Multimedia Tools and Applications (2020) 79:8287–8323
https://doi.org/10.1007/s11042-019-08412-4

Ring decomposition based video copy detection using global ordinal measure features and local features

Alongbar Wary[1] · Arambam Neelima[1]

Received: 20 December 2018 / Revised: 27 September 2019 / Accepted: 1 November 2019 /
Published online: 4 January 2020

Abstract

Visual hashing-based or fingerprinting-based video copy detection approach has been adopted numerously by the video search community due to significant escalation of manipulated copies of original videos over the Internet. Most of the existing video copy detection approaches are robust against the content-preserving distortions such as brightness enhancement and compression, but less robust against the geometric distortions such as rotation and scaling. To mitigate the problem of computation overhead is still challenging in video copy detection. Moreover, there exist a trade-off between discriminability and robustness properties in most of the existing copy detection approaches. In this paper, an effective and fast video copy detection method is presented by exploiting both spatial-temporal information to tackle the above-mentioned challenges. The novelty of proposed method lies in reducing the computation overhead by generating an intermediate candidate database that are similar to the query video using ring-based Ordinal Measure (OM). Then, distinct visual features based on Histogram of Oriented Gradient (HOG) and Singular Value Decomposition (SVD) are extracted from each key-frame of every scenes of the videos of an intermediate candidate database and a query video for copy detection. To avoid the creation of redundant key-frames, the video frames are grouped into different scenes based on Discrete Cosine Transform (DCT). To further preserve the spatial-temporal information, the Temporally Informative Representative Image (TIRI) is used to generate each key-frame of every scenes of videos. The experimental result shows that the proposed method is more efficient and robust against various distortions which outperforms the state-of-the-art copy detection approaches.

Keywords Visual hashing or fingerprinting · Video copy detection · Ring partition · Pre-classifier · Key-frame · OM · DCT · TIRI · HOG · SVD

✉ Alongbar Wary
alongwar56@gmail.com

Arambam Neelima
neelimaarambam@yahoo.co.in

[1] Department of computer science and engineering, NIT Nagaland, Chumukedima, Dimapur 797103, India

1 Introduction

Fostered by the perpetual evolution of Internet technology and the prevalence of digital products in the recent years, online activities related to videos such as uploading, downloading, modifying and viewing have attained notable swell of attention [31]. Since video content can be easily manipulated, disseminated and edited via the Internet, the escalation of video copies has become a critical issue. Moreover, it has become a critical challenge for the owner of the commercial video Web servers such as Netflix and YouTube to administer and detect the gigantic number of videos which are uploaded every day. Therefore, video copy detection approaches have triggered a significant deal of interest in the multimedia industry and the research community. Using such an approach, platform providers such as YouTube can delete similar copies uploaded by users; content owners such as Disney can trace specific videos with respect to copyright infringements and royalty payments. The original video content can be transformed by applying certain distortions such as geometric (e.g., scaling, rotation) and content-preserving (e.g., gamma correction, lossy compression) [36]. Due to the advancement of video editing apps or software, e.g., iMovie, Final Cut Pro 7 etc. and video navigation technology, the users can modify the content of a video by editing or combining identical versions of same video and can find or navigate any sequence of a TV show respectively. Video copy detection approach is indispensable in many real-time application areas such as monitoring of real-time TV commercial media content over multi-broadcast channels [24] and detection of duplicate Web videos [32].

Two main approaches, namely, digital watermarking-based and fingerprinting-based, are extensively used in video copy detection. Digital watermarking embeds identification codes or extra information of the content owner in original videos before it is distributed imperceptibly and can be extracted later to verify the integrity of video content [8]. Watermarking technique cannot detect contents without watermark such as legacy content which has already been distributed. On the other hand, the fingerprinting-based technique extracts perceptual features from the video content which are taken as a unique and compact signature to distinguish different video contents. The quality of video content is not affected as fingerprinting-based video copy detection scheme does not embed any extra information in original videos [25]. Moreover, the combination of two techniques can yield detection which is more robust. In this paper, we focus on the fingerprinting-based video copy detection technique. Figure 1 shows the general flowchart of video copy detection task and its challenges.

Accomplishing a fast and high detection accuracy is the utmost priority of research in the field of video copy detection. A good copy detection approach should be high discriminative as well as more robust against various distortions. It is yet a challenging task to yield more robustness against geometric distortions (e.g., rotation) as most of the recent state-of-the-art approaches are able to handle content-preserving distortions (e.g., contrast enhancement) only with little high accuracy comparably. The prime concerns for fingerprinting-based video copy detection researchers are, the computation overhead and feature representation for videos. The feature representation must be compact and robust against diverse video transformations made by an adversary, which is challenging. Motivated by this, researchers have developed the copy detection methods which extract the features from spatial-domain [20, 36, 38]. The features extracted from spatial-domain are classified into two categories as local features [36, 38] and global features [20, 21]. For instance, spatial-based local features [36] have high discriminative capability as well as robust against distortions as they extract the features from the Region of Interest (ROIs) of video frames, but can incur high computational cost if the features are

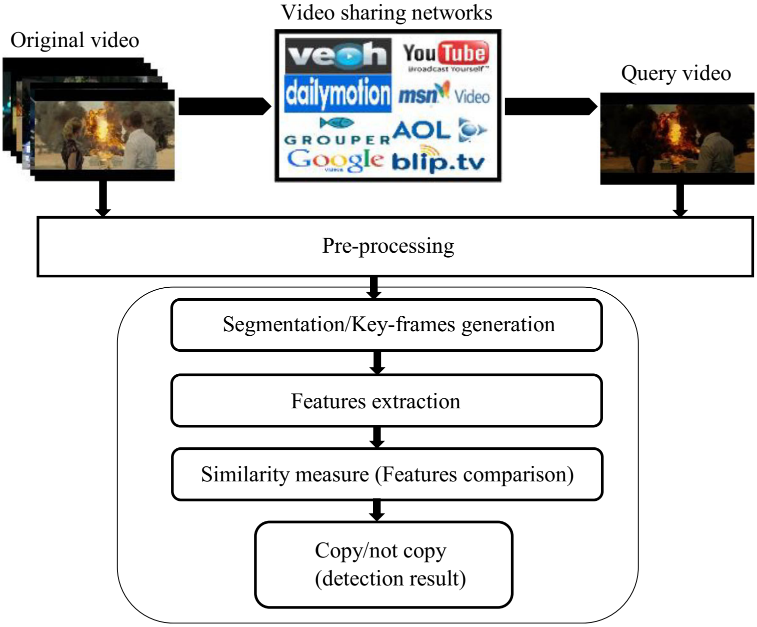

Original video
Video sharing networks
veoh
YouTube Broadcast Yourself
dailymotion
msn Video
GROUPER
AOL
Google
blip.tv
Query video
Pre-processing
Segmentation/Key-frames generation
Features extraction
Similarity measure (Features comparison)
Copy/not copy (detection result)

Fig. 1 General flowchart to show the task of video copy detection and its challenges

extracted from each frame; spatial-based global features [21] are advantageous for efficient and fast video copy detection, but have less discriminative capability. Therefore, some researchers have developed approaches to utilize the properties of both the local and global features [7, 16] to attain highly discriminative as well as robustness. The main disadvantage with this spatial-based approach is the ignorance of temporal information which is also an important property of a video sequence as the motion-based features are represented across time in videos. These approaches cannot handle the temporal transformations such as frame rate change. So, research has been developed to consider the temporal information of a video sequence [12, 51, 53] to solve the issue related with temporal context of the video. This temporal-based approach lacks in discriminative capability and robustness against spatial based video transformations as it ignored the spatial information. Moreover, this approach can cause a vital desynchronization problem between the copy and source video. To overcome the limitations of utilizing respective spatial-based and temporal-based approach, various spatial-temporal based [34, 40, 56] video copy detection approaches have been developed currently. These approaches have attained comparatively better performance in video copy detection field. Authors in [26] have exploited both the spatial and temporal information to achieve better copy detection result. They have extracted both the local and global features using HOG [45] and OM [21] to exhibit high discriminative capability as well as robustness. This approach has shown good performance in object detection and robustness against the content-preserving distortions, but less robustness against the geometric distortions. Moreover, this approach incurs high computational cost as the features were extracted from each frame of video sequence.

The prime issues related with the existing video copy detection approaches are: (1) They have focused mainly on attaining high detection accuracy and given less attention to computational cost which needs to be taken into account for fast and efficient copy detection. (2) Existence of trade-off between discriminative capability and robustness. Less focus on achieving high robustness against the geometric distortions (e.g., rotation). (3) Huge search time and memory space requirement to search and store the large features extracted from each video frame in large database. (4) There exists the formation of redundant key-frames in the key-frame generation process as the key-frames were generated by simply averaging the whole frames of video sequence. The identification of key-frames to represent the video precisely is an important issue.

To consider the above-mentioned issues and to facilitate fast and high detection accuracy, the proposed work focuses on the reduction of computational overhead by generating a compact and robust signature based on digital fingerprinting technique. The high searching time requirement in the giant video database and the issue in identifying precise key-frames that represent the video sequence in which there exists the formation of redundant key-frames in state-of-the-art approaches has mainly motivated to design the proposed pre-classifier technique and the generation of key-frames from each distinct scene. The HOG and OM features descriptor used in [26] is adopted in the proposed work in which they extracted the features from each video frame which leads to high dimensional feature vector and the extracted OM feature values from each rectangular block of frames will vary when geometric distortions specially rotation attack is applied. In the proposed work, the global OM features are extracted from each ring of video frames so that it can resist rotation attack without varying the feature values. The local HOG features descriptor is used because of its highly effectiveness in object detection which are robust against photometric and local geometric transformations [9, 41]. The global OM features descriptor is used because of its sequence ordering or ranking property and it has high robustness against content-preserving distortions such as color shifting and size variation [16, 21]. SVD approach used in [36, 46] is adopted in which they have utilized it for the dimensional reduction purpose only. But in the proposed work, the largest singular values (SVs) are extracted directly from each key-frame to utilize its highly stability properties as the SV do not change when there is a little disturbance or distortions [36, 39].

It is worthwhile to discuss the working mechanism of proposed approach briefly. The video copy detection framework is designed in such a way that an intermediate candidate database is generated to alleviate computation overhead, which is the foremost goal of the work. In the proposed work, firstly, the video frames are partitioned into rings based on masking technique. Then, the global OM [21] features are generated from each ring of the video frames to resist any kind of video transformations or distortions such as rotation and scaling. Therefore, these generated features of reference video database are compared to feature of query video for similarity measure to create an intermediate candidate database in which only the similar videos from reference video database compared to query video are stored. Secondly, HOG [45] and SVD [36, 39] features are generated from each key-frame of every scenes of the videos of an intermediate candidate database respectively, and compared to the features generated from each key-frame of a query video for similarity measure. The HOG features are used because of their high effectiveness in object detection, whereas, SVD features are used because of their stability and compact energy packing properties. Moreover, SVD features are robust against variations in the local statistics of a frame image. The first horizontal and vertical coefficients of DCT [40] are used to group the video frames into different scenes to avoid the formation of redundant keyframes. Only the similar video frames are grouped into a particular scene. Then, the key-frames are generated from each scene of videos using the concept called TIRI [34, 48], which is represented by weighted average of the frames that efficiently represents a short segment of the video. The number of key-frames generated is equal to the number of scenes.

Since, video represent motion-based information across time or temporal direction of video frames, it is requisite to consider the temporal context of a video along with spatial information which also plays a vital role in copy detection. This is procured by using TIRI which preserves both the spatial and temporal information of the video frames. In addition to the creation of an intermediate candidate database, the features extraction from each keyframe of videos helps in reducing the computation overhead to the large extent as it does not have to process the whole video frames. The similarity between the videos in an intermediate candidate database and the query video is computed using Canberra distance metric.

Last but not the least, the proposed method can tackle all the above summarized issues faced by state-of-the-art approaches because of the following reasons: (1) Consideration of both the global features which are generated specifically from each ring of the video frames and local features which are generated from each video key-frame leads to method which is highly discriminative as well as more robust against various distortions. (2) Features extraction from each key-frame of the pre-classified intermediate candidate database videos will avoid extraction of features from large number of dissimilar videos stored in reference video database unnecessarily. This will lessen the computational cost up to large extent. (3) Avoidance of the formation of redundant key-frames from each scene of video frame sequence based on TIRI transform in which different scenes are generated by grouping the similar frames based on the first horizontal and first vertical coefficients of DCT.

