Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Hierarchical Multi-Modal Signatures for OTT Content Identification and Redistribution Detection

Byeongchan Park

pbc866@ssu.ac.kr

Soongsil University

Seok-Yoon Kim

Soongsil University

Youngmo Kim

Soongsil University

Research Article

Keywords: OTT Streaming, Content Identification, Redistribution Detection, Hierarchical Signatures, Metadata-based Verification

Posted Date: March 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9183859/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Multimedia Systems on August 10th, 2026. See the published version at https://doi.org/10.1007/s00530-026-02565-7.

A Hierarchical Multi-Modal Signatures for OTT Content Identification and Redistribution Detection

Byeongchan Park[1], Seok-Yoon Kim[1], Youngmo Kim[1]

[1]Department of Computer Science and Engineering , Soongsil University, Seoul, Korea

Abstract As OTT services continue to expand, the distribution paths of premium video content have become increasingly diverse, making it essential to accurately identify and detect redistributed content under various transformations such as re-encoding, subtitle insertion, resolution changes, and aspect-ratio modification. Conventional hash-based methods often face limitations in such environments because low-level representations may change even when the semantic identity of the content remains intact. Therefore, a more robust content identification and verification method is required for OTT environments.

This study proposes hierarchical multi-modal signatures for OTT content identification and redistribution detection. The proposed method extracts representative frames from HLS-based streaming segments and refines visual features through preprocessing and postprocessing, which are then combined with hash values, metadata, and rights information to generate searchable content signatures. In addition, a hierarchical verification procedure is applied to improve both retrieval efficiency and identification accuracy. The proposed study extends conventional signature generation techniques for illegal streaming detection toward OTT-oriented content identification and redistribution detection, and it can serve as a basis for content protection, distribution monitoring, and rights verification in OTT platforms.

Keywords OTT Streaming, Content Identification, Redistribution Detection, Hierarchical Signatures, Metadata-based Verification

1 Introduction

The rapid growth of over-the-top (OTT) services has significantly changed the way video content is distributed and consumed. OTT platforms provide users with on-demand access to premium media content across various devices and network environments. However, the expansion of OTT services has also increased the risk of unauthorized redistribution, including illegal restreaming, re-uploading, and transformed reposting of copyrighted video content [4, 8, 9, 11].

In OTT environments, the same content is often redistributed after undergoing several modifications, such as re-encoding, subtitle insertion, resolution conversion, aspect-ratio modification, frame-rate reduction, and visual overlay. These

transformations preserve the semantic identity of the original content while changing its structural and visual representation. For this reason, direct file comparison or exact hash matching is not sufficient for practical identification in redistributed OTT environments [5–7, 10, 13].

Existing approaches to video content protection and verification have mainly relied on metadata matching, watermarking, hashing, and content-based fingerprinting. Metadata-based methods are computationally efficient, but they are vulnerable to modification, omission, and inconsistency in unauthorized distribution environments. Watermarking methods can provide persistent ownership information, but they usually require prior embedding during content production or distribution, and their robustness may be affected by retransmission, geometric distortion, and screen-shooting attacks [1, 2, 14–18]. Hash-based methods are also effective for exact matching, but they are sensitive to even small modifications in the underlying video representation [6, 12].

To overcome these limitations, content-based fingerprinting and video copy detection methods have been actively studied. Recent studies have proposed segment-based fingerprint generation, dual-level copy detection, similarity alignment for video segment matching, and large-scale partial video copy detection [3, 5–7]. In addition, several studies have addressed copy detection and verification under social-media redistribution, blockchain-based verification, and deep feature-based multimedia fingerprinting [9–12]. These studies show that robust video identification requires not only visual similarity analysis but also a structured integration of multiple information sources.

Recent multimedia protection studies have also emphasized the importance of robust multimodal verification. Audio fingerprinting, audio-visual alignment, and cross-modal matching have been used to improve robustness in multimedia retrieval and forgery detection tasks [8, 19–21]. At the same time, several recent surveys have reported that modern multimedia forensics should consider not only visual copy detection but also broader issues such as multimodal consistency, redistribution traces, and attack robustness [12, 22–25]. These observations suggest that a more integrated signature representation is required for OTT content identification and redistribution detection.

In this paper, hierarchical multi-modal signatures are proposed for OTT content identification and redistribution detection. The proposed method extracts representative frames from HLS-based OTT media segments, performs preprocessing to reduce unnecessary visual variation, and generates visual, hash, and descriptive signatures from the input content. These signatures are then integrated into a hierarchical structure for efficient retrieval and robust verification. In this way, the proposed method is designed to support practical OTT content identification under various redistribution conditions.

The main contributions of this paper are as follows. First, a hierarchical multi-modal signature generation method is proposed for OTT content identification and redistribution detection. Second, the proposed method combines representative frame selection, preprocessing, visual signature extraction, hash-based filtering, and descriptive signature construction in a unified process. Third, a hierarchical

verification and retrieval process is introduced to improve both matching efficiency and robustness in large-scale OTT environments. Finally, the effectiveness of the proposed method is evaluated under several OTT redistribution conditions, including subtitle insertion, resolution change, aspect-ratio modification, and frame-rate reduction.

The remainder of this paper is organized as follows. Section 2 reviews previous studies related to video fingerprinting, watermarking, and OTT content verification. Section 3 describes the proposed hierarchical multi-modal signature generation and verification method. Section 4 presents the experimental setup and performance evaluation results. Finally, Section 5 concludes this paper.

