International Journal of Networked and Distributed Computing (2023) 11:60–74
https://doi.org/10.1007/s44227-023-00010-z

RESEARCH ARTICLE

On-Chain Video Copy Detection Based on Swin-Transformer and Deep Hashing

Wenqian Shang[1] · Xintao Liu[1] · Miaoran Song[1]

Received: 10 June 2023 / Accepted: 8 August 2023 / Published online: 18 August 2023

Abstract
In recent years, short videos are spreading faster and become higher quality due to edge-cloud technology. People receive information gradually from graphic to video. At the same time as the number of videos spread rapidly, infringing videos are also flooding the Internet. The wild spread of infringing videos on the Internet has brought serious losses to video websites and original authors. Although video copy detection can solve such problems, the detection results are easy to be tampered with, and the detection results are hardly convincing. Based on this, this paper proposes an open, transparent and verifiable video copy detection method, which uses blockchain technology to ensure the transparency and openness of the results. In addition, this method adopts the combination of on-chain and off-chain methods to automatically perform copyright detection by invoking smart contracts on the chain. This mechanism can securely and immutably store video feature values on the blockchain, ensuring the originality of copyrighted works and the ability to verify detection results. Swin-Transformer and deep hashing are used to obtain video features off the blockchain, which can efficiently match the similarity of existing videos. The method of block comparison can greatly shorten the comparison time, which is 1/50 of the traditional comparison time. Experimental results show that this method is very effective in retrieving similar images and detecting the similarity between original and pirated videos.

Keywords Video copy detection · Blockchain · Deep hashing · Deep learning

1 Introduction

Digital copyright is the right of network publishing and dissemination of all kinds of information resources in the digital age. With the rapid development of network information technology and the continuous progress of the information media industry, people's increasing spiritual needs have driven the rapid growth of the digital rights industry. People seek to fulfill their spiritual needs through digital media and online platforms, including entertainment, knowledge acquisition, and information retrieval. Therefore, the digital rights industry needs to strengthen copyright protection efforts to ensure that individuals can obtain legal, high-quality, and verifiable spiritual fulfillment and information experiences in the digital environment. The market space is huge, and the development prospects are also very broad. According to the "Research Report on the Development of China's Online Audio-visual Industry" released on June 2, 2021, the market size of short video in the network audio-visual field accounted for the largest proportion in 2020, reaching 205.13 billion yuan, an increase of 57.5% year-on-year. By the end of 2020, the user utilization rate of short videos in China's online audio-visual users is 88.3%, accounting for nearly 90% of all Internet users [1]. It can be seen that short video has occupied a very key strategic position in the field of Internet audio-visual in China. At the same time, with the help of high and new technologies such as big data and artificial intelligence, short video has rapidly spread in the fast-paced life with its characteristics of fragmentation, and has rapidly occupied a large area of network video.

However, with the rapid increase of video sharing and publishing activities on the Internet, video processing tools are gradually popular, and video processing methods are diverse, which leads to video infringement acts continue to

Xintao Liu and Miaoran Song contributed equally to this work.

✉ Xintao Liu
1193781696@qq.com

[1] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China

occur. At present, the protection of original video content has become a high-risk and arduous challenge [2]. Video infringement act refers to packaging others' works as their own or taking others' works directly for use without the permission of copyright owners. Compared with the original video, the infringing video may undergo some processing, such as hiding it in a long video, changing its color, adding black box, etc., which brings great difficulties to the video copyright protection. With the edge-cloud technology, short videos can be transmitted faster, with less delay and higher quality, and people can browse short videos smoothly anytime and anywhere. This also makes the infringing videos spread more widely and in greater quantity, which brings more serious damage to the copyright owner. Therefore, there needs to be a way to detect infringing videos and prevent the spread of such videos.

Video copy detection is an important means to protect video copyright. It is used to ensure the uniqueness of video copyright. It can protect the legal rights and interests of the copyright owner, and prevent others from abusing the copyright content. Video copy detection technology is mainly used to detect the degree of similarity between the copyright video and the detected video, as a pre-screening for detecting pirated videos. If you want to get more accurate results, it is better to need manual final judgment. The detected video generally refers to the similar video obtained after processing the original copyright video. There are obvious processing methods such as scaling, color matching, out-of-order and mirroring, and there will also be some non-obvious processing methods, such as changing frame rate and resolution, which will bring certain difficulties to video copy detection [3].

Video copy detection technology also has a long history of development. The traditional video copy detection method is based on watermarking at first. Digital watermarking is to embed additional information into the original digital content for dissemination, and then the watermark information can be extracted from the video and the copy video can be detected [4]. However, this method cannot detect videos that have been released without watermarking, and digital watermarking will damage the video content to some extent, so digital watermarking cannot solve the problem of video copy detection.

Secondly, content-based detection methods begin to appear. The traditional method of extracting video features is used firstly. These methods use SIFT (Scale-invariant feature transform), HOG (Histogram of Oriented Gradient) and other techniques to extract features from video frames and images. Then, similar images are matched to obtain the video copy detection results. There is also the use of hash encoding to encode video, generate video fingerprints, and perform video copy detection [5].

However, traditional methods of extracting video features are not convenient and robust. With the development of deep learning, with the emergence of Convolutional Neural Networks (CNN) and Vision Transformer models (ViT), these networks have stronger ability to express images and are more convenient to extract video features. Based on this, many researchers use deep learning networks to process video frames and images, and then perform similarity retrieval to detect video copies [6–9]. In addition to extracting the spatial features of the frame image, the temporal features are further added for auxiliary detection, which has been frequently used in recent years [10].

These methods have high computational cost, high complexity and low retrieval efficiency. At present, deep hashing is widely used in image retrieval. It can easily generate compact binary fingerprints, and use inverted index to improve retrieval efficiency. Deep hashing is a combination of deep learning networks and hashing. In order to reduce the overhead of deep learning networks, this paper chooses the Swin-Transformer model [11], which is an improvement of ViT. It adopts a hierarchical structure, greatly reducing the computational complexity of high-resolution images.

Although these methods bring great convenience to video copy detection, they can solve some video copyright problems. However, because the detection results are important to the copyright owners and may involve important legal issues such as whether to file a lawsuit, the final detection results must be open, transparent and convincing. In recent years, blockchain technology has emerged, and information sharing and transparency are its characteristics. It relies on a global P2P network rather than a central trusted authority. Each node on the blockchain can store complete blockchain data locally. The users can view the data on the blockchain

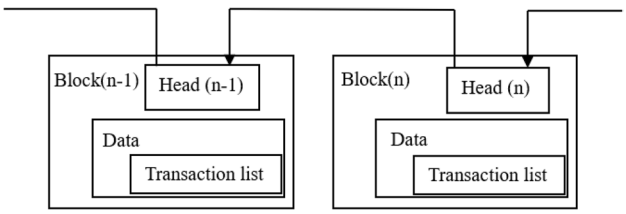

Fig. 1 Block structure of blockchain

at any time, and the transparency of operation is guaranteed [12]. At the same time, the information on the blockchain can also be protected from tampering. The structure of the blockchain is shown in Fig. 1. Its chain structure is formed by including the hash value of the previous block in the current block header. Since the data recorded on the blockchain contains time stamps, the data cannot be exactly the same, and the final hash value is different. If you want to modify the existing content in the block, the corresponding hash value will be changed. The block will not be connected to the original blockchain, and the users in the block will not be able to reach a consensus on the block. Unless there are more than half malicious nodes in the system, the data in the blockchain cannot be tampered with.

