International Journal of Multimedia Information Retrieval (2019) 8:61–78
https://doi.org/10.1007/s13735-018-0159-x

TRENDS AND SURVEYS

A review on robust video copy detection

Alongbar Wary[1] · Arambam Neelima[1]

Received: 22 May 2018 / Revised: 6 August 2018 / Accepted: 24 August 2018 / Published online: 31 August 2018

Abstract

The unprecedented escalation and proliferation of digital multimedia and Internet technology have triggered the enormous copyright infringement issues and tampering of digital content. Detection or localization of copy–paste forgery of digital content and distinguishing between original and manipulated video have become a weighty challenge at the present era of multimedia technology. Several distortions such as rotation, scaling and gamma correction are applied into an original video by an adversary to manipulate the original video for copyright infringement. Due to the emergence of ubiquitous digital videos on the Internet and to surpass the challenges, various copy detection schemes have been introduced by several researchers. Many real-time applications such as detection of duplicate Web videos and monitoring of real-time TV commercial media content over multi-broadcast channels require the robust copy detection approach for high security purpose. The other applications include the rapid advancement of video navigation and editing technology such as finding the opening sequence of a TV show and combining or editing similar versions of the same video for copyright infringement. This paper depicts a comprehensive overview of robust visual hashing to identify similar video contents for digital piracy detection, which overcomes the demerits of conventional cryptographic hash functions and watermarking. The paramount goal of this scheme is to generate the perceptual hash code of fixed size of length from video segments which are robust against distinct distortions or attacks such as scaling, rotation, compression, frame rate change, frame dropping, contrast enhancement, etc., made by an adversary. Besides, in this paper, distinct state-of-the-art schemes used for copy detection have been studied thoroughly and classified based on the methodology they have implemented.

Keywords Robust visual hashing · Video copy detection · Cryptographic hash · Watermarking

1 Introduction

The video sharing and publishing activities on the Internet are increasing tremendously due to the exponential upswing of multimedia technologies. Protections of original video content by content owners, distributors and publishers have become a high-risk and tough challenge. Detecting and administering the enormous amount of videos which are uploaded every day to the video sharing Web sites such as YouTube, Netflix, etc., is a critical challenge for the owner of the commercial video Web servers. Keyed digital information can be replicated and arbitrarily distributed by an adversary without the consent of the copyright holder. The original video content can be manipulated by an adversary by applying certain distortions such as content-preserving (e.g., lossy compression, contrast enhancement) and geometric (e.g., rotation, scaling) distortions into an original video. Protection and management of highly sensitive digital information have become a critical task. The term copy of a video is a manipulated or transformed video sequence which is similar or less similar but not identical compared to the source video [1]. Due to the advancement of video navigation technology, it has become easier for the users to navigate or find any sequence of a TV show such as finding the opening sequence of a show. Moreover, due to the advancement of video editing softwares or apps such as Final Cut Pro 7 and iMovie, the users can alter the content of a video by combining or editing similar versions of the same video as required, in which the quality of a video might be degraded or improved. It has become hard to detect the near-duplicate video copy of an edited version of the same video in which the quality is improved. Employing the

✉ Alongbar Wary
alongwar56@gmail.com

Arambam Neelima
neelimaarambam@yahoo.co.in

[1] Department of Computer Science and Engineering, NIT Nagaland, Chumukedima, Dimapur 797103, India

fast and robust method for accurate detection of illegal copy or manipulated version of an original video still remains a challenging task. There are many other real-time application areas such as detection of duplicate Web videos [2] and monitoring of real-time TV commercial [3, 4] media content over multi-broadcast channels, where robust video copy detection scheme is mostly required. Still, manual work is involved in such monitoring and real-time performance is very poor. It is indispensable to employ a copy detection scheme that is both discriminative and robust against various distortions such as picture-in-picture, region cropping, scaling, etc., which is yet a challenging research field.

Numerous researches have been done to cope up with the copyright issues. Watermarking-based [5] copy detection is extensively used. Extra information is embedded into the original content of media before it is distributed imperceptibly [6] and can be extracted later to acquire information of the original video content and to identify the copy of a video content to its original in the watermarking scheme. During the entire distribution process, the information is embedded with the media content and can be used to detect illegal distribution of the content. However, the watermarking scheme has some demerits: (1) Contents without watermark such as legacy content which has already been distributed cannot be traced or detected through watermarking; (2) even a minor alteration induced by the watermark degrades the quality of content which is not suitable for some applications such as detection of digital content involving medical images; (3) there is a trade-off between the imperceptibility and robustness. The watermark should be robust to diverse transformation operations on the digital content [7] which is still not adequate for copy detection. Moreover, a conventional cryptographic hash function [8, 9] is also used for the authentication of a digital signature in which message is identified by a constant and short length of bit feature vector uniquely. As the cryptographic hash function operates on the basis of a whole message it is impossible to obtain the identity and check the integrity of a part of the message of a digital signature. In addition, the feature vector generated by a cryptographic hash function changes substantially when the input message changes by a single bit [10].

Considering this entire pitfall, a robust visual hashing (also called digital fingerprinting) is introduced, for the digital rights management of multimedia data [11]. It is an alternative approach for copy detection and avoids embedding operation unlike watermarking. Generally, the perceptual hashing method is popularly used for the purpose of content-based image retrieval, image indexing and image authentication [12]. Later, this method is adopted as an approach for video copy detection [13, 14] which extracts its fingerprints, called the hash value, by analyzing the signal of a video sequence. This value could permit the unambiguous identification of the signal (e.g., human fingerprint with people). The foremost objective of visual hashing-based copy detection is to extract a compact feature or hash code of fixed size of length from video segments (it can be a key frame or whole frame of a video scene) for identification and differentiating the original video from the manipulated one for copyright management, tracking and organizing giant video databases. To preserve the properties of visual hashing, such as (1) uniqueness; (2) compactness; (3) robustness, a short robust hash code [15] is extracted from video segments and matched using a distance metric for identifying the pirated content. Coskun et al. [16] have come up with a new idea of visual hashing in which both spatial and temporal domain is considered as video frames contain motion information across time and is robust against certain distortions such as content-preserving (e.g., contrast enhancement) and geometric (e.g., frame dropping) distortions. Some literature review paper [17, 18] also elucidated the essence of various video copy detection schemes. Figure 1 roughly depicts the visual example of near-duplicate copy of an original video frame.

This paper is aligned as follows: In Sect. 2, detailed state of the art is discussed on visual hashing-based video copy detection. In Sects. 3 and 4, major challenges and current trends are analyzed and discussed, respectively. Conclusion is presented in Sect. 5.

2 State of the art

Various copy detection methods have been proposed for solving the piracy issues and managing the huge video databases. The visual hashing- or fingerprinting-based copy detection method is more preferable compared to the watermarking-based copy detection method because of their high discriminability and robustness property against various distortions. The visual hashing-based copy detection method extracts the compact hash code or fingerprint that can tell whether a suspicious piece of content matches a multimedia document registered in the fingerprint database. Moreover, unlike watermarking approach, hashing or fingerprinting approach can be applied to the legacy content (content that has already been distributed) of a digital media [14, 15]. Figure 2 represents an overview of the working principles of hashing- or fingerprinting-based video copy detection approach.

In this section, various existing visual hashing-based video copy detection methods are discussed. The methods are classified based on the domains on which the hash codes (digital fingerprints) or feature vectors are extracted, i.e., spatial domain, temporal domain and spatial–temporal domain.

2.1 Based on spatial domain

In spatial domain, the hash code or feature vector is extracted from each key frame [19] or each frame [20] of a video. Spa-

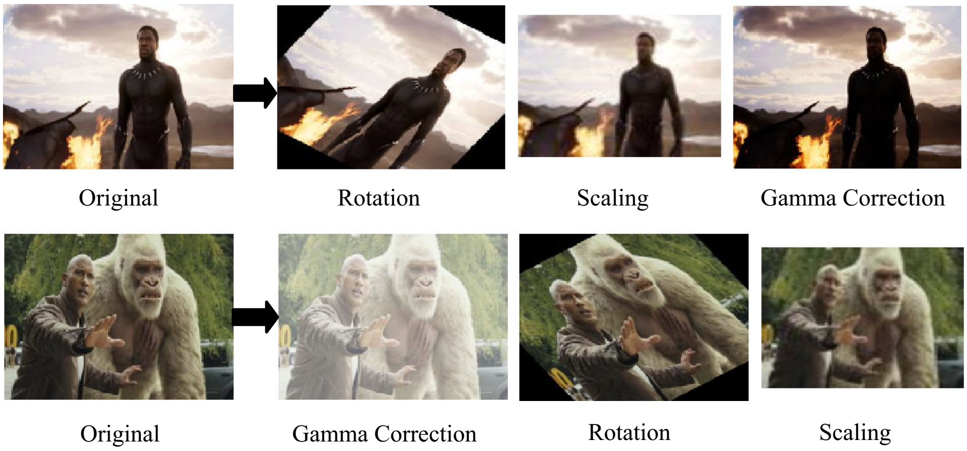

Fig. 1 Visual example of near-duplicate copy of an original video

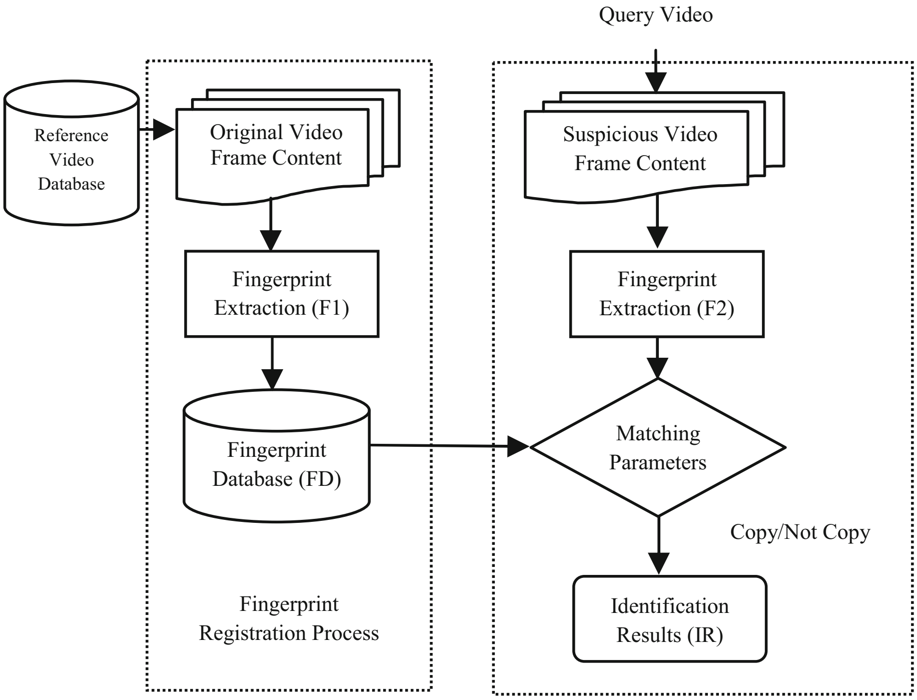
Query Video
Reference Video Database
Original Video Frame Content
Suspicious Video Frame Content
Fingerprint Extraction (F1)
Fingerprint Extraction (F2)
Fingerprint Database (FD)
Matching Parameters
Copy/Not Copy
Identification Results (IR)
Fingerprint Registration Process

