Fast Video Facial Expression Recognition by Deeply Tensor-compressed LSTM Neural Network on Mobile Device

Peining Zhen
Shanghai Jiao Tong University
Shanghai, China
zhenpn@sjtu.edu.cn

Bin Liu
Southern University of Science and Technology
Shenzhen, China
georgeokelly@foxmail.com

Yuan Cheng
Shanghai Jiao Tong University
Shanghai, China
cyuan328@sjtu.edu.cn

Hai-Bao Chen
Shanghai Jiao Tong University
Shanghai, China
haibaochen@sjtu.edu.cn

Hao Yu
Southern University of Science and Technology
Shenzhen, China
yuh3@sustc.edu.cn

ABSTRACT

Poster: Mobile devices usually suffer from limited computation and storage resource which seriously hinders them from deep neural network applications. In this paper, we introduce a deeply tensor-compressed LSTM neural network for fast facial expression recognition (FER) in videos on mobile devices. Firstly, a spatio-temporal FER LSTM model is built by extracting time-series feature maps from facial clips. The LSTM model is further deeply compressed with tensorization. Based on dataset of Acted Facial Expression in Wild (AFEW) 7.0, experimental results show that the proposed method achieves 55.60% classification accuracy; and significantly compresses the size of network model by 219×. Our work is further implemented on RK3399Pro IoT device with Neural Process Engine, and the runtime of feature extraction part can be reduced by 12.83× with only 7.73W power consumption.

KEYWORDS

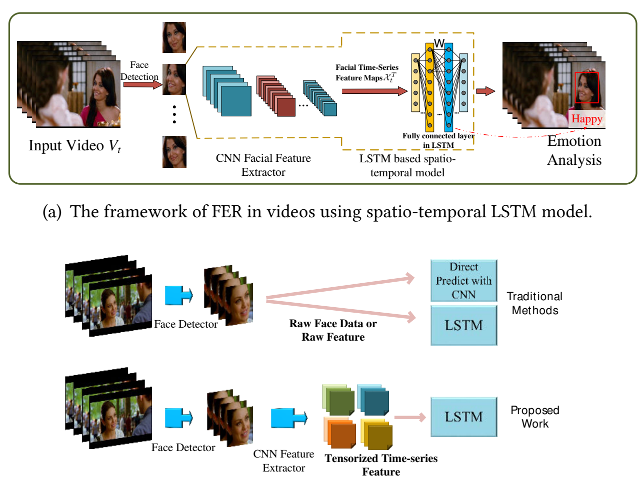
mobile device, deep learning, tensor decomposition, facial expression recognition

ACM Reference Format:
Peining Zhen, Bin Liu, Yuan Cheng, Hai-Bao Chen, and Hao Yu. 2019. Fast Video Facial Expression Recognition by Deeply Tensor-compressed LSTM Neural Network on Mobile Device. In *The Fourth ACM/IEEE Symposium on Edge Computing (SEC 2019), November 7–9, 2019, Arlington, VA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3318216.3363322

1 INTRODUCTION

Facial expression recognition (FER) is important with wide applications [6, 15], which can be classified into two categories: hand-crafted and deep learning based approaches. It however remains as a challenge to develop a fast FER in videos. The deep learning based approaches such as recurrent neural networks (RNN) can be utilized for FER with temporal information [4, 5]. However, RNN and its variations (e.g., IRNN [10] and BRNN [13]) usually cannot perform well due to their weak capability of capturing high dimensional facial features in videos. In [9], Kim *et al.* leveraged a normal CNN for spatial information and a plain RNN or LSTM for temporal information with good performance. Zhang *et al.* trained two parallel CNNs called PHRNN and MSCNN to make use of spatio-temporal information respectively, which can significantly boost the performance of FER [17]. All those methods are however limited by the huge number of redundant parameters in the neural network architecture with non-structured treatment of the spatial-temporal features.

In this paper, to achieve a fast and accurate FER in videos on terminal devices, we develop a spatio-temporal LSTM model with structured or tensorized time-series facial features, which is further deeply compressed by tensor decomposition. Fig. 1(a) shows the overall framework. It first extracts facial features from each frame of facial clips. A LSTM model is constructed based on the time-series feature maps, which can model facial temporal information for the sake of facial expression recognition.