The contributions of this proposed approach are highlighted as follows: (1) A novel ring decomposition based video copy detection approach is proposed to exhibit high discriminative capability as well as robustness against any transformations such as geometric and content preserving distortions. The global OM features extracted from each ring of the video frames can handle any kind of distortions such as rotation and scaling attacks made by an adversary to the original video. (2) The proposed method introduces a novel pre-classifier technique to generate an intermediate candidate database to reduce the computational cost up to a large extent which will make the approach faster. Here, the extracted global OM features of reference video database are compared to the OM feature of query video for similarity matching to generate an intermediate candidate database in which only the similar videos from reference video database compared to query video are pre-classified or stored for further processing. Further, the local HOG and SVD features are extracted from each key-frame of this pre-classified intermediate candidate database videos which will avoid extraction of features from large number of dissimilar videos stored in reference video database unnecessarily. Thus, this technique helps in reducing storage requirement as well as computational cost. (3) The proposed approach introduces the creation of different scenes by grouping the similar video frames based on the first horizontal and vertical coefficients of DCT and then TIRI transform technique is applied on each scene to generate key-frames. This process of key-frame generation will highly avoid the creation of redundant key-frames and the use of TIRI will preserve both the spatial as well as temporal information. This technique of feature extraction from each key-frame will also contribute in reducing computational cost as it does not have to process the whole video frames. (4) The proposed approach utilizes the highly stable local SVD features which are extracted directly from each key-frame of video sequence to resist local statistic variations rather than utilizing only for dimensional reduction purpose as in state-of-the-art approaches. (5) Experimental results are analysed using the TRECVID 2010 dataset [1] and the results shows that the proposed approach outperforms the state-of-the-art approaches.

The remaining of this paper is categorized as follows. Section 2 briefly summarizes related works. Section 3 illustrates the proposed method. Experimental results are discussed in section 4. Section 5 discusses about the work done. Section 6 concludes the presentation.

2 Related works

In this section, related topics are reviewed briefly: (1) spatial-based feature representation, (2) temporal-based feature representation, (3) spatial-temporal-based feature representation, (4) deep learning based video analysis, (5) long short-term memory based video analysis.

2.1 Spatial-based feature representation

The exponential growth of video piracy or copyright infringement issues caused by an adversary by manipulating the original videos has triggered the development of copyright detection techniques. Several spatial-based feature representation methods such as extraction of interest points from each keyframe [36] and from region of interest (ROIs) [27] have been exploited by researchers in the state-of-the-art to cop up the issues and problems related with the video copy detection. In spatial-based representation, the features are extracted either from every frame or from each key-frame of videos. It has the ability to locate the salient region of interest points either locally or globally in its spatial space. Typically, the methods can be categorized based on two distinct feature representations in spatial space, namely, local-feature and global-feature respectively. The authors in [36] used a local feature descriptor called scale-invariant feature transform (SIFT) to extract the invariant key points from each key-frame which was selected from video frame sequence. There exists the creation of redundant key-frames and the SIFT feature descriptor is partially invariant to affine transformation as well as illumination changes. The authors in [38] proposed the Speeded Up Robust Features (SURF) and oriented FAST rotated BRIEF (ORB) method to extract the local features. Li et al. [27] used Fast Retina Keypoint (FREAK) method to extract the interest points from ROIs of video frames. Zhang et al. [58] proposed a method to extract the local features from video frames based on SURF [38] for copy detection. These methods have high discriminative capability, but incurs high computation overhead by occupying large memory space for each interest points. Moreover, they are less robust against the global changes. On the other hand, the method proposed in [20] extracts the global features from every rings of each frame image based on binarized statistical image features (BSIF) in addition to invariant color descriptor (ICD). Hua et al. [21] proposed a copy detection method in which the OM feature (global feature) was generated from every blocks of each video frame in which each block was sorted based on their average gray (ranking) level. Himeur and Sadi [19] used BSIF [20] and local color descriptor descriptor (LCD) to generate the global features. The chance of redundant key-frames generation is very high. These methods based on global-feature representation have less discriminative capability and less sensitive to local changes which is its demerits. To improve the performance and to mitigate the demerits of using respective local-features and global-features based methods, several researchers have proposed the methods that combined both local and global features. The authors in [10, 16] proposed a method to integrate both the local and global features for video copy detection. However, they also incur high computation overhead. Moreover, all these methods focus only on spatial-based features representation and is not robust against the temporal-based video transformations such as frame rate change which needs to be enhanced. In contrast, the proposed work mainly focuses on reducing the computation overhead and exploiting both the spatial and temporal information in video copy detection.

2.2 Temporal-based feature representation

Since video represent the motion-based features across time or temporal direction typically, several state-of-the-art methods focused on considering the temporal-based feature representation. For example, Wang et al. [51] introduced a method in which temporal context of key-frames of videos was demonstrated as binary codes. Here, the binary code which represents temporal context of center key-frame was generated by clustering the surrounding frames of each key-frame of a video into two distinct groups based on their temporal relationship with center key-frame. Chen and Stentiford [5] introduced a temporal-based method which extracts the visual features between two consecutive video frames in the time axis or temporal direction for video copy detection. Wu et al. [54] extracted the anchor frames which represent video temporal structure (feature) using Cumulative Luminance Histogram Difference (CLHD) and the statistics gathered in a local window along with adaptive threshold after the temporal subsampling of video frames. Tasdemir and Cetin [47] proposed Mean of the Magnitudes of Motion Vectors (MMMV) and Mean of the Phase Angles of Motion Vectors (MPMV) methods to generate the motion vector along the temporal direction of a video. However, all these methods can cause a vital desynchronization problem between the copy and source video. Moreover, all these methods for copy detection has ignored the importance of spatial information and hence, it is not robust against the spatial video transformations such as brightness enhancement and rotation. These temporal based approaches are applicable only with long duration videos and are not feasible with short duration videos [35].

2.3 Spatial-temporal-based feature representation

To overcome the problems and issues facing by respective spatial-based and temporal-based features representation methods, several research works are focusing on combining both the spatial and temporal information of a video. Yuan et al. [56] introduced Shearlet-based Video Fingerprint (SBVF) method which was applied on TIRI that represent a short segment of video and preserves spatial and temporal information about video segment. The creation of redundant key-frames is the main demerit with this method and the method is less robust against the geometric distortions. Boukhari and Serir [2] also adopted the TIRI concept for generating key-frames in which the Weber Binarized Statistical Image Features (WBSIF) was applied to extract local textural descriptors for video copy detection. The main loop hole with these methods is the creation of redundant key-frames as the key-frames were generated by simply averaging the whole video frames. This method used only the global features which is less discriminative and ignored the importance of local features. Chen and Chiu [4] proposed a video copy detection method which used not only spatial interest points but also temporal interest points along the temporal direction of a video frames. This method incurs huge computational cost as the features were extracted from each video frame. Nie et al. [37] proposed a high order tensor-model based projection method which exhibits assistance and consensus among distinct features to exploit the spatial-temporal information. High computational cost is the main drawback of this method and is unable to handle the video with complex scenes. Lee et al. [26] proposed a video copy detection method in which histogram of oriented gradient (HOG) and ordinal measure (OM) descriptors were used to generate the compact features. The features were extracted from each frame which leads to high dimensional feature vector and the generated OM feature values from each rectangular block of frames cannot resist geometric distortions specially rotation attack. All these methods generate high dimensional feature vector representations which leads to computation overhead and the generation of redundant key-frames is the main demerits with all these methods. Moreover, robustness against certain distortions such as geometric distortions (e.g., rotation, scaling, frame dropping) by these state-of-the-art methods are not up to the peak point which needs to be improved further. Hence, the proposed work considered all these issues to enhance the efficiency and robustness of video copy detection.

2.4 Deep learning based video analysis

The exponential growth of gigantic multimedia database due to the advancement of video sharing Internet technology has triggered the study of deep learning for video analysis recently. The deep learning approaches have been utilized by researchers in video analysis in which most of the approaches focused on action recognition [22, 42], video retrieval [17, 30, 44], and video captioning [15, 43]. The authors in [22] exploited convolutional neural network (CNN) based approach for action recognition in which deep architecture with 2D convolutions were used to extract features from frames without considering temporal model. The deep CNN was trained by structuring a large-scale Sports-1 M dataset. To consider the temporal modelling, two-stream based CNN approach was introduced by authors in [42] where the CNN was applied on pre-trained optical flow features. 3D CNN was utilized by authors in [49] to exploit both the spatial and temporal model within deep architecture in which they have shown good result by training on Sports-1 M dataset. This model is not feasible for long range temporal modelling. Multi-size training and long-range temporal modelling is not supported by 3D convNets [52]. To achieve an efficient video searching in large database, Liong et al. [30] proposed a deep video hashing (DVH) in which they built an end-to-end deep CNN learning model which exploited spatial-temporal information after the stacked convolutional pooling layers to yield compact binary values or codes. The model was trained with a Siamese network discriminatively. Hao et al. [17] introduced a video search approach in which stochastic multi-view hashing (SMVH) [18] was extended to unsupervised hashing through student t-distribution matching scheme called t-USMVH and its deep hashing extension through neural network, so called t-UDH. Song et al. [44] proposed a self-supervised video hashing (SSVH) in which end-to-end hierarchical binary auto encoder and neighbourhood structure were used to encode temporal and visual information simultaneously. To improve the work done in [43] for video captioning, Gao et al. [15] proposed a unified encoder-decoder model called hierarchical Long Short-Term Memory (LSTM) with adaptive attention in which 2D CNN was used to extract frame-level features initially. Next, hierarchical LSTMs consisting of two layers were integrated to decode visual and semantic information to create video caption. Zhang et al. [59] used deep CNN features and graph-based segment matching technique for copy detection. Li et al. [29] utilized two-class 3D_CNN classifiers for copy detection. All of these discussed approaches require pre-defined or fixed size input as well as fixed length input. It requires a huge database storage for pre-trained known dataset as it increases when the network size increases.

2.5 Long short-term memory based video analysis

Since temporal information in video is indispensable for content analysis [52], Long Short-Term Memory (LSTM) network was proposed by Ng et al. [57] for video classification in which both the visual and temporal information were exploited. They introduced distinct feature fusion approaches after stacked convolution and the pooling layers, and five stacked LSTM layers networks was investigated for video classification. Wang et al. [52] proposed a method for action recognition in which they shown that two-stream 3D convNets fusion framework can handle the input of variable size and length which outperformed conventional 3D convNets [49]. They used spatial-temporal pyramid pooling (STPP) to encode video clips or learn fixed length descriptions from arbitrary sized video clip or shot initially. Then, LSTM or CNN-E model was trained on these time varying descriptions to recognize the action. Both the appearance and optical flow clips as input can be trained by this model. This recent development of LSTM networks and because of its effectiveness in sequence learning, LSTM networks are adopted by researchers in other research areas such as natural language processing [3, 28] and video captioning [14, 15, 43]. Lianli et al. [14] proposed an attention-based LSTM approach to bridge the semantic gap between video contents and corresponding language context. Here, attention mechanism utilized the dynamic weighted sum of local 2D CNN to capture the salient features at frame level initially, and then LSTM decoder learnt these features to generate representative words. Finally, the cross-view model was exploited to preserve the consistency between visual contents and the generated language context.

The common limitations of state-of-the-art approaches are: (1) redundant key-frames generation by simply averaging the whole video frames, (2) high dimensional feature vector generation by extracting features from each video frame which leads to high computational cost, (3) less robustness against the geometric distortions, (4) high searching time requirement and (5) existence of trade-off between discriminative capability and robustness. The superiorities of the proposed approach are: (1) scene generation based on the first horizontal and vertical coefficients of DCT leads to formation of key-frames which will represent the video more precisely, (2) extraction of features from each key-frame reduces the computational cost, (3) the extraction of global OM features from each ring of video frame makes the method more robust against the geometric distortions specially rotation attack, (4) the use of pre-classifier technique to generate intermediate candidate database makes the method more faster and efficient by reducing the searching time, (5) the exploitation of both local and global features makes the method highly discriminative as well as robustness against distortions and (6) the use of TIRI transform technique preserves both the spatial and temporal information. Choosing a good copy detection approach mainly depends on what we are focusing for and where we are focusing it.

3 Proposed method

The main objective of the proposed method is to reduce the computational overhead and achieve efficient video copy detection to overcome the limitations of state-of-the-art video

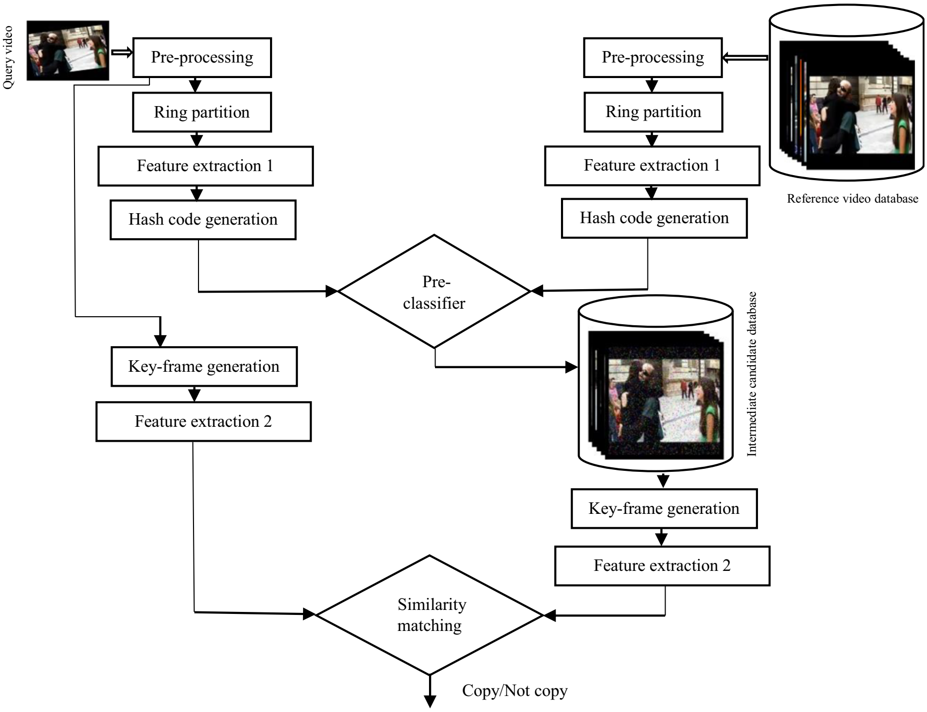


Fig. 2 Block diagram of the proposed video copy detection method

copy detection methods. Figure 2 depicts the basic steps of the proposed video copy detection method.

