FERL-YOLO: Facial Expression Recognition Model of Learners

Tao SUN[a*], Li CHEN[b], Sijie XIONG[a], Cheng TANG[a], Gen LI[a], Atsushi SHIMADA[a]
[a]*Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University, Japan*
[b]*Division of Math, Sciences, and Information Technology in Education, Osaka Kyoiku University, Japan*
*sun.tao.480@s.kyushu-u.ac.jp

Abstract: Facial expression recognition of learners plays a crucial role in optimizing educational strategies. However, variability in facial expressions and environments limit the accuracy of existing models. To address this challenge, we propose FERL-YOLO, a model based on YOLOv11, integrating Haar Cascade for face detection, SENet for adaptive feature enhancement, and Optuna for hyperparameter optimization. Experiments on FER-2013 demonstrate that FERL-YOLO achieves competitive performance. Furthermore, to evaluate the reliability of emotion recognition, we conducted a reading experiment using our model to recognize learners' emotions through facial expressions and learners' self-reported feelings. Our findings provide valuable insights into the refinement of our model in future real educational settings.

Keywords: Emotions of learners, Facial expression recognition, YOLO

1. Introduction

Research has demonstrated that emotions can influence cognitive functions such as attention, learning, and memory (Jung et al., 2014). By analyzing and recognizing learners' emotions, educators can better understand learners' needs. Among advanced approaches in deep learning, facial expression recognition (FER) is relatively more applicable. However, variability in facial expressions and environments limit FER's performance.

To address these issues, deep learning techniques have been leveraged by researchers. One representative of these methods is You Only Look Once (YOLO) (Redmon et al., 2016). However, YOLO still underperforms humans in terms of accuracy on benchmark datasets. The most concerning factor is the use of Spatial Pyramid Pooling Fast (SPPF) (He et al., 2015). Although SPPF can improve spatial invariance via aggregating multi-scale information, YOLO greatly demands an adaptive attention mechanism (Bhavana et al., 2024).

To overcome YOLO's deficiencies, we replace SPPF with Squeeze-and-Excitation Network (SENet) and propose an improved architecture, FERL-YOLO. Furthermore, we conducted a pilot experiment designed to assess emotion recognition performance in an educational setting. Our contributions are as follows:

- We modify the YOLOv11 architecture to better suit facial expression classification. Specifically, we integrate SENet into Backbone to enhance feature extraction, allowing our model to adaptively recalibrate feature maps and improve recognition accuracy.
- We improve facial detection accuracy by adopting Haar Cascade (Viola et al., 2001) as a preprocessing step to crop facial regions for classification. Furthermore, we incorporate Optuna (Akiba et al., 2019) for automated hyperparameter optimization during training.
- To validate the practical applicability of our model in real learning environments, we design a controlled reading scenario and conduct a pilot experiment.

2. Method

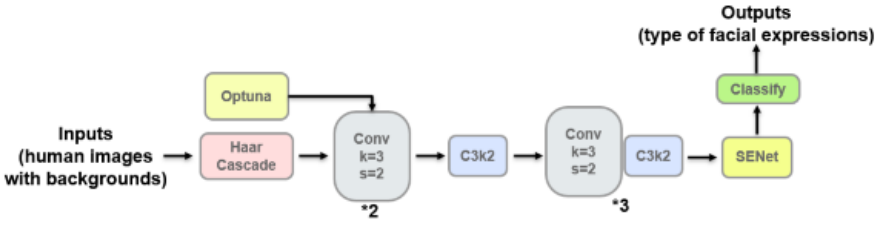

Figure 1. Architect of FERL-YOLO

As shown in Figure 1, firstly, we need to eliminate unnecessary background information and reduce computational complexity. We achieve this at a lower cost by abandoning Neck and Head of YOLOv11, while adopting Haar Cascade for initial face localization. Second, our model processes the input image through two Conv blocks with a kernel size of 3 and a stride of 2, followed by a C3K2 block to extract deep-level features efficiently. To further improve our model's ability to focus on key facial features, the original SPPF block is replaced with SENet.