Fig. 2 Flowchart of hashing- or fingerprinting-based video copy detection approach

tial feature plays an important role in video copy detection and identification as it can locate the salient points either locally or globally within its spatial space and is robust against common video processing steps [21] such as lossy compression, resizing, frame rate change, etc., as well as geometric attacks (e.g., scaling, rotation) [19]. However, the identification of key frames which will represent the video efficiently is an important issue in spatial domain [19] and it utilizes a large memory space for operating the vast video databases [22]. The time and frame difference (temporal information) [23] which are the salient properties of video frame are not considered in spatial domain.

Further, the methods can be classified based on local features [24], global features [25], coarse features [26] and local–global features [27], according to the features extracted in spatial domain.

2.1.1 Methods based on local features

Here, local descriptors are extracted to form a compact hash code vector such as extraction of the local interest points [24, 28] of the frame image. Local features exhibit high discriminative capability; however, it is less sensitive to global changes [29]. Neelima and Singh [19] introduced a scale-invariant feature transform (SIFT)-based local feature descriptor which is invariant to scaling, rotation and translation applied on the video frame sequence. The invariant key points were first extracted using SIFT descriptor from each key frame that was selected from video frame sequence and then clustered into 32 different clusters. Based on the cluster centroid, thirty-two distinct blocks of pixels of size $m \times n$ were generated. Finally, the maximum singular value was extracted as a feature vector by applying singular value decomposition (SVD) on each block. SIFT descriptor is partially invariant to illumination changes and affine transformation which is its drawback.

The interest points were detected using difference of Gaussians (DoG) method in [24], where Gaussian kernel is a scale-space kernel candidate. It detects the repeatable key points so that its pixel location is detectable even after geometric attacks such as scale change, rotation, etc. They have used a cascade of subsampled images (multi-resolution) and filters called octaves. This method is a good approximation of the normalized Gaussian Laplacian with respective variance $\kappa \cdot \sigma$ and σ:

$$G(u, v, k\sigma) - G(u, v, \sigma) \approx (k-1)\sigma^2\nabla^2 G. \quad (1)$$

The difference of Gaussians (DOG) image was obtained by convolution of (1) with the frame image as follows:

$$D(u, v, \sigma) = (G(u, v, k\sigma) - G(u, v, \sigma)) * I(u, v), \quad (2)$$

where $G(u, v)$ is the Gaussian image and $I(u, v)$ is the frame image of pixel coordinate (u, v). In each octave, they have used an initial scale factor σ of value 1.6 and a multiplicative factor κ of value 1.15. The location of the points of interest was represented by the extrema of the DOG. Only the points with good precise space localization and good contrast were kept [30]. To reduce the storage space and enhance the performance, the interest points were extracted from a key frame [31]. To compute the local descriptor for characterizing each key point, they have applied a circular neighborhood of radius $\Re$ (e.g., 20) which is invariant to rotation and computed the key point orientation by summing the gradient vectors in a small disk around the key point. The neighboring disk was divided into nine regions using this orientation, and the local histogram of sixteen bins was computed to represent the local descriptor. However, this method is not robust against the global changes such as color variation and incurs high computational cost.

In [28], the proposed method works in a similar way as in [24]. But, the only difference was that they have used an enhanced version of the Harris detector [32] based on a generalized random transform that is invariant to rotation and scale changes. For each interest point, the local description of the region of interest was computed and a distance similarity metric was used that fuses the geometric information and intensity to compare the key frames extracted using a scene detection algorithm. The geometry of interest points was captured as follows.

Let $p_i = (x_i, y_i)$ be the ith key point in the frame image, and a weighted average of the separation vectors $\xi_{ij} = p_i - p_j$ was calculated for each point p_i as follows:

$$\xi_i^m = \frac{1}{k-1} \sum_{\substack{j=1 \\ j \neq 1}}^{k} \omega\big(|\xi_{ij}|\big)\xi_{ij}, \quad (3)$$

where k is the total number of interest points in the frame image and $\omega(\cdot)$ is a monotonically decreasing function on $\Re^+$. However, the experimental result shows that the method is not invariant to rotation and global variations.

Li et al. [33] also came up with the concept of extracting the interest points from regions of interest (ROIs) using a new method called FREAK (Fast Retina Keypoint) which is robust against scaling, rotation and noise. Initially, ROIs were extracted using thresholding and morphological merging technique. Later, FREAK points were extracted from each ROI and were normalized to reduce the effects caused by slight inaccurate extraction of ROI as given below:

$$\text{nf} = F(\text{Glf}), \quad (4)$$

where Glf represents the feature vector of current frame, $F(\cdot)$ is a function to normalize the parameter, and nf is the normalized feature vector. Subsequently, FREAK points were clustered by spectral clustering to reduce the redundancies of features at the same shot and were used as fingerprints. In order to be invariant to rotation, Zhang et al. [34] have introduced a method based on speeded up robust features (SURF) which extracts the local features from the frame image and identified a reproducible orientation for the interest points. In addition, locality-sensitive hashing (LSH) was applied for indexing to enhance the performance of the copy detection which generates the compact hash code by projecting similar hash code into a hash bucket. The high computational cost is the drawback of this method and is also less sensitive to global changes such as color variation. Similarly, the SURF [34] method was also used in [35], in addition to oriented FAST and rotated BRIEF (ORB) [36] method to detect the pirated video content. This method cannot handle the illumination changes.

2.1.2 Methods based on global features

The global features of video sequence were extracted using centroid of gradient orientations (CGO) descriptor proposed in [21]. In this method, each resized frame was partitioned into a grid of $m \times n$ blocks and then, CGO was calculated for each block resulting in (mn)-dimensional feature vector. The proposed method is pair-wise independent (not dependent between unrelated video segments) and robust against certain distortions such as lossy compression, frame rate change, etc. However, the global feature descriptors are insensitive to local changes which lead to discriminability issue [29]. Moreover, the method is not robust against general geometric transformations such as rotation, shift, cropping, etc., which needs to be enhanced.

For computing the similarity between images, the ordinal measure (OM) was first introduced in [25] and then extended to video in [37] for copy detection in which the ordinal measure (global feature) was computed from all the N blocks of each frame image, and then each block was sorted according to their average gray level (ranking). The OM $M(t)$ of the tth frame was presented as:

$$M(t) = (R_0, R_1, \ldots, R_{N-1}), \tag{5}$$

where R_i is the rank of the ith block. Generally, the ordinal measure is robust against the transformations such as noise, filtering, recompression, etc., that are applied to the whole frame. However, it cannot survive the local transformations such as logo insertion, cropping, shifting, etc. The method proposed by Yang and Li [38] works in a similar way as in [21] in which visual features were extracted from each $m \times n$ block of each frame using the gradient orientations of luminance centroid. The method is robust against the common video processing steps, but not robust against the geometric distortions.

A copy detection scheme based on quadrant of luminance centroid was introduced in [39] by Uchida et al. in which each frame was divided into 4×4 blocks $b_i (1 \leq i \leq 16)$ and the coordinate of the luminance centroid $\left(x_i', y_i'\right)$ was calculated for each block as follows:

$$x_i' = \frac{\sum_{(x,y)\in b_i} x \cdot I(x,y)}{\sum_{(x,y)\in b_i} I(x,y)}, \quad y_i' = \frac{\sum_{(x,y)\in b_i} y \cdot I(x,y)}{\sum_{(x,y)\in b_i} I(x,y)}, \tag{6}$$

where $I(x, y)$ is the luminance of a frame image at coordinate (x, y). Subsequently, a block-level luminance centroid was binarized into a 32-bit quadrant feature and a stable key frame was selected to enhance the pair-wise independence between unrelated video segments. Finally, stable features were compared using adaptive mask. However, this method is not invariant to strong local variations such as cropping, frame shift, etc.

To improve the robustness against rotation and flipping attacks of the work proposed in [40], the Himeur and Sadi [41] have combined the binarized statistical image features (BSIF) local texture descriptor and local color descriptor using weighting parameters to obtain the global descriptor. BSIF histogram was computed from all the rings of each BSIF frame image, and the histogram over hue for every decomposed patch of each frame was computed from the corresponding RGB values of each pixel. This method is not robust against other transformations such as cropping, pattern insertion, etc. They came up again with a new approach in [42] to enhance the performance in which the same ring decomposition-based BSIF [41] method was adopted in addition to invariant color descriptor (ICD). To construct an invariant color description which is robust against the geometric attacks such as rotation and flipping, ICD was applied to the video frames. The method has less discriminative capability that needs to be improved further.

