Artificial Intelligence for Affective Computing

Student Name: Huiqi Zou Supervisor: Dr. Linqi Song

Department of Computer Science
香港城市大學
City University of Hong Kong

INTRODUCTION

Facial expression recognition (FER) decodes and estimates the intensity of specific emotions. Despite the high accuracy achieved by previous deep neural networks in this task, challenges remain in terms of computational complexity. Besides, few existing FER systems interact with multiple users simultaneously, restricting their potential for broader application scenarios.

To address these challenges, this project presents a lightweight system that accurately identifies the emotions of multiple people and maintains acceptable performance in real-world conditions.

METHODOLOGY

RetinaFace is a lightweight face detector based on feature pyramid network. Using a multi-task loss function to generate a face score, a face box, five facial landmarks and 3D face vertices, it achieves state-of-the-art and real-time performance on a single CPU core.

Swin Transformer is a versatile backbone that outperforms CNN-based benchmarks on various vision tasks and reduces the computational burden of Vision Transformer (ViT). SwinV2 addresses the inefficiency of Swin Transformer when transforming the pre-trained model across different window resolutions. The adopted model is pre-trained on FER2013.

The **fine-grained manifold distillation** method transfers transformer-based teacher information to students at batch and patch levels. The loss function has two terms: a **knowledge distillation (KD) loss** and an **MD loss**.

$$L_{kd} = (1-\lambda)H_{CE}(f_s(x), y) + \lambda\tau^2 H_{KL}(f_s(x)/\tau, f_t(x)/\tau)$$

where H_{CE} is the cross-entropy, H_{KL} is the Kullback-Leibler divergence function, τ is a label smoothing hyper-parameter, and λ is a balancing hyper-parameter.

$$L_{mf} = \| M(\psi(F_S)) - M(\psi(F_T)) \|^2_{F'}$$

$$M(\psi(F_S)) = (\psi(F_S))(\psi(F_S))^T, M(\psi(F_S)) = (\psi(F_T))(\psi(F_T))^T$$

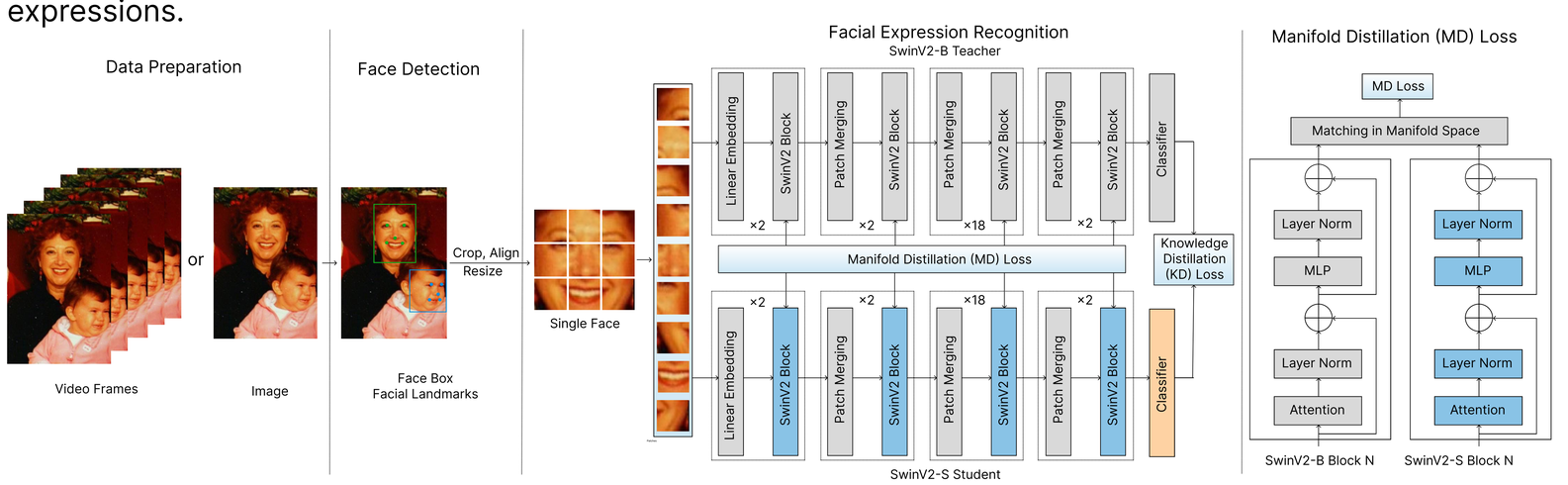
where F_S and F_T are features of the student and the teacher layers, and ψ is a reshaping operation $\mathbb{R}^{D_1 \times D_2 \times D_3} \rightarrow \mathbb{R}^{D_1 D_2 \times D_3}$.