SENet mainly has three key steps: Squeeze, Excitation, and Recalibration. It allows the network to focus on more informative features. Instead of attention-lacking SPPF, SENet can introduce channel-wise attention and is effective at recalibrating feature maps by strengthening informative channels.

After the process of SENet, feature maps are passed to our classification module, which predicts the final facial expression label. Our classification module first applies a convolutional layer to refine features, followed by an adaptive average pooling layer to aggregate global spatial information. Then, a dropout layer prevents overfitting, and a fully connected layer maps the features to the target expression classes.

3. Experiment on the Dataset and Simulation

FER2013 is a widely used dataset in FER. It is categorized into seven emotion classes: **angry, disgust, fear, happy, sadness, surprise**, and **neutral**. The 35,886 images are partitioned into three subsets: 60% of the images for training, 20% for validation and 20% for testing.

The FERL-YOLO model is trained under the cross-entropy loss function. After incorporating SENet and Optuna, our model reaches a peak accuracy of 75.4% over 500 training epochs (Figure 2), which represents an improvement of 4.7% compared to baseline accuracy. We have also compared our FERL-YOLO with other advancing FER-task methods (Table 1), including CNN+SVM, VGG, ResNet (Giannopoulos et al., 2018), and Siamese-like CNN (Zhang et al., 2015).

Table 1. *Comparison of Different Models in FER-2013*

Method	CNN+SVM	VGG	ResNet	Siamese-like CNN	Ours
Accuracy (%)	71.2	73.3	72.4	75.1	75.4

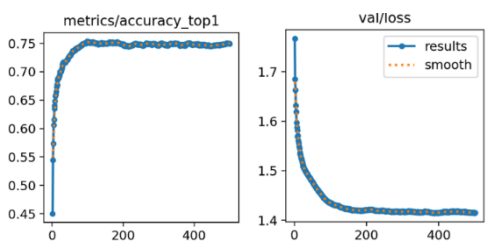

Figure 2. Performance of FERL-YOLO

Table 2. *Accuracy of our model in a simulated experiment (Without **Angry** and **Disgust**)*

Emotion	**neutral**	**surprise**	**fear**	**sadness**	**happy**	average
Accuracy (%)	73.2	75.3	33.4	50.6	90.4	66.3

To simulate a realistic educational setting, we have four postgraduate volunteers to participate in a simulated reading experiment and a total of 5,150 web-camera frames are collected. We observe significant variations in the model's accuracy across different emotions (Table 2). A closer examination reveals that **fear** and **sadness** are often mistakenly classified as **surprise** due to similar facial features, such as furrowed brows, wide-open eyes.

4. Conclusion

In this study, we propose FERL-YOLO, a FER model designed to detect learners' emotions in educational settings. The results on FER2013 show that FERL-YOLO achieves competitive accuracy. However, in our pilot experiment, there is inconsistent accuracy across different emotions. For future work, our model still requires further optimization for real educational settings. Integrating FERL-YOLO with learning analytics can provide educators with a promising tool to observe relationships between students' emotions and their behaviors.

Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR22D1 and JSPS KAKENHI Grant Number JP22H00551, JP24K16759, Japan.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631).

Bhavana, N., Kodabagi, M. M., Kumar, B. M., Ajay, P., Muthukumaran, N., & Ahilan, A. (2024). POT-YOLO: Real-Time Road Potholes Detection using Edge Segmentation based Yolo V8 Network. *IEEE Sensors Journal*.

Giannopoulos, P., Perikos, I., & Hatzilygeroudis, I. (2018). Deep learning approaches for facial emotion recognition: A case study on FER-2013. In *Advances in hybridization of intelligent methods: Models, systems and applications*, 1-16.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *37*(9), 1904-1916.

Jung, N., Wranke, C., Hamburger, K., & Knauff, M. (2014). How emotions affect logical reasoning: evidence from experiments with mood-manipulated participants, spider phobics, and people with exam anxiety. *Frontiers in psychology*, *5*, 570.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001* (Vol. 1, pp. I-I). IEEE.

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2015). Learning social relation traits from face images. In *Proceedings of the IEEE international conference on computer vision* (pp. 3631-3639).