2.1.3 Methods based on coarse features

In this category of methods, the coarse features are extracted to represent the video content, such as extraction of features by nonnegative matrix factorization from each key frame [31], attention region representation by saliency map [43, 44], discrete cosine transform coefficients [26, 45, 46], mean luminance comparison between two adjacent subregions of a ring [47], extraction of contourlet coefficients [48]. The accuracy of detection is not so good, since the coarse features can only identify an approximate representation of the video content [49]. The authors in [31] came up with a new scheme called nonnegative matrix factorization (NMF) which extracts the perceptual fingerprints from each key frame via Gaussian weighting. The transform-invariant NMF (T-NMF)-based video indexes were integrated with the proposed scheme to assure robustness and compactness against geometric attacks and global luminance changes. However, video copy detection based on only key frames cannot yield temporal localization [23] and the method can incur high computational cost.

In [43], the coarse representation of the feature vector was extracted from the visual attention regions which were represented by saliency map. The unique saliency map was formed by combining the normalized visual feature maps such as color maps, intensity maps and orientation maps which were computed from the input frame as given below:

$$S = \frac{1}{m} \sum_{i=1}^{m} N(X_i), \tag{7}$$

where $N(\cdot)$ is normalization function, X_i represents a feature map, and S is the combined map. Then, the saliency map was partitioned into a grid of m rows and n columns, resulting in $m \times n$ blocks, and the average saliency value of each block was calculated. Finally, the coarse representation of the saliency map was adaptively quantized to a binary vector as the proposed video feature vector or fingerprint. The bottom-up approach was used, which avoids the effect from the top level of human visual system (HVS). This method is robust against content-preserving distortions, but not robust against the geometric distortions. The method introduced in [44] was based on the similar concept used in [43], but the difference was in the use of self-information-based method to create the visual saliency map. In this method, the saliency of a location was quantified by the self-information of an $m \times n$ local image patch which was centered on that location and then introduced salient covariance matrix (SCM) descriptor as a robust and compact feature descriptor for video copy detection. The high computational cost and less discriminative capability are the main disadvantages of this method.

The discrete cosine transform (DCT) used in [26, 45] for extracting DC coefficients of luminance component in each block of frame was later adopted by the authors in [46] with a slight improvement. They have used color layout descriptor (CLD) which is a robust and compact frame-based descriptor that captures the frequency content in a highly coarse representation of the frame image. The CLD feature was obtained by converting the frame image to an 8×8 image along each (YC_bC_r) channel on average. DCT was computed for each frame image, and the DC along with the first five AC coefficients (in zigzag scan order) for each channel was selected to form an 18-dimensional CLD feature vector and was further encoded by vector quantization (VQ). Significant gamma variation and cropping can, however, distort the CLD adequately to cause errors, which is the main disadvantage of this method.

A texture descriptor called region binary pattern (RBP) was proposed by Kim et al. [47]. The method extracts the two complementary region binary patterns from subregions of several rings of a key frame to preserve the spatial structure and is robust against the rotation and flipping. A key frame was divided into several rings, and each ring was further divided into subregions from which the RBPs were extracted. The first (intra-type) RBP represents a binary pattern in a single ring, and the second (inter-type) RBP was computed from a relationship between adjacent rings by calculating the mean luminance of subregions in a ring. The spatial distribution is not considered because it is not invariant against the global changes and other transformations such as frame dropping, logo insertion, etc.

The contourlet transform hidden Markov tree (CHMT) model was proposed by Sun et al. [48]. In this method, each resized frame of a video was partitioned into a grid of $m \times n$ blocks. Subsequently, each block was transformed into contourlet coefficients, and then the standard deviation matrices of the CHMT model were extracted as the intermediate feature. Finally, SVD [19] was applied to reduce the dimension of the standard deviation matrices and the largest singular value of each matrix was taken as the feature vector. Using few parameters, the CHMT model can capture all inter-scale, inter-direction and inter-location dependencies of the counterlet coefficients and is robust against common content-preserving operations such as lossy compression, filtering, etc., but not robust against the geometric attacks (e.g., frame dropping, rotation, etc.).

2.1.4 Methods based on local and global features

In this class of methods, both the local and global features are extracted from the spatial domain of the video frames to preserve the robustness and discriminability properties of the feature descriptors, which is the main issue faced while using only local features and global features, respectively. To meet both the robustness and discriminability properties, the authors in [27] introduced a method in which they have used similarity-preserving multimodal hash learning (SPM^{2}H) for generating compact hash code. In this scheme SIFT [19] and pyramid histogram of oriented gradients (PHOG) were used to extract the local features and global features from each key frame, respectively, and then, SPM^{2}H was applied for combining both the features to generate low-dimensional compact hash code with good accuracy. The method used in [20] was also used by Ding and Nie [29] for copy detection with a slight difference. In this method, interest points were extracted using SURF [34] local feature descriptor from each key frame of the video to reduce the computational cost. Then, each key frame was divided into an equal-area circle ring based on the center point which can be found from the interest points and key frame boundaries. Further, each circle ring was divided into an equal-area sector, and from each sector, the ordinal measure (global feature) [37] vector was computed, which was taken as a fingerprint. The authors in [50] have adopted the SIFT [27] descriptor for local feature extraction to estimate the copy transformation, and then, ordinal measure [37] was used as a global feature to accelerate the copy detection subsequently. The random sample consensus (RANSAC) algorithm was used to estimate the affine transformation that was used to map the points in query frame to those in its reference frame. Subsequently, the mismatched local feature points were removed by the same algorithm. Chiu et al. [51] also adopted the same method as in [50] along with the segment-based similarity matching technique for copy detection. Computational complexity is the main demerit of these methods.

To detect the copy–move forgery (CMF), two texture descriptors known as cellular automata (CA) and local binary pattern (LBP) were introduced by Tralic et al. [52]. The core idea was that CA learns a set of rules for all the overlapping blocks of each frame which describes the intensity changes in every block appropriately, and then histogram of rules was created, which can be used as a feature vector for forgery detection. Binary representation of feature vector was obtained by using LBP locally to every neighborhood of each block which has led to a remarkable reduction in the number of possible rules before histogram was created. The method is unable to detect when a part of a frame is being copied and pasted to a different frame in the same sequence as it has considered the whole frame. In the method proposed in [53], ORB [36] descriptor was used to extract the local binary features vector from each key frame, whereas color correlation histogram and key frame thumbnail were introduced to extract the global features vector for copy detection. To select a matched video, the corresponding features vector similarity was evaluated in an intuitive voting system which requires at least two matched feature vectors. ORB has a good performance at low cost compared to SURF and SIFT [36]. However, ORB descriptor shows less resistance against image distortion, illumination changes and changes in scale.

2.2 Based on temporal domain

The methods based on temporal domain extract the visual features or hash values between two consecutive video frames in the temporal direction [34]. It is obligatory to consider the temporal information which is linked with video frames intrinsically to acquire the frame-level representation entirely, since videos represent motion-based features across time typically. Temporal localization is important for locating actions precisely in time, even when the surrounding frames are visually similar [23]. Some demerits arise due to the use of temporal information only for video copy detection: (1) It cannot be applied to the short-time duration video segments as it is feasible only with the long-duration videos and (2) it is not suitable for online applications that are of short-time duration videos [22]. A global descriptor in the temporal domain was extracted by the method proposed in [1] and used for the fingerprint. In this method, the feature value of tth frame was calculated as a weighted sum of per pixel squared differences of the corresponding t and $t-1$ frames as given below:

$$V(t) = \sum_{i=0}^{N-1} B(i)(I(i,t) - I(i,t-1))^2, \tag{8}$$

where $B(i)$ is a weight function to improve the significance of the central pixels, N is the number of pixels for each frame, and $I(i,t)$ $(i=0,1,\ldots,N-1)$ is the pixel's intensity of the tth frame. The fingerprint was computed around the frame with maximum temporal activity $V(t)$, and the spectral analysis by FFT leads to a 16-dimensional vector which was based on the phase of the temporal activity or feature value. This method uses only the content relation in the temporal domain and is not robust against local distortions such as region cropping, frame insertion, etc.

The ordinal measure [37] used for global feature descriptor in spatial domain has been extended to the temporal domain [54] by ranking the regions (blocks) along the temporal axis or time. If each frame was divided in N blocks and if λ^n the ordinal measure of the region n in a temporal window with the length T, the distance D between a reference video V_{rf} and a query video V_{q} at time t was given as follows:

$$D\left(V_{\mathrm{q}}, V_{\mathrm{rf}}^{\mathrm{s}}\right) = \frac{1}{N}\sum_{n=1}^{N} d\left(\lambda_{\mathrm{q}}^{\mathrm{n}}, \lambda_{\mathrm{rf}}^{\mathrm{s,n}}\right), \tag{9}$$

where

$$d\left(\lambda_{\mathrm{q}}^{\mathrm{n}}, \lambda_{\mathrm{rf}}^{\mathrm{s,n}}\right) = \frac{1}{C_T}\sum_{i=1}^{T}\left|\lambda_{\mathrm{q}}^{\mathrm{n}}(i) - \lambda_{\mathrm{rf}}^{\mathrm{s,n}}(s+i-1)\right|. \tag{10}$$

Here, s is the tested temporal shift and C_T is the normalizing factor. The best temporal shift s between two consecutive frames was selected. This measure is robust against certain transformations such as time shifting, recompression, etc., but cannot tolerate transformations that change a subset of the frames in the video clip such as frequent region cropping, insertion of large area, etc.

Radhakrishnan and Bauer [55] introduced a subspace projection scheme for extracting the fingerprints from the group of frames for each time interval in the video in which the basis vectors of a coarse representation were generated using SVD [19] first. Then, a subspace representation of the input video frames was obtained by projecting the coarse representation of the video frames onto a subset of the basis vectors. Finally, the fingerprint was generated by projecting a temporal average of these representations onto the pseudo-random basis vectors. The temporal average T_{a} of $(R_0^s, R_1^s, \ldots, R_{M-1}^s)$ was computed as given below:

$$T_{\mathrm{a}}(z) = \frac{1}{M}\sum_{i=0}^{M-1} R_i^s(z), \quad z = 0, 1, \ldots, M-1, \tag{11}$$

where $R_i^s(\cdot)$ is the coarse representation of the video frames. The top Z values of T_{a} are selected for the T_{t} time interval. This method is not robust to certain transformations such as illumination changes, region cropping, etc.