2 Related Work

2.1 Metadata- and Hash-Based Content Identification

Existing methods for video content identification have commonly relied on metadata matching and hash-based verification. Metadata-based methods use descriptive information such as title, file name, codec parameters, content category, and service-related attributes to identify media content. These methods are computationally efficient and can be useful when the metadata is complete and consistently maintained. However, in unauthorized redistribution environments, metadata can be modified, omitted, or intentionally falsified. For this reason, metadata alone is not sufficient for robust content identification in practical OTT redistribution environments [9, 12].

Hash-based methods have also been widely used because they provide compact and deterministic representations of digital content. These methods are effective for exact matching and integrity verification. However, even small changes in encoding structure, frame composition, or transmission format can produce different hash values. In OTT environments, content is frequently redistributed after re-encoding, resizing, subtitle insertion, or frame-rate change. Therefore, exact hash matching is limited in its ability to identify semantically identical but structurally modified OTT content [6, 12, 13].

For this reason, metadata and hash information are more suitable for preliminary filtering than for final verification. In practical OTT content identification, these methods can reduce the candidate search space efficiently, but additional content-based evidence is required for robust matching under redistribution conditions [6, 7].

2.2 Watermarking and Signature-Based Video Identification

Digital watermarking has long been studied as an effective method for copyright protection and ownership verification. Watermarking methods embed imperceptible information into media content and use the embedded information for ownership proof, distribution tracing, and copyright protection. Recent studies have proposed adversarial watermarking, angle-robust watermark restoration, screen-shooting-resistant watermarking, and temporally aware watermarking for improved robustness under practical attack conditions [1, 2, 14–18].

Although watermarking is useful for copyright protection, it generally requires prior embedding during content production, editing, or distribution. In addition, the robustness of the embedded watermark can be degraded by geometric distortion, retransmission, transcoding, and screen-shooting attacks [14–18]. For this reason, watermarking alone is not always sufficient for retrospective monitoring of redistributed OTT content.

To overcome these limitations, video fingerprinting and content-based copy detection methods have been actively studied. These methods extract representative features from the intrinsic characteristics of video content and use them for similarity-based identification. Recent studies have proposed segment-based fingerprint generation, dual-level copy detection, similarity alignment for segment matching, large-scale partial video copy detection, and robust video hashing [3, 5–7, 13]. These approaches are more suitable than exact hash matching in environments where redistributed content remains semantically identical but is structurally transformed.

2.3 Limitations of Existing Approaches in OTT Environments

Recent studies have shown that robust video content identification requires more than a single type of feature. In particular, audio fingerprinting, audio-visual matching, cross-modal regularization, and multimodal alignment have been used to improve robustness in multimedia detection and verification tasks [8, 19–21]. These studies indicate that multiple information sources can complement each other when one type of evidence becomes unstable under practical transformation conditions.

Recent research has also expanded video copy detection toward more realistic redistribution scenarios, such as social-media reposting, partial copy detection, blockchain-based verification, and attack-robust multimedia fingerprinting [5, 9–12]. In these environments, content may be redistributed after temporal reduction, geometric distortion, overlay insertion, or platform-specific compression. Therefore, robust OTT-oriented detection requires an integrated method that can jointly use content-derived visual information, structural information, and descriptive information.

Several recent survey studies have also emphasized that modern multimedia forensics should consider not only video copy detection but also multimodal consistency, redistribution traces, and attack robustness [12, 22–25]. These observations suggest that a hierarchical and multi-modal verification structure is needed for practical OTT content identification and redistribution detection.

Based on these observations, this paper proposes hierarchical multi-modal signatures for OTT content identification and redistribution detection. The proposed method combines representative frame-based visual signatures, hash signatures, and descriptive signatures in a unified hierarchical process. In this way, the proposed method is intended to improve both matching efficiency and robustness in practical OTT redistribution environments.

3 Hierarchical Multi-Modal Signature Generation

3.1 Overview of the Proposed Signature Generation Process

The hierarchical multi-modal signature generation process proposed in this paper for OTT content identification and redistribution detection is shown in Fig. 1.

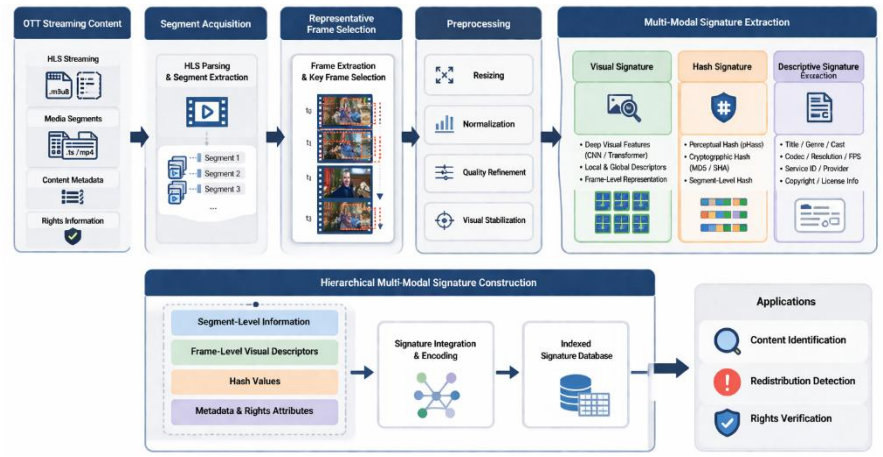


Fig. 1 Hierarchical multi-modal signature generation process

First, HLS playlists, media segments, content metadata, and rights information are collected from OTT streaming content. In this step, streaming content actually delivered in OTT environments is used as the input source.

