Detection of Augmented Reality Images with Facial Accessories

Natchaya Yimtanom
Faculty of Information Technology
Thai-Nichi Institute of Technology
Bangkok, Thailand
2463110110@tni.ac.th

Saprangsit Mruetusatorn*
Faculty of Information Technology
Thai-Nichi Institute of Technology
Bangkok, Thailand
email : saprangsit@tni.ac.th
*Corresponding author

***Abstract*—Technologies such as Artificial Intelligence (AI) and Augmented Reality (AR) have become deeply integrated into daily life. These technologies provide benefits in media creation, but also have risks regarding data forgery, such as deceptive imagery. The increasing realism of AR accessories makes it challenging to differentiate between digital enhancements and real accessories. This ambiguity leads to significant confusion and misinterpretation of digital content.**

This research presents various methods to distinguish between images of real accessories and those enhanced with AR. We propose a rule-based approach that identifies custom decision rules and criteria by converting images into the HSV color model to analyze luminance and calculate values based on defined features, such as Mean Absolute Deviation or Standard Deviation. Furthermore, we also use existing Machine Learning (Deep Learning) models to compare the performance of each methodology using Confusion Matrix. The custom dataset was made with 2,000 images, including real and AR versions of both types of accessory images. The results demonstrate that while the rule-based model is effective, achieving 75% accuracy for both glasses and facial masks, it confirms that the selected HSV features are sufficiently robust for distinguishing AR accessories. Nevertheless, this performance is surpassed by machine learning approaches, where the optimal model varies by accessory type. The Xception model achieved the highest accuracy for glasses at 88% and outperformed others for facial masks with a peak accuracy of 98%. The results suggest that Xception's depthwise separable convolutions are particularly well-suited for detecting the artifacts inherent in AR-modified images compared to other models.

Keywords—augmented reality, image forensics, deep learning, HSV color model, rule-based approach

I. INTRODUCTION

The internet is an important tool for searching and sharing information, alongside the fast-paced development of AI. Nowadays, AI can do much more, including the alteration of data such as facial images and videos. While there are many tools developed to detect AI-manipulated data, other types of image manipulation like Augmented Reality (AR) are also becoming more realistic. However, there are still very few tools specifically designed to detect these AR changes. Even when using current AI detectors or Large Language Models (LLMs), the results are often inconsistent and can be incorrect across different platforms. In the context of biometric security, detecting AR accessories is particularly critical because they specifically obscure or distort key facial landmarks. These subtle digital overlays can mislead facial recognition systems, creating significant vulnerability.

In this research, we propose a method for detecting images manipulated with Augmented Reality (AR), specifically focusing on facial images with added accessories such as glasses or facial masks. We introduce a rule-based model that utilizes custom detection rules by analyzing the brightness values of the accessories within the image. To achieve this, the HSV color model is employed, allowing for the inspection of brightness independent of color information, unlike the RGB model where the three channels are highly correlated and sensitive to lighting changes [1]. This approach evaluates both the overall brightness and the distribution of light across pixels to compare real physical accessories with AR-generated ones. Furthermore, we employ a transfer learning approach by utilizing pre-trained deep learning models [2]-[4] to demonstrate the effectiveness of machine learning when applied to case-specific datasets. Finally, we provide a comparative analysis of detection accuracy across different model types, using the confusion matrix to evaluate and showcase the performance and reliability of each approach.

II. RELATED WORKS

This section reviews related works to identify similarities and distinctions with this study. A key divergence is the feature extraction approach; while existing methods vary, this research focuses on the HSV color model to establish criteria for a rule-based classification model in AR forensics.

A. Augmented Reality (AR)

The use of Augmented Reality (AR) with human faces requires a high degree of accuracy in face detection. D. Kang and L. Ma [5] demonstrated this by using AR technology for 3D displays, proving that a system can still track facial positions effectively even when parts of the face are covered by accessories. In this study, we utilize MTCNN to detect the face as a first step in identifying AR facial accessories. Both projects share a common need to accurately locate a face despite the presence of physical or digital overlays. This shows that reliable face detection is a critical foundation for both the creation and the forensic analysis of AR images.