In [56], the authors have come up with an idea in which the video signature or compact hash value was extracted based on the temporal variation or shot change position of the video files. The anchor frames that represent video temporal structure (signature) were extracted using cumulative luminance histogram difference (CLHD) and statistics collected in a local window along with an adaptive threshold after temporal subsampling of the video frames. Later, to achieve fast matching of signatures, an efficient suffix array data structure was applied. The method does not work well for video contents with lots of gradual transitional effects and object movement.

Similarly, the motion vector along the temporal direction of a video was extracted using the combination of mean of the magnitudes of motion vectors (MMMV) and mean of the phase angles of motion vectors (MPMV) methods proposed in [57]. This method does not produce precise result when motion vectors are extracted from consecutive frames with a high capture rate. To solve the problem faced by methods that is based on key frames or frame-by-frame, Wang et al. [58] came up with a new concept in which the temporal context of key frames was expressed as binary codes. The surrounding frames of each key frame were clustered into two groups based on their temporal relationships with the center key frame which was then used for generating a binary code that represents the temporal context of center key frame. Before matching, the key frames were first projected into distinct buckets by locality-sensitive hashing [34] technique and the distance between the temporal context binary codes (TCB) of key frames that are in the same bucket was computed using hamming distance metric in the stage of sequence matching. The complexity of this method is in finding a robust key frame that uniquely represents the video sequence.

2.3 Based on spatial–temporal domain

The features that are extracted from the spatial and temporal domains play a crucial role in video copy detection, respectively, but there exist so many shortcomings which use respective spatial- and temporal-based methods for video copy detection. To overcome those shortcomings, several methods have been proposed by many researchers considering both spatial and temporal information of videos to yield better performance result. Taking into account all those shortcomings and challenges, Coskun et al. [16] came up with an idea in which video sequence's luminance component was transformed by 3-dimensional discrete cosine transform (3D-DCT) or 3-dimensional random bases transform (3D-RBT) methods. The low-pass transform coefficients were ordered and quantized using the median of the rank-ordered coefficients, generating $4 \times 4 \times 4$ binary bits for each 3D cube. This method can resist some temporal transformations such as frame rate change or frame dropping and be robust against certain spatial transformations such as recompression, contrast change, etc., but cannot tolerate the manipulations that destroy the spatial and temporal information such as picture-in-picture, frame insertion, etc.

A new approach called temporally informative representative image (TIRI) was introduced in [59–61] for copy detection that represents a short segment of the video and contains spatial–temporal information about the video segment. The pixels of TIRI for each video segment were generated as a weighted sum of the frames as follows:

$$I_{u,v} = \sum_{i=1}^{K} \alpha_i l_{u,v,i}, \tag{12}$$

where $l_{u,v,i}$ is the luminance value of the (u, v)th pixel on the ith frame in a segment of K frames and α_i is the weight associated with each frame. Then, the TIRIs were segmented into overlapping blocks of size $w \times w$ and the first vertical and the first horizontal DCT coefficients (features) were generated from each block using 2-dimensional discrete cosine transform (2D-DCT) [62, 63]. To enhance the similarity search performance, the inverted-file-based and cluster-based similarity search approaches were applied. Devi et al. [64] have adopted the same TIRI-DCT [59] method in addition to the low-pass band coefficients (features) that were extracted using discrete wavelet transform (DWT) [65] from each block of the TIRIs to enhance the performance result. The authors in [66] also adopted the same method in which the output of TIRIs was first transformed into R, G and B channels and was then partitioned into $s \times s$ blocks. Then, color correlation was extracted and the percentage of number of pixels belonging to a particular group was computed which was again normalized to obtain the color correlation histogram as a feature vector. Similarly, in [67], the key frames were generated by applying TIRI transform onto the preprocessed video to preserve the spatiotemporal information. The method also reduced the feature vector size as well as decreased the computing time. The local textural descriptors were extracted from each key frame using Weber binarized statistical image features (WBSIF), and histogram was computed for each key frame. The final feature vector was computed by concatenating the k number of WBSIF histograms. In [68], the authors have adopted the similar TIRI transformation of the video sequence in which the proposed Shearlet-based video fingerprint (SBVF) method was applied to generate the fingerprints that preserve both the spatial and temporal properties. The SBVF was built by the Shearlet coefficients in Scale-1 (lowest coarse-scale) for unveiling the spatial features and Scale-2 (second lowest coarse-scale) for unveiling the directional features. Inverted index file (IIF) hash searching approach was used for comparison and performance evaluation. However, the TIRI could not represent the video information effectively as the methods did not take the scene change into account. Moreover, the overlapping blocks generate a huge number of TIRIs, which leads to a large amount of redundant information.

The concept of generation of a TIRI [59] and representative saliency map (RSM) [69–71] for spatial–temporal-based video copy detection was replaced by generation of a temporally representative frame (TRF) [72] using temporally visual weighting (TVW) method based on visual attention [43] proposed in [73] by Liu et al. to generate a compact hash value that provides better performance and was further improved by the authors in [74]. Here, they have fused both the visual appearance and visual attention features using a deep belief network (DBN) to gain the compact hash value that represents the whole video. The visual appearance feature was extracted from each block of the TRFs directly, while the visual attention feature was extracted from each block of the RSMs of the video in which the Gaussian mixture model (GMM) was used to derive the dynamic attention model, whereas static attention model was created based on intensity, texture and color features to create a saliency map. TRF of a video segment was generated as given below:

$$F(x, y) = \sum_{i=1}^{K} \omega_i \cdot F(x, y, i), \tag{13}$$

where $F(x, y)$ is the intensity of the (x, y)th pixel of the ith frame of a video segment with K frames and ω_i is the temporally visual weight which was computed based on the strength of the visual attention shift. $F(x, y)$ is the intensity of the TRF. RSM was also generated in the same way as TRF:

$$\mathrm{RSM}(u, v) = \sum_{j=1}^{W} \alpha_j S(u, v, j), \tag{14}$$

where $S(u, v, j)$ is the luminance value of the (u, v)th pixel of the jth saliency map of the video segment that has K frames, α_j is the temporal visual weight, and $\mathrm{RSM}(u, v)$ represents the luminance value of pixels of the RSM. However, the frequent frame insertion and large area of region cropping will affect the method.

In [75, 76], a 3D-DWT-based method was proposed to overcome the limitations and inefficiency faced by the 2-dimensional discrete wavelet transform (2D-DWT) [65, 77] as video has a 3-dimensional vector form. In this method, a hash of group of frames was computed from the spatial–temporal low-pass (LLL) band obtained by applying the 3D-DWT on a video segment which serves as the spatiotemporally informative images (STIRIs) for the segment as the method involves weighted temporal averaging inherently. The STIRI was partitioned into overlapping blocks of size $b \times b$, and then, blocks were shuffled using a secret key k to derive a frame f. The DCT [62] was applied on the overlapping blocks of STIRI for decorrelation of the correlated wavelet coefficients, and then, the hash was computed from the DCT coefficients. However, this method is not robust against the geometric manipulations such as rotation.

Since the interest points can represent a video sequence's salient contents, the methods in [30, 78, 79] have used not only the spatial interest points [24], but also the temporal interest points along the time axis to achieve higher robustness against the content-preserving as well as geometric attacks. These spatial–temporal interest points correspond to points in which the image values have remarkable local variation in both the space and time. An improved version of the Harris interest point detector [80] was used for extracting the interest points, and a differential description of the local region around each interest point was created. The points that have significant corresponding eigenvalues were considered salient. However, this method incurs high computational cost and the synchronization between two salient points can easily be broken in geometric attacks as some points can be replaced by new ones. Chen and Chiu [81] also used the same methods as in [30], but the only difference was that spatial–temporal interest points were detected in visual attention region [73]. In order to remove the noisy feature points, the geometric constraint measurement was employed for bidirectional point matching. Similarly, the authors in [82–84] used the spatial–temporal interest points [30] for detecting the local interest point of regions. In [82, 83], the Kanade–Lucas–Tomasi (KLT) [85] feature tracker was used for tracking the Harris points to get the stable local feature points trajectory. In [84], the local fingerprints were extracted using contrast context histogram (CCH) in local regions around each interest point by evaluating the intensity differences between the center pixel and other pixels. These methods incur high-dimensional and computational complexity.

In video copy detection, the computational complexity that arises due to the high dimensionality of hash or feature vector plays a crucial role which affects the performance of the methods up to a great extent. To solve this issue, Nie et al. [86] introduced a high-order tensor model-based projection technique that exhibits assistance and consensus among different features, and then video tensor was decomposed via the Tucker model. This method outperforms the projection-based video hashing approach in [87–90]. Subsequently, the comprehensive feature was computed by the low-order tensor that was acquired from tensor decomposition, and finally, the video hash was generated using this feature. The tensor-based projections can give good robustness while capturing the spatiotemporal essence of the video effectively for discriminability [87]. However, the random frame insertion and large amount of illumination change can distort the robustness of this method.

The spatial ordinal measure [37] has been extended to the temporal domain [91, 92] by ranking the blocks along the temporal or time axis to generate the robust fingerprints for accurate matching between the original and pirated videos. This method cannot handle the certain transformations such as frequent region cropping, frame insertion, etc. Lee et al. [93] introduced a video copy detection method based on combined histogram of oriented gradients (HOG) descriptor and ordinal measure [92] representation of the frame. HOG descriptor was used for object detection and for describing the global feature of frames in video sequence. Ordinal measure histogram (OH) was used for generating the feature vector of entire video sequence as temporal feature which is robust against the color shifting and size variations. There is a trade-off between robustness and discriminability. In [94], the proposed method extracted the spatiotemporal compact feature ST_k from the key frames of a video by abrupt change of luminance as follows:

$$\mathrm{ST}_k = \left\{\Delta_q(k, 1), \Delta_q(k, 2), \ldots, \Delta_q(k, 9), D_k\right\}, \quad \text{for} \quad k = 1, \ldots, K, \tag{15}$$

where D_k is the temporal interval between the current key frame F_k and the prior selected key frame F_{k-1}, and $\Delta_q(k, m)$ is the luminance differences of 9 blocks in key frame. The complexity of this method lies on the selection of a robust key frame.