Next, in the segment acquisition step, HLS parsing is performed, media segments are collected, and the segments are organized in temporal order. In this step, the streaming content is divided into segment units and arranged in a form suitable for subsequent processing.

After that, in the representative frame selection step, frames are extracted from the segments, and representative frames are selected for content identification. In this step, not all frames are used directly. Instead, redundant frames are reduced, and visually informative frames are selected.

In the preprocessing step, resizing, normalization, quality refinement, and visual stabilization are applied to the selected representative frames. In this step, unnecessary visual variations caused by re-encoding, resolution changes, and aspect-ratio modification are reduced in order to improve the consistency of feature extraction.

Next, in the signature extraction step, visual signature information, hash signature information, and descriptive signature information are generated. The visual signature represents frame-level visual characteristics, while the hash signature provides compact identification information for fast comparison. In addition, the descriptive signature provides supplementary identification evidence using metadata and rights information.

Finally, in the hierarchical multi-modal signature construction step, segment-level information, frame-level visual features, hash values, metadata, and rights attributes are integrated into a unified signature. The generated signature is stored in the signature repository and is used for content identification, redistribution detection, and rights verification.

The proposed method is designed to cope with various transformations in OTT environments by hierarchically combining visual features, hash information, metadata, and rights information.

3.2 Representative Frame Selection

The representative frame selection process proposed in this paper is shown in Fig. 2.

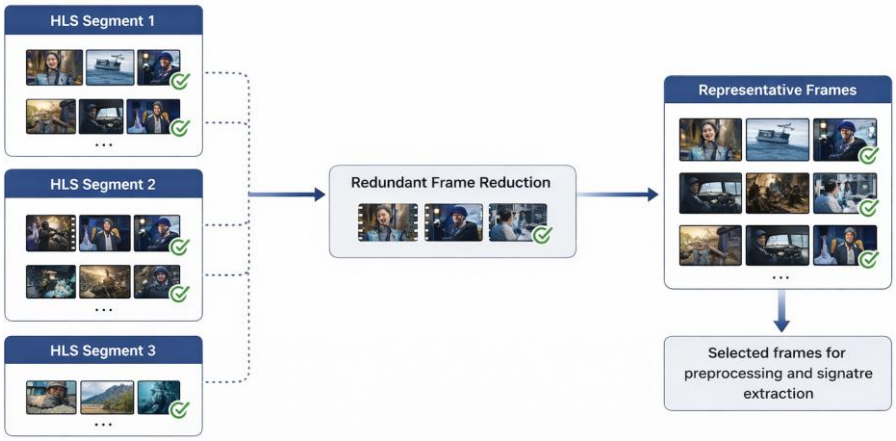


Fig. 2 Representative frame selection process from HLS media segments

First, video frames are extracted from HLS media segments. In this step, the OTT streaming content is divided into segment units, and frames are generated from each segment in temporal order. Since OTT content is delivered as segmented streaming data, frame extraction is performed on a segment basis rather than on the entire video at once.

Next, candidate frames are selected from each segment. In this step, all extracted frames are not used directly because consecutive frames often contain highly similar visual information. If all frames are used, redundant information increases and the computational cost of signature generation also increases. For this reason, a subset of representative frames is selected from each segment.

In this paper, representative frames are selected by considering temporal position and visual variation. The temporal position is used to ensure that frames are distributed across the segment, while the visual variation is used to exclude frames with little discriminative information. As a result, the selected frames can preserve the visual identity of the content while reducing redundant data.

Let S_i denote the i-th segment and let $F_{i,j}$ denote the j-th frame extracted from that segment. The frame set extracted from the i-th segment is defined as shown in Eq. (1).

$$S_i = \{F_{i,1}, F_{i,2}, \dots, F_{i,N_i}\} \quad (1)$$

Here, N_i denotes the number of extracted frames in the i-th segment.

Among the extracted frames, a subset of representative frames is selected. The representative frame set is defined as shown in Eq. (2).

$$R_i = \{F_{i,k} \mid k \in \Omega_i\}, R_i \subseteq S_i \quad (2)$$

Here, Ω_i denotes the index set of representative frames selected from the i-th segment.

To reduce redundant frames, the visual difference between adjacent frames is measured. A frame is selected as a representative frame when the visual distance exceeds a predefined threshold. This condition is expressed as shown in Eq. (3).

$$D(F_{i,j}, F_{i,j+1}) > \tau \quad (3)$$

Here, $D(\cdot)$ denotes the visual distance between two adjacent frames, and τdenotes the threshold for representative frame selection.

Finally, the representative frames selected from all segments are integrated into a single representative frame set. The final representative frame set is defined as shown in Eq. (4).

$$R = \bigcup_{i=1}^{M} R_i \quad (4)$$

Here, M denotes the total number of segments.

The main parameters used in the representative frame selection process are shown in Table 1.

Table 1 Main parameters for representative frame selection

Parameter	Description	Example
M	Number of segments	3, 5, 10
N_i	Number of extracted frames in segment S_i	30
K	Number of representative frames selected per segment	3
τ	Threshold for visual difference	0.15
$D(\cdot)$	Visual distance measure	Histogram distance or feature distance

Through this process, redundant frames are reduced and visually informative frames are selected for preprocessing and signature extraction. The representative frame selection step is important because it affects both the efficiency and the robustness of the proposed hierarchical multi-modal signature generation process.

