Weighted Class Loss for Single-Staged Facial Emotion Recognition

Jo Vianto[1], Hyung-Jeong Yang[1*], Seung-won Kim[1], Ji-eun Shin[2], Soo-Hyung Kim[1]
[1]Dept. of AI Convergence, Chonnam National University
[2]Dept. of Psychology, Chonnam National University*Corresponding Author

*Corresponding author: hjyang@jnu.ac.kr

Abstract

Facial emotion recognition (FER) is becoming crucial in fields like human-computer interaction and surveillance. Traditional FER systems rely on two-stage models with face alignment preprocessing, which increases complexity and inference time. In this research, we propose a single-stage approach using YOLOv6 combined with weighted class loss to address these inefficiencies. Our method improves computational efficiency while enhancing the detection of minority classes in imbalanced emotion datasets. The experiments demonstrate that although the weighted loss function helps with class detection, it slightly reduces overall accuracy. Nevertheless, the model shows promise for real-time FER applications, balancing accuracy and speed. This work not only introduces a more efficient approach but also highlights the potential of single-stage models in advancing emotion recognition tasks.

1. Introduction

A deep facial expression recognition model typically includes a face alignment pre-processing step to remove background and non-face areas. The Viola-Jones (V&J) [1] which is a conventional face detector algorithm that is popular and widely used for this stage[2]. A more modern and accurate model using deep learning methods is also employed. This type of model, which consists of a face alignment step and a classification step, is referred to as a two-stage model.

The main issue with two-stage models is their inference time and overall efficiency. Since the two stages are disjointed, they require separate learning and inferencing processes. Each stage comes with its own set of parameters, many of which could potentially be shared between the stages but aren't, leading to redundant computations. As a result, this separation causes inefficient resource utilization and slower inference times, making the model less optimal for real-time applications compared to single-stage models.

To address the issue above, we conducted experiments using a single-stage You Look Only Once(YOLO) model for the facial expression recognition task. The first version of YOLO design enables end-to-end training, with great average precision and real-time speed for object detection [3]. It proposed a unified model for bounding boxes and class prediction simultaneously. The YOLOv6 [4] model is employed in our experiments to evaluate its performance on the facial expression recognition task without the need for face alignment, relying only on a single-stage process.

Additionally, YOLOv6 faces challenges in detecting certain classes with a low number of training samples, leading to imbalanced performance across emotions. The model tends to perform well on frequently occurring classes but struggles with underrepresented ones due to its reliance on sufficient data for effective generalization. We mitigated this issue using weighted class loss. Overall, the main contribution can be summarized as follows:

- Exploring the YOLO model for single-stage facial expression recognition.
- Performing weighted class loss to mitigate imbalanced class dataset numbers.
- Conducting extensive experiments to analyze model performance with multiple hyper-parameter settings.
- Propose a novel approach to evaluate single-staged facial expression recognition model performances.

2. Related Work

There is not much research has been done on employing the YOLO model for the expression recognition task. Chaitanya et al. in [5] performed emotion recognition from thermal image input using the YOLO model. Background from the original thermal image was eliminated before being fed into the YOLO-Lite [6] model. Traditional emotion recognition methods often rely on visible light images and require face alignment and controlled lighting conditions. However, the authors propose using thermal images, which capture physiological heat patterns on the face, to bypass these limitations. Thermal imaging offers the advantage of being effective in various lighting environments and can reveal subtle changes related to emotions, such as stress or anxiety. YOLO, known for its real-time object detection capabilities, is applied here for detecting facial expressions in a single stage without face alignment. This approach leverages YOLO's

speed and accuracy, while addressing the drawbacks of conventional methods, making it a more efficient solution for real-time emotion recognition from thermal data. However, thermal imaging lacks the detail of traditional RGB images, making it difficult to capture subtle facial expressions. Additionally, A thermal camera is more expensive than an RGB camera, making it less cost-effective for production environments. Environmental factors like temperature fluctuations also affect thermal accuracy.

3. Methodology

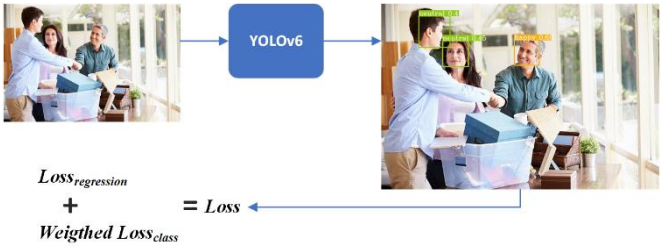

Figure 1. Our model framework could detect face and recognizes emotions simultaniously.