The problem of efficient searching for highly deformed videos in small datasets also affects the performance of the methods in video copy detection system. To address this problem, Douze et al. [95] introduced a spatiotemporal post-filtering scheme in which the matched frames were grouped into sequences and the matches which are not consistent in terms of scaling and rotation with the dominant hypothesis for database image were discarded using weak geometry consistency (WGC) strategy. In this model, temporal shift was first determined based on 1-dimensional Hough voting strategy and then, spatial component was determined by estimating 2-dimensional affine transformation between the matching video sequences, respectively. Here, the local patches or salient interest points were detected using Hessian affine region detector (HARD) firstly, and then the pattern of the surrounding local regions was described by SIFT [19] or center-symmetric local binary pattern (CS-LBP) descriptors. Subsequently, the descriptors were clustered to form a bag-of-features and the matched frames were computed based on Hamming embedding method. This method does not give importance toward the frequent frame deletion and region cropping.

The authors in [96] have proposed a method that fuses the spatial and temporal information of a video sequence. Here, the spatial fingerprint was extracted using the so-called method TIRI-DCT [61] and the temporal fingerprint was extracted using the temporal variances (differences) V. Subsequently, the temporal strength TS of V was extracted which was used to determine the importance of temporal fingerprints at the stage of modality fusion adaptively. This method overcomes the limitations of the previously developed methods which have used only the pre-specified weights for combining spatial and temporal information. One of the main issues related to this method is as follows: If the gap between the temporal strengths of the compared temporal fingerprints is big, then the temporal fingerprints can easily be distinguished from each other. Similarly, in [97], three spatiotemporal parameters, i.e., color space, frame partitioning and sampling frame rate, were evaluated for video copy detection based on normalized average luminance descriptors. This method is limited to the content-preserving distortions and is not robust against the geometric distortions such as frame deletion, rotation, scaling, etc. Moreover, reduction in the sampling frame rate and increasing the number of frame partitions can lower the efficiency as well as the performance of the method. Several methods such as that based on video tomography and bag-of-visual-word [98], histogram of oriented gradients (HOG) and compression properties [99], identifying shot-based semantic concepts along the temporal axis [100] and self-similarity matrix (SSM) [101] also have been proposed by many researchers, respectively, which exploits both spatial and temporal information in a video clip or sequence to yield the better performances for robust video copy or forgery detection.

2.4 Other methods

2.4.1 Learning-based approaches

Ye et al. [102] introduced a new learning-based hashing called structure learning for indexing the large-scale multimedia data. The idea behind this approach was to leverage data properties and human supervision based on some known training datasets to derive a compact and accurate hash code. This method was based on supervised learning in which structure information exploits both the discriminative local visual patterns occurring in video frames that are connected with the same semantic class and temporal consistency over successive frames. The idea of this method was further improved by Chen et al. [103], where they have developed a multilayer neural network to learn discriminative and compact hash codes. This methodology exploits both the nonlinear relationship between video samples and the structure information between distinct frames within a video. In addition, the intra-video similarity was also taken into consideration. To further improve the performance, a subspace clustering method was employed to cluster the frames into distinct scenes. The motion information such as optical flow is not considered in this method which can degrade the performance.

2.4.2 Deep neural network-based approaches

Another learning-based hashing method called deep video hashing (DVH) was proposed by authors in [104], which learns binary codes for the entire video in a deep network to exploit both discriminative and temporal information of videos. The method was designed for scalable video search in a large multimedia database which works based on convolution neural network (CNN) learning framework. As the method uses supervised information based on deep learning network, there may exist ambiguity between the label information that can degrade the performance. Hao et al. [105] proposed an unsupervised hashing extension of stochastic multi-view hashing (SMVH) [106] through Student *t*-distribution matching scheme, the so-called *t*-USMVH and its extension of deep hashing through neural network called *t*-UDH. The aim of this method was also to increase the scalable search performance in large video databases. Hu and Lu [107] introduced a deep learning-based method in which the CNN [104] and recurrent neural network (RNN) were used jointly to achieve better copy detection accuracy. This method has overcome the limitations faced by the proposed methods in [108, 109]. In this method, the features or fingerprints were extracted initially from video frames using residual convolutional neural network (ResNet) and then Siamese Long Short-Term Memory (SiameseLSTM) architecture was trained for fusing both the spatial and temporal properties and sequence matching of video. Lastly, the graph-based neural network was used for identification of copied segments of a video. However, this method can incur high computational cost because of large number of trained datasets. Moreover, robustness against the geometric and content-preserving attacks is not analyzed properly.

Li et al. [110] proposed a parallel 3-dimensional convolutional neural network (3D-CNN) approach for video classification which relaxes 3D-CNN to two-class classification task from the multi-class classification task to reduce the data requirement on training. Features were extracted from the video input streams directly using 3D-CNN to obtain the local motion information from video. The parallel 3D-CNN classification model was built by a number of 3D-CNNs. As each 3D-CNN is a two-class video classifier, the number of 3D-CNNs is equal to the number of video classes. Finally, the decision was obtained by concatenating the classification results of all 3D-CNN classifiers. However, high computational cost can be incurred by this method which needs to be enhanced and robustness against the different distortions is not analyzed properly. The authors in [111] introduced a data-driven approach that uses deep neural network to learn robust video fingerprint or descriptor from a raw video. The task of learning video descriptor was broken down into subproblems, and then neural network was trained to tackle each of them by this proposed method. The conditional restricted Boltzmann machine (CRBM) was used as one of the prominent components for building deep feature learning network (conditional generative model) and was trained to capture the intrinsic visual characteristics as well as the spatiotemporal correlations among visual contents of video which were represented as an intermediate descriptor. A nonlinear encoder called denoising auto-encoder was then trained using pairs of intermediate descriptors extracted from manipulated and original videos to learn a compressed yet robust representation of intermediate descriptor. To preserve the optimal balance between robustness and discriminative capability of the output descriptor, the top layers of the network were trained. However, the challenge with this method lies in computational cost as training dataset will get increased by increasing the size of network. Nie et al. [112] came up with an idea that combined both the handcrafted visual features and semantic features of videos for near-duplicate video detection purpose. Firstly, low-level representation fingerprint (LRF) was generated from handcrafted visual features using a tensor-based approach which can preserve the mutual relations among various visual features. Secondly, CNN [104] approach was used to learn deep semantic features for generating deep representation fingerprint (DRF) to give heterogeneity assistance to LRF. This approach will also incur high computational cost which needs to be taken into consideration for better performance.

2.4.3 Miscellaneous approaches

Singh and Aggarwal [113] introduced a method to detect upscale-crop (frame-level) and splicing (region-level) forgeries that were performed using an image processing operation called resampling in digital videos. The detection operation of resampling artifacts (compression and noise) was carried out based on the pixel-covariance correlation and noise-inconsistency analysis whose outcomes are later combined to give better performance. The modified Gallagher (MG) detector and fractional modified Gallagher (F-MG) detector were used for the pixel-covariance correlation analysis. Analysis was performed using MG detector based on fast Fourier transform (FFT) and discrete cosine transform (DCT) domains, and the analysis was performed using F-MG detector based on discrete fractional Fourier transform (DFrFT) domain. Similarly, the noise-inconsistency analysis was performed based on wavelet denoising filter. For splicing (region-level) forgery detection, the methods were applied into regions of interest (ROIs) in video frames. The main challenge with this method lies in estimation of parameters used for analysis purpose such as scaling factors and interpolation filter for forgery detection. They came up again with a new idea of detection and localization of copy–paste forgeries [114], which alters the content of particular region of a frame in digital videos. Sensor pattern noise (SPN), Hausdorff distance-based clustering and color filter array (CFA) methods were used for copy–paste forgery detection and localization. As this approach considers the frame-to-frame and region-based matching, it can incur high computational complexity.

Multimodal visual–audio fingerprints-based video copy detection approach was proposed by Roopalakshmi et al. [115] in which both the visual and audio features were combined to detect illegal copies. Initially, the 1-D motion feature vector was generated by computing the average of differences between region-wise motion vector magnitudes of consecutive frames and the 1-D acoustic feature was generated using mel-frequency cepstral coefficients (MFCCs). In this approach, the DCT [62] was applied, in which the DCT coefficients of log powers of mel-frequency cepstrum (short-term power spectrum of a sound or audio) were considered for generating MFCCs. Secondly, sliding-window-based dynamic programming approach was applied to achieve accurate frame-to-frame matching. Subsequently, both the features were combined to generate a 1-D feature vector for copy detection. The performance of the proposed method was improved compared to the reference methods [116–118]. However, computational complexity can degrade the performance of this method.

A key parameter-dependent heat kernel signature (HKS)-based 3D model hashing was proposed in [119, 120] by Lee et al. This methodology was mainly developed for video authentication and is robust against the isometric modifications. The local and global HKS coefficients were obtained through timescales by computing the eigenvalues and eigenvectors of a mesh Laplace operator. Then, these HKS coefficients were clustered into 2D square cells with variable bin sizes and the feature values were extracted from the weighted distance of HKS coefficients based on n-order Butterworth function. The binary hash was generated through binarization of the intermediate hash values that were obtained by projecting the feature values onto random values. Further, to improve the robustness, uniqueness, security and spaciousness the two parameters called bin-center points and cell amplitudes were used. Choosing a robust key and parameters is the main challenge with this methodology. Many other methods [121–124] were introduced by several researchers, where they have explored the significance of visual hashing or fingerprinting in the field of video copy detection system.

3 Major challenges

The hashing- or fingerprinting-based copy detection is more preferable compared to watermarking-based copy detection for illegal video copy detection as multimedia content is often transformed or manipulated before being uploaded on the video sharing Web sites [1]. Still, there remain some challenges with the existing methods. Based on the works found in the state of the art, the main challenges that should be taken into consideration in order to enhance the performance of copy detection system are as follows:

- To acquire a better trade-off between discriminability and robustness against the geometric as well as content-preserving distortions.
- To lessen the computational cost of fingerprint extraction and matching.
- To enhance the efficiency of fingerprint database search.
- To lessen the storage space requirement for each fingerprint.
- To incorporate the semantic concepts to lessen the semantic gap between the high-level and low-level feature representation of frame images.
- To integrate fingerprinting-based and watermarking-based copy detection techniques in order to yield content identification as well as user authentication for high security.