B. Digital Image Processing

T. Kusnandar et al. [1] presented an enhanced color selection technique within the HSV color space by utilizing the Heaviside step function to optimize the identification of specific target hues. Their results showed that isolating the hue

channel from intensity and saturation ensures more stable color values and improves computational efficiency. This research supports our methodology of employing the HSV color model to analyze luminance independently of color data, allowing for the extraction of brightness information without affecting other color components. This provides a robust foundation for AR accessory detection to remain stable across inconsistent lighting conditions.

P. Duszejko et al. [6] conducted a comprehensive review of passive and active digital image forensic methods, focusing on the evolution of deep learning-based detection. Their findings demonstrated that deep learning models outperform traditional statistical approaches by autonomously identifying intricate patterns and subtle pixel-level inconsistencies that indicate digital tampering. This study is relevant to our study as it validates the necessity of using advanced models like Xception to capture fine-grained AR artifacts, which often exceed the detection capabilities of rule-based models.

C. *Rule-based Model*

S. Polat et al. [7] presented the development of classification rules which are defined by analyzed image features. By examining the spectral curve characteristics and intensity levels across different intervals, they developed a method to distinguish between plastic and plant image data based on spectral shapes. This approach related with our study in the use of image data analysis to establish rule-based systems for classifying two categories of data.

D. *Machine Learning Model*

S. Yadav and A. Kumar [8] compared conventional forensic methods with deep learning for image manipulation detection. Their findings highlighted that while traditional forensic rules provide high interpretability, they are often surpassed by deep learning models, such as CNNs, which can identify intricate pixel-level anomalies that manual rules fail to capture. This supports the integration of transfer learning in our study to address the complexity of AR-generated artifacts.

III. METHOD

In this study, we aim to distinguish between images with real facial accessories, as in glasses and facial masks, and those featuring AR accessories. We curated a new dataset by selecting images from existing datasets and analyzed them to identify features that can distinguish between images with real and AR accessories. These features were then used in the development of our rule-based model. Furthermore, we conducted experiments using existing machine learning models to compare the performance of each approach, using confusion matrix, in classifying both image types.

A. *Dataset preparation*

We compiled a custom dataset of 2,000 images by merging CelebA [9] and Real-Time-Medical-Mask-Detection [10] sources. The data features a symmetrical distribution of glasses and facial masks (1,000 each), equally split between real and AR-augmented versions (500 images per class). AR variations were generated by applying Snapchat filters, which consist of 2D digital overlays (such as glasses or facial masks) that are anchored to facial landmarks, onto CelebA [9] images. Detailed distribution is summarized in Table I.

TABLE I. DISTRIBUTION OF THE CUSTOM DATASET

Accessory type	Real accessories	AR accessories	Total
Glasses	500	500	1,000
Facial masks	500	500	1,000
Total	**1,000**	**1,000**	**2,000**

In this study, the established rule-based thresholds and trained machine learning models may not fully generalize AR accessories from other platforms, such as TikTok, or Instagram. Because different platforms utilize distinct rendering engines and digital blending techniques, which may result in different luminance patterns and texture characteristics in the HSV color model. Examples of the images in the dataset are shown in Fig. 1 and Fig. 2.

Fig. 1. Example of mask-wearing images in dataset. (A) Real mask-wearing images. (B) AR mask-wearing images.

Fig. 2. Example of glasses-wearing images in dataset. (A) Real glasses-wearing images. (B) AR glasses-wearing images.

From completed custom dataset, images were partitioned into functional subsets for model development. The data allocation was divided into training, validation, and test sets with a ratio of 7:1:2, respectively. This standard partitioning results in:

- Training Set (1,400 images): Used as the primary data source for both models, equally balanced between real and AR categories (700 each), with 350 glasses and 350 facial masks per class.
- Validation Set (200 images): Utilized exclusively by the machine learning models to monitor for overfitting. Equally balanced between real and AR categories (100 each), with 50 glasses and 50 facial masks per class.
- Test Set (400 images): A shared evaluation baseline used to compare the final performance of both the rule-based and machine learning approaches. Equally balanced between real and AR categories (200 each), with 100 glasses and 100 facial masks per class.