A. Model Framework

In **Fig 1** we employed YOLOv6 v.3.0[4], a state-of-the-art object detection algorithm, for our task due to its superior performance in real-time detection and accuracy. YOLOv6 v.3.0 will be referred to as YOLOv6 hereafter for simplicity. YOLOv6's architecture, optimized for speed and precision, allowed us to achieve high detection rates without compromising computational efficiency. EfficientRep is employed as the backbone, before being fed into the "neck" layer called Rep-PAN. The neck of YOLOv6 is responsible for multi-scale feature integration, a critical aspect of object detection. It adopts the PAN (Path Aggregation Network) structure from previous YOLO versions but modifies it to include RepBlock. By effectively aggregating low-level physical features with high-level semantic features, the neck builds feature maps at different scales, enhancing the model's ability to detect objects of varying sizes and complexities.

Finally, YOLOv6's decoupled head architecture boosts performance by incorporating anchor-aided training (AAT), which improves accuracy while maintaining speed. During training, auxiliary anchor-based branches are added to the classification and regression heads, providing additional guidance and stability. These branches are removed during inference, preserving the model's speed while benefiting from the performance improvements gained during training[4].

B. Weighted Loss Function

In our experiment, we used the same loss function with YOLOv6. The only difference is we introduced a weighted class loss function to overcome the class imbalance problem in our training set, we employed the weighted cross-entropy loss approach for classification loss. The weighted cross-entropy loss function is defined below:

$$\mathcal{L}_{weighted} = -\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \cdot y_{i,c} \log(\hat{y}_{i,c})$$

$$w_c = \frac{N}{N_c}$$

where N is the number of samples, N_c is the number of samples in class C, w_c is the weight for a class, $y_{i,c}$ and $\hat{y}_{i,c}$ are true value and predicted value of a class respectively.

C. Multi-scale aligned image metrics.

Figure 2. Visualization of multi-scale augmented images results.

We proposed a multi-scaled aligned image to test our model on the FER2013 dataset[7], which primarily contains images of faces without any background. This default configuration could introduce bias into the results if tested directly without any preprocessing on a single-staged model. We implemented an augmentation step to mitigate this issue and reduce potential bias. Specifically, the images in FER2013 were resized to multiple sizes by M x M, then placed onto a black background where the overall size matched the input size of the model. This augmentation process helps simulate more diverse real-world conditions.

The M x M resizing was treated as a hyperparameter, adjustable for benchmarking and optimization, enabling us to evaluate the model's performance across a range of image scales shown in **Fig 2**. This approach allowed us to assess not only the model's capability in classifying emotions and detecting faces. Face detection accuracy is calculated as the number of face images correctly detected is divided by the total number of images. This preprocessing step helped ensure that the model was robust when applied to real-world data with varying scale, backgrounds and image characteristics.

4. Experiment & Result

A. Experiment Detail

Our model is trained and evaluated using Exp-W datasets. Exp-W[8], [9][8], [9] the dataset that we used contains 7 classifications of human emotions. This dataset doesn't have a balanced number of images per class, All images were scaled to 640x640, and random crop, rotation, and recoloring were done as a preprocessing phase. Our model is evaluated with 10 epoch intervals. Facial alignment isn't required because our model will automatically predict bounding boxes and class simultaneously. Our models are trained and measured using PyTorch library and NVIDIA GeForce RTX 2080 Ti GPU.

Model	Class (mAP@0.5) %								Params	Inference Time
	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	All		
YOLOv6 Small	48.00	**12.20**	12.30	65.80	47.90	55.40	48.90	41.50	**18.50M**	**6.57 ms**
YOLOv6 Medium	**50.10**	12.00	11.70	65.60	**49.40**	**55.90**	48.60	41.90	34.81M	7.05 ms
YOLOv6 Large	49.20	11.30	**12.40**	**67.10**	48.90	55.60	**49.40**	**42.00**	59.54M	12.97 ms

Table 1. Performance of our model on Exp-W validation set.