3.3 Preprocessing and Signature Extraction

The preprocessing and signature extraction process proposed in this paper is shown in Fig. 3.

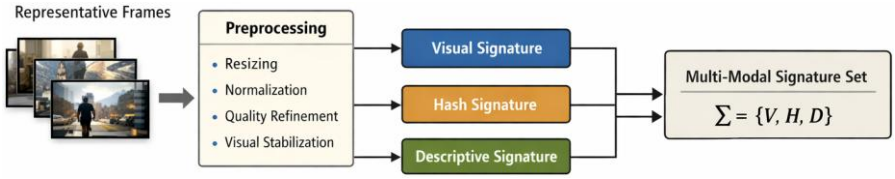


Fig. 3 Preprocessing and signature extraction process

First, preprocessing is performed on the representative frames selected in the previous step. In this step, resizing, normalization, quality refinement, and visual stabilization are applied to the input frames. Since redistributed OTT content may include various visual distortions such as re-encoding, resolution change, subtitle insertion, and aspect-ratio modification, preprocessing is used to reduce unnecessary variation and to improve the consistency of subsequent signature extraction.

Let $R = \{R_1, R_2, \ldots, R_n\}$ denote the representative frame set. The preprocessed frame set is defined as shown in Eq. (5).

$$P = \phi(R) \qquad (5)$$

Here, $\phi(\cdot)$ denotes the preprocessing function, and P denotes the set of preprocessed representative frames.

In this paper, the preprocessing function consists of four major operations. The resized frame is obtained by mapping the input frame into a fixed spatial resolution. This process is expressed as shown in Eq. (6).

$$P_r = \psi_r(R) \qquad (6)$$

Here, $\psi_r(\cdot)$ denotes the resizing operation, and P_r denotes the resized frame set.

Next, normalization is applied to reduce variation caused by brightness and contrast changes. The normalized frame is defined as shown in Eq. (7).

$$P_n = \psi_n(P_r) \qquad (7)$$

Here, $\psi_n(\cdot)$ denotes the normalization operation, and P_n denotes the normalized frame set.

After that, quality refinement and visual stabilization are performed to improve the robustness of feature extraction. The final preprocessed frame set is defined as shown in Eq. (8).

$$P = \psi_s(\psi_q(P_n)) \qquad (8)$$

Here, $\psi_q(\cdot)$ denotes the quality refinement function, and $\psi_s(\cdot)$ denotes the visual stabilization function.

After preprocessing, three types of signature information are extracted from the input content. These are visual signature information, hash signature information, and descriptive signature information. In this paper, the three signatures are generated independently and then used together in the hierarchical multi-modal signature construction step.

First, the visual signature is extracted from the preprocessed frames. In this step, frame-level visual descriptors are generated to represent the intrinsic visual characteristics of the content. The visual signature set is defined as shown in Eq. (9).

$$V = \{v_1, v_2, \dots, v_n\} = f_v(P) \qquad (9)$$

Here, $f_v(\cdot)$ denotes the visual feature extraction function, and v_i denotes the visual descriptor extracted from the i-th preprocessed frame.

Next, the hash signature is generated to support compact identification and fast comparison. In this step, hash values are extracted from the representative frames or corresponding segment information. The hash signature set is defined as shown in Eq. (10).

$$H = \{h_1, h_2, \dots, h_n\} = f_h(P) \qquad (10)$$

Here, $f_h(\cdot)$ denotes the hash generation function, and $h_i$denotes the hash value corresponding to the i-th frame or segment.

In addition, descriptive signature information is generated using metadata and rights information associated with the OTT content. This information includes title, genre, service identifier, codec information, copyright information, and license-related attributes. The descriptive signature is defined as shown in Eq. (11).

$$D = f_d(M, C) \qquad (11)$$

Here, $f_d(\cdot)$ denotes the descriptive signature generation function, M denotes metadata information, and C denotes rights-related information.

Finally, the extracted visual, hash, and descriptive signatures are used as the input elements for hierarchical multi-modal signature construction. The signature tuple generated in this step is defined as shown in Eq. (12).

$$\Sigma = \{V, H, D\} \qquad (12)$$

Here, Σ denotes the final multi-modal signature set.

The main components used in preprocessing and signature extraction are shown in Table 2.

Table 2 Main components of preprocessing and signature extraction

Component	Description	Output
Resizing	Converts frames into a fixed spatial resolution	P_r
Normalization	Reduces brightness and contrast variation	P_n
Quality refinement	Improves frame quality for feature extraction	Intermediate refined frames
Visual stabilization	Reduces visual inconsistency	P
Visual signature extraction	Generates frame-level visual descriptors	V
Hash signature generation	Produces compact hash values	H
Descriptive signature generation	Uses metadata and rights information	D

Through this process, the proposed method generates multiple complementary forms of signature information from representative frames and associated descriptive data. The preprocessing step improves the consistency of input frames, and the signature extraction step provides visual, hash, and descriptive evidence for subsequent content identification and redistribution detection.

3.4 Hierarchical Multi-Modal Signature Construction

The hierarchical multi-modal signature construction process proposed in this paper is shown in Fig. 4.