While balanced, the 2,000-image dataset may be insufficient for broad generalization of complex deep learning models. Additionally, relying exclusively on Snapchat filters restricts the findings to a single rendering engine, potentially limiting performance on other platforms like TikTok or Instagram. Future validation with larger, more diverse datasets is required to ensure reliability in real-world scenarios.

B. *Development of Rule-based Model*

The primary objective of the rule-based model is to provide an explainable method for distinguishing between real

and AR accessories. As illustrated in the development workflow (Fig. 3), the process begins by analyzing images from the dataset to identify key discriminative features. While different accessories like glasses and facial masks possess unique characteristics, this study focuses on establishing a universal set of rules to facilitate a straightforward performance comparison across accessory types.

To demonstrate data efficiency, we sampled a calibration subset of 120 images, consisting of 60 glasses and 60 facial masks, equally balanced between real and AR versions from the training set. These images were used to calculate optimal decision thresholds and conduct a preliminary evaluation to determine how effectively the defined criteria could differentiate between real and digital enhancements.

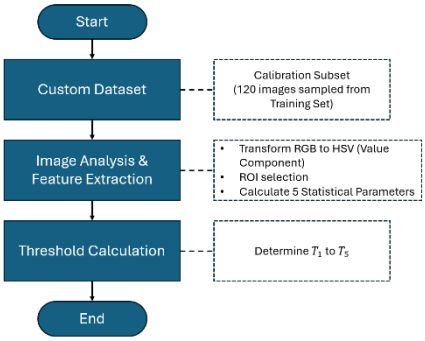

Fig. 3. Development workflow for the rule-based model

To extract the discriminative features, the input images are first transformed from the RGB to the HSV model. Following the transformation method described by T. Kusnandar et al. [1], the Value (V) component which represents the brightness intensity of the image, is computed as follows:

$$V = max(R, G, B) \quad (1)$$

While the Value component measurement is applied across all accessory types, the region of interest (ROI) selection varies by category. A 15×15 pixel window is centered on a selected coordinate within the accessory region. In the case of facial masks, all pixels within this window are sampled directly as they exclusively represent the mask surface. However, for glasses, the window often encompasses adjacent skin pixels. To isolate the accessory from the background skin, a color similarity test is performed based on the Euclidean distance metric [11]. Pixels exceeding a 15% color deviation from the reference point are excluded to generate a binary mask. This threshold was experimentally optimized for effective color segmentation and mask evaluation. This segmentation ensures that only the luminance properties of the accessory are extracted for analysis. Within this ROI, three distinct points are selected to ensure a broad spatial distribution of the sampled data, thereby avoiding localized bias. To analyze the intensity distribution within the ROI, five statistical parameters are calculated as follows:

- Mean Box Intensity: The average of the summed intensities across the three selected 15×15 regions.
- Mean Intensity: The arithmetic mean value of all individual pixels pooled from the combined sample areas.
- Mean Absolute Deviation (Mean AD): The average distance between each pixel and the mean, providing a robust measure of variability against outliers.
- Median Absolute Deviation (Median AD): The median of absolute deviations from the median, capturing the core brightness consistency while ignoring anomalies.
- Standard Deviation (SD): The measure of intensity dispersion relative to the mean, indicating the overall contrast and texture variation.

These values serve as the primary features for the classification process. Upon deriving the five statistical parameters, each metric was calculated across a dataset of 120 sample images. This analysis was conducted to identify distinct patterns between both accessories. The rule-based model employs five statistical parameters calculated from the ROI to identify distinct patterns between real and AR accessories. While the Value (V) component is consistently measured across all types, the classification criteria and optimal thresholds vary depending on the accessory category. The consolidated classification rules and calibrated threshold values (T_1 to T_5) are summarized in Table II.