The decentralized and tamper-proof features of blockchain can meet the requirements of video copyright detection, make transactions traceable and tamper-proof, while also ensuring authenticity and security. The smart contract on the blockchain can be used to detect the infringement of video copyright after uploading it to ensure the uniqueness and legitimate rights and interests of video copyright. The video copy detection results are open and transparent, making them more credible and verifiable. Therefore, it is an effective solution to combine blockchain technology for video copyright detection.

Based on this, this paper proposes a video copy detection method combined with blockchain, which consists of on-chain and off-chain operations. On-chain refers to operations that occur on the blockchain. The smart contract is an automated contract that contains predefined rules and logic. Once deployed on the blockchain, it becomes an immutable program, increasing the credibility of copyright detection. In this paper, the logic executed by the smart contract is as follows: when a user uploads a video, the smart contract automatically detects the similarity between that video and the videos in the library. If no similar video is found, the authentication is passed, and the video's feature information is permanently stored on the blockchain. If a similar video is detected, the authentication fails, and the video is discarded, preventing it from being uploaded. Due to the limited resources that can be stored on-chain, the stored information is a collection of video frame feature fingerprints extracted by the proposed model in this paper. This part of the work is completed through off-chain operations, using the method proposed in this paper, which combines Swin-Transformer and deep hashing. This not only ensures the openness and transparency of the detection results, but also save storage space. The contributions of this paper are as follows:

(1) A video copy detection model combining blockchain technology and deep learning technology is proposed;
(2) In the process of video copy detection, an on-chain comparison process is designed, which can not only realize automatic detection but also record the results on the blockchain.
(3) Finally, the similarity results between the original video and the tampered video are compared to prove the effectiveness of the proposed method in this paper.

2 Related Work

(1) Video copy detection

Video copy detection technology has a long history of development, and many scholars are studying it. In the early stage of traditional video copy detection, the watermarking methods are mainly used to embed additional information before content distribution, which can detect illegal distribution of content. Chongtham et al. propose an invisible video watermarking algorithm based on discrete wavelet transform. The binary watermarks are embedded into high-frequency coefficients of video frames, and the traditional manual features are used to resist rotation attacks [4].

However, watermarking technology can reduce the quality of content and its robustness is insufficient. Therefore researchers later use a more robust video fingerprint method, which has higher discriminability and robustness against various distortions, and has been widely studied. Video fingerprints are extracted mainly in the time domain, spatial domain and spatial–temporal domain. The spatial domain includes local and global features. Ozbula et al. propose an improved method for detecting pirated video content by combining traditional features with ORB feature descriptors. This method extracts local features from frame images to obtain a compact and effective representation, and then performs further query matching. However, this method cannot recognize brightness changes [13]. Himeu et al. combine Invariant Color Descriptor (ICD) and Binarized Statistical Image Features (BSIF) methods based on ring decomposition to construct a global invariant color descriptor that is robust to geometric attacks such as rotation and flipping, and applied to video frames [14]. However, this method has

no distinguishing ability and needs further improvement. Wang et al. combine content and temporal information to analyze video key frames and convert them into binary code that is easy to store and compare. According to the temporal relationship, the images near the key frame are divided into two groups. The similar information is extracted as the main information of the key frame. Based on this, a binary code representing the effective information in the key frame is obtained. Before matching, key frames are first projected into different buckets using position sensitive hashing technology. Hamming distance metric is used to calculate the distance between binary codes in the same bucket at the sequence matching stage [15]. However, it is difficult to find effective frames that can represent key video information. Lee et al. introduce a video copy detection method based on combined histogram of orientation gradient features and an ordinal metric representation of frames. The Histogram of Oriented Gradient (HOG) feature is used to describe the global features of frames in a video sequence. The ordinal measure histogram is used to generate the feature vector of the whole video sequence as the temporal feature, which is robust to color shifts and size changes [16]. However, there is a trade-off between robustness and discriminability in this method.

In recent research, there has been a tendency to combine deep neural networks, such as CNN and Transformer models. Zhang et al. perform key frame retrieval by extracting deep CNN features [6]. Firstly, deep CNN features are used to encode the image content, then Euclidean distance is used to search for video copy candidates. Finally, they use a graph-based sequence matching method to process the detection and localization of copied video. However, the detection effect of this scheme is not very good when adding black edges to the original image in addition to using CNN to encode video key frames to obtain a similarity matrix. Han et al. use their proposed CNN structure to learn two matrices from the similarity matrix, which are used as similarity measures for the time dimension and the indication map of the segments of the time sequence. They adopt a self-supervised learning method to obtain good results [7]. Tan et al. use CNN feature extraction and KNN to select candidate videos, and then perform video copy detection in these videos to improve detection efficiency [8]. He et al. use attention mechanism of transformer to enhance features and improve the accuracy of video copy detection by capturing the temporal correlation [9].

(2) Blockchain and video copy detection

In recent years, more and more scholars have begun to study the copyright protection technology based on blockchain. Guo et al. have built a blockchain-based digital rights management system for online education multimedia resources. This system combines public chain and private chain, mainly designing registration, secure storage system of digital certificates and non-intermediary verification [17]. Cerba et al. have improved the blockchain structure, and combined digital watermarking technology with an extensible blockchain model to construct a media transaction framework for distributed digital rights management, allowing only authorized users to use online content and provide original multimedia content [18]. Zhai et al. propose a blockchain-based digital rights certificate storage system model, achieving privacy protection for users' real identities [19]. Zhang et al. apply blockchain technology to digital music copyright management, providing copyright proof and originality proof for music copyright [20]. Hu et al. extend the block structure, mainly carry out copyright review and authentication in the field of text works, and can fully store 100,000 levels of text work content, which has confirmed the feasibility[21]. Yang and Yu propose a video copyright storage system that combines blockchain and facial expression recognition, using CNN to recognize facial expression in videos, and recording these features to represent videos to save storage space [22].

At present, most researchers focus on the authorization management of the system, the expansion of blockchain architecture, and the exploration of privacy, information security and other aspects. In all the scientific research content, there is rarely a key link, namely, the copyright detection part of copyright protection. Li et al. further protect the copyright of original works by encrypting and extracting the feature values of digital works and storing them on the blockchain, using the tamper-proof characteristics of the blockchain as a standard for copyright detection to be compared [23]. Mehta et al. conduct image copyright detection using perceptual hashing algorithms to obtain image fingerprints and further use smart contracts to automatically detect infringing images, but did not consider video copyright [24]. Zheng et al. propose a video copy detection method that combines blockchain and dual watermarks. The dual watermark algorithm improves robustness while ensuring that the watermark is invisible, and achieves tamper localization [25].

(3) Deep hashing and video copy detection

The deep hash algorithm is originally designed to solve the problem of waste of storage space and low retrieval

efficiency in image retrieval. It represents images as hash codes of a specific length for efficient comparison and retrieval. This paper applies it to video copy detection, using deep hashing to generate fingerprints on key frames of the video to obtain a fingerprint set of the video, thereby obtaining a representation of the video.

The purpose of deep hashing algorithm is to extract the deep level features of an image through the deep neural network, and then map these features from the original space to the Hamming space, finally obtaining a shorter binary code. Deep Supervised Hashing (DSH) is one of the classical deep hashing algorithms that uses a regularizer to generate discrete binary hash codes on the output of a real-valued network after CNN. [26]. HashNet smoothly converts real-valued features into binary codes by using an extension method based on Tanh function. At the same time, it uses weighted cross-entropy loss to maintain the similarity between data when learning sparse data [27]. The GreedyHash method uses the Sign function in the hash layer, and passes the gradient as the identity map of the hash layer to avoid the gradient disappearance phenomenon [28]. The Improved Deep Hashing Network (IDHN) uses cross-entropy loss and mean square error loss to deal with the "hard similarity" and "soft similarity" in multi-label image retrieval, respectively [29]. Central Similarity Quantization (CSQ) optimizes the central similarity between data points according to their hash centers to further increase the saliency of hash codes used for image retrieval [30]. Deep Polarized Network (DPN) uses the polarization loss as a bit-hinge loss, which causes different output channels to be away from zero and increases the high separability between different types of hash codes [31].