Most of the video copy detection approaches are robust against the common content-preserving distortions such as contrast enhancement, blurring, frame rate change, frame resizing, etc., but robustness against the geometric distortions such as rotation, scaling, frame dropping, flipping and picture-in-picture still poses specific challenges to the problem of fingerprinting- or hashing-based video copy detection. The problem of copyright infringement of original video by an adversary still remains a big issue as the multimedia technology has increased. Many researchers are trying to find a solution for geometric distortions, and a large number of solutions are being proposed, but the robustness and percentage of detection accuracy are not up to the peak point, which needs to be enhanced further.

4 Current trends and discussion

Since the emergence of copyright infringement or piracy issues of multimedia, various approaches have been proposed by several researchers to tackle the issues. Fingerprinting-based copy detection approach has been adopted mostly because of its discriminability and robustness property compared to watermarking-based copy detection system [15]. Most of the existing methods are robust against the content-preserving distortions, so the researchers are currently working hard to achieve the high robustness against the geometric distortions, which is still a challenging task. Besides robustness, the discriminative capability also plays an important role in video copy detection system. It can be observed from the state of the art that there exists a trade-off between discriminability and robustness. So, currently the researchers are also working on acquiring a better trade-off between discriminability and robustness for optimal performance. Some state-of-the-art approaches have used local feature descriptors such as SIFT [19] for discriminability, while some others have used global feature descriptors such as OM [37] for robustness to common distortions and some have used combination of both the local and global feature descriptors to improve the performance. It can be seen that both the local and global features are extracted from the spatial domain where temporal domain is ignored, which is also an important property of a video. To overcome this limitation, some approaches have been proposed based on both

Table 1 Summary of the classification of some important existing hashing-based video copy detection methods based on different characteristics

Methods	Feature type	Domain	Robustness	Discriminability	Computational cost	Citations
SIFT + SVD	Local	Spatial	High	High	High	[19]
Harris detector	Local	Spatial	High	High	High	[32]
FREAK	Local	Spatial	Middle	High	High	[33]
SURF	Local	Spatial	Middle	High	High	[34]
SURF + ORB	Local	Spatial	High	High	High	[35]
OM	Global	Spatial	Middle	Middle	Low	[37]
BSIF + LCD	Global	Spatial	High	Middle	Low	[41]
BSIF + ICD	Global	Spatial	High	Middle	Low	[42]
NMF	Coarse	Spatial	Middle	High	High	[31]
CHMT	Coarse	Spatial	Middle	High	High	[48]
SIFT + PHOG	Local + Global	Spatial	High	High	High	[27]
SIFT + OM	Local + Global	Spatial	Middle	High	High	[50]
Temporal OM	Global	Temporal	Middle	Low	Low	[54]
CLHD	Global	Temporal	Middle	Low	Low	[56]
3D-DCT + 3D-RBT	Coarse	Spatial + temporal	Middle	Middle	High	[16]
TIRI-DCT	Coarse	Spatial + temporal	Middle	Middle	Low	[59]
TIRI-WBSIF	Local	Spatial + temporal	High	High	High	[67]
TIRI-SBVF	Coarse	Spatial + temporal	High	Middle	High	[68]
CNN + RNN	Semantic	Deep learning	High	Middle	High	[107]
3D-CNN	Semantic	Deep learning	Middle	Middle	High	[110]

the spatial and temporal domains such as TIRI-DCT [59]. Recently, the researchers have been focusing on the utilization of the importance of deep neural network-based learning approaches such as CNN [104] in the field of video copy detection. This approach incurs a high computational cost as it requires a large amount of database storage for pre-trained known dataset which increases as the network size increases. Moreover, there may exist an ambiguity between the label information which uses supervised information-based deep learning network. As this approach has recently been adopted by researchers for video copy detection, still there exists a huge scope for analyzing the shortcomings broadly and a fast and optimal solution for copy detection can be achieved in future (Table 1).

How to choose a better copy detection method firmly depends on what we are seeking for and where we are seeking it. No universal description and no single approach seem to be optimal to various applications that require video copy detection. Some application cases for finding copies are identified below:

- Finding exact copies in a stream for statistics on commercials.
- Finding transformed full movie with possible decrease in quality (camcording) and no postproduction.
- Finding short segments on TV stream with possible large postproduction transformation.
- Finding short videos on the Internet with various transformations.

For the first case, local feature descriptors such as Harris detector [32] and SIFT [19] will work better for describing the precise interest point to detect the exact copies. For finding transformed full movie, as the length of video sequence is important, global feature descriptors such as OM [37] are probably efficient and faster than the local feature descriptors. For the third case, finding short segments in a video stream is a critical issue, and Harris detector [32] will probably give better result compared to global feature descriptor. For the fourth case, multiple difficulties are mixed for videos on the Internet and the solutions depend on the quality required. The method that combines both the local and global features which preserve both the spatial and temporal properties seems more promising for solving various transformations. It can be observed clearly that various distortions such as rotation, scaling, cropping, gamma correction, etc., are applied to the original video to get information by an adversary in a large extent. The choice of method is still open, but the combination of both the handcrafted visual features (local and global features based on spatiotemporal domain) and deep semantic features based on deep neural network constituting both the discriminability and robustness properties seems to be more promising for video classification and accurate detection of illegal copies. Almost all of the methods are being used by various video sharing Web sites such as YouTube, Netflix, etc., for copy detection purpose. Still they are facing large number of copyright infringement issues which need to be analyzed deeply and implement the robust method for fast and accurate copy detection.

All of the top methods in fingerprinting- or hashing-based video copy detection follow the paradigm of computing the compact signatures or hash codes from the content of a digital media without altering the content which is important for various multimedia applications. The generated compact hash or fingerprint can tell whether a dubious piece of content matches a multimedia document registered in the fingerprint database; thus, it can detect content replication of an original video robustly. Unlike watermarking approach, the fingerprinting approach can be applied to legacy content of a media that has already been distributed. The fingerprinting approach is more discriminative as well as robust against various content transformations compared to the watermarking approach.

5 Conclusion

The objective of this paper is to provide a detailed summary of the existing visual hashing- or fingerprinting-based video copy detection system. Most of the existing video copy detection methods were based on spatial, temporal and combination of both spatial and temporal domains according to the extracted features, and many other techniques have been used. Methods that were based on extracting the local features such as regions of interest (ROIs) points in spatial domain of a video have more discriminating power as compared to the other extracted features, but less robust against geometric attack. The methods considering only the spatial domain are not sufficient enough to survive the temporal attacks such as frame rate change when motion information is considered along the time or temporal axis of a video sequence. To solve the issues, many researchers have developed the methods that exploit both spatial and temporal information of a video. Still, most of the methods are not robust against both the content-preserving and geometric attacks as they were mainly focused on extracting the local features in grayscale form, where global features such as color information are ignored, which is also an important property when color pictures of a video come into play. There is a trade-off between discriminability and robustness properties in most of the existing methods. In recent decades, detection of copied or pirated version of an original video content has become more complex as multimedia technology has emerged tremendously. So, employing the methods that have both the discriminability and robustness properties against various content-preserving as well as geometric attacks such as lossy compression, resizing, rotation, scaling, etc., has become the most challenging in video copy detection system.

For tackling the problems and issues faced in video copy detection system, many researchers are currently working in this field and trying to improve the performance and efficiency of copy detection system based on the robust visual hashing or fingerprinting techniques.