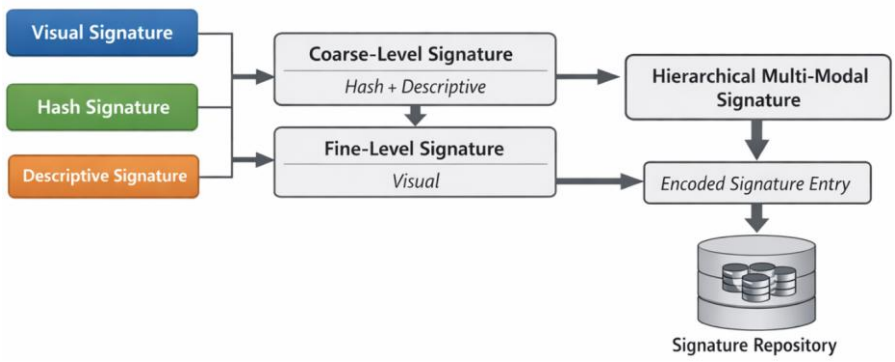


Fig. 4 Hierarchical multi-modal signature construction process

First, the visual signature, hash signature, and descriptive signature extracted in the previous step are collected as the input elements for final signature construction. In this step, the proposed method does not use a single type of signature independently. Instead, multiple types of signature information are combined in a hierarchical manner in order to improve both identification robustness and retrieval efficiency.

Let V, H, and D denote the visual signature set, hash signature set, and descriptive signature set, respectively. The hierarchical multi-modal signature is defined as shown in Eq. (13).

$$\Sigma_h = \{V, H, D\} \qquad (13)$$

Here, Σ_h denotes the hierarchical multi-modal signature.

In this paper, the visual signature provides the main evidence for content-level similarity, since it reflects the intrinsic visual characteristics of the OTT content. The hash signature is used as auxiliary information for fast comparison and candidate filtering. The descriptive signature is used as supplementary information for metadata-based verification and rights-aware identification. The role of each signature component is shown in Table 3.

Table 3 Roles of signature components in hierarchical multi-modal signature construction

Component	Role	Description
V	Primary visual evidence	Represents frame-level visual characteristics for robust content identification
H	Fast filtering evidence	Supports efficient comparison and candidate reduction
D	Supplementary descriptive evidence	Provides metadata and rights-related information for practical verification

Next, the proposed method organizes the signature components into a hierarchical structure. In this step, descriptive information and hash information are first used for coarse-level filtering, and visual signature information is then used for fine-level verification. This hierarchical structure is defined as shown in Eq. (14).

$$\Sigma_h = \{\Sigma_c, \Sigma_f\} \qquad (14)$$

Here, Σ_c denotes the coarse-level signature, and Σ_f denotes the fine-level signature.

The coarse-level signature consists of hash and descriptive information. The coarse-level signature is defined as shown in Eq. (15).

$$\Sigma_c = \{H, D\} \qquad (15)$$

The fine-level signature consists of visual signature information. The fine-level signature is defined as shown in Eq. (16).

$$\Sigma_f = \{V\} \qquad (16)$$

Through this hierarchical structure, the proposed method first reduces the candidate search space using hash and descriptive information and then performs robust visual verification using frame-level visual features. This structure is effective in OTT environments because it improves efficiency in large-scale content repositories while preserving robustness against various transformations.

Finally, the hierarchical multi-modal signature is encoded and stored in the signature repository. The encoded signature entry is defined as shown in Eq. (17).

$$E_i = \{\mathrm{ID}_i, \Sigma_{h,i}, M_i, C_i\} \qquad (17)$$

Here, ID_i denotes the content identifier of the i-th OTT content, $\Sigma_{h,i}$ denotes the hierarchical multi-modal signature, M_i denotes metadata information, and C_i denotes rights-related information.

The stored signature entries are used for indexed retrieval, content identification, redistribution detection, and rights verification. In this paper, the hierarchical multi-modal signature construction step is important because it integrates multiple complementary types of evidence into a unified structure for practical OTT content identification.

3.5 Hierarchical Verification and Retrieval

The hierarchical verification and retrieval process proposed in this paper is shown in Fig. 5.

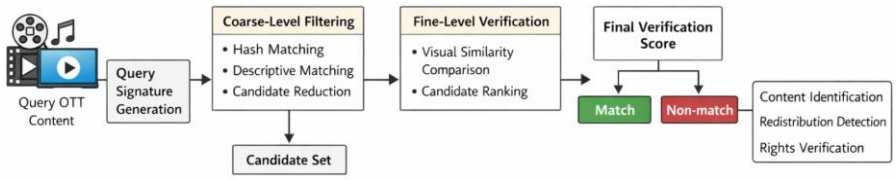


Fig. 5 Hierarchical Verification and Retrieval

First, the query OTT content is converted into a hierarchical multi-modal signature using the same procedure described in the previous steps. In this step, representative frames are selected, preprocessing is performed, and visual, hash, and descriptive signatures are extracted. The generated query signature is then used as the input for retrieval and verification.

Let Σ_q denote the hierarchical multi-modal signature of the query content, and let $\Sigma_{h,i}$ denote the stored signature of the i-th reference content in the signature

repository. The verification process is performed by comparing Σ_q with the reference signatures stored in the repository.

In this paper, the verification process is performed in a hierarchical manner. First, coarse-level filtering is applied using hash signature information and descriptive signature information. In this step, the search space is reduced by removing clearly irrelevant candidates before visual comparison is performed. The coarse-level candidate set is defined as shown in Eq. (18).

$$C_c = \{\Sigma_{h,i} \mid \delta(H_q, H_i) \leq \alpha \text{ or } \gamma(D_q, D_i) = 1\} \quad (18)$$

Here, C_c denotes the coarse-level candidate set, H_q and H_i denote the hash signatures of the query and reference contents, D_q and D_i denote the descriptive signatures, $\delta(\cdot)$ denotes the hash distance function, $\gamma(\cdot)$ denotes the descriptive matching function, and α denotes the threshold for coarse-level filtering.

