Abstract

Images and videos play a vital role in communication and information dissemination, with extensive applications in fields such as the film industry, advertising, journalism, surveillance, and law enforcement. However, the increasing accessibility of advanced digital devices and editing tools has made it effortless to manipulate or tamper with visual content, even without technical expertise. As a result of rapid technological advancements, the visual quality of synthetic media has become nearly indistinguishable from real content, making manual detection extremely challenging. This growing threat undermines the credibility of visual information and highlights the urgent need for developing robust and reliable forgery detection techniques to prevent the spread of misinformation and maintain public trust.

Digital forgeries refer to the intentional alteration or fabrication of visual content to misrepresent reality. Earlier, manual editing tools such as Adobe Photoshop were used for basic manipulations including cropping, cloning, and splicing, which required time and skill. In videos, forgeries traditionally involved frame insertion, deletion, or reordering to fabricate events. The emergence of Artificial Intelligence, particularly Generative Adversarial Networks (GANs), has revolutionized forgery creation, enabling the automatic synthesis of highly realistic faces and scenes with minimal human intervention. These AI-generated manipulations, commonly known as deepfakes, are extremely difficult to distinguish from authentic content and pose serious challenges to media authenticity, personal privacy, and social stability. In response, researchers have developed a range of detection methods based on machine learning, deep learning, attention mechanisms, and multi-modal feature analysis. Nevertheless, issues such as adversarial robustness, cross-domain generalization, and real-time applicability remain open challenges. This thesis aims to explore, analyze, and propose advanced techniques for detecting image and video forgeries, thereby contributing to the ongoing global effort to safeguard the integrity of digital visual media.

Image forgery can be broadly categorized into three main types: copy-

move, splicing, and retouching. In copy-move forgery, a portion of an image is duplicated within the same image to conceal or replicate objects, making detection difficult due to consistent lighting and texture. Splicing combines content from multiple images to fabricate a scene that never existed, whereas retouching subtly alters image attributes for enhancement or deception. Similarly, video forgeries can be classified into intra-frame, inter-frame, and spatio-temporal tampering. Intra-frame forgery modifies visual content within individual frames, inter-frame forgery manipulates the temporal sequence by inserting, deleting, or reordering frames, and spatio-temporal forgery combines both spatial and temporal alterations, producing highly convincing results such as deepfakes.

This thesis first presents an extensive review of existing forgery detection techniques, encompassing traditional handcrafted-feature-based, statistical, and deep learning-based approaches. Conventional methods analyze inconsistencies in pixels, compression artifacts, or noise residuals, while modern convolutional and transformer-based architectures extract high-level semantic representations. Hybrid approaches integrating both handcrafted and deep features have also shown promise in enhancing robustness and interpretability. Based on this analysis, the study emphasizes three significant categories of forgery that have major societal implications: copy-move forgery in images, deepfake manipulations in images and videos, and inter-frame video forgeries involving frame duplication and deletion.

A customized MultiResUNet architecture is proposed for detecting copy-move forgeries, designed to leverage multi-resolution feature extraction for improved detection accuracy. The model captures both fine-grained textures and high-level contextual information essential for identifying duplicated regions. To enhance computational efficiency, standard convolutional layers are replaced with separable convolutions, reducing trainable parameters from 7.27 million to 3.26 million without sacrificing performance. Residual connections are incorporated to mitigate the vanishing gradient problem and preserve spatial details during decoding. The method achieves precise localization of manipulated areas and is evaluated on four benchmark

datasets: CoMoFoD, COVERAGE, CASIA TIDE v2.0, and MICC-F600, which include diverse post-processing conditions such as compression, illumination variations, and low contrast. The proposed model attains F1-scores and accuracies of 84.51% and 99.34% on CoMoFoD, 79.81% and 98.57% on COVERAGE, and 93.96% and 96.28% on MICC-F600. Even on the challenging CASIA v2 dataset, it achieves a competitive F1-score of 74.97%. These results outperform traditional block-matching and keypoint-based techniques, confirming the efficiency and practical applicability of the proposed lightweight model for real-world forensic analysis.