Most of these methods are deep hash algorithms combined with CNN. In recent years, Vision Transformer (ViT) [32] has become a hot topic in the field of computer vision. Many researchers have applied vision transformer to computer vision tasks. They have achieved comparable or even better results than convolutional neural networks. Dubey et al. propose a model combining ViT and hash for image retrieval, which achieves better performance than CNN as the backbone network [33]. However, ViT suffers from high complexity and low efficiency when dealing with high-resolution images. Swin-Transformer [11] is an improvement of ViT. Swin-Transformer adopts a hierarchical structure, starting from small image regions and gradually merging neighboring image regions into deeper Transformer layers to build hierarchical representations and extract hierarchical feature maps. This allows Swin-Transformer to better deal with high-resolution images and intensive vision tasks [11].

Based on these, in order to improve the efficiency of video similarity retrieval, this paper adopts a deep hash method to generate compact binary fingerprint features for inverted indexing. Deep learning network chooses to use Swin-Transformer to extract video features efficiently and effectively. In addition, combined with blockchain technology, the copied video can not only be detected, but also the results are permanently stored on the chain to ensure the openness and verifiability of the detection results.

3 Blockchain-Based Video Copy Detection Model

For copyrighted content in the form of video, before applying for copyright authentication, it is necessary to ensure the uniqueness of the copyrighted work, otherwise it may cause infringement. Therefore, in the video copy detection method based on blockchain, a combination of on-chain and off-chain is adopted. Firstly, key frames are extracted from the video content offline, and the video is converted into images. Then, key information of the images is extracted to generate fingerprints. Finally, a smart contract is called on the chain to compare the similarity between video key frames,

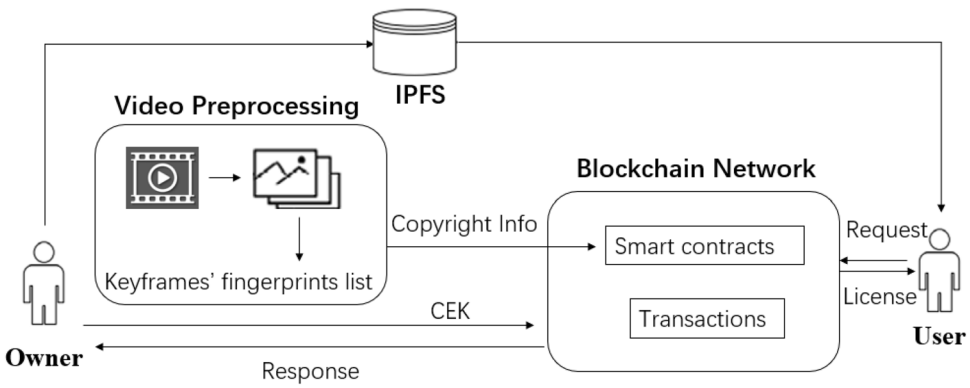

Fig. 2 Blockchain-based video copy model

and then the similarity between the videos is obtained. This model is shown in Fig. 2.

The main components are as follows:

Copyright owner The owner of digital copyright, who can personally control the uploading and trading of copyrighted content, and also needs to respond to user requests;

User The user can query the existing copyright information on the blockchain, request transactions based on requirements, and directly trade with the owner through a smart contract to obtain permission.

Video processing This part is to extract key frames of video, extract features from key frame images, and form key frame fingerprint sequences. This operation is off-chain.

IPFS This part stores original works and feature fingerprints, a decentralized peer-to-peer hypermedia distribution protocol that can be addressed based on content. It uses encrypted hashing to set a unique fingerprint for each file, eliminating redundancy on the network, and uses this fingerprint for information retrieval. By using IPFS, storage space on the blockchain can be saved.

Smart contracts The comparison process of copyright detection mainly relies on smart contracts, which provide operability for blockchains. It is needed for copyright detection when the copyright owner uploads content, and it automatically compares whether the uploaded content is similar to the copyright content of the blockchain.

Blockchain network There are three kinds of nodes in it:

(1) Authentication node: It is mainly used to verify the identity and user requests. After the verification, a series of operations can be displayed;
(2) Recording node: It records video copyright information, authenticated requests, and transaction information that has passed similarity detection.
(3) Agent node: It mainly displays the information about copyrighted works for users to choose and trade, and also serves as promoter to expand the dissemination of copyrighted works.

Video copy detection can be mainly divided into two parts, one is on-chain and the other is off-chain. The off-chain mainly involves video processing, extracting key frames from videos, extracting features from key frames, forming key frame sequences, and then adding them to copyright information and uploading them to the block chain. On the chain, this information needs to be verified by nodes and smart contracts to detect video copyright. If the similarity detection is successful, it will be published on the blockchain. Otherwise, it will be regarded as pirated video and prohibited from being stored on the chain, and feedback will be provided to the publisher. The structure diagram of the

