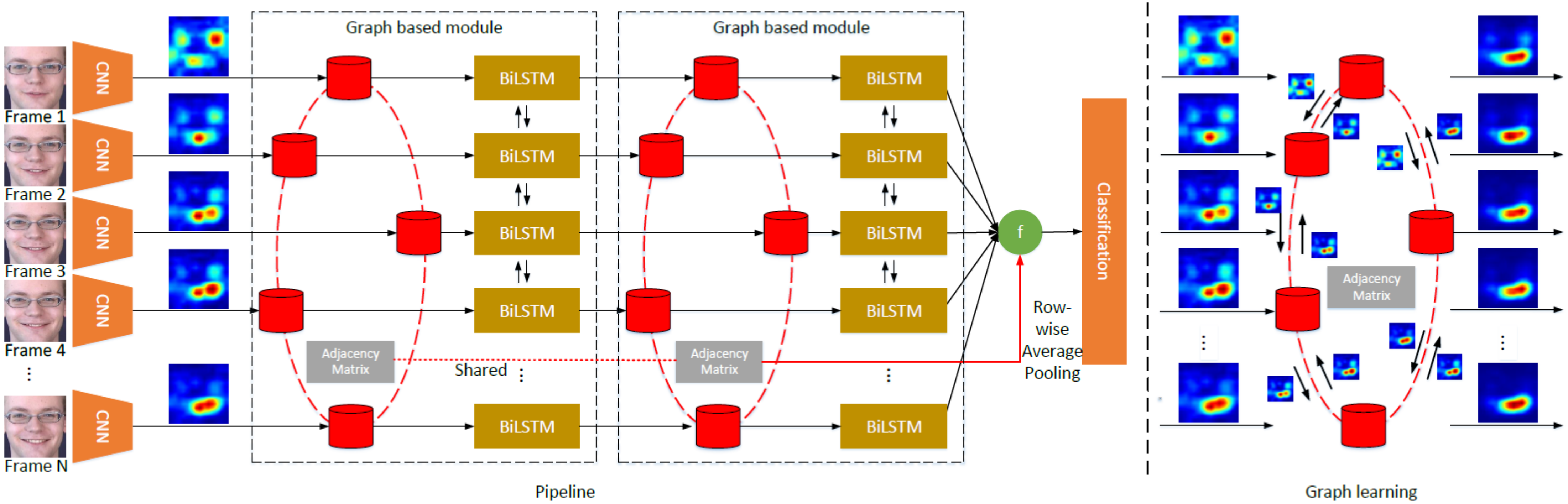


"Video-based Facial Expression Recognition using Graph Convolutional Networks"
D Liu, H Zhang, P Zhou, **ICPR 2020**

Motivation

- Existing methods directly utilize CNN-RNN or 3D CNN to extract the spatial-temporal features from different facial units, instead of concentrating on a certain region during expression variation capturing. We utilize GCN layer to learn more significant facial expression features which concentrate on certain regions.
- The learned features of the peak frames have more informative expressional representations than those of non-peak frames and should be considered more for final recognition. We design a weight assignment mechanism to weight the output of different nodes for final classification by characterizing the expression intensities in each frame.

Pipeline



“Video-based Facial Expression Recognition using Graph Convolutional Networks”
D Liu, H Zhang, P Zhou, **ICPR 2020**

Quantitative Comparison

TABLE I: Average accuracy on the CK+, Oulu-CASIA and MMI datasets respectively.