In the domain of deepfake detection, the study investigates both global and local facial inconsistencies to improve detection accuracy in image and video deepfakes. Unlike traditional approaches that focus only on specific facial regions such as the eyes or mouth, this work integrates global facial context with localized features from facial patches. An initial experiment evaluates the influence of different color spaces RGB, HSV, and YCbCr using separate Mesoinception-4 models, and demonstrates that fusing color-space-specific features improves classification performance. Building upon this, a refined model employs cropped face images processed through the Xception network for global feature extraction, combined with a soft attention mechanism that highlights local inconsistencies caused by manipulation. The attention-weighted features are classified using dense layers, enhancing sensitivity to subtle artifacts. Further improvements are achieved through an ensemble strategy that combines features from multiple deep pretrained models and refines them using ranking-based feature selection, followed by multilayer perceptron classification, resulting in improved robustness across datasets.

To further strengthen deepfake detection, a novel architecture named ViXNet is proposed, combining patch-wise self-attention and Vision Transformer (ViT) mechanisms for local inconsistency modeling with Xception-based global feature extraction. Facial images are divided into patches, where self-attention assigns adaptive weights to manipulated regions. These local representations are integrated with global features extracted by Xcep-

tion and passed through a dense classifier for final prediction. ViXNet achieves AUC scores of 98.57%, 99.26%, and 98.93% on FaceForensics++, Celeb-DF (V2), and DFID datasets in intra-dataset evaluations, and 83.60%, 74.78%, and 75.13% respectively in inter-dataset tests. On the DFDC dataset, it attains an AUC of 86.32% and an F1-score of 79.06%, demonstrating strong generalization and robustness against unseen manipulations.

An additional approach is proposed to combine deep and handcrafted features to capture both global and local inconsistencies. Here, the Transformer-based local feature extractor is replaced with a handcrafted feature pipeline using DAISY descriptors extracted from 25 facial patches. Each face generates a 2600-dimensional descriptor vector, and Euclidean distances between patches yield a 300-dimensional representation of intra-patch inconsistencies. These handcrafted features are concatenated with deep Xception features to form a unified vector, which is optimized using a hybrid hierarchical feature selection method based on Grey Wolf Optimization (GWO) and Vortex Search (VS). This reduces redundancy while preserving discriminative information. Classification is performed using a Support Vector Machine (SVM) with an RBF kernel, achieving AUC scores of 99.35% on Celeb-DF (V2), 99.16% on FaceForensics++, and 85.67% on DFDC, using only 10–13% of the original features. The method outperforms several state-of-the-art approaches, offering a compact and efficient solution for deepfake detection.

For video forgery detection, two primary inter-frame manipulations are considered: frame duplication and frame deletion. To detect frame duplication, an ensemble-based method is proposed that integrates three individual detectors through a majority voting strategy. Each detector extracts distinct texture-based features using two variants of the Gray Level Co-occurrence Matrix (GLCM) and Local Binary Pattern (LBP). Frame features are sorted lexicographically to identify similar frame groups, enabling the detection of duplicated sequences. A post-processing step merges partial detections to handle fragmented duplications. The method effectively detects both single and multiple duplication instances and is resilient to common post-processing operations such as noise addition, blur, and brightness variations. Experi-

ments on an in-house dataset of 300 videos show superior accuracy compared to existing techniques.

To further enhance duplication detection, a statistical method is developed using the Structural Similarity Index Measure (SSIM) and Longest Common Subsequence Comparison (LCSC) algorithm. SSIM values are computed between consecutive frames to identify temporal anomalies, followed by shot segmentation to partition the video into coherent segments. Within each segment, LCSC identifies repeated frame patterns corresponding to duplicated sequences. This two-stage process improves precision and reduces false detections, performing robustly in both static and dynamic scenes. Evaluation on videos from Urban Tracker, DERF, and REWIND datasets demonstrates an average accuracy of 98.90%, validating its robustness for practical applications.

Frame deletion detection is addressed using a Mean Squared Error (MSE)-based approach. Consecutive frame differences are converted into a spatial representation, termed an MSE curve, which is input to a modified MesoNet CNN to classify videos as real or tampered. Upon detection, a second-order derivative analysis of the MSE curve accurately localizes deleted frame positions. The method performs reliably under distortions such as blur, noise, and illumination variations, achieving an average accuracy of 96.00% for deletion detection and 98.30% for localization on a custom dataset.

In summary, this thesis presents a comprehensive framework for detecting forgeries in both images and videos. The proposed methods combine advanced feature extraction, statistical modeling, and optimization-based learning to address various types of manipulation, including copy-move forgeries, deepfakes, and inter-frame tampering. The approaches are computationally efficient, highly accurate, and robust against common post-processing operations, making them suitable for real-world applications such as digital forensics, surveillance, and media verification. Collectively, these contributions form a strong foundation for future research aimed at preserving the authenticity and reliability of visual media in increasingly complex digital environments.