TABLE II. CONSOLIDATED CLASSIFICATION THRESHOLD VALUES.

Statistical Parameters	Symbol	AR Condition (Glasses)	AR Condition (Facial Masks)
Mean Box Intensity	T_1	$V \leq T_1$	$V > T_1$
Mean Intensity	T_2	$V \leq T_2$	$V > T_2$
Mean AD	T_3	$V \leq T_3$	$V \leq T_3$
Median AD	T_4	$V \leq T_4$	$V \leq T_4$
Standard Deviation	T_5	$V \leq T_5$	$V \leq T_5$

In this study, texture analysis and edge detection were excluded to maintain a universal set of criteria across different accessory types. Given that glasses and facial masks possess different structural characteristics, applying a singular texture or edge-based rule would likely lead to inconsistent results.

C. Machine Learning Model Training and Fine-tuning

This study leverages transfer learning to achieve high-performance classification between real and AR accessories. The training workflow is illustrated in Fig. 4. Consistent with the dataset partitioning described in Section III.A, the deep learning models utilize the Training Set and Validation Set to perform feature extraction. This contrasts with the rule-based approach, which relies on a smaller calibration subset.

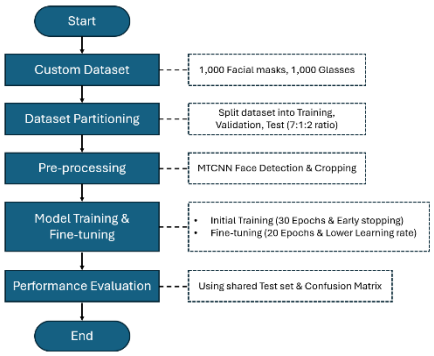

Fig. 4. Workflow for training and fine-tuning machine learning models

Each image was cropped to encompass the entire facial structure. This preprocessing step ensures that all images are standardized in size and maintain a consistent ROI across the entire dataset. Then, three machine learning (deep learning) models were selected to represent different computational priorities: Xception [2] for its ability in extracting fine-grained spatial features, MobileNetV3 [3] for its optimized performance on resource-constrained devices, and EfficientNetV2S [4] for its balanced approach between model parameter efficiency and high classification accuracy [12]. All deep learning models were optimized using the Adam optimizer with a batch size of 32. For the machine learning models, data augmentation was applied to the training set to enhance generalization and prevent overfitting. This included random transformations: rotation, shifting, shearing, zooming, brightness adjustment, and horizontal flipping. For the initial training phase, an initial learning rate of 0.001 was applied. During the fine-tuning stage, the learning rate was reduced to 0.0001 (1e-4) to prevent the loss of pre-trained feature weights. The training process is divided into two phases:

- Initial Training: Each selected model (Xception [2], MobileNetV3 [3], and EfficientNetV2S [4]) is trained for 30 epochs. To prevent overfitting, an early stopping mechanism terminates training if the validation loss fails to improve for five consecutive epochs.
- Fine-tuning: The best-performing weights are saved, and the later layers of the models are unfrozen. We reduce the learning rate to ensure stable convergence and prevent new gradients from overwriting established features. The models are then retrained for an additional 20 epochs using the same early stopping criteria.

Finally, the model with the lowest validation loss is saved and evaluated against the shared 400 image test set. Although performance is evaluated separately for glasses and facial masks, both categories follow this identical experimental procedure to ensure a fair and consistent comparison.

D. Model Performance Evaluation

For the model training phase, Google Colab was utilized, running Python 3.10 (or higher) with the NVIDIA T4 GPU as the hardware accelerator. The performance of each model is analyzed and compared using a Confusion Matrix. This comparison focuses on various analytical aspects, such as the proportion of correct predictions relative to the ground truth. To quantify the effectiveness of the models, four standard metrics are employed: Accuracy, Precision, Recall, and F1-Score. Finally, the comprehensive results will be summarized in a tabular format to facilitate a clear and comparative understanding of each model's performance.

