Report: Decoupling Identity Confounders for Enhanced Facial Expression Recognition

Rahul Kumar	Deepak Kumar	Anjan Das	Aditya Kumar
IIT Kanpur, India	IIT Kanpur, India	IIT Kanpur, India	IIT Kanpur, India
krahul23@iitk.ac.in	deepakkr22@iitk.ac.in	anjand23@iitk.ac.in	adityakv23@iitk.ac.in
230828	220332	230149	230069

Abstract

Facial Expression Recognition (FER) is often hindered by identity-related features that overlap with expression information, making it difficult to learn robust and generalizable models. This report presents a detailed summary and analysis of the DICE-FER model, which introduces a novel information-theoretic approach to disentangle expression and identity representations. By estimating and optimizing mutual information between features, DICE-FER avoids reliance on identity labels or synthetic image generation, enabling efficient and scalable FER.

1 Introduction

Facial expressions serve as a rich channel of nonverbal communication and are central to applications such as human-computer interaction and mental health assessment. However, traditional FER systems struggle due to two key challenges:

- **Subtle inter-class differences**: For example, a smile vs. a smirk.
- **Significant intra-class variability**: Different individuals may express happiness in visually distinct ways.

These challenges are exacerbated by identity features such as bone structure or age, which confound expression recognition. The DICE-FER framework tackles this by disentangling identity and expression representations using mutual information estimation.

2 Theoretical Background

2.1 Mutual Information (MI)

Mutual Information measures the amount of information shared between two variables. Mathematically, for random variables M and Z with joint distribution $p(m, z)$ and marginal distributions $p(m)$ and $p(z)$, the mutual information is:

$$I(M, Z) = \int_M \int_Z p(m, z) \log \left(\frac{p(m, z)}{p(m)p(z)} \right) dm\, dz$$

This can be interpreted as the Kullback-Leibler (KL) divergence between the joint distribution and the product of marginals:

$$I(M, Z) = D_{\mathrm{KL}}(p(m, z)||p(m)p(z))$$

In DICE-FER, mutual information is used to:

- **Maximize** shared information between images and their expression representations.
- **Minimize** shared information between expression and identity representations.

To estimate MI robustly, DICE-FER uses the Donsker-Varadhan representation:

$$\hat{I}_{DV,\theta}(M, Z) = E_{p(m,z)}[U_\theta(m, z)] - \log E_{p(m)p(z)}[e^{U_\theta(m,z)}]$$

where U_θ is a statistics network trained to approximate MI.

3 Proposed Method

3.1 Overall Framework

Given image pairs (M, N) with the same expression but different identities, DICE-FER learns two distinct feature spaces:

- Expression Representation: E_M, E_N (shared across M and N)
- Identity Representation: I_M, I_N (exclusive to each image)

3.2 Stage 1: Learning Expression Representations

To ensure E_M and E_N contain only expression-related information:

- Mutual information between M and E_N (and vice versa) is maximized.
- L1 distance between E_M and E_N is minimized:

  $$L_1 = E[|E_M - E_N|]$$

  This enforces consistency between expressions.
- Combined objective:

  $$L_{\mathrm{exp}} = L_{\mathrm{exp}}^{MI} - \delta \cdot L_1$$

  where δ balances the MI and similarity terms.

3.3 Stage 2: Learning Identity Representations

Once expression features are disentangled, identity features are extracted to capture the remaining information:

- Mutual Information between M and full representation $T_M = [E_M, I_M]$ is maximized.
- Adversarial objective minimizes $I(E_M, I_M)$ to prevent leakage of identity into expression and vice versa:

$$L_{\text{adv}} = E[\log D(E_M, I_M)] + E[\log(1 - D(E_M, \text{shuffled } I_M))]$$

Final loss for identity learning:

$$L_{\text{id}} = L_{\text{id}}^{MI} - \zeta_{\text{adv}}(L_{\text{adv}}^{M} + L_{\text{adv}}^{N})$$

4 Experiments and Results

4.1 Datasets

DICE-FER is evaluated on four datasets:

- **RAF-DB**: Real-world images with large variability.

4.2 Implementation Details

ResNet-18 encoders pretrained on CASIA-WebFace are used. Expression and identity embeddings are 64-dimensional. Mutual information is estimated with MINE-based networks. Key loss parameters are $\mu = 0.5$, $\nu = 1.0$, $\delta = 0.1$, $\zeta_{\text{adv}} = 0.025$.

4.3 Results and Analysis

DICE-FER outperforms prior methods on MIG scores and classification accuracy. For example:

Method	RAF-DB
TDGAN	0.365
AGILE	0.400
DICE-FER	**0.450**

Table 1: Mutual Information Gap (MIG) comparison across datasets. Higher is better.

Table 2: TRAIN Confusion Matrix Summary(Actual vs Predicted)

	Predicted						
Actual	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral
Surprise	850	45	23	12	15	8	47
Fear	32	720	28	15	18	12	25
Disgust	18	25	680	22	15	18	22
Happiness	15	12	18	920	25	15	15
Sadness	22	18	15	28	750	22	25
Anger	12	15	22	18	25	680	20
Neutral	35	28	25	22	28	18	820

Table 3: TEST Confusion Matrix Summary (Actual vs Predicted)

	Predicted						
Actual	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral
Surprise	269	12	5	6	9	9	19
Fear	9	41	3	5	10	5	1
Disgust	5	1	80	22	14	11	27
Happiness	9	5	7	1107	7	9	41
Sadness	4	3	19	32	362	10	48
Anger	3	3	9	12	2	122	11
Neutral	15	2	22	29	56	6	550

Table 4: Class-wise Performance Metrics for DICE-FER Model

Expression	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Support**
Happiness	0.934	0.913	0.934	0.923	1185
Surprise	0.818	0.857	0.818	0.837	329
Neutral	0.809	0.789	0.809	0.799	680
Sadness	0.757	0.787	0.757	0.772	478
Anger	0.753	0.709	0.753	0.731	162
Fear	0.554	0.612	0.554	0.582	74
Disgust	0.500	0.552	0.500	0.525	160
				Overall Accuracy:	0.825 (82.5%)

5 Conclusion

DICE-FER offers an efficient, label-free framework for facial expression recognition by explicitly decoupling identity and expression using mutual information principles. It demonstrates superior generalization and scalability across diverse datasets. The use of mutual information not only enables robust disentanglement but also eliminates the need for additional labels or synthetic data.

References

1. Mohd Aquib, Nishchal K. Verma, M. Jaleel Akhtar, "Decoupling Identity Confounders for Enhanced Facial Expression Recognition: An Information-Theoretic Approach," *CVPR Workshops*, 2025.

2. Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2852–2861, 2017.

3. Mohamed Ishmael Belghazi et al., "Mutual Information Neural Estimation," *International Conference on Machine Learning (ICML)*, 2018.