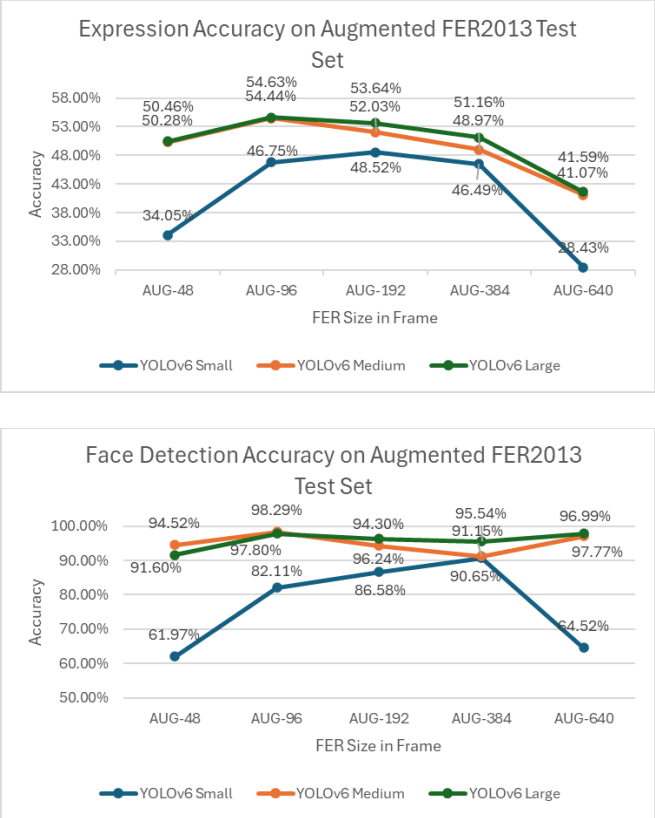

Figure 3. Our model's classification (top) and face detection (bottom) accuracy on augmented FER2013 test set with various

B. Model Performance

In **Table 1**, we evaluated our model using mean average precision(mAP) metrics. We used to intersect of union (IoU) threshold of 0.5. The number of images per class and number of predicted in that class are also present in the table. It also presents performance on Exp-W Validation with various YOLOv6 model sizes. The results show that the YOLOv6 models perform reasonably well in facial emotion recognition without face alignment preprocessing, with the "Happy" and "Surprise" classes achieving the highest accuracy across all model sizes. YOLOv6 Large has the best overall mAP (42.00%) but also the slowest inference time (12.97 ms), while YOLOv6 Small offers the fastest inference time (6.57 ms) with a slightly lower overall mAP (41.50%). Despite the moderate performance for most emotions, the models struggle with recognizing "Disgust" and "Fear," which have the lowest

Model	Inference Time	Accuracy
VGG-Net + MTCNN	55ms	**73.28%**
Inception + MTCNN	49ms	71.60%
ResNet + MTCNN	42ms	72.40%
YOLOv6 Small-unweighted class-640	**6.57 ms**	42.07%
YOLOv6 Small-unweighted class-Mean	**6.57 ms**	47.08%
YOLOv6 Small-640(Ours)	**6.57 ms**	28.43%
YOLOv6 Small-Mean(Ours)	**6.57 ms**	40.85%
YOLOv6 Medium-640(Ours)	7.05 ms	41.07%
YOLOv6 Medium-Mean(Ours)	7.05 ms	49.36%
YOLOv6 Large-640(Ours)	12.97 ms	41.59%
YOLOv6 Large-Mean(Ours)	12.97 ms	50.29%

Table 2. Model comparison on FER2013.

mAP values (~12%). For real-time applications, YOLOv6 Small offers the best trade-off between speed and accuracy, while YOLOv6 Large is more suitable for scenarios where accuracy is the priority.

To compare with other methods, we tested our model using the FER2013 dataset, a common benchmark for facial emotion recognition models. **Table 2** presents a comparison of our model with state-of-the-art (SOTA) models. Our models demonstrate significantly faster inference times, with speeds as low as 6.57 ms, compared to traditional models like VGG-Net and ResNet, which have inference times of 55 ms and 49 ms, respectively. This speed advantage is because our models bypass the face alignment step using MTCNN, streamlining the process for real-time applications. However, this speed comes at the cost of lower accuracy, with our models achieving accuracies ranging from 28.43% to 47.08%, highlighting a trade-off between speed and performance. In **Table 2**, "640" refers to the image resizing during augmentation, and "Mean" represents the average accuracy for different size parameter settings shown in **Fig 3**. "640" indicates that we used the original FER2013 images, as our model's input is 640x640, meaning no padding was applied. **Fig 3** highlights the model's ability to classify emotions across different face scales. Notably, there is a significant difference in performance between the YOLOv6 Small and Medium