IV. Research Result and Evaluation

A. Threshold Values for Rule-based Classification

The threshold values for rule-based classification were determined by analyzing the luminance patterns within the training dataset for each accessory type. This process involved calculating feature metrics and calibrating the optimal thresholds for both the average ROI intensity and global pixel luminance. These parameters were refined based on observed data distributions to maximize classification accuracy. Mean Box Intensity is measured in intensity units, whereas all other parameters are normalized to a range of [0, 1]. The threshold values for both accessory are presented in Table III.

TABLE III. Classification criteria and threshold values for glasses and facial masks.

Statistical Parameters	Symbol	Glasses Threshold	Facial Masks Threshold
Mean Box Intensity	T_1	5.000	190.000
Mean Intensity	T_2	0.200	0.85
Mean AD	T_3	0.070	0.070
Median AD	T_4	0.051	0.003
Standard Deviation	T_5	0.084	0.056

Following this calibration, the model's performance was evaluated against the 120-image calibration subset (60 per accessory type). For glasses, the model achieved a 73.3% overall accuracy with correct predictions for 63% of real accessories, and 83% of AR accessories. For facial masks, the overall accuracy reached 71.7%, showing 56% accuracy for real accessories, and 87% for AR accessories. These results, detailed in Table IV, confirmed the efficacy of the selected features before proceeding with the full test evaluation.

TABLE IV. Preliminary calibration performance (calibration subset)

Accessory category	Glasses Accuracy (%)	Facial Masks Accuracy (%)
Real accessories	63 (19/30)	56 (17/30)
AR accessories	83 (25/30)	87 (26/30)
Total	**73.3 (46/60)**	71.7 (43/60)

Rule-based model performance was evaluated by applying classification criteria and threshold values from previous section to the test dataset. The detection performance was analyzed separately for glasses and facial masks to ensure a detailed assessment of each category, using test dataset of 200 images for each accessory type (100 real accessories and 100 AR accessories for each accessory type).

B. Image Detection Results using the Rule-based Model

Rule-based model performance was evaluated by applying classification criteria and threshold values from previous section to the test dataset. The detection performance was analyzed separately for glasses and facial masks to ensure a detailed assessment of each category, using test dataset of 200 images for each accessory type (100 real accessories and 100 AR accessories for each accessory type).

a) Detection Performance for Glasses: Applying the parameters and threshold values for glasses images to the test dataset of glasses images to evaluate the performance in accessories classification and detection. The detection results show that both real and AR accessories achieved a correct prediction accuracy of 75% (75 out of 100 images for each category)

b) Detection Performance for Facial Masks: Applying the parameters and threshold values for facial masks images to the test dataset of facial masks images to evaluate the performance in accessories classification. The detection results show that real accessories achieved a correct prediction accuracy of 51% and AR accessories achieved a correct prediction accuracy of 98%. The bias in facial mask detection exists because AR filters produce consistent luminance patterns that align closely with the calibrated thresholds (T_1 *to* T_5), leading to 98% accuracy. Conversely, the natural lighting and texture variations of real masks often fall outside these manual rules. High-intensity lighting on real masks can overlap with AR luminance profiles, causing false-positive augmented detections.

These outcomes were also further analyzed using a confusion matrix, as illustrated in Fig. 5.

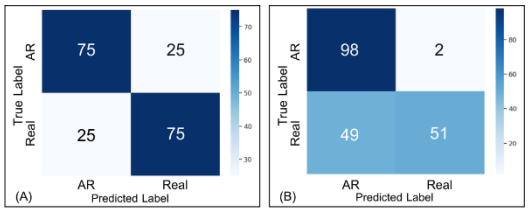

Fig. 5 Confusion matrix for accessories detection using the rule-based model. (A) Glasses detection (B) Facial mask detection

Based on both matrices, the four performance metrics were calculated, and the results are presented in Table V.

TABLE V. CLASSIFICATION METRICS FOR GLASSES AND FACIAL MASKS ACCESSORY USING RULE-BASED MODEL.