After coarse-level filtering, fine-level verification is performed using visual signature information. In this step, the visual similarity between the query content and each candidate content is measured. The visual similarity score is defined as shown in Eq. (19).

$$S_v(\Sigma_q, \Sigma_{h,i}) = \frac{1}{n}\sum_{j=1}^{n} \text{sim}(v_{q,j}, v_{i,j}) \quad (19)$$

Here, $S_v(\cdot)$ denotes the visual similarity score, $v_{q,j}$ denotes the visual descriptor of the j-th representative frame in the query content, $v_{i,j}$ denotes the corresponding visual descriptor in the i-th candidate content, and $\text{sim}(\cdot)$ denotes the similarity function.

Next, the final verification score is computed by combining the visual similarity score with the coarse-level evidence. The final verification score is defined as shown in Eq. (20).

$$S_f = w_v S_v + w_h S_h + w_d S_d \quad (20)$$

Here, S_f denotes the final verification score, S_v denotes the visual similarity score, S_h denotes the hash-based similarity score, S_d denotes the descriptive similarity score, and w_v, w_h, and w_d denote the weighting factors of each component.

The weighting factors used in the final verification process are shown in Table 4.

Table 4 Weighting factors for final verification

Component	Description	Weight
w_v	Weight of visual similarity	0.6
w_h	Weight of hash similarity	0.2

Component	Description	Weight
w_d	Weight of descriptive similarity	0.2

Finally, the query content is determined as matched content when the final verification score exceeds a predefined threshold. The final decision rule is defined as shown in Eq. (21).

$$\text{Match} = \begin{cases} 1, & \text{if } S_f \geq \beta \\ 0, & \text{otherwise} \end{cases} \qquad (21)$$

Here, β denotes the final decision threshold.

Through this process, the proposed method first reduces the candidate set using hash and descriptive signatures and then performs fine-level visual verification on the filtered candidates. This hierarchical retrieval strategy improves both computational efficiency and matching robustness in large-scale OTT environments. In this paper, the final retrieval result is used for content identification, redistribution detection, and rights verification.

4 Experimental Results

4.1 Hierarchical Verification and Retrieval

The experimental environment used in this paper is shown in Table 5.

Table 5 Experimental environment

Component	Description
Operating system	Windows 11 / Ubuntu 22.04
CPU	Intel Core i7 / AMD Ryzen 7 class
GPU	NVIDIA RTX 3060 or higher
Memory	32 GB
Storage	SSD 1 TB
Development environment	Python 3.10, OpenCV, NumPy
Database environment	Relational DBMS and indexed signature repository
Streaming format	HLS-based OTT media segments

In this paper, the experiment was performed in an environment capable of processing OTT streaming content, extracting representative frames, generating multi-modal signatures, and retrieving candidate signatures from the repository.
The input data consisted of OTT streaming content delivered in HLS format. In this step, HLS playlists, media segments, metadata, and rights-related information were collected and used for signature generation and verification.

The experimental environment was configured to evaluate both the efficiency and the robustness of the proposed method. In particular, the experiment focused on representative frame selection, preprocessing, multi-modal signature generation, and hierarchical verification under OTT redistribution conditions.

The main experimental procedure used in this paper is as follows. First, original OTT content was collected and converted into hierarchical multi-modal signatures. Next, transformed OTT content was generated under several redistribution conditions. After that, the transformed content was used as query data, and hierarchical verification and retrieval were performed against the stored signature repository. Finally, the retrieval accuracy, verification score, and detection performance were measured and analyzed.

4.2 Dataset Construction

The dataset construction used in this paper is shown in Table 6.

Table 6 Dataset construction for OTT redistribution detection

Component	Description
Original OTT contents	Movies, dramas, entertainment, and animation
Number of original contents	1,000
Streaming format	HLS
Segment unit	3 segments per content unit
Representative frames	3 frames selected per segment
Redistribution transformations	Re-encoding, subtitle insertion, resolution change, aspect-ratio modification, brightness change, frame-rate reduction
Query data	Transformed OTT content
Reference data	Original content signatures stored in repository

In this paper, the dataset was constructed using original OTT content and transformed OTT content. The original content was used to build the reference signature repository, and the transformed content was used as query data for verification and retrieval.

The original dataset consisted of OTT video content in several categories, including movies, dramas, entertainment programs, and animation. Each content item was converted into HLS streaming format, and segment-level data were collected for representative frame extraction and signature generation.

To evaluate the robustness of the proposed method, transformed content was generated under several redistribution conditions. In this paper, the main transformation conditions included re-encoding, subtitle insertion, resolution change, aspect-ratio modification, brightness change, and frame-rate reduction. These transformations were selected because they commonly appear in unauthorized redistribution environments.

The transformation conditions used in this paper are shown in Table 7.

Table 7 Transformation conditions for redistributed OTT content

Transformation type	Description
Re-encoding	Bitrate and codec configuration changed
Subtitle insertion	Additional subtitle overlay inserted
Resolution change	Spatial resolution modified

Aspect-ratio modification	Screen ratio changed
Brightness change	Brightness adjusted
Frame-rate reduction	Temporal sampling reduced

Through this dataset construction, the experiment was designed to evaluate whether the proposed method can preserve identification performance under common OTT redistribution conditions.

4.3 Evaluation Metrics

The evaluation metrics used in this paper are shown in Table 8.