(a) The framework of FER in videos using spatio-temporal LSTM model.

(b) Comparison of the proposed work in this paper with previous method.

Figure 1: Framework and method comparison.

Fig. 1(b) shows the comparison of the proposed work in this paper with other previous methods, which directly handle non-structured raw data or features with no compression. Though the construction of a spatio-temporal LSTM model can adequately provide a FER in videos, it becomes computationally expensive due to the large size of the resulting network. There are too many network parameters to be identified in fully connected layers, which hinders the use of LSTM in previous methods [7, 12, 14] from

SEC 2019, November 7–9, 2019, Arlington, VA, USA

ACM ISBN 978-1-4503-6733-2/19/11…$15.00
https://doi.org/10.1145/3318216.3363322

solving practical FER in videos. In order to solve this problem, a tensor based compression method is introduced in this paper to further optimize the spatio-temporal LSTM model. The example of computing one element of a 3-dimensional tensorized tensor is graphically shown in Fig. 2.

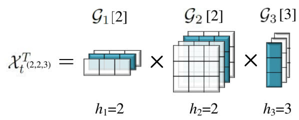


Figure 2: An example of computing one element of 3-dimensional tensor.

The proposed LSTM-based spatio-temporal model with tensorization for FER is shown in Fig. 3. We assume that $V_t \in \mathbb{R}^{l_1 \times l_2}$ are the input video frames, $F_t \in \mathbb{R}^{l_1 \times l_2}$ are the detected faces and $\mathcal{X}_t^T \in \mathbb{R}^{l_1 \times l_2 \times l_3}$ are facial time-series features, where t indicates the time sequence, l_k represents the size of this dimension and k is the dimensionality. All faces in the video frames are firstly detected and this procedure can be expressed as $D(V_t) = F_t$. Then we adopt a CNN based feature extractor, $E(F_t) = \mathcal{X}_t^T$, to extract facial features. The entire process can be expressed as: $\mathcal{X}_t^T \xleftarrow{E(F_t)} F_t \xleftarrow{D(V_t)} V_t$.

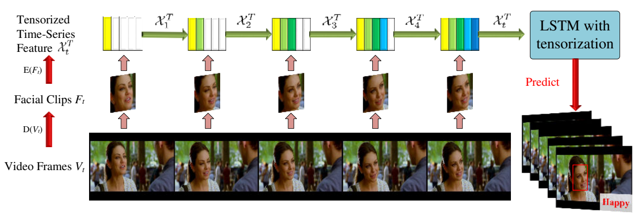


Figure 3: Tensorization of the spatio-temporal LSTM model.

Our framework is then carefully implemented on RK3399Pro IoT board with Neural Process Unit (NPU). NPU is customized for accelerating operations in neural network, which supports 1920 Int8 MAC and 64 FP16 operations per cycle. The most time-consuming parts of our framework are the feature extraction part and LSTM prediction part, and our goal is to reduce the inference time of these two parts on mobile device. Since the LSTM network has been speed up by tensor decomposition, we use the NPU to accelerate the feature extraction part. Verification platform is shown in Fig. 4, and detailed experimental results are reported in section 2.2.

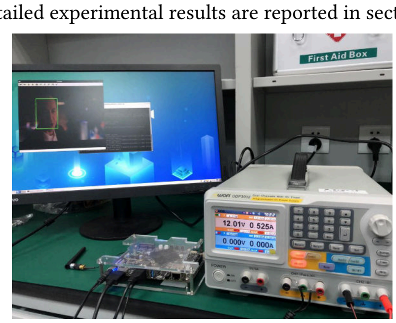


Figure 4: Verification platform for our proposed method.

2 EXPERIMENTAL RESULTS

2.1 Performance Analysis

Table 1 shows the overall classification accuracy obtained by our work and other techniques on the AFEW 7.0 dataset. The method presented in [12] applies transfer learning and audio-visual approach for emotion recognition, but the result is still 2.7% lower than ours. Yan *et al.* use a complex fusion of facial textures, facial landmark action and audio signal to recognize expressions [16], and its classification accuracy is 6.38% lower than ours. A multimodal approach is proposed in [14] for facial expression recognition in videos and our result is 7% higher than that on validation set. As shown in Table 1, our framework achieves highest accuracy using time-series feature maps and tensor decomposition.