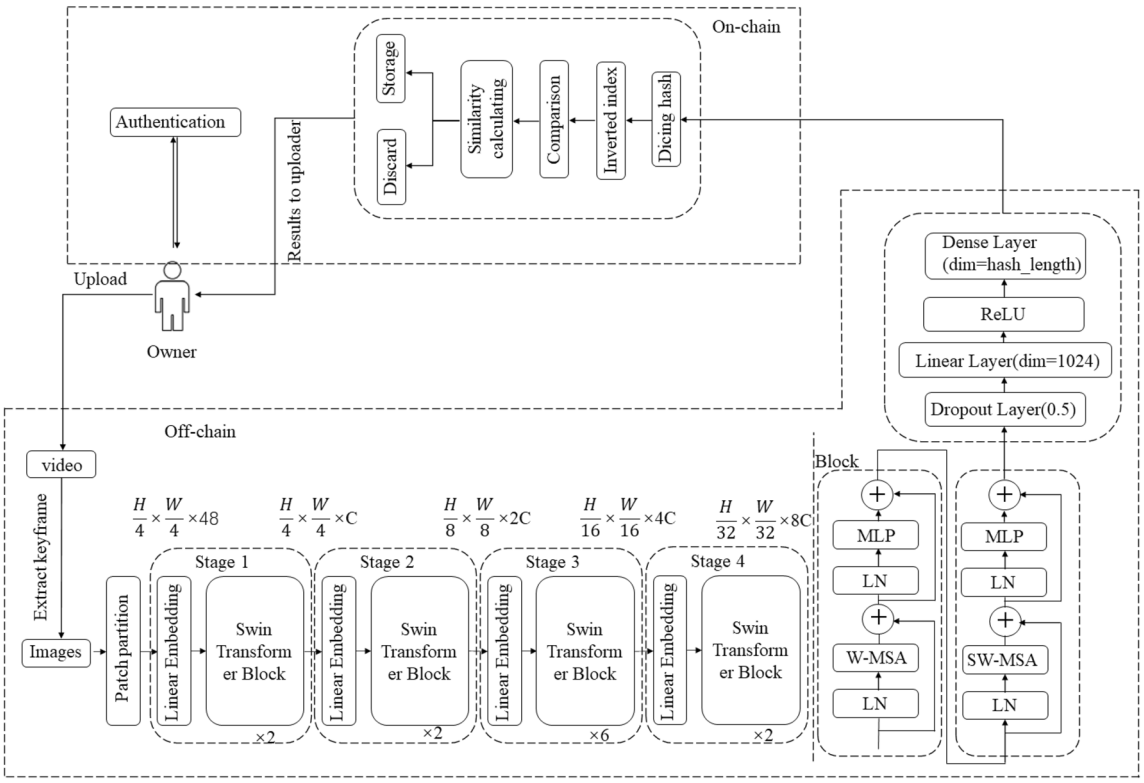

Fig. 3 Video copy detection model

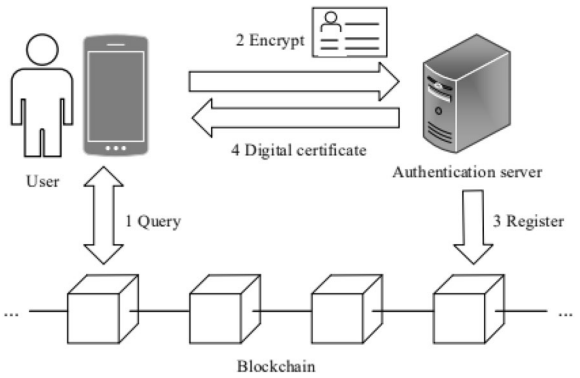

Fig. 4 User identity initialization process

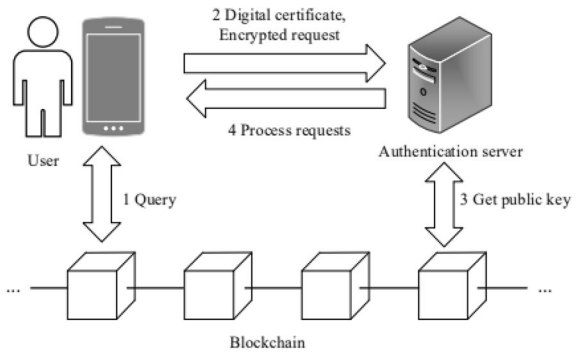

Fig. 5 User identity authentication flow chart

video copy detection model proposed in this paper is shown in Fig. 3. This model will be introduced separately in the following sections.

3.1 On-Chain Detection

3.1.1 Authentication

In order to ensure the security of copyright data, the user's identity needs to be verified before conducting any operations related to blockchain data. Only users who pass the verification can perform other operations.

(1) User identity initialization

The verification nodes in the system are a kind of node on the blockchain, which is not fixed, but will change frequently. The system will store the authentication server ID and its corresponding public key on the blockchain in real time, so as to provide the authentication server ID for users to authenticate and ensure the security of the authentication server.

User initialization is equivalent to the user registering their identity in the system. They need to provide a user name and password to log in to the offline system, ensuring that they are logged in by themselves. Additionally, a series of authentication information is required, and the user generates a public and private key pair locally.

Firstly, the user queries its public key on the blockchain based on the authentication server's ID, and then encrypts the user's identity information using the authentication server's public key. The user then generates a registration request and sends it to the authentication server, which mainly includes the user's public key and encrypted identity information. After receiving the registration request, the server first decrypts it using its own private key, and then verifies the authenticity and validity of the identity information. After the verification is successful, the identity information will use the hash algorithm to get the digital fingerprint of the identity, which is used as the user's identity ID. The identity ID and corresponding public key are stored on the blockchain. At the same time, the verification server uses its own private key to sign the user's identity ID and obtain a digital certificate, which is returned to the user. The user identity initialization process is shown in Fig. 4.

(2) User identity authentication

User identity authentication is equivalent to logging in after user registers. Firstly, users need to send identity authentication requests to the authentication server, mainly including request information signed with the user's private key and digital certificates. Then, after the verification server receives the user's request, it uses the public key to verify the user's digital certificate and obtain the user's identity ID. Then check whether there is a corresponding ID and the user authenticated. Finally, the user's public key is used to decrypt the request information, proving that the information is a request made by the authenticated user. The user identity authentication process is shown in Fig. 5.

3.1.2 Video Similarity Comparison

After the user identity is verified, when the user uploads the video, copyright detection will be performed on the video to determine whether it is a copied video. What needs to be done on the chain is to use smart contracts on the blockchain to achieve automatic matching. The smart contract is publicly stored on the blockchain platform, and the data and rules involved can be viewed, making the information public and transparent. At the same time, all transaction information on the blockchain is publicly recorded, and there will be no other problematic or potential transactions. Therefore, video comparison in the smart contracts can be monitored and the detection results are also authentic and trustworthy.

Algorithm 1: Video similarity comparison algorithm based on dicing comparison

Input：F_A : the video fingerprint set of video A, F_B : the video fingerprint set of video B, *m*: the number of slices, *n*: Hamming distance threshold;

Output：$S_{A,B}$: video similarity.

Cut each fingerprint in FA and FB to m blocks;

F_B ← the more fingerprints，F_A← the less fingerprints;

Build an inverted index to F_B;
Count: *cont*←0;
for *i*←0 to *m*×len（F_A）do
    f_a← get the *i*th fingerprint F_{Ai};
    f_b← find the fingerprint block through the inverted index;
    if f_b can be found do
        if hanming（f_a, f_b）< *n do*
            cont← *cont*+1;
        end if
    end if
end for

Calculate video fingerprint similarity：$S_{A,B} \leftarrow \frac{cont}{\min(len(F_A), len(F_B))}$;

For the fingerprint method of extracting key frames using the previously proposed hash algorithm, achieving similarity comparison requires calculating the hamming distance between perceived hash values. An appropriate threshold is set in advance. If the hamming distance is less than or equal to this threshold, two images are considered to be similar. The fingerprint feature of video keyframes forms the fingerprint feature set of the video, representing that particular video. Then, the video fingerprint set is detected in pairs to calculate the similarity of the two videos. However, the violent comparison method will lead to low efficiency of global comparison. Therefore, in order to improve the comparison efficiency, this paper uses the fingerprint block [34] method to quickly calculate the similarity of two videos.

Firstly, each fingerprint in the video fingerprint set is evenly divided into *m* blocks. If the threshold of hamming distance is defined as *n*, at most *n* small blocks will be different when two fingerprints are determined to be similar, so at least *m*–*n* (*m* > *n*) blocks will be identical. Therefore, according to this rule, *m*–*n* blocks can be randomly selected from the segmented fingerprint blocks in the video to be compared, and an inverted index table can be created to record the position of the corresponding fingerprint blocks. When searching the video fingerprint blocks of copyright videos, if there are exactly the same fingerprint blocks, then further compare the whole fingerprint and finally compare all the fingerprints. If the hamming distance of the final result is not greater than the threshold value n, it is considered to be similar fingerprint; if it does not exist, the comparison of this fingerprint block is skipped and the next fingerprint block is continued. In this way, pre-screening can be achieved after comparing a small amount of information,

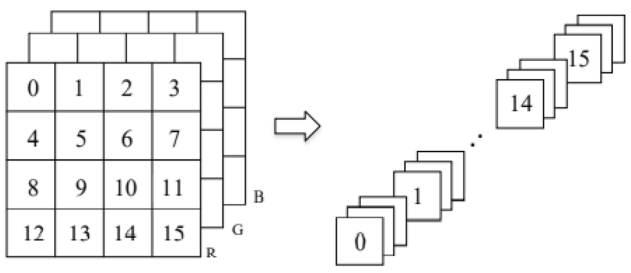

Fig. 6 Transformation of the Patch Partition layer

reducing unnecessary comparison. The final video similarity calculation is obtained from the number of similar frames of the two videos as a percentage of the total number of key frames of the shorter video. The pseudo-code of the video similarity detection algorithm for slice alignment is shown in Algorithm l.