References

1. Law-To J, Chen L, Joly A, Laptev I, Buisson O, Gouet-Brunet V, Boujemaa N, Stentiford F (2007) Video copy detection: a comparative study. In: Proceedings of the 6th ACM international conference on image and video retrieval. ACM, pp 371–378
2. Liu L, Lai W, Hua XS, Yang SQ (2007) On real-time detecting duplicate web videos. In: 2007 IEEE international conference on acoustics, speech and signal processing (ICASSP), vol 1. IEEE, pp I–973
3. Kitanovski V, Taskovski D (2010) Real-time TV commercial monitoring based on robust visual hashing. In: 2010 2nd European workshop on visual information processing (EUVIP). IEEE, pp 140–143
4. Lee JY, Kim HG (2014) Video Fingerprinting for real-time TV commercial advertising identification. In: 2014 international conference on information science and applications (ICISA). IEEE, pp 1–2
5. Kunhu A, Nisi K, Sabnam S, Majida A, Saeed AM (2016) Index mapping based hybrid DWT-DCT watermarking technique for copyright protection of videos files. In: 2016 Online international conference on green engineering and technologies (IC-GET). IEEE, pp 1–6
6. Chongtham C, Khumanthem MS, Chanu YJ, Arambam N, Meitei D, Chanu PR, Singh KM (2018) A copyright protection scheme for videos based on the SIFT. Iran J Sci Technol Trans Electr Eng 42(1):107–121
7. Barni M, Bartolini F (eds) (2004) Watermarking systems engineering: enabling digital assets security and other applications. CRC Press, Boca Raton
8. Singh TR, Singh KM, Roy S (2013) Video watermarking scheme based on visual cryptography and scene change detection. AEU Int J Electron Commun 67(8):645–651
9. Mishra M, Pandit S (2014) Image encryption technique based on chaotic system and hash function. In: 2014 International conference on computer communication and systems. IEEE, pp 063–067
10. Katz J, Menezes AJ, Van Oorschot PC, Vanstone SA (1996) Handbook of applied cryptography. CRC Press, Boca Raton
11. Nikolaidis N, Pitas I (2006) Image and video fingerprinting for digital rights management of multimedia data. In: 2006 International symposium on intelligent signal processing and communications (ISPACS). IEEE, pp 801–807
12. Tang Z, Huang L, Zhang X, Lao H (2016) Robust image hashing based on color vector angle and canny operator. AEU Int J Electron Commun 70(6):833–841
13. Mucedero A, Lancini R, Mapelli F (2004) A novel hashing algorithm for video sequences. In: 2004 International conference on image processing (ICIP), vol 4. IEEE, pp 2239–2242
14. De Roover C, De Vleeschouwer C, Lefebvre F, Macq B (2005) Robust video hashing based on radial projections of key frames. IEEE Trans Signal Process 53(10):4020–4037
15. Coskun B, Sankur B (2004) Robust video hash extraction. In: 2004 12th European signal processing conference. IEEE, pp 2295–2298
16. Coskun B, Sankur B, Memon N (2006) Spatio–temporal transform based video hashing. IEEE Trans Multimed 8(6):1190–1208
17. Roopalakshmi R, Reddy GR (2010) Recent trends in content-based video copy detection. In: 2010 IEEE international conference on computational intelligence and computing research (ICCIC). IEEE, pp 1–5
18. Shinde SR, Chiddarwar GG (2015) Recent advances in content based video copy detection. In: 2015 International conference on pervasive computing (ICPC). IEEE, pp 1–6
19. Neelima A, Singh KM (2017) Collusion and rotation resilient video hashing based on scale invariant feature transform. Imaging Sci J 65(1):62–74
20. Yang G, Chen N, Jiang Q (2012) A robust hashing algorithm based on SURF for video copy detection. Comput Secur 31(1):33–39
21. Lee S, Yoo CD (2008) Robust video fingerprinting for content-based video identification. IEEE Trans Circuits Syst Video Technol 18(7):983–988
22. Mao J, Xiao G, Sheng W, Hu Y, Qu Z (2016) A method for video authenticity based on the fingerprint of scene frame. Neurocomputing 173:2022–2032
23. Mohan R (1998) Video sequence matching. In: Proceedings of the 1998 IEEE international conference on acoustics, speech and signal processing (ICASSP), vol 6. IEEE, pp 3697–3700
24. Massoudi A, Lefebvre F, Demarty CH, Oisel L, Chupeau B (2006) A video fingerprint based on visual digest and local fingerprints. In: 2006 IEEE international conference on image processing (ICIP). IEEE, pp 2297–2300
25. Bhat DN, Nayar SK (1998) Ordinal measures for image correspondence. IEEE Trans Pattern Anal Mach Intell 20(4):415–423
26. Nie X, Qiao J, Liu J, Sun J, Li X, Liu W (2010) LLE-based video hashing for video identification. In: 2010 IEEE 10th international conference on signal processing (ICSP). IEEE, pp 1837–1840
27. Peng H, Deng C, An L, Gao X, Tao D (2013) Learning to multimodal hash for robust video copy detection. In: 2013 20th IEEE international conference on image processing (ICIP). IEEE, pp 4482–4486
28. Maani E, Tsaftaris SA, Katsaggelos AK (2008) Local feature extraction for video copy detection in a database. In: 2008 15th IEEE international conference on image processing (ICIP). IEEE, pp 1716–1719
29. Ding G, Nie R (2010) Ring fingerprint based on interest points for video copy detection. In: 2010 IEEE international symposium on multimedia (ISM). IEEE, pp 347–352
30. Laptev I (2005) On space-time interest points. Int J Comput Vis 64(2–3):107–123
31. Cirakman O, Gunsel B, Sengor NS, Gursoy O (2010) Key-frame based video fingerprinting by NMF. In: 2010 17th IEEE international conference on image processing (ICIP). IEEE, pp 2373–2376
32. Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Trans Pattern Anal Mach Intell 27(10):1615–1630
33. Li J, Guo X, Yu Y, Tu Q, Men A (2014) A robust and low-complexity video fingerprint for multimedia security. In: 2014 International symposium on wireless personal multimedia communications (ISWPMC). IEEE, pp 97–102
34. Zhang Z, Cao C, Zhang R, Zou J (2010) Video copy detection based on speeded up robust features and locality sensitive hashing. In: 2010 IEEE international conference on automation and logistics (ICAL). IEEE, pp 13–18
35. Özbulak 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 (ICDSP). IEEE, pp 124–128

36. Rublee E, Rabaud V, Konolige K, Bradski G (2011) ORB: An efficient alternative to SIFT or SURF. In: 2011 IEEE international conference on computer vision (ICCV). IEEE, pp 2564–2571
37. Hua XS, Chen X, Zhang HJ (2004) Robust video signature based on ordinal measure. In: 2004 International conference on image processing (ICIP), vol 1. IEEE, pp 685–688
38. Yang L, Li Y (2010) Research of robust video fingerprinting. In: 2010 International conference on computer application and system modeling (ICCASM), vol 12. IEEE, pp V12–V43
39. Uchida Y, Hashimoto M, Kawada R (2010) Fast and robust content-based copy detection based on quadrant of luminance centroid and adaptive feature comparison. In: 2010 17th IEEE international conference on image processing (ICIP). IEEE, pp 1021–1024
40. Himeur Y, Ait-Sadi K, Oumamne A (2014) A fast and robust key-frames based video copy detection using BSIF-RMI. In: 2014 International conference on signal processing and multimedia applications (ICSIGMAP). IEEE, pp 40–47
41. Himeur Y, Sadi KA (2015) Joint color and texture descriptor using ring decomposition for robust video copy detection in large databases. In: 2015 IEEE international symposium on signal processing and information technology (ISSPIT). IEEE, pp 495–500
42. Himeur Y, Sadi KA (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
43. Su X, Huang T, Gao W (2009) Robust video fingerprinting based on visual attention regions. In: 2009 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1525–1528
44. Zheng L, Qiu G, Huang J, Fu H (2011) Salient covariance for near-duplicate image and video detection. In: 2011 18th IEEE international conference on image processing (ICIP). IEEE, pp 2537–2540
45. Nie X, Liu J, Sun J (2010) Robust video hashing for identification based on MDS. In: 2010 IEEE international conference on acoustics speech and signal processing (ICASSP). IEEE, pp 1834–1837
46. Sarkar A, Singh V, Ghosh P, Manjunath BS, Singh A (2010) Efficient and robust detection of duplicate videos in a large database. IEEE Trans Circuits Syst Video Technol 20(6):870–885
47. Kim S, Lee SH, Ro YM (2014) Rotation and flipping robust region binary patterns for video copy detection. J Vis Commun Image Represent 25(2):373–383
48. Sun R, Yan X, Gao J (2017) Robust video fingerprinting scheme based on contourlet hidden Markov tree model. Opt Int J Light Electron Opt 128:139–147
49. Lian S, Nikolaidis N, Sencar HT (2010) Content-based video copy detection–a survey. In: Intelligent multimedia analysis for security applications. Springer, Berlin, pp 253–273
50. Gu X, Zhang D, Zhang Y, Li J, Zhang L (2013) A video copy detection algorithm combining local feature's robustness and global feature's speed. In: 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1508–1512
51. Chiu CY, Tsai TH, Hsieh CY (2013) Efficient video segment matching for detecting temporal-based video copies. Neurocomputing 105:70–80
52. Tralic D, Grgic S, Zovko-Cihlar B (2014) Video frame copy-move forgery detection based on Cellular Automata and Local Binary Patterns. In: 2014 X international symposium on telecommunications (BIHTEL). IEEE, pp 1–4
53. Guzman-Zavaleta ZJ, Feregrino-Uribe C (2016) A simple approach towards efficient partial-copy video detection. In: 2016 IEEE 18th international workshop on multimedia signal processing (MMSP). IEEE, pp 1–6
54. Chen L, Stentiford FW (2008) Video sequence matching based on temporal ordinal measurement. Pattern Recogn Lett 29(13):1824–1831
55. Radhakrishnan R, Bauer C (2008) Robust video fingerprints based on subspace embedding. In: 2008 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2245–2248
56. Wu PH, Thaipanich T, Kuo CC (2009) A suffix array approach to video copy detection in video sharing social networks. In: 2009 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 3465–3468
57. Tasdemir K, Cetin AE (2010) Motion vector based features for content based video copy detection. In: 2010 20th International conference on pattern recognition (ICPR). IEEE, pp 3134–3137
58. Wang RB, Chen H, Yao JL, Guo YT (2016) Video copy detection based on temporal contextual hashing. In: 2016 IEEE second international conference on multimedia big data (BigMM). IEEE, pp 223–228
59. Malekesmaeili M, Fatourechi M, Ward RK (2009) Video copy detection using temporally informative representative images. In: 2009 international conference on machine learning and applications (ICMLA). IEEE, pp 69–74
60. Esmaeili MM, Ward RK (2010) Robust video hashing based on temporally informative representative images. In: 2010 Digest of technical papers International conference on consumer electronics (ICCE). IEEE, pp 179–180
61. Esmaeili MM, Fatourechi M, Ward RK (2011) A robust and fast video copy detection system using content-based fingerprinting. IEEE Trans Inf Forensics Secur 6(1):213–226
62. Xu Z, Ling H, Zou F, Lu Z, Li P, Wang T (2009) Fast and robust video copy detection scheme using full DCT coefficients. In: 2009 IEEE international conference on multimedia and expo (ICME). IEEE, pp 434–437
63. Setyawan I, Timotius IK (2014) Spatio-temporal digital video hashing using edge orientation histogram and discrete cosine transform. In: 2014 International conference on information technology systems and innovation (ICITSI). IEEE, pp 111–115
64. Devi S, Vishwanath N, Pillai SM (2012) A robust video copy detection system using TIRI-DCT and DWT fingerprints. Int J Comput Appl 51(6):29–34
65. Garboan A, Mitrea M, Prêteux F (2011) DWT-based robust video fingerprinting. In: 2011 3rd European workshop on visual information processing (EUVIP). IEEE, pp 216–221
66. Thomas RM, Sumesh MS (2015) A simple and robust colour based video copy detection on summarized videos. Procedia Comput Sci 46:1668–1675
67. Boukhari A, Serir A (2016) Weber binarized statistical image features (WBSIF) based video copy detection. J Vis Commun Image Represent 34:50–64
68. Yuan F, Po LM, Liu M, Xu X, Jian W, Wong K, Cheung KW (2016) Shearlet based video fingerprint for content-based copy detection. J Signal Inf Process 7(02):84
69. Sun J, Wang J, Zhang J, Nie X, Liu J (2012) Video hashing algorithm with weighted matching based on visual saliency. IEEE Signal Process Lett 19(6):328–331
70. Wang J, Sun J, Liu J, Nie X, Yan H (2012) A visual saliency based video hashing algorithm. In: 2012 19th IEEE international conference on image processing (ICIP). IEEE, pp 645–648
71. Wang W, Sun J, Liu J (2015) A memorability based method for video hashing. In: 2015 IEEE 16th international conference on communication technology (ICCT). IEEE, pp 309–313
72. Liu X, Sun J, Liu J (2013) Shot-based temporally respective frame generation algorithm for video hashing. In: 2013 IEEE international workshop on information forensics and security (WIFS). IEEE, pp 109–114