3.1 Pre-processing

In the initial step, the input sequence of a video is converted to a standard signal in terms of the number of frames or frame dimensions via resampling and smoothing. The input video sequence is resampled at fixed frame-rate R frames per second (fps) to handle frame rate change. Then, each frame is downsized into fixed dimensions to normalize the width and height into fixed values w and h respectively. This step strengthens the robustness of proposed method against frame resizing as different videos may have different frame size. In the next step, each frame is transformed to grayscale so that the proposed method become robust against color variations and also applicable to grayscale videos. Then, Gaussian low-pass filter is applied to each resized video frame image. This operation reduces the high-frequency components that is easily influenced by minor transformations, e.g., filtering and noise contamination. Generally, symmetric convolution mask can be used for this purpose. Let $E^{(g)}(l, k)$ be the pixel element in the l^{th} row and k^{th} column of the mask, where l and k are the distances from the origin in x-axis and y-axis of the frame image respectively. Thus, it can be represented as follows.

$$E^{(g)}(l,k) = \frac{E^{(1)}(l,k)}{\sum_l \sum_k E^{(1)}(l,k)} \qquad (1)$$

in which $E^{(1)}(l, k)$ is presented as

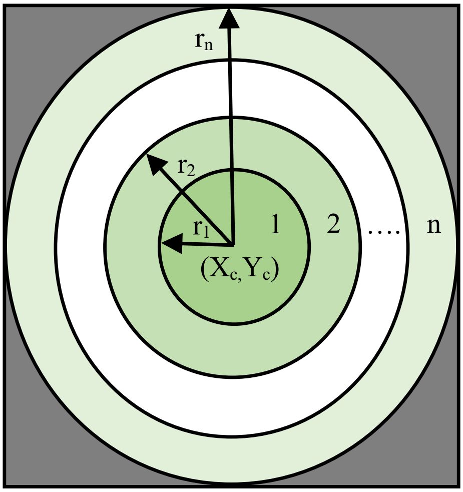


Fig. 3 Schematic representation of ring partition of a frame image based on masking

$$E^{(1)}(l,k) = e^{\frac{-(l^2+k^2)}{2\sigma^2}} \tag{2}$$

where, σ is the standard deviation of Gaussian distribution.

3.2 Ring partition

To make the proposed method invariant against rotation attack, each pre-processed frame image is partitioned into equal rings based on masking technique. Generally, this concept was used for image hashing in [23], in which a raw image was partitioned into equal rings based on masking. Since, it has good rotation invariant property, this concept is adopted in the proposed method. The region in the inscribed circle of a frame image remains same even after rotation operation as the frame image generally rotates with respect to the center of frame image. Thus, the information in the rings of the original video frame and rotated video frame image does not change. This property strengthens the robustness of the proposed method against rotation attack by extracting salient features from each ring of the frame image. An illustrative partition with four rings is shown in

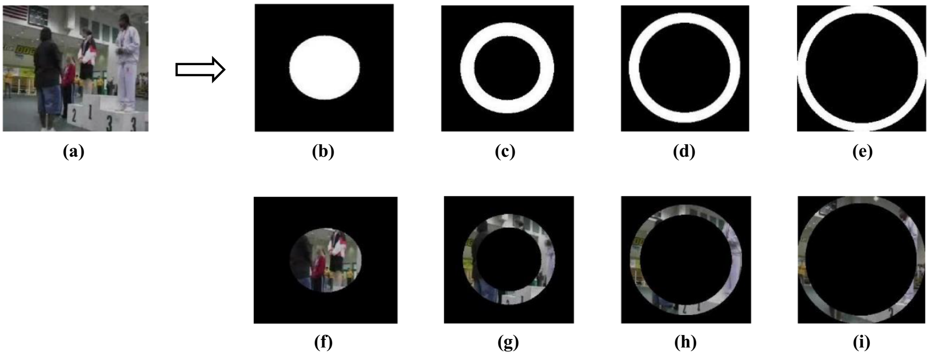


Fig. 4 Visual representation of generating the ring frame image based on masking **a** Video frame image **b** Mask 1 **c** Mask 2 **d** Mask 3 **e** Mask 4 **f** Ring-1 frame image **g** Ring-2 frame image **h** Ring-3 frame image **i** Ring-4 frame image

Fig. 3. The masking technique has been used to generate the secondary ring frame image using convolution operation. The pixels whose distances with respect to the center coordinate pixel of the frame image are less than or lies within the range of radius of a particular concentric circle are masked with value 1 (white), while other pixels of the frame image are masked with value 0 (black) which does not lie within the range of radius of a particular concentric circle. The Fig. 4a-i depicts the visual representation of generating the ring frame image based on masking technique. The idea behind the consideration of four equal ring partitions is to generate the low dimensional feature vector. As the number of rings increases the dimension of feature vector increases. However, consideration of rings less than four will result in loss of information. So, the proposed method adopted the standard four equal ring partitions to mitigate computation overhead. This technique overcomes the limitations of traditional block-based partition of the frames as block partition is not good enough to resist the geometrical attacks such as resizing and rotation, since block size as well as contents of block can get changed with a small angle orientation [20]. The algorithm 1 illustrates the implementation details.

Algorithm 1. Pseudocode of ring partition based on masking technique

```
Input: The number of video frames, $K$
Output: The secondary ring frame image, $S_i$
An arbitrary size of $K$ is converted to a standard size of $W \times W$ using bilinear interpolation
Each frame $K$ is converted to grayscale form
Initialize $N = 4$, $R_N = \frac{L}{2}$, $(X_c, Y_c) = \begin{cases} X_c = \frac{L}{2} + 0.5, \quad Y_c = \frac{L}{2} + 0.5\,; & \text{if } L \text{ is even} \\ X_c = \frac{(L+1)}{2}, \quad Y_c = \frac{(L+1)}{2}; & \text{if } L \text{ is odd} \end{cases}$
where, $N$ is the number of rings, $R_N$ is the radius of rings, $(X_c, Y_c)$ is the center of the frame image $K$
$A_i = \pi R_N^2$    // Area of the inscribed circle of frame image
$A_c = \lfloor A_i / N \rfloor$  // Area of each concentric circle of frame image
$R_1 = \sqrt{A_c / \pi}$  // Radius of the first ring
for $i = 2$ to $N - 1$ do
    $R_i = \sqrt{(A_c + \pi R_{i-1}^2)/\pi}$    // To determine the radius of the $i^{th}$ ring
end for
For $i = 1$ to $N$ do
    $M_i = \begin{cases} 1, & D_{XY} \le R_i \\ 0, & D_{XY} > R_i \end{cases}$
    where, $D_{XY} = \sqrt{(X - X_c)^2 + (Y - Y_c)^2}$  // Mask $M_i$ of the radius $R_i$ is created with dimension $W \times W$
end for
$S_i = M_i * K$  // Convolution of mask $M_i$ with video frames $K$ of size $W \times W$ to generate ring frame image $S_i$
```

3.3 Feature extraction 1

In the first stage of feature extraction, global ordinal-measure [21, 26] feature representation is used to generate the feature vector from each ring of every frames of the whole video sequence. This OM feature preserves both the spatial-temporal information which is robust against color shifting and size variation. Further, to improve the robustness of the proposed method against rotation attack, the feature vector is extracted from each ring of every frames. The extracted feature information will still remain same even after the frame is rotated. Here, the feature vector is extracted from four equal rings of every frames by computing the means of every sub-rings. Then, these 4 mean intensities are sorted based on their ranking to get the ordinal information. The result of each possible combination of ordinal measure can be treated

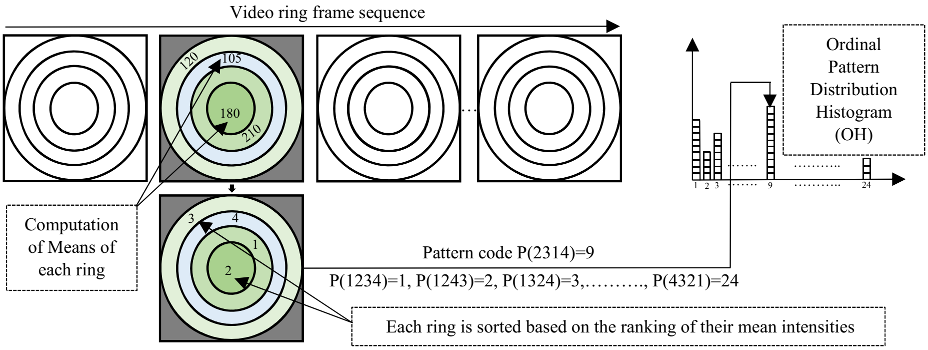

Video ring frame sequence
120
105
180
210
Computation of Means of each ring
3
4
1
2
Pattern code P(2314)=9
P(1234)=1, P(1243)=2, P(1324)=3,………., P(4321)=24
Each ring is sorted based on the ranking of their mean intensities
Ordinal Pattern Distribution Histogram (OH)
1 2 3
9
24

Fig. 5. Pictorial representation of ordinal pattern distribution histogram (OH) generation

as a distinct pattern, and then each frame will be quantized to a pattern code. The Ordinal Histogram (OH) is generated by accumulating all the patterns along the temporal axis. Therefore, OH of 24 (4 ! = 24) dimension is generated that represent the video segment compactly. For instance, let N_r be the number of rings of each frame of video sequence. The OM feature is extracted from all the N_r rings of each frame, and then each ring is sorted according to their mean intensities or average gray level based on ranking. The OM $O_m(t)$ of the t^{th} frame is described as follows:

$$O_m(t) = (R_0, R_1, \ldots, R_{N_r}) \tag{3}$$

where R_k is the rank of the k^{th} ring. Figure 5 illustrates the generation of OH. Table 1 shows each possible combination of pattern code for generating OH of 24 dimensions (4! = 24).

3.4 Hash code generation

The compact hash code is generated from ordinal pattern distribution histogram (OH) of video sequences. Here, the length of hash code will always range up to 24 as size of the pattern code generated based on OM feature extraction technique. Generally, the purpose of using hash code of fixed length is to reduce the computation overhead.

3.5 Pre-classifier

The main novelty of the proposed method lies in the use of pre-classifier which is used to construct an intermediate candidate database. In this step, the generated compact hash code of query video and the reference database videos are compared for similarity using Canberra distance metric. The basic idea is that, if the distance is less than a pre-defined threshold value, then the video is considered as similar and vice-versa. Only the similar videos are retrieved

Table 1 Pattern code table for generation of ordinal histogram (OH) of size 24 (4!)

Pattern code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Ordinal pattern	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4
	2	2	3	3	4	4	1	1	3	3	4	4	1	1	2	2	4	4	1	1	2	2	3	3
	3	4	2	4	2	3	3	4	1	4	1	3	2	4	1	4	1	2	2	3	1	3	1	2
	4	3	4	2	3	2	4	3	4	1	3	1	4	2	4	1	2	1	3	2	3	1	2	1

from the reference video database and stored in a newly constructed intermediate candidate video database. All the subsequent processing will now be done on the videos of an intermediate candidate database rather than processing on the whole reference database videos unnecessarily. This technique will help in reducing the computation time up to large extent and thus, will enhance the overall performance of the proposed method compared to state-of-the-art methods. The Canberra distance metric between two hash codes $Hc1$ and $Hc2$ of length T is describe as follows:

$$dis(Hc1, Hc2) = \sum_{i=1}^{T} \frac{|Hc1_i - Hc2_i|}{|Hc1_i| + |Hc2_i|} \tag{4}$$

in which, the resultant distance will always range between 0 to 1.

For setting the above-mentioned threshold value to pre-classify only similar videos of the reference database to an intermediate candidate database with respect to the query video, the intra-hash and inter-hash distance between perceptually similar and dissimilar videos are calculated using the same Canberra distance metric based on the OH respectively. Initially, only five videos are selected from the database for this evaluation. The details about the collection of videos from the dataset and the attacks that are applied to the original videos for distance calculation are studied properly in section 3. Figure 6 shows the intra-hash and inter-hash distance between perceptually similar and dissimilar videos of the database.