Due to the fact that calculating the similarity between two videos is automatically compared on a smart contract, the first step after uploading video content is to call the smart contract to calculate the similarity between the uploaded video and the video in the copyright database. If the similarity is too high, it needs to be further reviewed manually to improve the accuracy. If the final result is determined to be infringement of the uploaded video, the video cannot enter the copyright management platform, and even after the later design is sound, it is possible to carry out cross-chain operation and submit it to the Internet court—"Tianping chain" for filing.

3.2 Off-Chain Detection

The basic process of video copy detection is to extract the key frames of the video, use Swin-Transformer to extract the features of the frames to obtain the key frame feature sequence. Then use the deep hash algorithm to generate the corresponding binary fingerprints. Finally, the video fingerprint set is compared with the similarity between them by the cut alignment method to obtain the similarity results with other videos.

3.2.1 Video Preprocessing

For video copyright detection, the first step is to extract key frames. In this paper, an inter-frame difference method based on local maximum is used to extract video key frames. The first step is to process the video, divide it into all image frames, and then calculate the difference value of all adjacent images in it, and further calculate the mean of inter-frame difference. Finally, the image representing the local maximum value of the inter-frame difference intensity is found in the set and extracted as the final key frame sequence of the video. All the key frames extracted by this method can contain the key information of the whole video to the greatest extent, and better represent the entire video.

3.2.2 Feature Extraction of the Key Frames

The next step is to extract the features of key frames. In previous work, perceptual hashing was directly used to extract the features of key frames, which can realize video copyright detection. However, this method is sensitive to the content and location of the picture, especially when the picture is mirror transformed, the detection result is not ideal. The model in deep learning can extract the high-level semantic features of the picture, so the deep learning model is used to extract the features of the picture. In this paper, we use the Swin-Transformer model, an improvement of ViT, to extract features of key frames.

The pre-trained Swin-Transformer model is used to extract the effective features of video key frames. The Swin-Transformer model mainly divides image feature extraction into several stages, and each stage gradually reduces the resolution of the input feature map and expands the range of extracted features. Firstly, the input image is divided into blocks in the Patch Partition layer, and these non-overlapping sub-image blocks are used as the input of the stage in the model. The feature of the image block is composed of the concatenated RGB values of the original pixel. The original images are generally $H \times W \times 3$ (high × width × channel number). According to the 4×4 size of each sub-image block, the final feature dimension of all sub-image blocks is $\frac{H}{4} \times \frac{W}{4} \times 48$, which is equivalent to $\frac{H}{4} \times \frac{W}{4}$ sub-image blocks. Each sub-image block feature vector length is $4 \times 4 \times 3$. The conversion process is shown in Fig. 6 ($16 \times 16 \times 3$ input image, for example). The number of images has increased, the size of the images has decreased, resulting in a much smaller amount of computation when calculating self-attention based on image size.

In stage 1, the features of the image after the initial patch partition are mapped through the linear embedding layer to obtain the image features of any channel number C. Then, getting $\frac{H}{4} \times \frac{W}{4} \times C$ image feature dimension specified by the Swin-Transformer block processing. In the following stage, the Patch Merging layer performs two-fold downsampling, that is, the image block will be expanded. Then merging the original sub-image patch with the adjacent sub-image patch to obtain the image patch with double size, and doubling the number of output channels each time, so the output feature dimensions are $\frac{H}{8} \times \frac{W}{8} \times 2C$, $\frac{H}{16} \times \frac{W}{16} \times 4C$ and $\frac{H}{32} \times \frac{W}{32} \times 8C$. In this way, the hierarchical structure of Swin-Transformer is gradually constructed, which can better deal with multi-scale vision tasks.

In the Swin-Transformer block, W-MSA calculates self-attention in specified non-overlapping windows. Assuming that each window is M in width and height, the feature map can be divided into $\frac{H}{M} \times \frac{W}{M}$ windows. According to the original method, the multi-head self-attention is calculated in a single window. But it needs to be calculated $\frac{H}{M} \times \frac{W}{M}$ times in the end, and the multi-head self-attention results are obtained in each window. In order to let the windows interact with each other, using SW-MSA to calculate multi-head self-attention of all the windows. The two improved Transformer blocks in which W-MSA and SW-MSA are connected, which not only reduces the amount of calculation, but also interacts the information between Windows. Then, the calculation relationship between them is shown in formula (1)–(4).

$$\hat{t}^{l} = W - MSA\left(LN\left(t^{l-1}\right)\right) + t^{l-1} \tag{1}$$

$$t^{l} = MLP\left(LN\left(\hat{t}^{l}\right)\right) + \hat{t}^{l} \tag{2}$$

$$\hat{t}^{l+1} = SW - MSA\left(LN\left(t^{l}\right)\right) + t^{l} \tag{3}$$

$$t^{l+1} = MLP\left(LN\left(\hat{t}^{l+1}\right)\right) + \hat{t}^{l+1}, \tag{4}$$

where $\hat{t}^{l}$ and t^{l} represent the output features of W–MSA/SW-MSA and MLP module of the lth block. LN is LayerNorm layer. W-MSA and SW-MSA represent two different multi-head self-attention modules.

In order to further solve the issues of storage space waste and low retrieval efficiency, a depth hash algorithm is used to generate easily comparable fingerprints, and then combined to generate a video fingerprint set, thereby improving the detection efficiency at the video level.

In order to convert the video frame features obtained through the Swin-Transformer model into hash code fingerprints, a hash module needs to be added to the output of the Swin-Transformer model, which can also facilitate the learning of hash codes. Firstly, the image features of the final output of the Swin-Transformer model are dropped by the Dropout layer with a parameter of 0.5. Then, the features are converted to 1024 dimension, then use the ReLU activation function layer. Finally, linear projection is used again to generate the final hash features. The dimension size is the same as the length of the hash bit.

4 Experiment Results and Analysis

4.1 Dataset and Baseline Model

Video keyframe similarity retrieval is the same as image retrieval, so the commonly used CIFAR-10 dataset [35] is used to verify the effectiveness of Swin-Transformer and deep hashing for image fingerprint extraction. We mainly collect 60000 images from 10 categories, of which each category has 6000 images. When training the model, it is divided into 5 training sets and 1 test set. 5000 images are randomly selected from the training set, and there are 500 images for each category. Then, the query set with 1000 images is randomly and evenly sampled from the 10 categories. The remaining 59000 images are used as the database.

Table 1 The parameter configuration of Swin-T

Parameter	Parameter meaning	Configuration
hash_length	The length of output hash feature	64 bit
img_size	Input image size	224 × 224
window_size	Sub-window size(the number of pixel in the window)	7 × 7
patch_size	Pixel patch size	4 × 4
layers	The number of Transformer block per stage	[2, 2, 2, 6]
headers	The number of self-attention headers	[3,6,12,24]
channels	The number of input channels	3
downscaling	Downscaling per stage	[2,2,2,2]