Table 8 Evaluation metrics

Metric	Description
Accuracy	Ratio of correctly identified query contents
Precision	Ratio of correctly matched contents among detected contents
Recall	Ratio of correctly detected contents among all relevant contents
F1-score	Harmonic mean of precision and recall
Verification time	Time required for hierarchical verification
Retrieval time	Time required to retrieve candidates from repository

In this paper, identification performance was evaluated using accuracy, precision, recall, and F1-score. In addition, the computational performance of the proposed method was evaluated using verification time and retrieval time.

Accuracy is defined as shown in Eq. (22).

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \qquad (22)$$

Here, N_{correct} denotes the number of correctly identified query contents, and N_{total} denotes the total number of query contents.

Precision is defined as shown in Eq. (23).

$$\text{Precision} = \frac{TP}{TP + FP} \qquad (23)$$

Recall is defined as shown in Eq. (24).

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (24)$$

F1-score is defined as shown in Eq. (25).

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (25)$$

Here, TP, FP, and FN denote true positive, false positive, and false negative, respectively.

The average verification time is defined as shown in Eq. (26).

$$T_v = \frac{1}{N}\sum_{i=1}^{N} t_i \qquad (26)$$

Here, $T_v$denotes the average verification time, $t_i$denotes the verification time for the i-th query, and Ndenotes the total number of queries.

These evaluation metrics were used to measure both the identification effectiveness and the computational efficiency of the proposed method.

4.4 Experimental Results and Analysis

The experimental results obtained in this paper are shown in Table 9.

Table 9 Identification performance under OTT redistribution conditions

Condition	**Accuracy (%)**
Original content	97
Subtitle insertion	96
Image insertion	93
Brightness change	92
Grayscale conversion	91
Aspect-ratio modification	83
Mirroring	85
Rotation	87
Frame-rate reduction	76
Average	88.80

The accuracy comparison under different redistribution conditions is shown in Fig. 6.

In this paper, the proposed method showed stable identification performance under several OTT redistribution conditions. In particular, the method achieved relatively high accuracy for subtitle insertion, image insertion, brightness change, and grayscale conversion. This result indicates that the proposed hierarchical multi-modal signature can preserve the visual identity of content under moderate visual transformations.

On the other hand, the performance decreased under aspect-ratio modification, mirroring, rotation, and frame-rate reduction. This result indicates that geometric transformation and temporal variation have a larger effect on representative frame consistency and visual signature matching. Among the tested conditions, frame-rate reduction showed the lowest accuracy. This result suggests that temporal sparsity may reduce the stability of representative frame-based identification.

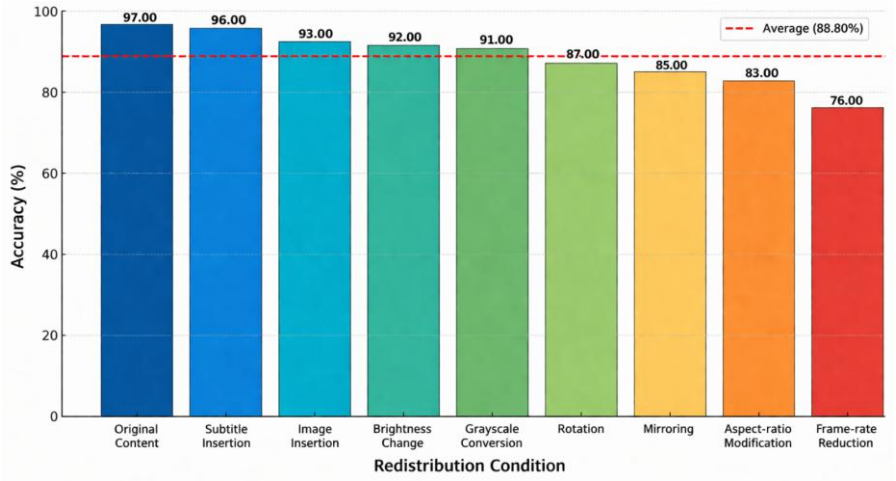


Fig. 6 Accuracy comparison under OTT redistribution conditions

The average identification accuracy obtained in this paper was 88.80%. The average verification time was 8.6258 s. The main performance results are shown in Table 10.

Table 10 Overall performance of the proposed method

Metric	Result
Average accuracy	88.80 %
Average verification time	8.6258 s
Best condition	Original content
Most robust transformation	Subtitle insertion
Most challenging transformation	Frame-rate reduction

The experimental results indicate that the proposed method is effective for OTT content identification and redistribution detection under several practical transformation conditions. In particular, the combination of visual, hash, and descriptive signatures improved the robustness of retrieval and verification compared with a single signature type.

However, the results also show that geometric distortion and temporal reduction remain challenging conditions. Therefore, additional improvement is needed in future work to improve robustness against severe structural transformation in OTT redistribution environments.

5 Conclusion

In this paper, hierarchical multi-modal signatures were proposed for OTT content identification and redistribution detection. The proposed method was designed to improve the robustness and efficiency of content verification in OTT environments by combining representative frame selection, preprocessing, visual signature extraction, hash signature extraction, descriptive signature generation, and hierarchical verification.

First, representative frames were selected from HLS media segments in order to reduce redundant visual information and to preserve the main characteristics of OTT content. Next, preprocessing was performed to reduce unnecessary visual variation caused by re-encoding, resolution change, and aspect-ratio modification. After that, visual, hash, and descriptive signatures were extracted and integrated into a hierarchical multi-modal signature structure. Finally, hierarchical verification and retrieval were performed by combining coarse-level filtering and fine-level visual verification.