It can be observed from Fig. 6 that the maximum intra-hash distance between perceptually similar videos is 0.25 and the minimum inter-hash distance between perceptually dissimilar videos is 0.1. If we set the 0.25 as the threshold value, then all the perceptually similar videos will be pre-classified into an intermediate candidate database but the discriminative capability will be compromised. That means some dissimilar videos might also be pre-classified. If we set the mean of the maximum intra-hash distance and minimum inter-hash distance as the threshold value (i.e., 0.17), then both the robustness between similar videos and discriminative capability between dissimilar videos will be compromised. That is, all the perceptually similar videos will not be pre-classified and some of the dissimilar videos might be pre-classified. Here, to pre-classify all the perceptually similar videos, 0.25 is considered as the threshold value. Thus, if the distance between hash code of all the original videos in reference database and the query video is less than 0.25 threshold value, then the corresponding videos will be

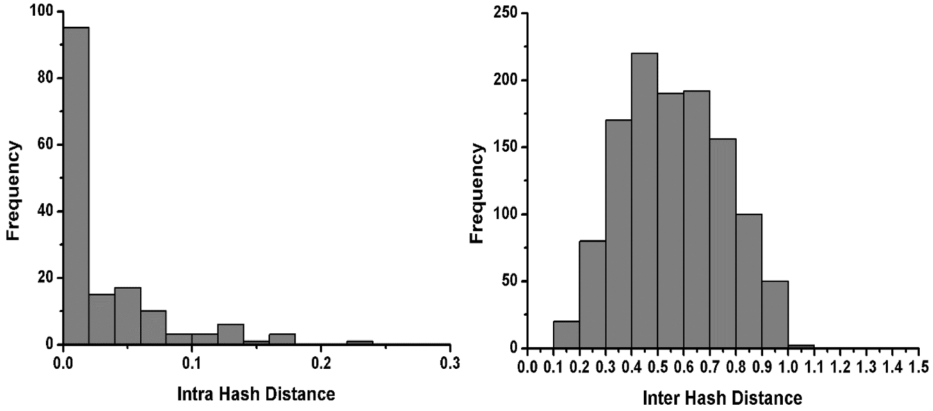


Fig. 6 Intra-hash and inter-hash distance between perceptually similar and dissimilar videos

considered as perceptually similar and will be pre-classified into an intermediate candidate database for further processing.

3.6 Scene formation and key-frame generation

The key-frames are generated using TIRI transform [2, 34] from each scene of video sequence. This technique calculates the weighted average of all the frames of each scene of the videos to construct a representative image (key-frame) which is a single blurred image basically. The purpose of using TIRI transform for the generation of key-frames is to preserve both the spatial and temporal information as well as to alleviate the computation time by reducing the size of the feature vector. The proposed method adopted the same TIRI-based key-frames generation technique as in [2], the only difference is that the proposed method generates the key-frames from each scene of a video sequence. In other word, the number of generated key-frames is equal to the number of scenes. Therefore, the key-frames are obtained as follows: let $Y_{p,\,q,\,r}$ be the luminance value of the (p, q)th pixel of the rth frame in a set of H frames in each scene. The pixels of key-frames are then generated as a weighted average of the frames as shown below.

$$Y^{'}_{p,q} = \sum\nolimits_{r=1}^{H} \omega_r Y_{p,q,r} \tag{5}$$

where, ω_r is the weight corresponding to the rth frame which is describe as follows:

$\omega_r = 1$	for simple averaging,
$\omega_r = j$	for linear weight,
$\omega_r = 1 - e^{\frac{(j-\mu)}{\sigma}}$	for Gaussian weight,
$\omega_r = \gamma^j$	for exponential weight,

where, μ, σ and γ are constants used in Gaussian weighting and exponential weighting function respectively. The exponential weighting function with value $\gamma = 0.6$ is used in the proposed work as according to [2], this exponential weight value could give the optimum result.

The above-mentioned scenes of video sequence are generated based on the 2D-DCT [6, 34] transform coefficients. The DCT transform is applied to each frame of the query video as well as the videos of pre-classified intermediate candidate database. Here, the DCT transform is applied on overlapping (50% overlap) blocks of size $2u \times 2u$ of each frame of video sequence to obtain the coefficients or features. Only the first vertical and the first horizontal DCT coefficients are extracted from every block of each frame as they have the high energy or information compactness properties. The features of each frame which are generated from each block are concatenated to obtain a compact feature vector of each frame respectively. Then, each feature is compared to a specified threshold value, i.e., median value of the feature vector and the binary hash or fingerprint of each frame is obtained. These generated binary fingerprints of consecutive frames of a video sequence are then compared for similarity measure to create the distinct scenes. Here, the same Canberra distance metric in Eq. (4) is used for the similarity measure. Initially, the first frame of a video is assigned to the first scene. Then, the fingerprint of second consecutive frame is compared to the fingerprint of first frame in the first scene. If the distance is within the certain pre-specified threshold value, then the second frame will be assigned to the first scene, otherwise the frame will be assigned to the newly created scene or second scene. Likewise, the next consecutive frames are compared to the last frame of each scene for similarity measure and then assigned to the respective scenes. Only the similar

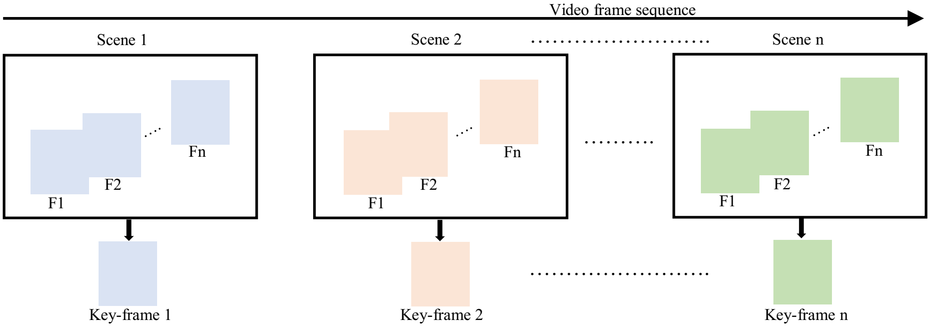


Fig. 7 Schematic illustrations of generation of key-frames from each scene of a video sequence

frames are grouped into the particular scene. Hence, this technique of key-frames generation from each scene of a video sequence will avoid the formation of redundant key-frames. Moreover, this technique will help in reducing the computation overhead to the large extent. Figure 7 shows the schematic illustrations of generation process of key-frames from each scene. The process of generating the key-frames from each scene are summarized as follows:

Let $f_{s,\,t}$ be the frame images at spatial coordinates (s, t) of a video sequence. Firstly, each video frame is segmented into overlapping blocks of size $2u \times 2u$ as described below:

$$B^{l,z} = \{f_{s,t} | s\in lu \pm u, t\in zu \pm u\} \tag{6}$$

where, $l \in \{0, 1, 2, 3, \ldots, W/u-1\}$; $z \in \{0, 1, 2, 3, \ldots, H/u-1\}$. Secondly, the first vertical and the first horizontal DCT coefficients are extracted from each block of every frames. The first vertical coefficient $\alpha_{l,\,z}$ can be derived for $B^{l,\,z}$ as:

$$\alpha_{l,z} = V_c^T B^{l,z} 1 \tag{7}$$

where, $V_c = \left[\cos\left(\frac{0.5\pi}{2u}\right), \cos\left(\frac{1.5\pi}{2u}\right), \ldots, \cos\left(\frac{\pi-0.5\pi}{2u}\right)\right]^T$ and the column vector of all ones is signified by 1. The first horizontal coefficient $\beta_{l,\,z}$ can be derived similarly for $B^{l,\,z}$ as:

$$\beta_{l,z} = 1^T B^{l,z} V_c \tag{8}$$

All the extracted coefficients of each frame are concatenated to generate single vector g of each frame. Then, the median m_d value of each frame is calculated, i.e., the median of the features of vector g. Next, the binary fingerprint or hash f of each frame is generated from vector g as follows:

$$f_i = \begin{cases} 1, & g_i \geq m_{di} \\ 0, & g_i < m_{di} \end{cases} \tag{9}$$

This generated binary hash f of each frame is then compared to create distinct scenes of a video sequence. Initially, the first frame of a video is assigned to the first scene s_1. Subsequently, the binary hash f of second consecutive frame is compared to the f of the first frame in the first scene of a video sequence using Eq. (4). If the distance is less than or within the range of pre-specified threshold value t, then the second frame of a video sequence will be assigned to the first scene s_1, otherwise, the frame will be assigned to the second scene s_2 or next consecutive scene s_i. Likewise, all the consecutive frames of a video sequence are compared to the last frame of each scene and are assigned to the respective scenes based on the similarity measure. Lastly, the key-frames k_i are generated from each consecutive scene s_i of a video sequence based on the TIRI [2] transform using Eq. (5).

3.7 Feature extraction 2

In the second stage of feature extraction, the HOG [9, 13, 26, 41, 45] and SVD [11, 36, 39, 55] methods are used for feature representation. Firstly, the same local HOG descriptor used in [9, 41] is adopted for feature extraction which is similar to the Lowe's SIFT [33, 36] feature descriptor. The HOG feature descriptor has shown significant performance in object detection (e.g., human detection) and robustness against the photometric and local geometric transformations [9]. The HOG features are extracted from each key-frame of a query video and the videos of pre-classified intermediate candidate database. In brief, the HOG descriptor divides each key-frame image window into a dense grid of cell (spatial regions), where each cell contains or accumulates a local histogram of edge orientations and gradient directions over orientation bins (i.e., 9 bins in 0–180 degrees). The key-frame image gradient magnitude and the angle are computed at each pixel of the cell, and the votes weighted by the magnitude of gradient vector are accumulated into the corresponding orientation bin over the pixels of each uniformly spaced cell. Then, these small cells of each key-frame image are grouped into distinct larger overlapping blocks of size 16 × 16 of four 8 × 8 pixel cells and a local contrast normalization process is done on each block to yield better illumination and size invariance. Therefore, concatenation of the normalized histograms of all the blocks of each key-frame will give the final histogram feature descriptor of a video sequence. During voting, the Gaussian mask and the spatial-angular linear interpolation is used for each block to reduce aliasing [9]. For instance, let S be the video sequence with key-frames $K(j)$, where $j = 1, 2, \ldots, M$, and let $K(j)_u$ and $K(j)_v$ be the gradients of the key-frame image $K(j)$ in the u and v directions respectively. The gradients are computed using 1-dimensional filters and its transpose as follows:

$$K(j)_u = K(j)*[-1\ 0\ 1] \tag{10}$$

$$K(j)_v = K(j)*[-1\ 0\ 1]^T \tag{11}$$

where, $*$ indicates a convolution operation between the key-frames and the filter mask. The magnitude R and orientation angle θ of the gradients of each pixel in all the key-frame image are calculated as follows:

$$R = \sqrt{K(j)_u^2 + K(j)_v^2} \tag{12}$$

$$\theta = \arctan\left(K(j)_u / K(j)_v\right) \tag{13}$$

Secondly, the Singular Value Decomposition (SVD) is applied on each key-frame of a query video and the videos of the pre-classified intermediate candidate database for low-level feature extraction. SVD is used to diagonalize and decompose matrices optimally in numerical analysis, which results in maximum signal energy packing into few coefficients. Let $D_{k,\,l,\,p}$

be the standard deviations matrices at *k*th row and *l*th column in the *p*th key-frame of a video sequence. SVD of $D_{k,\,l,\,p}$ is defined as follows:

$$D_{k,l,p} = USV^T \quad 1 \le k \le N, 1 \le l \le M \tag{14}$$

where S is the singular diagonal matrix, U and V are the left and right orthogonal matrices which represent the eigenvector of DD^T and D^TD respectively. The singular values (*SV*s) of a diagonal matrix represent the intrinsic algebraic properties and have good stability, that is, *SV*s do not change when there is a little disturbance in the key-frame image [11, 39]. The feature vector or fingerprint f_p of the *p*th key-frame is generated by the largest *SV* $S_{k,\,l,\,p}$ of each matrix $D_{k,\,l,\,p}$ as follows.

$$f_p = \left[S_{1,1,p}, S_{1,2,p}, \dots, S_{N,M,p}\right] \tag{15}$$

The fingerprints f_p extracted from consecutive F key-frames are concatenated to generate the video fingerprint. Finally, the histogram of 32 bin size is generated as the SVD feature vector or fingerprint of a video.

3.8 Similarity matching

The HOG and SVD feature vector are concatenated together to generate the final feature vector. The same Canberra distance metric in Eq. (4) is applied to the feature vector or fingerprint for similarity matching between a query video and the intermediate candidate database videos. If the distance is less than or within a certain pre-defined threshold value, then the video will be considered as a copy or pirated version of an original video.