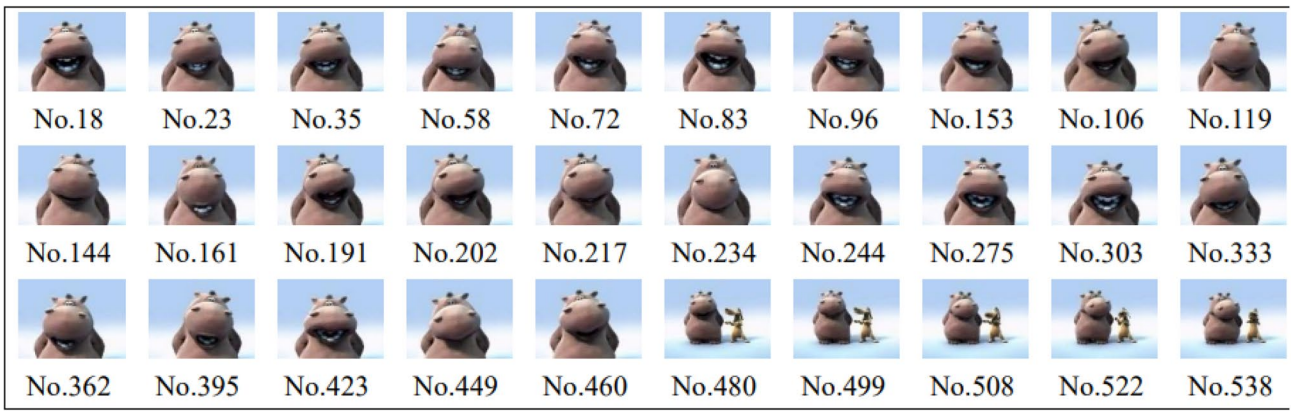

Fig. 7 The first 30 key frames of the original video A

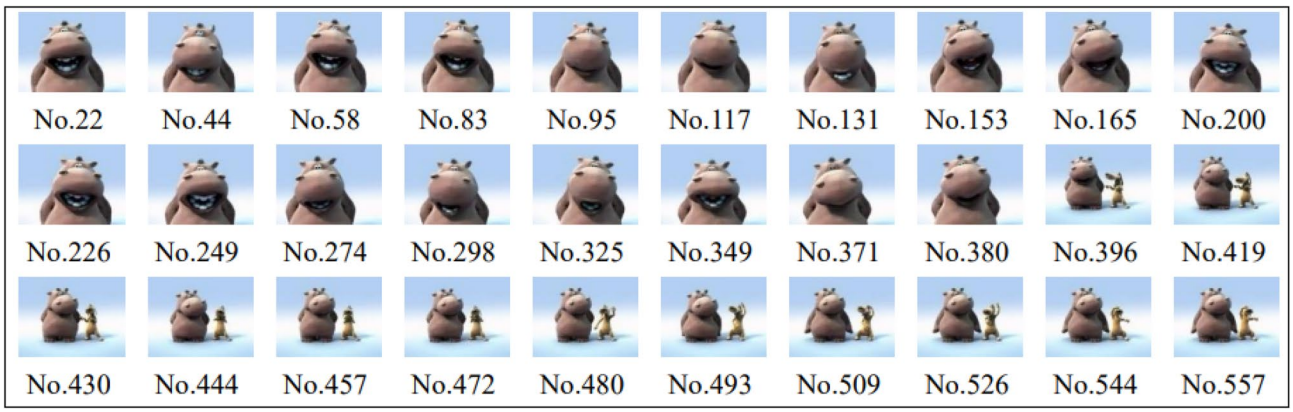

Fig. 8 Processing the first 30 key frames of video A1

Table 2 Hash value comparison

Image	Hash value comparison
Frame 508 of Video A	1,011,000,100,000,000,110,010,000,010,000,010,000,100,001,000,000,000,000,000,000,000
Frame 419 of Video A1	1,011,000,100,000,000,110,001,000,010,000,010,000,100,001,000,000,000,000,000,000,000

AlexNet [36] and ResNet50 [37] as the backbone networks are selected for experiments to show the improvement effect of the proposed scheme.

4.2 Evaluation Metrics and Experimental Configurations

In copy detection experiments, common performance evaluation metrics include Precision, Recall, Mean Average Precision (mAP), and $F1$ score. Precision and recall are a pair, and they can be calculated to produce the F1 score. They are calculated as shown in formula (5), formula (6), and formula (7)

$$\text{Precision} = \frac{N_{\text{positive}}}{N_{\text{detected}}} \tag{5}$$

$$\text{Recall} = \frac{N_{\text{positive}}}{N_{\text{pos-total}}} \tag{6}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{7}$$

where N_{position} represents the right number of results detected, N_{detected} represents the number of all the results detected, $N_{\text{pos-total}}$ represents the number of all positive results of predicted results.

The commonly used mAP indicators from PR curve. PR curve is a 2d curve with Precision as the ordinate and Recall as the abscissa. The area under the PR denotes the Average Precision (AP, Average Precision). It can also be seen as the average value of Precision corresponding to all values of Recall from 0 to 1. As you can see, mAP is simply a global measure of precision obtained by averaging the AP for each class. In summary, in this experiment, the mAP metric and PR plot are used to analyze the performance of different methods.

In this experiment, the miniature pre-trained Swin-T module is used for training, and its corresponding parameters are mainly shown in Table 1. The experiment is trained by the RMSprop optimizer with a learning rate of 1e–3.

In the on-chain part of the experiment, the Hyperledger Fabric blockchain is used to verify the feasibility of the proposed on-chain and off-chain combined video copy detection method. It is mainly to build the environment

Table 3 A Similarity of processed videos

Video A	Zoom B	Tone C	Mirror D	Disorder E	Cut F
Traditional similarity (%)	93.75	90.41	0.0	96.55	97.5
Cutting similarity (%)	97.92	85.71	0.0	97.92	97.5
Traditional comparison time (s)	12.84	11.47	16.65	11.13	9.29
Cutting comparison time (s)	**0.204**	**0.223**	**0.220**	**0.218**	**0.226**

The cutting comparison time is greatly shortened compared with traditional comparison method as shown in bold

and write the chain code of the complete on-chain comparison fingerprint based on native Fabric-SDK-Go to implement a video copyright detection. It is necessary to call the function in the chain code through the business layer for operating on the data state. The experiment is implemented under the software environment of go1.10.3 and python3.7, and the Ubuntu18.04 Linux operating system is set up on the experimental platform.