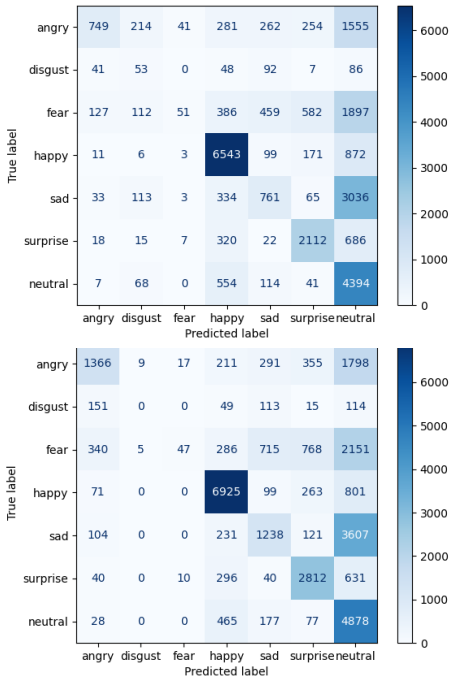

Figure 4. Confusion matrices between YOLOv6 small without weighted class loss(top) and with weighted class loss(bottom).

models, with the "Small" model struggling to classify small-scale and full-scale facial images. From Medium to Large, there is no significant improvement. Our model struggles particularly with detecting "Disgust" and "Fear" classes due to class imbalance in the Exp-W dataset, but it performs well in detecting the "Happy" emotion.

C. Weighted class loss effectivity

Fig 4 presents the comparison between the unweighted and weighted class loss models shows that the weighted model improves detection of minority classes ('disgust' and 'fear') by reducing false negatives and increasing true positives. While the overall accuracy is similar, the weighted model has However w. Overall, the weighted class loss better handles class imbalances, leading to improved F1-scores for the minority classes. However, as shown in **Table 2**, our weighted class loss introduces a trade-off between improving the accuracy of minority classes and maintaining overall accuracy. This trade-off could be addressed in future experiments by adding more sample images for the "Disgust" and "Fear" classes.

5. Conclusion

In conclusion, our model demonstrated strong capability in detecting facial bounding boxes and serves as a robust single-stage model, showcasing the potential of YOLOv6 for facial emotion recognition. Implementing weighted class loss effectively improved the accuracy of imbalanced classes, particularly for emotions like "Disgust." However, this improvement comes at the cost of a slight decrease in overall accuracy. Future work could focus on further addressing this trade-off to enhance both class-specific and overall performance.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government(MSIT).

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2023-00219107).

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program(IITP-2023-RS-2022-00156287) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation).

References

[1] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," 2001.

[2] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Trans Affect Comput*, vol. 13, no. 3, pp. 1195–1215, 2022, doi: 10.1109/TAFFC.2020.2981446.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," Jun. 2015, [Online]. Available: http://arxiv.org/abs/1506.02640

[4] C. Li *et al.*, "YOLOv6 v3.0: A Full-Scale Reloading," Jan. 2023, [Online]. Available: http://arxiv.org/abs/2301.05586

[5] Chaitanya, S. Sarath, Malavika, Prasanna, and Karthik, "Human Emotions Recognition from Thermal Images using Yolo Algorithm," in *2020 International Conference on Communication and Signal Processing (ICCSP)*, IEEE, Jul. 2020, pp. 1139–1142. doi: 10.1109/ICCSP48568.2020.9182148.

[6] R. Huang, J. Pedoeem, and C. Chen, "YOLO-LITE: A Real-Time Object Detection Algorithm Optimized for Non-GPU Computers," *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pp. 2503–2510, 2019, doi: 10.1109/BigData.2018.8621865.

[7] L. Wang, S. Xu, X. Wang, and Q. Zhu, "Eavesdrop the Composition Proportion of Training Labels in Federated Learning," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.06044

[8] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning Social Relation Traits from Face Images," Sep. 2015, [Online]. Available: http://arxiv.org/abs/1509.03936

[9] Z. Zhang *et al.*, "From Facial Expression Recognition to Interpersonal Relation Prediction." [Online]. Available: www.rferl.org