Table 1: Overall accuracy comparison on the AFEW 7.0 dataset.

Method	Accuracy
Multiple Temporal Models[12]	53.90%
Decision Fusion[16]	49.22%
Unidirectional LSTM on layer[14]	48.60%
Hybrid Feature II[3]	45.20%
ELM[8]	44.20%
Ours (Tensorized LSTM with features)	**55.60%**

Since we store the tensor cores rather than large scale weight matrices of fully connected layers, the input-to-hidden mapping parameters can be greatly reduced. This means that we can train our model with fewer parameters which leads to significant reduction in the storage space and training time. It can be seen from Table 2 that the LSTM model size can be compressed by 219×. Besides, compared with other advanced methods, our method also shows the advantages in storage size as shown in this table. These key features demonstrate the broad prospects of our method on mobile devices since they usually have limited storage space and computing resources.

Table 2: Storage size comparison.

Method	Storage size	Compression
deepnn[1]	41.8 MB	-
3D-CNN[11]	232.57 MB	-
ExpNet[2]	659.98 MB	-
Plain LSTM with raw data	678.06MB	-
Ours (Tensorized LSTM with features)	**3.10MB**	**219×**

2.2 Mobile Device Evaluation

We realize our work on RK3399Pro mobile device, and the total storage size of the proposed framework is 60.7MB, which only occupies 0.09% of total 64GB ROM on RK3399Pro. Furthermore, when processing a 720×576 video, our method only cost 92.53MB memory which is far from the total 6GB RAM. Finally, we verified the performance of acceleration with NPU on the RK3399Pro board. Table 3 shows that the accelerated feature extractor is 12.83× faster than original extractor without NPU.

Table 3: Run time comparison of feature extraction part.

Method	Runtime	Speed up
CPU only	93.58s	-
NPU Accelerated	7.29s	12.83×

REFERENCES

[1] [n.d.]. https://github.com/xionghc/Facial-Expression-Recognition.

[2] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. 2018. ExpNet: Landmark-free, deep, 3D facial expressions. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 122–129.

[3] Junkai Chen, Zenghai Chen, Zheru Chi, and Hong Fu. 2018. Facial expression recognition in video with multiple feature fusion. *IEEE Transactions on Affective Computing* 9, 1 (2018), 38–50.

[4] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 467–474.

[5] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 445–450.

[6] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S Lew. 2016. Deep learning for visual understanding: A review. *Neurocomputing* 187 (2016), 27–48.

[7] Sarasi Kankanamge, Clinton Fookes, and Sridha Sridharan. 2018. Facial analysis in the wild with LSTM networks. In *IEEE International Conference on Image Processing*.

[8] Heysem Kaya and Albert Ali Salah. 2016. Combining modality-specific extreme learning machines for emotion recognition in the wild. *Journal on Multimodal User Interfaces* 10, 2 (2016), 139–149.

[9] D. H. Kim, W. Baddar, J. Jang, and Y. M. Ro. 2018. Multi-Objective based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition. *IEEE Transactions on Affective Computing* (2018), 1–1. https://doi.org/10.1109/TAFFC.2017.2695999

[10] Quoc V Le, Navdeep Jaitly, and Geoffrey E Hinton. 2015. A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941* (2015).

[11] Behzad H Mohammad Mahoor et al. 2017. Facial expression recognition using enhanced deep 3D convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 30–40.

[12] Xi Ouyang, Shigenori Kawaai, Ester Gue Hua Goh, Shengmei Shen, Wan Ding, Huaiping Ming, and Dong-Yan Huang. 2017. Audio-visual emotion recognition using deep transfer learning and multiple temporal models. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 577–582.

[13] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[14] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. 2017. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 569–576.

[15] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27, 12 (2009), 1743–1759.

[16] Jingwei Yan, Wenming Zheng, Zhen Cui, Chuangao Tang, Tong Zhang, and Yuan Zong. 2018. Multi-cue fusion for emotion recognition in the wild. *Neurocomputing* 309 (2018), 27–35.

[17] Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. 2017. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing* 26, 9 (2017), 4193–4203.