Cross-Cultural Deepfake Detection: A Comparative Study Using TWHD and FaceForensics++

溫翎媛 [1] 紀懿佳 [2] 林詠章 [3]
國立中興大學資訊管理學系 [1`2`3]
wenjolin0124@gmail.com[1]
annacat1211@gmail.com[2]
iclin@nchu.edu.tw[3]

Abstract

As deepfake technologies advance, the widespread dissemination of manipulated visual content has emerged as a serious digital threat to societies worldwide. To address this challenge, we conducted experiments using deep learning models to detect deepfake facial images. In particular, we adopted two culturally distinct datasets: the Taiwan High-quality Deepfake Dataset (TWHD), representing East Asian features, and FaceForensics++ (FF++), a widely used dataset in Western contexts. These datasets were used to train and evaluate three representative neural network models—Xception, MobileNetV3, and Vision Transformer (ViT)—to investigate their effectiveness and cross-cultural generalizability. Our findings highlight the promising performance of TWHD in deepfake detection within its cultural domain and reveal significant domain shifts when models are applied across datasets. This study contributes to the development of culturally robust AI systems for combatting the growing global deepfake crisis.

Keywords: Deepfakes, Deep Learning, Cross-Cultural Datasets, Image Forensics, Model Generalization

1. Introduction

With the rapid development of deepfake technology, the identification of fake digital content has become increasingly difficult, posing a significant threat to public trust and information security. However, existing deepfake detection techniques largely rely on publicly available Western face datasets, such as FaceForensics++ (FF++). The facial features in these datasets predominantly come from Western populations, which may limit the models' ability to generalize across different cultural backgrounds. Based on this, this study aims to compare deepfake detection models trained on the Taiwan High-quality Deepfake Dataset (TWHD) and the FF++ dataset, and explore how regional differences in dataset sources affect model performance. We selected three representative deep learning models: Xception, MobileNetV3, and Vision Transformer (ViT), and through cross-dataset testing, we delve into the challenges and potential directions for improvement in their application across multicultural contexts.

2. Related Works

Deepfake detection aims to identify highly realistic forged content, particularly facial videos, which are difficult to distinguish with the naked eye, making accurate and automated detection techniques essential. Common methods in the literature include GAN-based models[10], such as enhanced versions of DCGAN, which combine data augmentation and label smoothing to improve discriminative power. Another approach is multi-domain feature extraction[12], like the Mixformer model, which integrates spatial, noise, and frequency domain features to boost accuracy.

Hybrid architecture[9] is also gaining increasing attention. For example, Inception Transformer combines the strengths of CNNs and Transformers, while the CSTAN model integrates Channel-Spatial-Triplet attention and omni-dimensional dynamic convolution (ODConv) to enhance generalization and detail recognition. Dual-stream networks and multi-task learning further improve detection stability.

However, most existing models still struggle with generalization when facing unknown forgery techniques and heterogeneous data. Common datasets such as DFDC, FF++, and Celeb-DF offer diversity but are predominantly composed of Western faces, limiting their adaptability to Asian faces. Therefore, this study uses the Taiwan High-quality Deepfake Dataset (TWHD), focused on Asian facial features, to promote the practical development of deepfake detection technologies in multicultural contexts.

3. Methods

To evaluate deepfake detection performance across different cultural contexts, we adopted two datasets from distinct regions and trained three representative deep learning models on each. This section outlines the datasets and models used, the data preprocessing and splitting strategy, as well as the training procedures and evaluation metrics.

3.1 Datasets

We selected two primary datasets as the foundation of our experiments:

3.1.1 Taiwan High-quality Deepfake Dataset