Performance Metrics	Value for glasses	Value for facial masks
Accuracy	0.75	0.75
Precision	0.75	0.67
Recall	0.75	0.98
F1-Score	0.75	0.80

C. Detection Results using the Machine Learning Model

For the machine learning approach, to ensure a fair comparison with the rule-based model, all three selected models were trained, validated, and tested using the same standardized datasets. Specifically, the test set is identical to the one used in the rule-based evaluation, providing a consistent baseline for performance comparison. Each model was trained separately for glasses and facial masks to analyze category-specific detection capabilities.

a) Detection Performance for Glasses: The performance of the deep learning models for glasses detection is visualized through the confusion matrices in Fig. 6. The results show distinct classification strengths for each model. Xception demonstrated the highest sensitivity in identifying AR accessories, correctly detecting 95 out of 100 samples, though it exhibited a higher rate of false positives by misidentifying real accessories as AR accessories. EfficientNetV2S showed the balanced performance, achieving the high correct identification count for real accessories. While MobileNetV3 maintained consistent performance across both categories, its overall correct classification count was lower than the other two models.

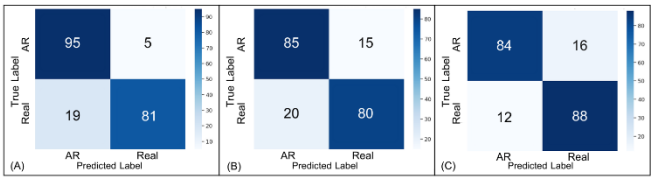

Fig. 6 Confusion matrix for glasses detection using selected machine learning model. (A) Xception (B) MobileNetV3 (C) EfficientNetV2S

b) Detection Performance for Facial Masks: The performance of the deep learning models for facial masks detection is visualized through the confusion matrices in Fig. 7. The results reveal distinct behavioral patterns among the selected models when handling larger accessory surfaces. Xception demonstrated exceptional precision in recognizing real accessories, correctly identifying all 100 samples with zero false positives. MobileNetV3 showing a highly balanced performance by correctly flagging 98 AR facial masks and 97 real facial masks. While EfficientNetV2S matched the high AR detection rate of 98, it exhibited a noticeable tendency to over-classify real facial masks as AR accessories, resulting in 21 misidentifications. These matrices indicate that for facial masks, which cover a larger facial area than glasses, autonomous feature extraction is highly effective, with Xception and MobileNetV3 significantly outperforming the rule-based model.

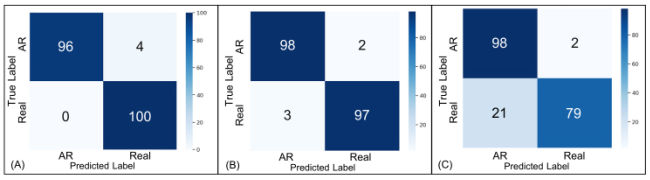

Fig. 7 Confusion matrix for facial masks detection using selected machine learning model. (A) Xception (B) MobileNetV3 (C) EfficientNetV2S

D. Performance Comparison between both Models

Based on the detection results for both real and AR accessories images, the performance varies across the rule-based and machine learning models. Each approach demonstrates distinct capabilities in identifying both types of images, as well as differences in effectiveness between glasses and facial masks. An overall performance comparison for rule-based model and machine learning model for glasses image is presented in Table VI, followed by comparison for facial masks image in Table VII.

TABLE VI. OVERALL PERFORMANCE COMPARISON FOR RULE-BASED MODEL AND MACHINE LEARNING MODEL FOR GLASSES IMAGES.

Performance Metrics	Rule-based Model	Machine Learning (Deep Learning) Models		
		Xception	MobileNetV3	EfficientNetV2S
Accuracy	0.75	**0.88**	0.82	0.86
Precision	0.75	0.83	0.80	**0.87**
Recall	0.75	**0.95**	0.85	0.84
F1-Score	0.75	**0.88**	0.83	0.86