73. Liu X, Sun J, Liu J (2013) Visual attention based temporally weighting method for video hashing. IEEE Signal Process Lett 20(12):1253–1256
74. Sun J, Liu X, Wan W, Li J, Zhao D, Zhang H (2016) Video hashing based on appearance and attention features fusion via DBN. Neurocomputing 213:84–94
75. Saikia N, Bora PK (2011) Robust video hashing using the 3D-DWT. In: 2011 National conference on communications (NCC). IEEE, pp 1–5
76. Saikia N (2015) Perceptual hashing in the 3D-DWT domain. In: 2015 International conference on green computing and internet of things (ICGCIoT). IEEE, pp 694–698
77. Pandey RC, Singh SK, Shukla KK (2014) Passive copy-move forgery detection in videos. In: 2014 International conference on computer and communication technology (ICCCT). IEEE, pp 301–306
78. Zhao YX, Liu GJ, Dai YW, Wang ZQ (2008) Robust hashing based on persistent points for video copy detection. In: 2008 International conference on computational intelligence and security (CIS), vol 1. IEEE, pp 305–308
79. Li YN, Lu ZM (2009) Video identification using spatio-temporal salient points. In: 2009 Fifth international conference on information assurance and security (IAS), vol 2. IEEE, pp 79–82
80. Harris C, Stephens M (1998) A combined corner and edge detector. Alvey Vis Conf 15(50):10–5244
81. Chen DY, Chiu YM (2013) Visual attention guided video copy detection based on feature points matching with geometric-constraint measurement. J Vis Commun Image Represent 24(5):544–551
82. Wu X, Zhang Y, Wu Y, Guo J, Li J (2008) Invariant visual patterns for video copy detection. In: 2008 19th International conference on pattern recognition (ICPR). IEEE, pp 1–4
83. Chen J (2010) Detection of video copies based on robust descriptors. In: 2010 International conference on apperceiving computing and intelligence analysis (ICACIA). IEEE, pp 303–306
84. Deng C, Zhang Y, Gao X (2012) Robust video fingerprinting using local spatio-temporal features. In: 2012 International conference on computing, networking and communications (ICNC). IEEE, pp 350–353
85. Shi J (1994) Good features to track. In: Proceedings of the 1994 IEEE computer society conference on computer vision and pattern recognition (CVPR). IEEE, pp 593–600
86. Nie X, Yin Y, Sun J, Liu J, Cui C (2017) Comprehensive feature-based robust video fingerprinting using tensor model. IEEE Trans Multimed 19(4):785–796
87. Li M, Monga V (2012) Robust video hashing via multilinear subspace projections. IEEE Trans Image Process 21(10):4397–4409
88. Sandeep R, Sharma S, Bora PK (2017) Perceptual video hashing using 3D-radial projection technique. In: 2017 Fourth international conference on signal processing, communication and networking (ICSCN). IEEE, pp 1–6
89. Sandeep R, Bora PK (2013) Perceptual video hashing based on the Achlioptas's random projections. In: 2013 Fourth national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG). IEEE, pp 1–4
90. Nie X, Liu J, Wang Q, Zeng W (2015) Graph-based video fingerprinting using double optimal projection. J Vis Commun Image Represent 32:120–129
91. Chiu CY, Chen CS, Chien LF (2008) A framework for handling spatiotemporal variations in video copy detection. IEEE Trans Circuits Syst Video Technol 18(3):412–417
92. Kim C, Vasudev B (2005) Spatiotemporal sequence matching for efficient video copy detection. IEEE Trans Circuits Syst Video Technol 15(1):127–132
93. Lee F, Zhao J, Kotani K, Chen Q (2017) Video copy detection using histogram based spatio-temporal features. In: 2017 10th International congress on image and signal processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE, pp 1–5
94. Kim J, Nam J (2009) Content-based video copy detection using spatio-temporal compact feature. In: 2009 11th International conference on advanced communication technology (ICACT), vol 3. IEEE, pp 1667–1671
95. Douze M, Jégou H, Schmid C (2010) An image-based approach to video copy detection with spatio-temporal post-filtering. IEEE Trans Multimed 12(4):257–266
96. Kim S, Choi JY, Han S, Ro YM (2014) Adaptive weighted fusion with new spatial and temporal fingerprints for improved video copy detection. Sig Process Image Commun 29(7):788–806
97. Rouhi AH (2015) Evaluating spatio-temporal parameters in video similarity detection by global descriptors. In: 2015 International conference on digital image computing: techniques and applications (DICTA). IEEE, pp 1–8
98. Min HS, Kim SM, De Neve W, Ro YM (2012) Video copy detection using inclined video tomography and bag-of-visual-words. In: 2012 IEEE international conference on multimedia and expo (ICME). IEEE, pp 562–567
99. Subramanyam AV, Emmanuel S (2012) Video forgery detection using HOG features and compression properties. In: 2012 IEEE 14th international workshop on multimedia signal processing (MMSP). IEEE, pp 89–94
100. Min HS, Choi J, De Neve W, Ro YM (2009) Near-duplicate video detection using temporal patterns of semantic concepts. In: 2009 11th IEEE international symposium on multimedia (ISM). IEEE, pp 65–71
101. Wu Z, Huang Q, Jiang S (2009) Robust copy detection by mining temporal self-similarities. In: 2009 IEEE international conference on multimedia and expo (ICME). IEEE, pp 554–557
102. Ye G, Liu D, Wang J, Chang SF (2013) Large-scale video hashing via structure learning. In: Proceedings of the IEEE international conference on computer vision (ICCV). IEEE, pp 2272–2279
103. Chen Z, Lu J, Feng J, Zhou J (2017) Nonlinear structural hashing for scalable video search. IEEE Trans Circuits Syst Video Technol 28:1421–1433. https://doi.org/10.1109/TCSVT.2017.2669095
104. Liong VE, Lu J, Tan YP, Zhou J (2017) Deep video hashing. IEEE Trans Multimed 19(6):1209–1219
105. Hao Y, Mu T, Goulermas JY, Jiang J, Hong R, Wang M (2017) Unsupervised t-distributed video hashing and its deep hashing extension. IEEE Trans Image Process 26(11):5531–5544
106. Hao Y, Mu T, Hong R, Wang M, An N, Goulermas JY (2017) Stochastic multiview hashing for large-scale near-duplicate video retrieval. IEEE Trans Multimed 19(1):1–14
107. Hu Y, Lu X (2018) Learning spatial-temporal features for video copy detection by the combination of CNN and RNN. J Vis Commun Image Represent 55:21–29
108. Wang L, Bao Y, Li H, Fan X, Luo Z (2017) Compact CNN based video representation for efficient video copy detection. In: 2017 International conference on multimedia modeling. Springer, pp 576–587
109. Jiang YG, Wang J (2016) Partial copy detection in videos: a benchmark and an evaluation of popular methods. IEEE Trans Big Data 2(1):32–42
110. Li J, Zhang H, Wan W, Sun J (2018) Two-class 3D-CNN classifiers combination for video copy detection. Multimed Tools Appl. https://doi.org/10.1007/s11042-018-6047-9
111. Li YN, Chen XP (2017) Robust and compact video descriptor learned by deep neural network. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2162–2166
112. Nie X, Jing W, Ma LY, Cui C, Yin Y (2017) Two-layer video fingerprinting strategy for near-duplicate video detection. In: 2017 IEEE international conference on multimedia and expo workshops (ICMEW). IEEE, pp 555–560

113. Singh RD, Aggarwal N (2017) Detection of upscale-crop and splicing for digital video authentication. Digit Investig 21:31–52
114. Singh RD, Aggarwal N (2017) Detection and localization of copy-–paste forgeries in digital videos. Forensic Sci Int 281:75–91
115. Roopalakshmi R, Venkatesh R, Rahul KM (2015) Robust temporal registration scheme for video copies using visual-audio features. Procedia Comput Sci 57:385–394
116. Roopalakshmi R, Reddy GR (2011) A novel approach to video copy detection using audio fingerprints and PCA. Procedia Comput Sci 5:149–156
117. Roopalakshmi R, Reddy GR (2013) A novel spatio-temporal registration framework for video copy localization based on multimodal features. Signal Process 93(8):2339–2351
118. Saracoglu A, Esen E, Ates TK, Acar BO, Zubari U, Ozan EC, Ozalp E, Alatan AA, Ciloglu T (2009) Content based copy detection with coarse audio-visual fingerprints. In: 2009 Seventh international workshop on content-based multimedia indexing (CBMI). IEEE, pp 213–218
119. Lee SH, Kwon KR, Hwang WJ, Chandrasekar V (2013) Key-dependent 3D model hashing for authentication using heat kernel signature. Digit Signal Process 23(5):1505–1522
120. Lee SH, Kwon KR, Kim DK, Kwon OJ (2015) Hash function for 3D mesh model authentication. In: 2015 21st Korea–Japan joint workshop on frontiers of computer vision (FCV). IEEE, pp 1–5
121. Zhuang N, Ye J, Hua KA (2016) Dlstm approach to video modeling with hashing for large-scale video retrieval. In: 2016 23rd International conference on pattern recognition (ICPR). IEEE, pp 3222–3227
122. Koz A, Lagendijk RL (2010) Perceptual video hashing in P2P networks. In: 2010 IEEE international conference on acoustics speech and signal processing (ICASSP). IEEE, pp 1842–1845
123. Lee S, Yoo CD, Kalker T (2009) Robust video fingerprinting based on symmetric pairwise boosting. IEEE Trans Circuits Syst Video Technol 19(9):1379–1388
124. Cho HJ, Lee YS, Sohn CB, Chung KS, Oh SJ (2009) A novel video copy detection method based on statistical analysis. In: 2009 IEEE international conference on multimedia and expo (ICME). IEEE, pp 1736–1739