4 Experimental results

In this section, the performance of the proposed method is evaluated using a TRECVID 2010 video dataset [1]. In general, a video copy detection approach should be discriminative for the distinct videos but robust for the perceptually similar videos under various distortions or common signal processing transformations. We have selected the 20 original videos of different sizes and maximum length of 3 min and 33 s from the TRECVID 2010 dataset [1] for the purpose of performance evaluation. Then, the proposed method is applied to each original video of the database to generate the fingerprint database. Next, various attacks or

Table 2 Parameter values of each attack

Attack	Description	Parameter value
Frame Swapping	Percentage of frames to be swapped	10, 20, 30, 40, 50
Frame Cropping	Percentage of frames to be cropped	10, 20, 30, 40, 50
Impulse Noise	Impulse noise ratio	0.05, 0.10, 0.15, 0.20, 0.25
Gaussian Noise	Variance	0.01, 0.02, 0.03, 0.04, 0.05
Median Filter	Window size	2 × 2, 3 × 3, 4 × 4, 5 × 5, 6 × 6
Scaling	Scaling factor	2 × 2, 3 × 3, 4 × 4, 5 × 5, 6 × 6
Rotation	Angle in degree	−10°, −5°, 5°, 10°, 20°
Blurring	Amount of blur (Sigma)	1, 2, 3, 4, 5
Sharpening	Sharpening value (Alpha)	0.1, 0.2, 0.3, 0.4, 0.5
Gamma Correction	Gamma value	0.6, 0.7, 0.8, 0.9, 1.0

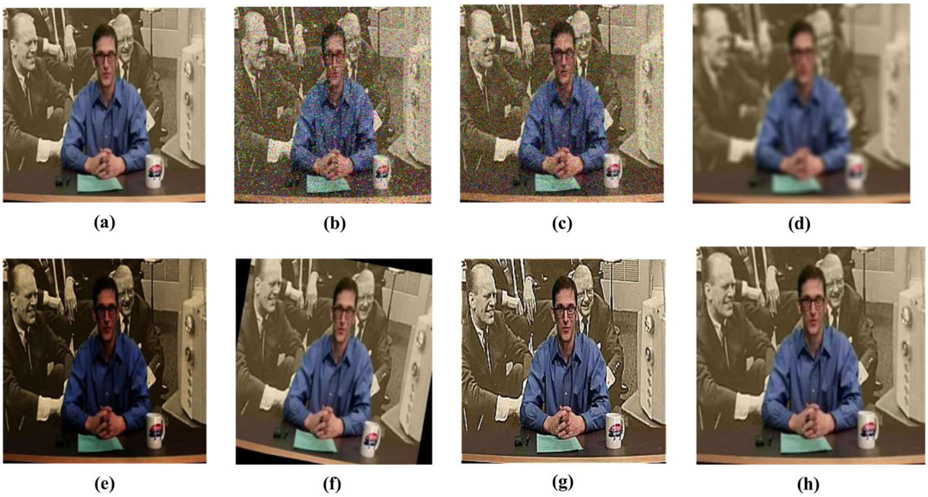


Fig. 8 Example of attacks: **a** Original video frame **b** Impulse noise **c** Gaussian noise **d** Blurring **e** Gamma correction **f** Rotation **g** Sharpening **h** Scaling

transformations are applied to the original videos in a database to create query or test videos. Each test video is compared to the original videos in the database for evaluating the efficiency of the proposed method. Here, ten different attacks with five distinct parameter values are used namely frame swapping, frame cropping, impulse noise, Gaussian noise, median filter, scaling, rotation, blurring, sharpening and Gamma correction. Table 2 shows the detailed corresponding parameter values of the attacks. Figure 8 depicts an example of distinct attacks or transforms used for evaluation to test the performance or efficiency of the proposed video copy detection method.

4.1 Robustness and discriminative capability testing

4.1.1 Robustness testing

To test the robustness of the proposed method against various attacks, simulations are done on a small dataset of 5 original videos out of 20 original videos namely Politicallunch, Misha_Williams, Brandon, Cgw_trailer and JesseGriffithsBwpIntro collected from TRECVID 2010 dataset [1]. Then, ten different transforms or attacks with five distinct parameter values which are mentioned above in Table 2 are applied to the sequences of these original videos, thus resulting in 5 × 10 × 5 = 250 distinct test videos in database for testing. These transformed or test videos are compared to the corresponding original videos in the database. For a method to be robust, the pairs of videos that is the original videos and their attacked videos should be perceptually similar. The intra-distances between the fingerprints extracted from the original videos and the attacked videos are computed using Canberra distance metric (Eq. 4) for similarity measure. The intra-distance close to 0 indicates that the original videos and their attacked videos are perceptually similar and vice-versa.

Figure 9 depicts the validation of robustness on five original videos for ten different attacks with each having five distinct parameter values. The detailed histogram of intra-distance between the original and their attacked videos is shown in Fig. 10a. In Fig. 9a and 9b, the proposed method is showing hundred percent robustness against the frame swapping and frame cropping attacks with different parameter values of 10–50 respectively, except that the frame swapping attack shows slight increase in distance for 50 parameter value. In Fig. 9c, the results of applying impulse noise attack shows that there is a change in distance between the parameter values 0.15 to 0.25. The distance of Cgw_trailer and JesseGriffithsBwpIntro videos are much less than 0.20 threshold value for 0.20 and 0.25 noise ratios. Fig. 9d shows the results of applying Gaussian noise attack on different videos for distinct variance values of 0.01–0.05 respectively in which distance increases with increase in the values of variance, except for video Brandon whose distance is much lower than 0.20 for 0.04 and 0.05 variance values. The Fig. 9e and 9f shows the results of applying median filter and scaling attacks with window size and scaling factor of 2–6 respectively. It can be noticed that distance changes slightly in both the attacks from 4 to 6 window size and scaling factor respectively. But the distance is very less for video Cgw_trailer for scaling attack in Fig. 9 f. The results of applying rotation attacks for distinct angles of -10° to 20° in Figure 9 g shows lesser distance in -10° and 10° angles compared to other angles, except only for video Misha_Williams shows an increase in distance for -10° angle. Figure 9 h depicts the results of applying blurring attack for different Sigma values of 1–5 in which the distance increases with increase in the Sigma values. The distance is much less for sigma value of 1. In Figure 9i, the results of applying sharpening attack for distinct Alpha values of 0.1–0.5 shows that distance for video Cgw_trailer is much less for Alpha value nearer to 0.1 compared to other videos. Figure 9j depicts the results of applying the Gamma correction attack for distinct Gamma values of 0.6–1. Here, all of the videos are showing almost same changes in distance with increase in the Gamma values and are less than 0.20, except for video Cgw_trailer whose distance is much less for 0.7–0.9 Gamma values compared to other videos. It can be clearly studied from Figure 9 that all of the intra-distances between the original and their attacked videos lies within the 0.20 value.

4.1.2 Discriminative capability

To test the discriminative capability, the proposed method is applied to various perceptually distinct videos in the database. Two perceptually distinct videos should not be similar for a good video copy detection method. The more discriminative capability a method has the more precisely we can distinguish between perceptually different videos. Here, we have calculated the inter-distance statistics using the Canberra distance metric (Eq. 4) between the fingerprints of perceptually distinct videos in the database to demonstrate the discriminative capability. Figure 10b illustrates the histogram of the frequencies of inter-distance obtained from pairs of perceptually dissimilar videos. There should be a significant difference between the minimum inter-distance and maximum intra-distance for the perceptually distinct videos to be distinguished correctly. The threshold value can be set within the range of 0.2 (maximum intra-distance) to 0.3 (minimum inter-distance) by clearly observing the Figure 10 to distinguish between perceptually similar and distinct videos. It can be noticed that no two different videos will be falsely identified as similar videos if 0.30 is selected as a threshold value. Here, 0.25 i.e. mean value of the threshold range is set as the maximum threshold value. The distance values that lies below the threshold value will be considered as perceptually similar videos and above it is considered as perceptually distinct videos. It can be analysed that the histogram of intra-distance lies below the range of threshold value completely whereas the histogram of inter-distance lies above the threshold range completely.

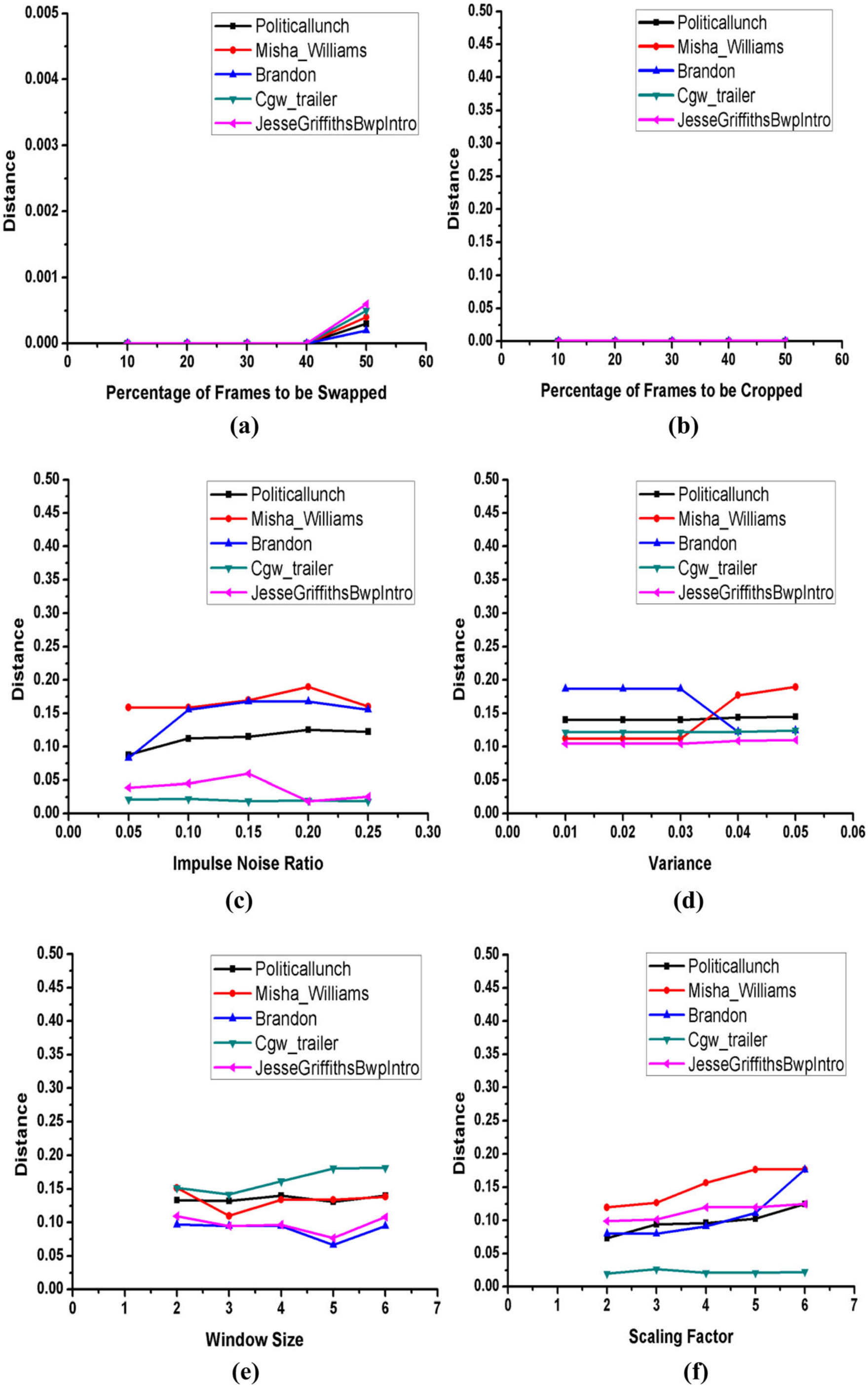
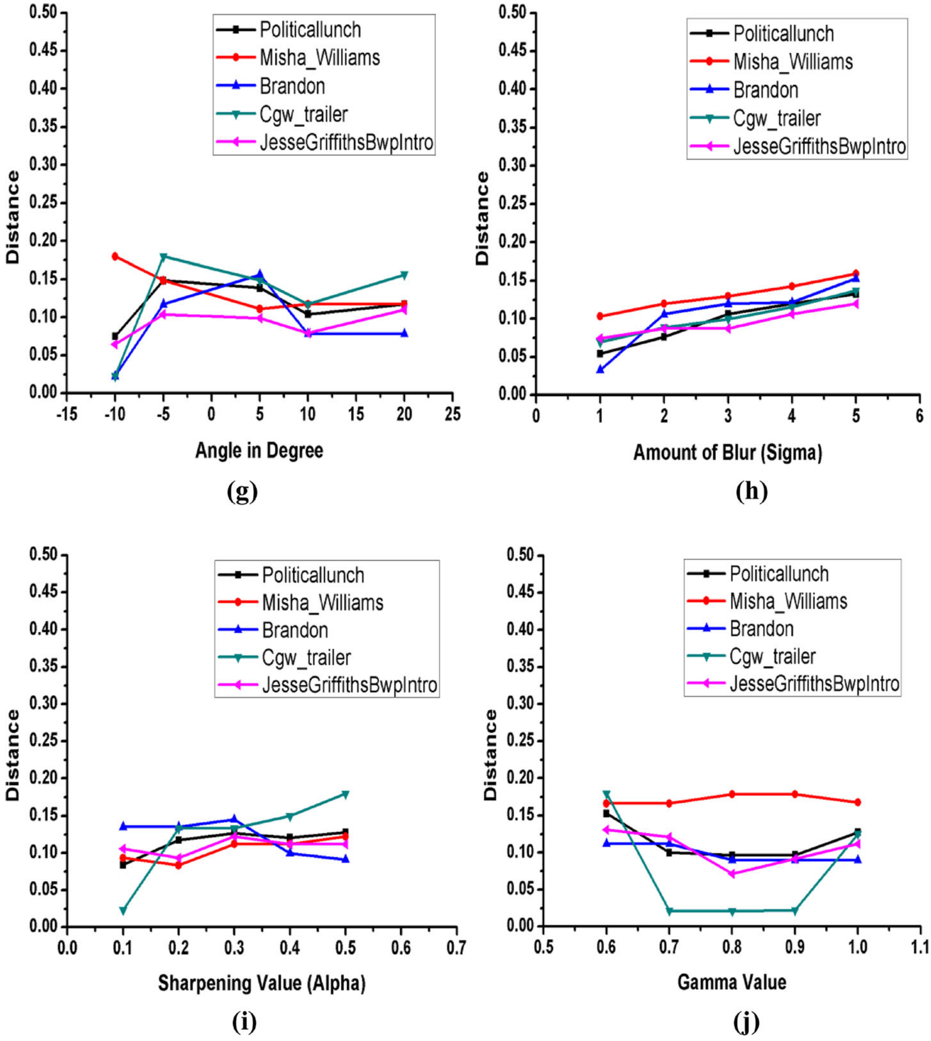