Method	CK+	Oulu	MMI	Feature
Inception [13]	93.20%	-	77.60%	static
IACNN [22]	95.37%	-	71.55%	static
DLP-CNN [23]	95.78%	-	-	static
FN2EN [24]	96.80%	87.71%	-	static
DeRL [25]	97.30%	88.00%	73.23%	static
PPDN [15]	99.30%	84.59%	-	static
3DCNN [14]	85.90%	-	53.20%	Dynamic
ITBN [26]	86.30%	-	59.70%	Dynamic
HOG 3D [27]	91.44%	70.63%	60.89%	Dynamic
TMS [28]	91.89%	-	-	Dynamic
3DCNN-DAP [14]	92.40%	-	63.40%	Dynamic
STM-ExpLet [29]	94.19%	74.59%	75.12%	Dynamic
LOMo [30]	95.10%	82.10%	-	Dynamic
3D Inception-Resnet [31]	95.53%	-	79.26%	Dynamic
Traj. on S+(2, n) [32]	96.87%	83.13%	79.19%	Dynamic
DTAGN [33]	97.25%	81.46%	70.24%	Dynamic
GCNet [34]	97.93%	86.11%	81.53%	Dynamic
PHRNN-MSCNN [6]	98.50%	86.25%	81.18%	Dynamic
Ours	**99.54%**	**91.04%**	**85.89%**	Dynamic

TABLE VI: Recognition accuracy of each single model on the validation dataset of AFEW 8.0.

Method	Accuracy
Emotiw2018 (baseline) [37]	38.81%
HoloNet [39]	46.50%
DSN-VGG-Face [40]	48.04%
Resne50-LSTM [38]	49.31%
DenseNet161-pool5 [41]	51.44%
VGG-Face-LSTM [38]	53.91%
Ours	**55.67%**

“Video-based Facial Expression Recognition using Graph Convolutional Networks”
D Liu, H Zhang, P Zhou, **ICPR 2020**

Ablation Study

TABLE V: Ablation study on the individual components.

Experiment model	CK+	Oulu-CASIA	MMI
VGG16	97.78%	85.83%	80.75%
VGG16 + graph based spatial-temporal module×1	98.39%	88.33%	84.37%
VGG16 + graph based spatial-temporal module×2	99.09%	89.79%	84.64%
VGG16 + graph based spatial-temporal module×3	99.00%	87.71%	83.07%
VGG16 + graph based spatial-temporal module×2 + weighted feature fusion	**99.54%**	**91.04%**	**85.89%**

“Video-based Facial Expression Recognition using Graph Convolutional Networks”
D Liu, H Zhang, P Zhou, **ICPR 2020**