KD utilizes the predicted label of a large high-performance model to guide a smaller one. The **MD loss** encourages each student block to generate a manifold relation map that shares the same patch-level structure with its paired teacher.

Therefore, the student can inherit the knowledge hidden in the predicted label and intermediate features.

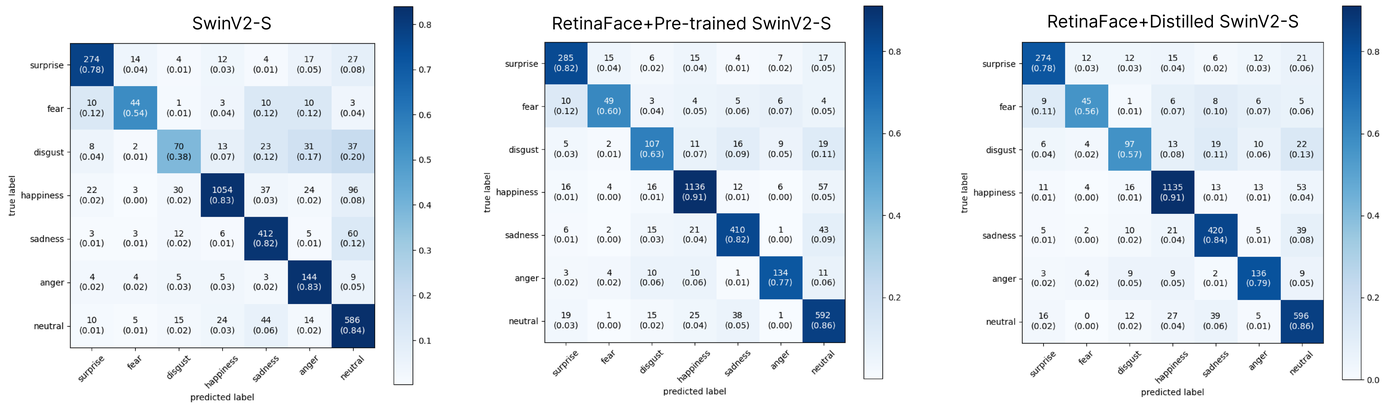
FRAMEWORK

The system consists of two components: **a face detection module** and **a FER module**. To detect faces in extracted video frames or a single image, RetinaFace is utilized to predict the facial bounding boxes and align them with detected facial landmarks. This approach allows the system to differentiate between multiple faces in a single image. For FER, the system employs Swin Transformer V2 (SwinV2) which has been trained with manifold distillation (MD) techniques. This model is adapted to FER, allowing it to accurately recognize facial expressions.



RESULTS

To validate the effectiveness of RetinaFace, pre-training and MD, ablation studies are conducted on RAF-DB using two metrics: **accuracy** and **unweighted accuracy (UAR)**.



The results indicate that **RetinaFace** and **pre-training** improves accuracy and UAR by 4.20% and 5.57%, respectively. RetinaFace enables the system to focus on essential facial features, while pre-training facilitates the transfer of prerequisite knowledge across datasets. However, the imbalanced sample distribution leads to a discrepancy between the two metrics.

For **MD**, when maintaining model size, the SwinV2-S learned from Swin-B with pre-training outperforms the original model by 3.88% and 4.14% in terms of the two metrics. This also indicates that directly transferring knowledge through pre-training is more efficient than using distillation.

References:
[1] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, 'RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2020.
[2] Z. Liu et al., 'Swin Transformer V2: Scaling Up Capacity and Resolution', in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11999–12009. doi: 10.1109/CVPR52688.2022.01170.
[3] Z. Hao et al., 'Learning Efficient Vision Transformers via Fine-Grained Manifold Distillation', in Advances in Neural Information Processing Systems, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Curran Associates, Inc., 2022.
[4] I. J. Goodfellow et al., 'Challenges in Representation Learning: A Report on Three Machine Learning Contests', in Neural Information Processing, M. Lee, A. Hirose, Z.-G. Hou, and R. M. Kil, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 117–124.
[5] S. Li, W. Deng, and J. Du, 'Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild', in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2584–2593. doi: 10.1109/CVPR.2017.277.