Fig. 9 Validation of robustness: **a** Frame swapping **b** Frame cropping **c** Impulse noise **d** Gaussian noise **e** Median filter **f** Scaling **g** Rotation **h** Blurring **i** Sharpening **j** Gamma correction



(g)

(h)

(i)

(j)

Fig. 9 (continued)

Thus, it can be observed from both the Figure 9 and Figure 10 that the proposed method is robust against various geometrical attacks or transforms under consideration as well as capable of discriminating or distinguishing perceptually distinct videos.

4.2 Performance evaluation and comparison

4.2.1 Evaluation metric

The performance of video copy detection method is measured by the percentage of how precisely the pirated version of an original video is detected. Implementation of the proposed method was run on Matlab R2018a with 64-bit PC, 2.4 GHz CPU, intel core i7 processor and

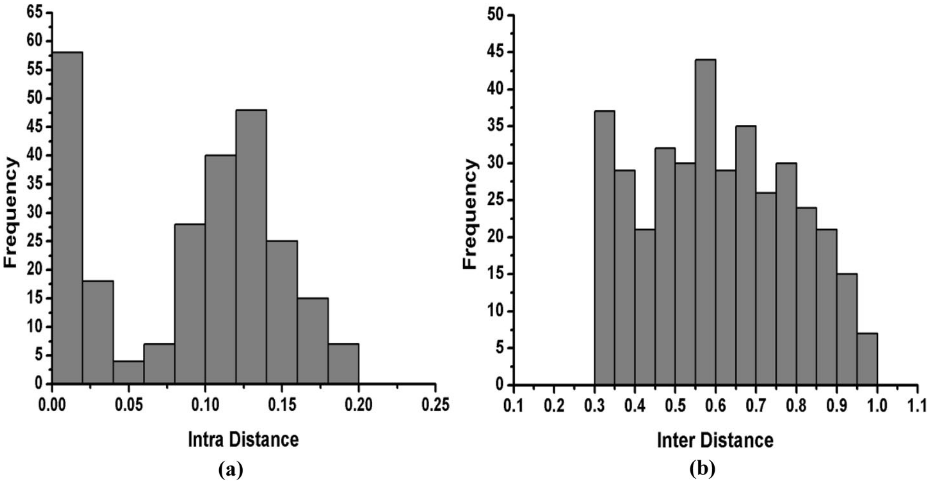


Fig. 10 Histograms of the frequencies of intra and inter distance: **a** Intra-distance for perceptually similar videos **b** Inter-distance for perceptually distinct videos

8 GB RAM. To evaluate the performance of the system or approach, the receiver operating characteristics (ROC) curve is used which is a plot of True Positive Rate (TPR) and False Positive Rate (FPR). Also, the F-score (F_α) [50] is adopted in our work for evaluation and comparative study. The TPR, FPR and F_α are defined as follows:

$$\mathrm{R_{TPR}}=\frac{\mathrm{TP}}{\mathrm{N_S}},\mathrm{R_{FPR}}=\frac{\mathrm{FP}}{\mathrm{N_D}}\,\mathrm{F}_\alpha=\left(1+\alpha^2\right)\frac{\text{Precision.Recall}}{\alpha^2.\text{Precision}+\text{Recall}} \tag{16}$$

where, TP, FP, $\mathrm{N_S}$ and $\mathrm{N_D}$ represents true positive (video sequence that matched precisely in the positive set; hit); false positive (video sequence that matched with a distinct video); total pairs of perceptually similar videos and total pairs of perceptually distinct videos respectively. The α represents how much weightage should be given to precision versus recall. The recall is same as the true positive rate (TPR) which is a measure of robustness of the method. Whereas, precision is defined as the percentage of precise hits within all of the detected video copies and is a measure of discriminative capability of the system. The F-score of an approach is a number ranging from 0 and 1. The value 1 represents the correct classification of a system, that is

Table 3 F-score of the proposed method for threshold values of 0.15, 0.2, 0.25 and 0.3

Attack	F-score			
	0.15-Th	0.2-Th	0.25-Th	0.3-Th
Frame swapping	1.00	1.00	1.00	1.00
Frame cropping	1.00	1.00	1.00	1.00
Impulse noise	0.96	0.99	1.00	1.00
Gaussian noise	0.97	0.99	1.00	1.00
Median filter	0.97	0.99	1.00	1.00
Scaling	0.97	0.99	1.00	1.00
Rotation	0.96	0.99	1.00	1.00
Blurring	0.97	0.99	1.00	1.00
Sharpening	0.97	0.99	1.00	1.00
Gamma correction	0.96	0.99	1.00	1.00

Table 4 Performance comparison of different methods for threshold value of 0.15

Attack	TPR (%)							FPR (%)			
	RBSIF-ICD [8]	Colour-based [24]	WBSIF [34]	Histogram-based [18]	CNN + Graph [47]	3D-CNN [48]	Proposed Method	RBSIF-ICD [8]	Colour-based [24]	WBSIF [34]	Histogram-based [18]
Frame swapping	100	100	100	100	98	86	100	0.00	0.00	0.00	0.00
Frame cropping	100	100	100	100	96	79	100	0.00	0.00	0.00	0.00
Impulse noise	79	71	71	79	100	100	83	0.00	0.05	0.00	0.05
Gaussian noise	79	54	77	71	100	87	87	0.00	0.15	0.00	0.05
Median Filter	77	73	72	64	100	100	87	0.00	0.1	0.00	0.36
Scaling	75	71	75	67	98	79	86	0.00	0.05	0.00	0.1
Rotation	75	61	73	64	75	73	84	0.00	0.05	0.00	0.15
Blurring	77	59	67	59	98	84	89	0.00	0.1	0.00	0.21
Sharpening	73	63	77	67	98	89	89	0.00	0.1	0.00	0.15
Gamma correction	71	52	63	63	100	83	85	0.00	0.21	0.00	0.05

Attack	FPR (%)			F-score						
	CNN + Graph [47]	3D-CNN [48]	Proposed Method	RBSIF-ICD [8]	Colour-based [24]	WBSIF [34]	Histogram-based [18]	CNN + Graph [47]	3D-CNN [48]	Proposed Method
Frame swapping	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.99	0.97	1.00
Frame cropping	0.73	0.05	0.00	1.00	1.00	1.00	1.00	0.98	0.94	1.00
Impulse noise	0.00	0.00	0.00	0.95	0.93	0.93	0.95	1.00	1.00	0.96
Gaussian noise	0.00	0.00	0.00	0.95	0.85	0.94	0.93	1.00	0.97	0.97
Median Filter	0.00	0.00	0.00	0.94	0.93	0.93	0.89	1.00	1.00	0.97
Scaling	0.52	0.21	0.00	0.94	0.93	0.94	0.91	0.99	0.94	0.97

Table 4 (continued)

Attack	FPR (%)			F-score						
	CNN + Graph [47]	3D-CNN [48]	Proposed Method	RBSIF-ICD [8]	Colour-based [24]	WBSIF [34]	Histogram-based [18]	CNN + Graph [47]	3D-CNN [48]	Proposed Method
Rotation	0.26	0.26	0.00	0.94	0.88	0.93	0.89	0.93	0.92	0.96
Blurring	0.89	0.00	0.00	0.94	0.87	0.91	0.87	0.98	0.96	0.97
Sharpening	0.52	0.15	0.00	0.93	0.89	0.94	0.91	0.99	0.97	0.97
Gamma correction	0.00	0.00	0.00	0.93	0.84	0.89	0.89	1.00	0.96	0.96

Table 5 Performance comparison of different methods for threshold value of 0.2

Attack	TPR (%)							FPR (%)			
	RBSIF-ICD [8]	Colour-based [24]	WBSIF [34]	Histogram-based [18]	CNN + Graph [47]	3D-CNN [48]	Proposed Method	RBSIF-ICD [8]	Colour-based [24]	WBSIF [34]	Histogram-based [18]
Frame swapping	100	100	100	100	98	87	100	0.00	0.00	0.00	0.00
Frame cropping	100	100	100	100	96	81	100	0.00	0.00	0.00	0.00
Impulse noise	86	77	84	81	100	100	98	0.00	0.1	0.00	0.15
Gaussian noise	87	61	88	79	100	98	99	0.00	0.36	0.00	0.05
Median filter	84	79	82	72	100	100	99	0.00	0.21	0.00	0.47
Scaling	84	79	84	79	98	83	98	0.00	0.15	0.00	0.21
Rotation	83	67	81	73	75	75	98	0.00	0.31	0.00	0.26
Blurring	86	65	77	68	98	89	99	0.00	0.36	0.00	0.36
Sharpening	81	67	87	77	98	96	99	0.00	0.42	0.00	0.31
Gamma correction	79	57	73	72	100	98	98	0.00	0.47	0.00	0.1

Attack	FPR (%)			F-score						
	CNN + Graph [47]	3D-CNN [48]	Proposed Method	RBSIF-ICD [8]	Colour-based [24]	WBSIF [34]	Histogram-based [18]	CNN + Graph [47]	3D-CNN [48]	Proposed Method
Frame swapping	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.99	0.97	1.00
Frame cropping	0.73	0.15	0.00	1.00	1.00	1.00	1.00	0.98	0.95	1.00
Impulse noise	0.00	0.00	0.00	0.97	0.94	0.96	0.95	1.00	1.00	0.99
Gaussian noise	0.00	0.00	0.00	0.97	0.88	0.97	0.94	1.00	0.99	0.99
Median filter	0.00	0.00	0.00	0.96	0.95	0.95	0.92	1.00	1.00	0.99
Scaling	0.52	0.57	0.00	0.96	0.95	0.96	0.94	0.99	0.95	0.99
Rotation	0.26	0.89	0.00	0.96	0.91	0.95	0.92	0.93	0.93	0.99
Blurring	0.89	0.00	0.00	0.97	0.90	0.94	0.91	0.98	0.97	0.99
Sharpening	0.52	0.68	0.00	0.95	0.91	0.97	0.94	0.99	0.98	0.99
Gamma correction	0.00	0.00	0.00	0.95	0.86	0.93	0.93	1.00	0.99	0.99

Table 6 Performance comparison of different methods for threshold value of 0.25

Attack	TPR (%)							FPR (%)			
	RBSIF-ICD [8]	Colour-based [24]	WBSIF [34]	Histogram-based [18]	CNN + Graph [47]	3D-CNN [48]	Proposed Method	RBSIF-ICD [8]	Colour-based [24]	WBSIF [34]	Histogram-based [18]
Frame swapping	100	100	100	100	98	91	100	0.00	0.00	0.00	0.00
Frame cropping	100	100	100	100	96	87	100	0.00	0.00	0.00	0.00
Impulse noise	95	82	89	87	100	100	100	0.00	0.31	0.47	0.57
Gaussian noise	97	68	92	83	100	100	100	0.00	1.00	0.52	0.42
Median Filter	97	85	87	81	100	100	100	0.00	0.57	0.68	0.68
Scaling	97	83	91	85	98	96	100	0.00	0.57	0.42	0.47
Rotation	95	75	87	79	75	79	100	0.00	0.89	0.73	1.15
Blurring	96	72	82	71	98	94	100	0.00	0.94	0.73	1.1
Sharpening	95	79	93	83	98	97	100	0.00	0.89	0.57	0.94
Gamma correction	93	68	78	81	100	100	100	0.00	1.1	0.89	0.89