Visualization

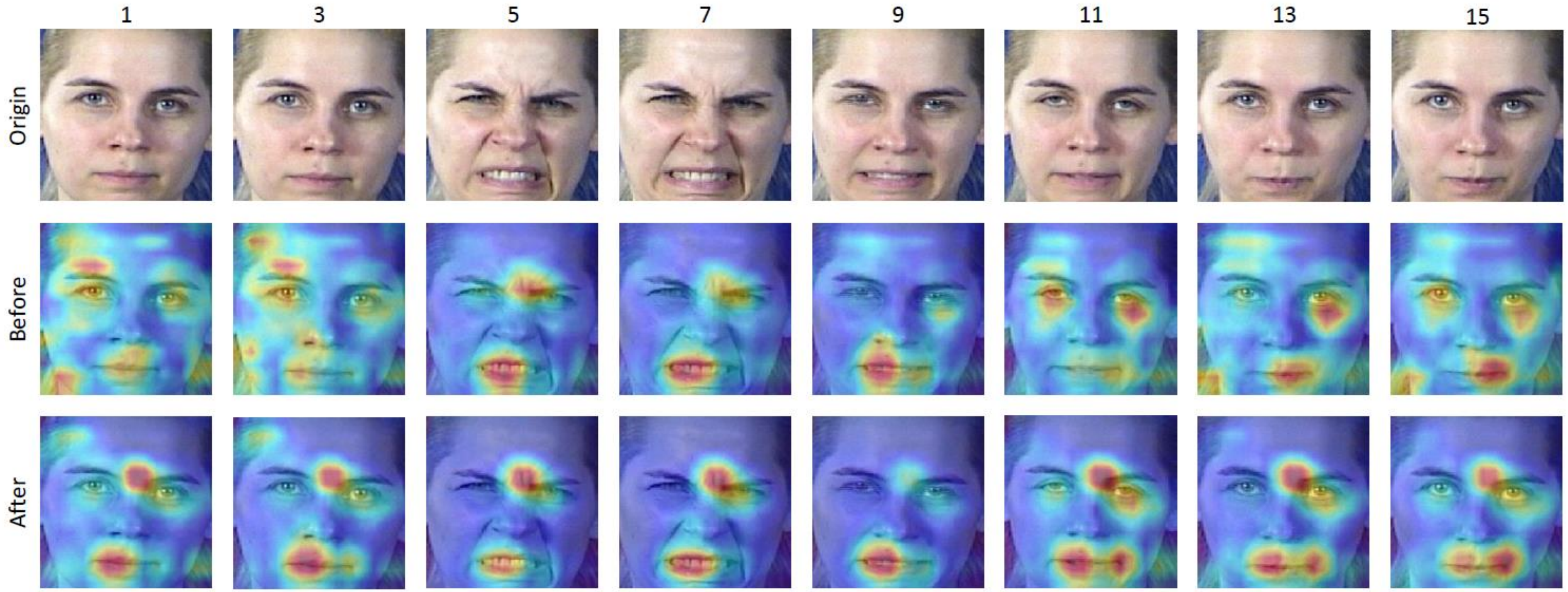
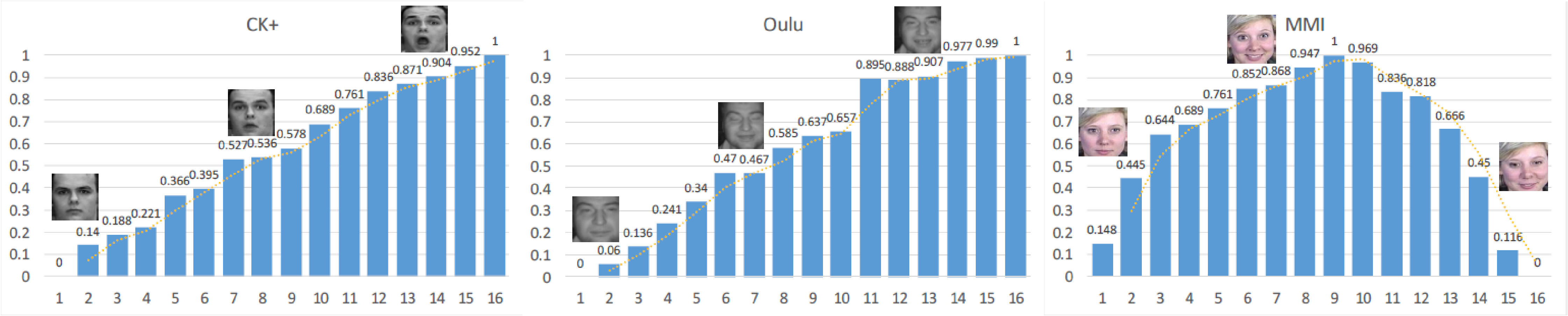


Fig. 3: Example of the feature reconstruction in our GCN layer. First row: Origin facial images of ”Disgust” in MMI dataset; Second row: input features of GCN layer; Third row: output features of GCN layer. It clarifies that our GCN layer shares most contributing expression features among frames to helps model focus more on the corresponding expression regions (such as mouth and nose here).

“Video-based Facial Expression Recognition using Graph Convolutional Networks”
D Liu, H Zhang, P Zhou, **ICPR 2020**

Visualization



(a) CK+ (b) Oulu-CASIA (c) MMI

Fig. 4: Visualization of expression intensity weights for 16 steps on three datasets respectively. The horizontal axis represents the step number in each video sequence. The values of temporal weighs are given in the vertical axis through a sigmoid function, which refer to the expression intensity of each frame in the dynamic expression variation.

“Video-based Facial Expression Recognition using Graph Convolutional Networks”
D Liu, H Zhang, P Zhou, **ICPR 2020**

Thanks!

Email: dzliu@hust.edu.cn