The experimental results showed that the proposed method achieved stable identification performance under several OTT redistribution conditions. In particular, the method showed relatively high accuracy for subtitle insertion, image insertion, brightness change, and grayscale conversion. The average identification accuracy was 88.80%, and the average verification time was 8.6258 s. These results indicate that the proposed method can provide practical identification and redistribution detection performance in OTT environments.

However, the results also showed that the identification performance decreased under frame-rate reduction, aspect-ratio modification, mirroring, and rotation. This result indicates that temporal reduction and geometric transformation still remain challenging conditions for representative frame-based verification. Therefore, additional improvement is needed to enhance robustness under severe structural transformation conditions.

This study is meaningful in that it extends conventional signature generation methods for illegal streaming analysis to OTT-oriented content identification and redistribution detection. The proposed method can be used as a practical basis for OTT content protection, redistribution monitoring, and rights verification. In future work, it will be necessary to improve robustness against temporal and geometric distortion and to expand the verification framework to large-scale OTT service environments.

Acknowledgment

This research project supported by Ministry of Culture, Sport and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sport and Tourism in 2025(Project Name: Development of Copyright Technology(+Law) for OTT Contents Copyright Protection Technology Development and Application, Project Number: RS-2023-00225267, Contribution Rate: 100%)

References

Kim JH, Hong SM, Shin YT (2024) A comparative study and performance analysis of adversarial watermarking techniques. *Journal of Software Forensics* 20(4):81–89

Lee JE, Shin YT (2024) Performance evaluation of feature point extraction algorithms for watermark restoration at various angles. *Journal of Software Forensics* 20(4):91–101

Kang SY (2023) Transformer-based segment fingerprint generation method for partial video copy detection. *Journal of KIISE* 50(3):257–262

Wang Y, Meng L, Li S et al (2026) Transformer-based audio-visual features for video copy detection. *Journal of King Saud University Computer and Information Sciences*

Luo Z, Zhang L, Lai J, Wang X, Tang C (2023) An efficient partial video copy detection for a large-scale video database. In: *Proceedings of BIGCOM 2023*, pp 117–125

Wang T, Ma F, Liu Z, Rao F (2023) A dual-level detection method for video copy detection. *CoRR* abs/2305.12361

Liu Z, Ma F, Wang T, Rao F (2023) A similarity alignment model for video copy segment matching. *CoRR* abs/2305.15679

Uikey A, Bedi AK, Choudhary P, Ooi WT, Saini M (2024) A highly robust deep learning technique for overlap detection using audio fingerprinting. *Multimedia Tools and Applications* 83(10):29119–29137

Shang W, Liu X, Song M (2023) On-chain video copy detection based on Swin-Transformer and deep hashing. *International Journal of Networked and Distributed Computing* 11:1–15

Amerini I, Anagnostopoulos A, Maiano L, Ricciardi Celsi L (2022) Learning double-compression video fingerprints left from social-media platforms. *CoRR* abs/2212.03658

Amerini I, Anagnostopoulos A, Maiano L, Ricciardi Celsi L (2021) Learning double-compression video fingerprints left from social-media platforms. In: *Proceedings of ICASSP 2021*, pp 2530–2534

Chen W, Gan W, Yu PS (2024) Digital fingerprinting on multimedia: a survey. *CoRR* abs/2408.14155

Yu Z, Liang X, Chen L, Zhang X, Tang Z (2025) Robust video hashing with DWT and tensor SVD for copy detection. *Expert Systems with Applications* 286:130669

Li Y, Liao X, Wu X et al (2024) Screen-shooting resistant watermarking with grayscale deviation simulation. *IEEE Transactions on Multimedia*

Fu L, Liao X, Guo J, Dong L, Qin Z (2024) WaveRecovery: screen-shooting watermarking based on wavelet and recovery. *IEEE Transactions on Circuits and Systems for Video Technology*

Liu L, Xu P, Xue Q (2025) Screen shooting resistant watermarking based on cross attention. *Scientific Reports* 15:1–17

Hu R, Zhang J, Li Y et al (2025) VideoShield: regulating diffusion-based video generation models via watermarking. *CoRR* abs/2501.14195

Teymoorianfard M, Ma S, Houmansadr A (2025) VIDSTAMP: a temporally-aware watermark for ownership and integrity in video diffusion models. *CoRR* abs/2505.01406

Zou H, Shen M, Hu Y, Chen C, Chng ES, Rajan D (2024) Cross-modality and within-modality regularization for audio-visual deepfake detection. *CoRR* abs/2401.05746

Du Y, Wang Z, Luo Y et al (2025) CAD: a general multimodal framework for video deepfake detection via cross-modal alignment and distillation. *CoRR* abs/2505.15233

Altwlkany K, Delalic S, Skeja A et al (2024) Application of audio fingerprinting techniques for real-time scalable speech retrieval and speech clusterization. *CoRR* abs/2410.21876

Wary A, Neelima A (2019) A review on robust video copy detection. *International Journal of Multimedia Information Retrieval* 8(2):61–78

Singh S, Dhumane A (2025) Unmasking digital deceptions: an integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges. *MethodsX* 14:103632

Khan AA, Asif M, Khan MI et al (2025) A survey on multimedia-enabled deepfake detection. *Information Systems Frontiers*

Carlson A et al (2023) Understanding and improving video fingerprinting attack robustness. In: *Proceedings of ACM CCS 2023 Workshops*