Attack	FPR (%)			F-score						
	CNN + Graph [47]	3D-CNN [48]	Proposed Method	RBSIF-ICD [8]	Colour-based [24]	WBSIF [34]	Histogram-based [18]	CNN + Graph [47]	3D-CNN [48]	Proposed Method
Frame swapping	0.00	0.42	0.00	1.00	1.00	1.00	1.00	0.99	0.97	1.00
Frame cropping	0.73	0.57	0.00	1.00	1.00	1.00	1.00	0.98	0.96	1.00
Impulse noise	0.00	0.00	0.00	0.99	0.95	0.97	0.96	1.00	1.00	1.00
Gaussian noise	0.00	0.00	0.00	0.99	0.91	0.98	0.96	1.00	1.00	1.00
Median Filter	0.00	0.00	0.00	0.99	0.96	0.96	0.95	1.00	1.00	1.00
Scaling	0.52	0.68	0.00	0.99	0.95	0.97	0.96	0.99	0.98	1.00
Rotation	0.26	0.89	0.00	0.99	0.93	0.96	0.94	0.93	0.94	1.00
Blurring	0.89	0.00	0.00	0.99	0.92	0.95	0.91	0.98	0.98	1.00
Sharpening	0.52	0.73	0.00	0.99	0.94	0.98	0.95	0.99	0.98	1.00
Gamma correction	0.00	0.00	0.00	0.98	0.91	0.94	0.95	1.00	1.00	1.00

Table 7 Performance comparison of different methods for threshold value of 0.3

Attack	TPR (%)							FPR (%)			
	RBSIF-ICD [8]	Colour-based [24]	WBSIF [34]	Histogram-based [18]	CNN + Graph [47]	3D-CNN [48]	Proposed Method	RBSIF-ICD [8]	Colour-based [24]	WBSIF [34]	Histogram-based [18]
Frame swapping	100	100	100	100	98	99	100	0.00	0.00	0.00	0.00
Frame cropping	100	100	100	100	96	96	100	0.00	0.00	0.00	0.00
Impulse noise	99	94	98	91	100	100	100	0.05	0.57	0.52	0.63
Gaussian noise	99	71	98	89	100	100	100	0.1	1.42	0.57	0.57
Median filter	100	96	97	87	100	100	100	0.05	0.68	0.73	0.73
Scaling	100	96	98	89	98	99	100	0.31	0.73	0.47	0.68
Rotation	99	85	98	83	75	93	100	0.68	1.1	0.89	1.21
Blurring	99	83	95	79	98	98	100	0.57	1.21	0.94	1.42
Sharpening	98	89	96	88	98	98	100	0.78	1.15	0.68	1.1
Gamma correction	98	73	93	88	100	100	100	0.89	1.47	0.94	0.94

Attack	FPR (%)			F-score						
	CNN + Graph [47]	3D-CNN [48]	Proposed Method	RBSIF-ICD [8]	Colour-based [24]	WBSIF [34]	Histogram-based [18]	CNN + Graph [47]	3D-CNN [48]	Proposed Method
Frame swapping	0.00	0.57	0.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00
Frame cropping	0.73	0.59	0.00	1.00	1.00	1.00	1.00	0.98	0.98	1.00
Impulse noise	0.00	0.00	0.00	0.99	0.98	0.99	0.97	1.00	1.00	1.00
Gaussian noise	0.00	0.00	0.00	0.99	0.92	0.99	0.97	1.00	1.00	1.00
Median filter	0.00	0.00	0.00	0.99	0.98	0.98	0.96	1.00	1.00	1.00
Scaling	0.52	0.68	0.00	0.99	0.98	0.99	0.97	0.99	0.99	1.00
Rotation	0.26	0.94	0.00	0.99	0.96	0.98	0.95	0.93	0.97	1.00
Blurring	0.89	0.00	0.00	0.99	0.95	0.98	0.93	0.98	0.99	1.00

Table 7 (continued)

Attack	FPR (%)			F-score						
	CNN + Graph [47]	3D-CNN [48]	Proposed Method	RBSIF-ICD [8]	Colour-based [24]	WBSIF [34]	Histogram-based [18]	CNN + Graph [47]	3D-CNN [48]	Proposed Method
Sharpening	0.52	0.78	0.00	0.99	0.97	0.98	0.96	0.99	0.99	1.00
Gamma correction	0.00	0.00	0.00	0.99	0.92	0.97	0.96	1.00	1.00	1.00

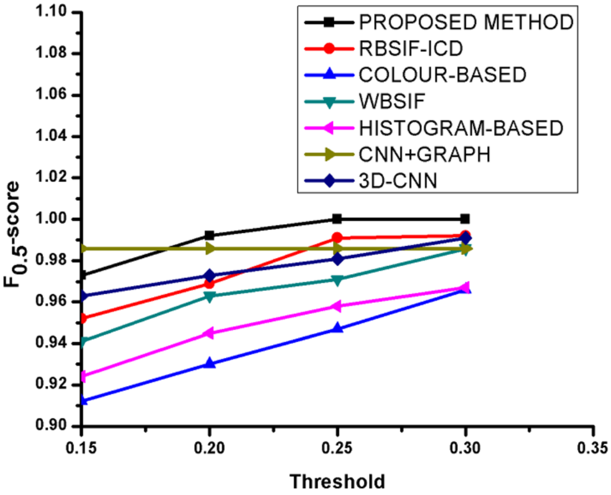


Fig. 11 Average F-score of the proposed method and state-of-the-art methods

completely discriminant and robust (100% recall and 100% precision) and vice-versa. Thus, F-score can be a single valuable measure to demonstrate the detection performance of a method with proper choice of α. To minimize the number of human interactions required, a good video copy detection method should have high precision. So, the precision is given twice the importance of recall by choosing α = 0.5 in our proposed work.

To achieve a good detection result, every video in the database is down-sampled to a fixed frame-rate of 4 frames per second and frame size of 320 × 240 before extracting the fingerprints for evaluation. The detailed proposed work is studied in section 3. To examine if a test or query video is a copied version of an original video, the similarity between the fingerprint of the test video and the original videos in the fingerprint database are calculated based on Canberra distance metric (Eq. 4). If the distance is lower than certain predefined threshold value (0.25), then the corresponding video will be declared as pirated version. Table 3 shows the F-score of the proposed method against distinct attacks for different threshold values to evaluate the performance. Finally, the performance is evaluated using 20 original videos collected from TRECVID 2010 dataset [1] in which ten different attacks, each having five distinct parameter values are mounted independently on each video in the database. It can be clearly seen from the Table 3 that an average F-score value is 1 in both the threshold values of 0.25 and 0.3 respectively. That is, the proposed method is completely robust and discriminant within the specified threshold value of 0.25 in our work. However, the F-score in 0.15 and 0.2

Table 8 Average execution time of different methods (in seconds)

Methods	Execution time (s)
RBSIF-ICD [20]	10.3
Colour-based [48]	7.6
WBSIF [2]	8.4
Histogram-based [26]	8.7
CNN + Graph [59]	18.2
3D-CNN [29]	16.9
Proposed Method	6.1

threshold values shows a slight decrease in the performance. Overall, the proposed method shows good performance against distinct attacks for all the threshold values and the best outcome can be achieved under 0.25 threshold value.

4.2.2 Comparative analysis

The proposed method is compared to state-of-the-art methods used for video copy detection such as RBSIF-ICD [20], Colour-based [48], WBSIF [2], Histogram-based [26], CNN + Graph [59] and 3D-CNN [29]. In [20], each key-frame of a video was partitioned into several rings and the generated histograms from different rings were concatenated to create a new matrix. BSIF transformation was then applied to this new matrix to extract the feature vector. Set of filters which were learnt from a set of 13 natural images were used by BSIF technique. Next, the ICD was applied to generate global vector of 125 bins by employing five color bins in all of the three color channels. This method is less discriminative as only the spatial global features were extracted for copy detection and ignored the importance of spatial local features. The chance of redundant key-frames generation is very high. Moreover, this method ignored the importance of temporal information too. In [48], initially the video sequence was divided into shots and its table of contents (TOC) was constructed. Temporally informative representative image was then generated by selecting one shot from the summarized form and its output was initially transformed into the R, G and B color channels which was then segmented into c × c blocks. Then, the color correlation was extracted and the color correlation histogram was plotted based on the feature vectors. This method used only the global color features which is less discriminative and less robust against the geometric distortions. There exists formation of redundant key-frames. In [2], first the TIRI transform was applied to the video frames to create key-frames in order to preserve spatio-temporal information, and then the WBSIF technique was applied on the selected key-frames to generate 2D histogram. WBSIF technique was calculated based on differential excitation and BSIF. The differential excitation at any pixel considered as a central pixel of key-frames was obtained by the sum of intensity difference between this central pixel and its neighbouring pixels. Then, BSIF was applied that used set of filters learnt from the set of 13 natural images to compute the convolution of the image patch. Finally, the 2D histogram was generated by regrouping the feature vector descriptor present on the excitations according to the local BSIF patterns. This method used only the global features which is less discriminative and the formation of redundant key-frames is very high. In [26], first the DC image sequence was generated, and then the HOG and OM were applied to generate the orientation histogram of 9 bins and ordinal histogram of total 720 combinations of separate pattern code respectively. To generate the ordinal histogram, first the frame image was partitioned into 6 equal regions and then means of every sub-region were computed. This method used both the local and global features to preserve both the discriminative as well as robustness properties. However, this method incurs high dimensional cost as the features were extracted from each video frame and the extracted OM feature values from each rectangular block of video frame will vary when the geometric distortion specially rotation attack is applied. In [59], AlexNet was used to train the network in which the deep CNN features were used for encoding visual content. Subsequently, the graph-based sequence matching technique was employed to preserve the temporal consistency in which only the top 5 matching results of generated frame level matching matrix were utilized for final outcomes. In [29], parallel 3D-CNNs were employed for classifying multi-class in which each 3D-CNN was utilized as a two-class classifier for a particular class of video. During the training of 3D-CNN network, each video was downsampled into multiple sub-videos having the same number of frames (i.e., 7 frames here) in which videos can be classified based on their sub-videos. This method incur high computational cost as the parallel structure can increase with the formation of new class.

The performance comparison with different approaches is carried out for four threshold values (i.e. 0.15, 0.2, 0.25 and 0.3) using 20 original videos in the database collected from TRECVID 2010 dataset [1]. As earlier mentioned, that ten different attacks with each having five distinct attack parameter values are applied to each original video in the database for evaluation. Table 4, 5, 6 and 7 shows the performance comparison between the proposed method and state-of-the-art methods against different attacks or distortions. It reports the F-score, average true positive rate (TPR) and false positive rate (FPR) of each attack on the original videos for four threshold values respectively. Figure 11 shows the performance comparison between the proposed method and state-of-the-art methods based on average F-score for different threshold values. Figure 11 and Table 4-7 clearly shows that the proposed method outperforms the existing stat-of-the-art methods with repect to the TPR, FPR and F-score results against different attacks. Tables 6 and 7 shows that the proposed method maintains high performance against all the attacks for 0.25 and 0.3 threshold values with an average F-score value of 1, and Tables 4 and 5 shows a slight degradation of performance of the proposed method for 0.15 and 0.2 threshold values with an average F-score of 0.97 and 0.99 repectively. The proposed method shows complete robustness and discriminative capability against distinct attacks within the threshold values of 0.25 and 0.3 with an average F-score of 1. It can be clearly studied from the Tables that the proposed method and most of state-of-the-art methods are performing well against the frame-swapping and frame-cropping attacks with 100% TPR, FPR and F-score results for all the threshold values except the CNN + Graph [59] and 3D-CNN [29] methods. That means most of the methods shows complete robustness and discriminative capability against these two attacks for all the threshold values. For different threshold values, FPR of the RBSIF-ICD [20], WBSIF [2] and the proposed methods are much lower which shows high discriminative capability compared to the Colour-based [48] and Histogram-based [26] methods.

Generally, the methods with lower FPR and higher TPR are considered to provide better performance. But, discriminative capability of RBSIF-ICD [20] and WBSIF [2] methods are degraded with increase in the FPR for 0.25 and 0.3 threshold values in Tables 6 and 7. The proposed method shows higher TPR for all the threshold values compared to state-of-the-art methods which results in higher robustness against all the attacks. Colour-based [48] and Histogram-based [26] methods shows poor TPR, FPR and F-score values against all the attacks for different threshold values which results in poor performance. In Table 7, RBSIF-ICD [20] method shows 100% robustness against the median filter and scaling attacks with 100% TPR for 0.3 threshold value but less discriminative capability with high FPR as compared to the proposed method. Comparatively, CNN + Graph [59] and 3D-CNN [29] methods shows high robustness against impulse noise, Gaussian noise, median filter and Gamma correction attacks but less robustness against cropping and rotation attacks specially. The results of F-score demonstrates that the proposed method is able to resist the geometrical attacks or transforms studied here and provides better video copy detection performance. Finally, it can be well observed from the Figure 11 and Tables that the proposed method is more robust and discriminative against all the attacks under consideration compared to the state-of-the-art methods.