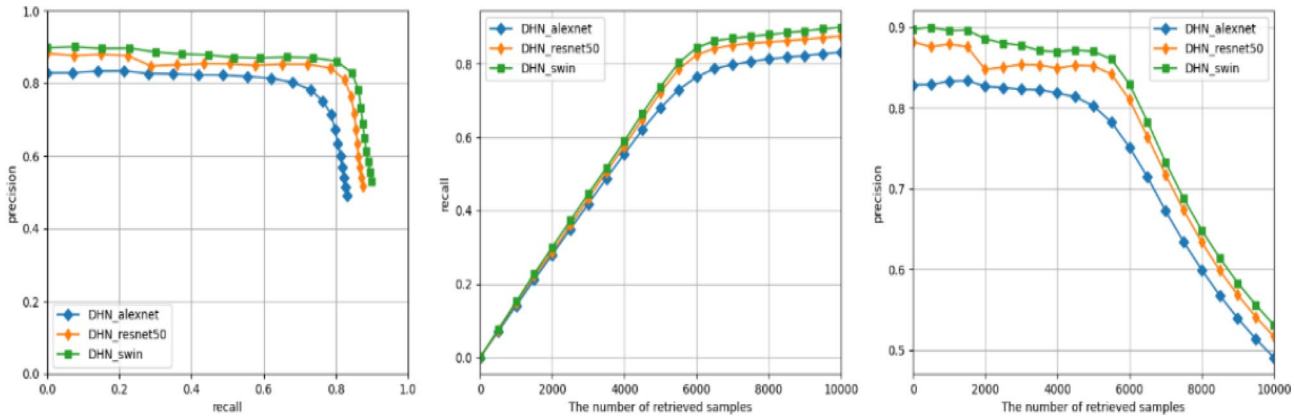

Fig. 9 PR of HashNet on three different backbone networks

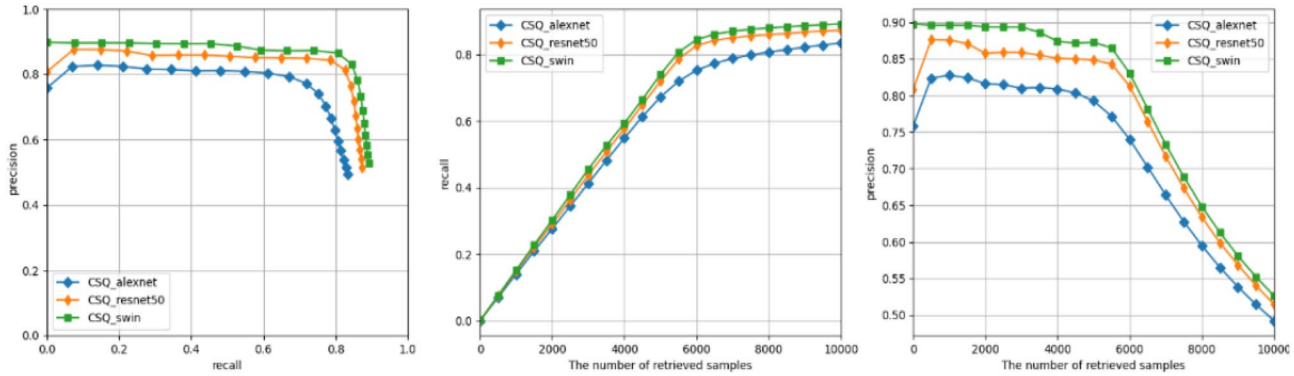

Fig. 10 PR of CSQ on three different backbone networks

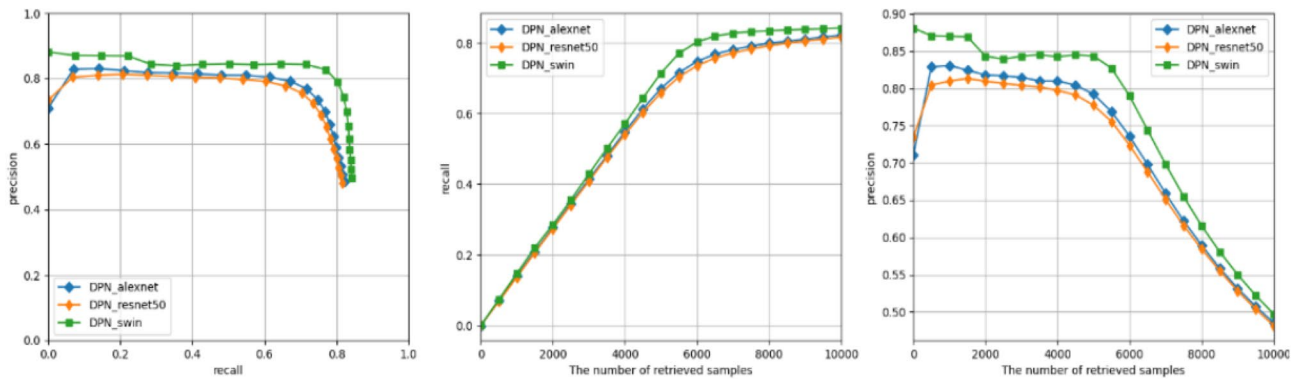

Fig. 11 PR of DPN on three different backbone networks

Table 4 mAP Values of different Deep Hash Algorithms (%)

Backbone	DSH	HashNet	GreedyHash	IDHN	CSQ	DPN
Swin-T	**81.8**	**86.1**	**89.2**	**88.2**	**86.2**	**81.8**
AlexNet	79.2	78.6	79.7	77.9	78.3	77.7
ResNet50	76.9	83.7	87.9	85.6	83.9	78.2

The model Swin-T's performence is all best with different Deep Hash as shown in bold

4.3 Experiment Results and Analysis

4.3.1 On-Chain Video Copy Detection

The original video A with a duration of about 1 min and its' processed video including common processing such as scaling, color matching, mirroring, out-of-order arrangement and cutting into short videos. Video A is captured for 40 s, and its image size is enlarged and its frame rate is reduced to obtain video A1.Video B, C, D, E and F are obtained by zooming, toning, mirroring, scrambling and cutting from the original video A, respectively.

The video is processed off-chain to obtain the key frames of the original video and the processed video. The first 30 frames are taken for similarity detection. The key frame extraction of video A and video A1 is shown in Figs. 7 and 8. The digital fingerprint of each image is calculated by perceptual hash, which is mainly generated according to the low-frequency information of each image. The Hamming distance of the perceptual hash values of similar images is relatively small, which can be used to determine the similar images. For example, it is obvious that frame 508 in Fig. 7 and frame 419 in Fig. 8 are similar images, and their perceptual hashes are shown in Table 2, respectively.

The Hamming distance between two images is 2. Generally, two images can be judged to be similar if the Hamming distance is less than 5. Similarly, the similarity of two videos can be determined by the proportion of the similar number of all key frames to the shorter video. If the number of key frames of two videos is n, the time complexity of the traditional comparison method is O(n2). Using block matching can greatly shorten the comparison time while ensuring the similarity. This comparison method is more suitable on the blockchain. The comparison results of the processed videos are shown in Table 3.

Since the perceptual hashing algorithm is generated according to the content of key frames in the video. The Hamming distance between the perceptual hashes of key frames will be larger than the set Hamming distance threshold in general after the video is mirrored, which leads to the inaccuracy of the final calculated similarity results. It can be seen that the perceptual hashing algorithm cannot detect similar frames well in the face of image content changes.

In the experiment combined with blockchain, in order to obtain accurate results during comparison, the two columns of video copyright ID and fingerprint value are associated in the Couchdb database on the chain to form a composite key. The video ID can be accurately located when the comparison results are obtained. It is convenient for finding the corresponding infringement video after similarity sorting, and further prohibiting it from being added to the copyright library in the blockchain. Although it will be relatively time-consuming to process on the blockchain, this comparison method takes less time, and the comparison process is open and transparent. The final detection result is credible and can withstand the test of other users.

4.3.2 Verification of the Effectiveness of Off-Chain Fingerprint Extraction

This experiment is mainly to test and analyze the performance of this method under six frameworks of Deep Hash. The better performance of deep hash algorithm is selected by comparing the mAP value, which lays a good foundation for the next generation of effective video fingerprint set. The Swin-Transformer model as the backbone of the method is replaced to the same function model AlexNet and ResNet50. The experimental results are analyzed and compared with the two models on PR images to prove the effectiveness of the proposed method.

On the CIFAR-10 dataset, experiments is conducted to compare a variety of scenarios of combining different backbone networks with different deep hashing frameworks. The PR value results and PR maps generated using different backbone networks under the frameworks of HashNet, CSQ and DPN are shown in Figs. 9, 10 and 11.

It can be seen from the above figures that the methods with Swin-Transformer backbone all perform better than AlexNet and ResNet50. This is due to the effective operation of the Transformer architecture for self-attention calculation in the region, which takes more global information into account than the convolution operation, so that it can achieve better results.

On the CIFAR-10 dataset, the mAP results obtained by combining Swin-Transformer, AlexNet and ResNet50 as the backbone to generate 64-bit fingerprint Hash values under six frameworks in Deep Hash are shown in Table 4.

From the results shown in Table 4, firstly, better results are achieved under the same deep hashing framework when the backbone model is the Swin-Transformer. Among them, the maximum average precision of 89.2 is achieved on the CIFAR-10 dataset. Compared with the AlexNet and ResNet50 backbone networks, the mAP values are increased, so the scheme is proved to be effective and excellent.