4.3 Time complexity analysis

The complexity or cost in execution time of state-of-the-art methods and the proposed method are analyzed and compared in this section. Table 8 shows an average execution time of proposed method and state-of-the-art methods in seconds.

From the Table 8, it can be clearly visualized that the proposed method outperforms the state-of-the-art methods in terms of an average execution time with only 6.1 s. Deep features based methods such as CNN + Graph [59] and 3D-CNN [29] are taking much time compared to handcrafted features based methods with 18.2 s and 16.9 s respectively. Thus, the proposed method performs better with less computational complexity.

5 Discussion

An important property of a video copy detection method lies in its ability to detect whether a query video is a pirated version of an original video or not within a huge database efficiently and reliably. To make the copy detection system faster and more efficient, the pre-classifier is used in our proposed work. The function of pre-classifier is to select only the perceptually similar videos from reference database with respect to the query video and store it into a newly generated intermediate candidate database in an initial stage. Further processing will be done only on the videos of intermediate candidate database that are pre-classified instead of processing on the whole videos of reference database for piracy detection. To make the proposed method more robust against rotation attack, initially, the video frames are partitioned into equal rings and then ordinal histogram (OH) is generated from the rings. Traditional block-based partition of the video frame is less robust against geometrical attacks such as rotation and resizing. A small angle orientation can change the block size which will give different extracted features. Our ring-based partition technique overcomes the limitations of traditional block-based partition technique and provides optimum solution to the copy detection issues.

The generation of keyframes based on the TIRI transform from each scene of a video sequence will also make the proposed method fast and efficient. DCT transform is applied to generate the scenes in which only the similar video frames are grouped. The extraction of features from the video keyframes rather than extracting from whole frames of a video sequence will also reduce the dimension and will enhance the performance. Moreover, the use of TIRI transform for keyframe generation preserves both the spatial and temporal information of a video sequence, since it is an important property for video copy detection system. Further, the features are extracted using HOG and SVD techniques. The HOG technique is well adapted for local object detection like SIFT and the singular values extracted using SVD technique have good stability and do not change when there is a small manipulation in the video frame image. The experimental result clearly shows that the proposed video copy detection method is more efficient compared to state-of-the-art copy detection methods with an average F-score of 1 within the specified threshold value (i.e., 0.25). Moreover, the proposed method is more robust against the geometric attacks under consideration and have high discriminative capability compared to existing state-of-the-art methods.

6 Conclusion

A fast and novel video copy detection method has been proposed to detect illegal copies of original videos in a huge database. The main novelty of proposed method lies on using the pre-classifier which classifies only the similar videos in reference database with respect to the query video and store them into a newly generated intermediate candidate database for further processing. This technique will boost the efficiency as well as reduce the computation overhead. Initially, videos in the database are pre-processed to make the proposed method well adapted to resist certain manipulations during performance evaluation. Moreover, the video frames are partitioned based on ring-decomposition technique and features are extracted from each ring based on ordinal measure approach in an initial stage to overcome the limitations of traditional block-based partition technique. This technique makes the proposed method more robust against geometric attacks specially rotation attack. Based on this global OM feature vector only the similar videos are pre-classified as earlier mentioned above. Further, to make the copy detection method faster, the local HOG and SVD feature vectors are generated from the keyframes instead of whole frames of a video sequence. Keyframes are generated from each scene of a video sequence using TIRI transform which preserves both the spatial and temporal information of a video. A robust descriptor is created by fusing the generated feature vectors which can resist the geometrical attacks. The experimental and analytical studies carried out on TRECVID 2010 dataset signifies that the proposed method is fast and more efficient compared to state-of-the-art methods. The experimental results also demonstrate that the proposed method is more robust against geometric attacks under consideration as well as have high discriminative capability compared to state-of-the-art methods. In future, both the audio and video features can be combined to enhance the performance as audio is also an integral part of a video.

References

1. Awad G, Over P, Kraaij W (2014) Content-based video copy detection benchmarking at TRECVID. ACM Trans Inf Syst 99(9):1–36
2. Boukhari A, Serir A (2016) Weber binarized statistical image features (WBSIF) based video copy detection. J Vis Commun Image Represent 34:50–64
3. Britz D, Goldie A, Luong T, Le Q (2017) Massive exploration of neural machine translation architectures. arXiv preprint arXiv:1703.03906
4. 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
5. Chen L, Stentiford FW (2008) Video sequence matching based on temporal ordinal measurement. Pattern Recogn Lett 29(13):1824–1831
6. Chen Q, Tian J, Yang L, Wu D (2011) A robust video hash scheme based on 2D-DCT temporal maximum occurrence. Secur Commun Netw 4(12):1369–1377
7. Chiu CY, Tsai TH, Hsieh CY (2013) Efficient video segment matching for detecting temporal-based video copies. Neurocomput 105:70–80
8. 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

9. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE computer society conference on computer vision and pattern recognition (CVPR), pp 886-893
10. Ding G, Nie R (2010) Ring fingerprint based on interest points for video copy detection. In: IEEE international symposium on multimedia (ISM), pp 347–352
11. Farajzadeh K, Zarezadeh E, Mansouri J (2017) Concept detection in images using SVD features and multi-granularity partitioning and classification. Inf Syst Telecommun 5(3):172–182
12. Farnebäck G (2003) Two-frame motion estimation based on polynomial expansion. In: Scandinavian Conference on Image analysis, Springer, pp 363–370
13. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645
14. Gao L, Guo Z, Zhang H, Xu X, Shen HT (2017) Video captioning with attention-based LSTM and semantic consistency. IEEE Trans Multimed 19(9):2045–2055
15. Gao L, Li X, Song J, Shen HT (2019) Hierarchical LSTMs with adaptive attention for visual captioning. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2019.2894139
16. 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: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1508–1512
17. 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
18. 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
19. Himeur Y, Sadi KA (2015) Joint color and texture descriptor using ring decomposition for robust video copy detection in large databases. In: IEEE international symposium on signal processing and information technology (ISSPIT), pp 495–500
20. 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
21. Hua XS, Chen X, Zhang HJ (2004) Robust video signature based on ordinal measure. In: IEEE international conference on image processing (ICIP), pp 685–688
22. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li F (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 1725–1732
23. Karsh RK, Laskar RH, Richhariya BB (2016) Robust image hashing using ring partition-PGNMF and local features. SpringerPlus 5(1):1–20
24. Kitanovski V, Taskovski D (2010) Real-time TV commercial monitoring based on robust visual hashing. In: IEEE 2nd European workshop on visual information processing (EUVIP), pp 140–143
25. 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, pp 371–378
26. Lee F, Zhao J, Kotani K, Chen Q (2017) Video copy detection using histogram based spatio-temporal features. In: IEEE10th international congress on image and signal processing, BioMedical engineering and informatics (CISP-BMEI), pp 1–5
27. Li J, Guo X, Yu Y, Tu Q, Men A (2014) A robust and low complexity video fingerprint for multimedia security. In: IEEE international symposium on wireless personal multimedia communications (ISWPMC), pp 97–102
28. Li J, Monroe W, Shi T, Ritter A, Jurafsky D (2017) Adversarial learning for neural dialogue generation. arXiv preprint arXiv:1701.06547
29. Li J, Zhang H, Wan W, Sun J (2018) Two-class 3D-CNN classifiers combination for video copy detection. Multimed Tools Appl 1:1–3
30. Liong VE, Lu J, Tan YP, Zhou J (2017) Deep video hashing. IEEE Trans Multimed 19(6):1209–1219
31. Liu J, Huang Z, Cai H, Shen HT, Ngo CW, Wang W (2013) Near-duplicate video retrieval: current research and future trends. ACM Comput Surveys (CSUR) 45(4):1–23
32. Liu L, Lai W, Hua XS, Yang SQ (2007) On real-time detecting duplicate web videos. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 973-976
33. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
34. Malekesmaeili M, Fatourechi M, Ward RK (2009) Video copy detection using temporally informative representative images. In: IEEE international conference on machine learning and applications (ICMLA), pp 69–74
35. Mao J, Xiao G, ShengW HY, Qu Z (2016) A method for video authenticity based on the fingerprint of scene frame. Neurocomput 173:2022–2032

36. Neelima A, Singh KM (2017) Collusion and rotation resilient video hashing based on scale invariant feature transform. Imaging Sci J 65(1):62–74
37. 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
38. Özbulak G, Kahraman F, Baykut S (2016) Robust video copy detection in large-scale TVstreams using local features and CFAR based threshold. In: IEEE international conference on digital signal processing (ICDSP), pp 124–128
39. Sadek RA (2012) SVD based image processing applications: state of the art, contributions and research challenges. arXiv preprint arXiv:1211.7102
40. Setyawan I, Timotius IK (2014) Spatio-temporal digital video hashing using edge orientation histogram and discrete cosine transform. In: IEEE international conference on information technology systems and innovation (ICITSI), pp 111–115
41. Setyawan I, Timotius IK (2014) Digital image hashing using local histogram of oriented gradients. In: IEEE 6th international conference on information technology and electrical engineering (ICITEE), pp 1-4
42. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, pp 568–576
43. Song J, Guo Y, Gao L, Li X, Hanjalic A, Shen HT (2018) From deterministic to generative: multi-modal stochastic RNNs for video captioning. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2018.2851077
44. Song J, Zhang H, Li X, Gao L, Wang M, Hong R (2018) Self-supervised video hashing with hierarchical binary auto-encoder. IEEE Trans Image Process 27(7):3210–3221
45. Subramanyam AV, Emmanuel S (2012) Video forgery detection using HOG features and compression properties. In: IEEE 14th international workshop on multimedia signal processing (MMSP), pp 89–94
46. Sun R, Yan X, Gao J (2017) Robust video fingerprinting scheme based on contourlet hidden Markov tree model. Optik 128:139–147
47. Tasdemir K, Cetin AE (2010) Motion vector based features for content based video copy detection. In: IEEE 20th international conference on pattern recognition (ICPR), pp 3134–3137
48. Thomas RM, Sumesh MS (2015) A simple and robust colour based video copy detection on summarized videos. Procedia Comput Sci 46:1668–1675
49. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of IEEE International Conference on Computer Vision, pp 4489–4497
50. van Rijsbergen CJ (1979) Information retrieval. Butterworths, London
51. Wang RB, Chen H, Yao JL, Guo YT (2016) Video copy detection based on temporal contextual hashing. In: IEEE second international conference on multimedia big data (BigMM), pp 223–228
52. Wang X, Gao L, Wang P, Sun X, Liu X (2017) Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length. IEEE Trans Multimed 20(3):634–644
53. Wang H, Oneata D, Verbeek J, Schmid C (2016) A robust and efficient video representation for action recognition. Int J Comput Vis 119(3):219–238
54. Wu PH, Thaipanich T, Kuo CC (2009) A suffix array approach to video copy detection in video sharing social networks. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 3465–3468
55. Yang JF, Lu CL (1995) Combined techniques of singular value decomposition and vector quantization for image coding. IEEE Trans Image Process 4(8):1141–1146
56. 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):84–97
57. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 4694–4702
58. Zhang Z, Cao C, Zhang R, Zou J (2010) Video copy detection based on speeded up robust features and locality sensitive hashing. In: IEEE international conference on automation and logistics (ICAL), pp 13–18
59. Zhang X, Xie Y, Luan X, He J, Zhang L, Wu L (2018) Video copy detection based on deep CNN features and graph-based sequence matching. Wirel Pers Commun 103(1):401–416

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Alongbar Wary is currently pursuing PhD in the Department of Computer Science and Engineering in National Institute of Technology (NIT) Nagaland, India (July 2017 to present). He has received the M.Tech degree in Computer Science and Engineering from NIT Nagaland, India in the year 2017, B.E degree in Computer Science and Engineering from Sree Sastha Institute of Engineering and Technology, Chennai, India in the year 2013. His current area of interest includes: Image and Video Processing, Multimedia Applications, Mutilmedia Security and Multimedia Retrieval.

Arambam Neelima is currently working as an Assistant Professor in Department of Computer Science and Engineering in National institute of Technology Nagaland (September 2013 to present). She has obtained her B.Tech degree In Information Technology from North- Eastern Hill University, Meghalaya, India in the year 2010, M.Tech degree in Information Technology from Tezpur University, Assam, India in the year 2012 and Ph.D. Degree from the Department of Computer Science of National Institute of Technology Manipur, Manipur in the year 2016. Her current area of interest includes: Information Security, Content Based Image Retrieval, Content Based Video Retrieval, Image Hashing, Image and Video Copy Detection, Multimedia Secret Sharing.