In the next step, Hamming distance between fingerprints needs to be compared when the similarity of video level is compared. Therefore, CSQ algorithm under the deep hash framework which can better combine with Hamming distance is selected. Therefore, the final video fingerprint generation method is obtained by combining CSQ algorithm with Swin-Transformer. In summary, the hash fingerprint generation method combined with Swin-Transformer is effective in detecting similar images. Therefore, it is further explained that the fingerprint set generated by this method in processing video key frames can also effectively detect similar videos.

5 Conclusion

This paper proposes an efficient video copy detection method combined with blockchain technology, which can effectively detect whether there is a video infringement problem. Blockchain technology is used to obtain credible results and ensure the transparency and openness of the results. The smart contract is invoked on the chain to automatically detect the copyright, which ensures the originality of the copyright work and the verifiability of the detection results. The video feature values are permanently stored on the blockchain without tampering. The Swin-Transformer and deep hash are used to obtain the features of the video off the chain, and the block comparison method is used for the similarity comparison on the chain. Experiments show that the proposed method can retrieve similar images, effectively detect the original video and processed video, and can greatly save storage space and improve the detection efficiency.

In the future, video retrieval can also be combined with the temporal information of the video, and the copy detection can be performed from the video level, not only from the image level, which may improve the accuracy of video copy detection.

Acknowledgements This work is partly supported by “the Fundamental Research Funds for the Central Universities”.

Author Contributions All authors have participated in conception and design, or analysis and interpretation of this paper. LX wrote the first draft. SW and SM read and revise it to the final draft. All authors read and approved the final manuscript.

Funding No funding in any form is received for this manuscript. Also, there are no financial or non-financial competing interests.

Availability of Data and Materials We have used CIFAR-10 dataset [35] which is publicly available.

Declarations

Conflict of Interest The authors declare that they have no competing interests.

Ethics Approval and Consent to Participate The work has not been published elsewhere nor is it currently under review for publication elsewhere.

Consent for Publication Authors provide consent for publication.

References

1. Yuyuan Z (2021) The Ninth China Network Audio-Visual Conference: deepening high-quality innovative development theme discussion. China Radio Film Televis 12:24–27
2. Lina L, Yongming Li (2020) Opportunities, Challenges and Development paths of digital rights protection under Blockchain Technology. Rule Law Res 04:127–135
3. Ling W, Yu B, Li H, et al (2017) Compact CNN based video representation for efficient video copy detection. In: International Conference on multimedia modeling, 2017, pp 576–587
4. Chongtham C, Khumanthem M, Chanu YJ et al (2018) A copyright protection scheme for videos based on the SIFT. Iran J Sci Technol Trans Electric Eng 42(1):107–121
5. Mucedero A, Lancini R, Mapelli F (2004) A novel hashing algorithm for video sequences. In: International Conference on Image Processing. IEEE, 2004, pp 2239–2242
6. Zhang X, Xie Y, Luan X et al (2018) Video copy detection based on deep CNN features and graph-based sequence matching. Wireless Pers Commun 103(1):401–416
7. Han Z, He X, Tang M, et al (2021) Video similarity and alignment learning on partial video copy detection. In: Proceedings of the 29th ACM International Conference on multimedia (MM '21), 2021, pp 4165–4173
8. Tan W, Guo H, Liu R (2022). A fast partial video copy detection using KNN and global feature database. In: Proceedings of

the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp 2191–2199
9. He S, He Y, Lu M, Jiang C, Yang X, Qian F, Zhang X, Yang L, Zhang J (2022) TransVCL: attention-enhanced video copy localization network with flexible supervision. American Aerobics Association International (AAAI), p 2023
10. Wary A, Neelima A (2019) A review on robust video copy detection. IntJ Multimed Inform 8(2):61–78
11. Liu Z, Lin Y, Cao Y, et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp 10012–10022
12. Zheng Z, Xie S (2018) Blockchain challenges and opportunities: a survey. Int J Web Grid Serv 14(4):352–375
13. Ozbulak G, Kahraman F, Baykut S (2016) Robust video copy detection in large-scale TV streams using local features and CFAR based threshold. In: 2016 IEEE International Conference on digital signal processing (DSP), 2016, pp 124–128
14. Himeur Y, Sadi KA et al (2018) Robust video copy detection based on ring decomposition based binarized statistical image features and invariant color descriptor (RBSIF-ICD). Multimed Tools Appl 77(13):17309–17331
15. Rong BW, Hao C, Jin LY, et al (2016) Video copy detection based on temporal contextual hashing. In: 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), 2016, pp 223–228
16. Lee F, Zhao J, Kotani K, et al (2017) Video copy detection using histogram based spatiotemporal features. In: 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2017, pp 1–5
17. Guo J, Li C, Zhang G et al (2020) Blockchain-enabled digital rights management for multimedia resources of online education. Multime Tools Appl 79(7):9735–9755
18. Garba A, Dwivedi AD, Kamal M, Srivastava G, Tariq M, Hasan, et al (2020) A digital rights management system based on a scalable blockchain. Peer-to-Peer Netw Appl 14:2665–2680
19. Zhai S, Chen S, Wang Y (2020) Research on digital copyright storage system model based on blockchain. Comput Eng Appl 56(19):13–21
20. Zhang G, Tang H, Chen J, Shen R, He Q, Huang B (2021) Digital music copyright management system based on blockchain. J Comput Appl 41(04):945–955
21. Hu D, Li Z, Zhou W, Wang J (2021) Digital Copyright authentication model based on blockchain. Computer Appl Softw 38(02):311–317
22. Yang Y, Yu D (2022) Short video copyright storage algorithm based on blockchain and expression recognition. Int J Digit Multimed Broadcast 2022:88278151–882781511
23. Li C, Dai B, Wang H, Wang X (2018) Digital copyright protection and trading system based on blockchain technology. Modern Computer 10:80–84
24. Mehta R, Kapoor N, Sourav S, et al (2019) Decentralised Image sharing and copyright protection using blockchain and perceptual hashes. In: 2019 11th International Conference on Communication Systems & Networks (COMSNETS). IEEE, 2019
25. Zheng J, Teng S, Li P, Ou W, Zhou D, Ye J (2021) A novel video copyright protection scheme based on blockchain and double watermarking. In: Security and communication networks, 2021
26. Liu H, Wang R, Shan S, et al (2016) Deep supervised hashing for fast image retrieval. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2016, pp 2064–2072
27. Cao Z, Long M, Wang J, et al (2017) Hashnet: Deep learning to hash by continuation. In: Proceedings of the IEEE International Conference on computer vision, 2017, pp 5608–5617
28. Shupeng Su, Chao Zhang, Kai Han, and Yonghong Tian (2018) Greedy hash: towards fast optimization for accurate hash coding in CNN. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18) 2018, pp 806–815
29. Zhang Z, Zou Q, Lin Y et al (2019) Improved deep hashing with soft pairwise similarity for multi-label image retrieval. IEEE Trans Multimed 22(2):540–553
30. Yuan L, Wang T, Zhang X, et al (2020) Central similarity quantization for efficient image and video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp 3083–3092
31. Fan L, Ng K, Ju C, et al (2020) Deep polarized network for supervised learning of accurate binary hashing codes. In: IJCAI, 2020, pp 825–831
32. Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations, 2021
33. Dubey SR, Singh SK, Chu WT (2022) Vision transformer hashing for image retrieval. In: 2022 IEEE International Conference on Multimedia and Expo (ICME), 2022, pp 1–6
34. Yue Wu, Jiangtao L, Rui L (2021) Video similarity detection method based on perceptual hash and slicing. Comput Appl 41(07):2070–2075
35. Krizhevsky A, Hinton G (2009) Learning multiple layers of features from tiny images. In: Handbook of systemic autoimmune diseases 2009, 1(4)
36. AK, IS, EHG (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems 2012, 25(2)
37. He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2016, pp 770–778