SIViP (2013) 7:665–677
DOI 10.1007/s11760-013-0482-x

ORIGINAL PAPER

An attack invariant scheme for content-based video copy detection

Debabrata Dutta · **Sanjoy Kumar Saha** ·
Bhabatosh Chanda

Received: 31 May 2012 / Revised: 15 September 2012 / Accepted: 1 December 2012 / Published online: 3 May 2013

Abstract Due to the advancement in the field of multimedia technology and communication, it has become easier to access, store, and edit video data. Easy manipulation of video data and its rapid distribution have made content-based video copy detection (CBVCD) an active area of research. In a CBVCD system, reference video sequence and query sequence are compared to detect whether the query video sequence is a copy of reference video sequence. Thus, the generation of fingerprint of a video sequence and sequence matching technique are the core tasks of such system. In order to evade such detection process, a copied version may undergo different kinds of transformations like photometric and post-production attack. So the detection system must be robust enough against such attacks. In this work, fingerprint is generated from the sub-bands of wavelet decomposed intensity image and localized intensity gradient histograms of low sub-band. The fingerprint thus obtained reflects considerable discriminating capability and robustness against the attacks. Furthermore, to cope up with the attacks, we have adopted simple pre-processing technique, which enhances the robustness of the system further. A robust sequence matching technique based on multivariate Wald–Wolfowitz test is proposed here. An experiment has been carried out with a database consisting of distinct 642 shots and 1,485 query sequences representing different attacks. Proposed methodology achieves high copy detection rate (99.39 %) and very low false alarm rate (0.14 %) and performs better than other spatio-temporal measure based systems.

WTF

↳ que outros?

Keywords Video copy detection · Video fingerprinting · Sequence matching · Hypothesis test · Photometric attack · Post-production attack

1 Introduction

With the rapid development in the field of multimedia technology, it has become easier to access and store video data of huge volume. It is well reflected in the availability of such data on various sites like video blogs and Web TV. Sharing and distribution of video over all such sites have led to exponential growth of data volume. Furthermore, the technology has enabled editing and duplication of video data that may lead to violation of digital rights. Thus, copyright protection becomes a crucial issue, and the enormous volume of video data makes the task further difficult. This has led to the emergence of video copy detection as an active area of research.

In case of content-based video retrieval (CBVR) system, the goal is to retrieve similar videos in the same category, whereas in a content-based video copy detection (CBVCD) system, it is to be detected whether a query sequence is a copied version of reference sequence or not. Instead of being an identical or near-replicated video sequence, a copy may be a transformed video sequence [26]. As a result, a copy may be visually less similar. In the context of copy detection, a CBVR system may give rise to high false alarm rates [20] which is undesirable. Unlike a CBVR system, response time is not too critical for a CBVCD system. A copy detection system acts on a complaint focusing on a reduced search space, and more often, it may be an offline process.

D. Dutta
Tirthapati Institution, Kolkata, India
e-mail: deababratadutta2u@gmail.com

S. K. Saha (✉)
CSE Department, Jadavpur University, Kolkata, India
e-mail: sks_ju@yahoo.co.in

B. Chanda
ECS Unit, Indian Statistical Institute, Kolkata, India
e-mail: chanda@isical.ac.in

There are two basic approaches to address the issue of copyright protection—watermarking and content-based copy detection. In the first approach, watermark/non-visible information is embedded into the content, and later, if required, this embedded information is used for establishing the ownership. But watermarking is not applicable for the video sequences already in circulation without any such embedded information. On the other hand, in content-based approach, no additional information is inserted. It is said that ”Video itself is the watermark” [14]. So in this approach, unique signatures (features) are derived from the content of the video itself. Such signatures are also extracted from the questioned video and are compared with those of the original media stored in the database [4,14,21,22,31]. Apart from protecting the right, copy detection may also help in media tracking [13] that is, how many times a particular media is being used.

Performance of a video copy detection scheme relies on the suitable signature of the frame sequence and also on the sequence matching scheme. The system must be robust to the presence of various distortions adopted by the copier. In this work, we have focused on handling of photometric and post-production attacks and the sequence matching technique. The paper is organized as follows. After this introduction, Sect. 2 presents a brief review of the video copy detection techniques. Section 3 describes the proposed methodology. Experimental results are presented in Sect. 4, and finally, the concluding remarks are presented in Sect. 5.

2 Past work

A content-based video copy detection system consists of two major modules namely: *fingerprint generation* and *sequence matching technique*. Fingerprint can be defined as perceptual features for short summaries of a multimedia object [39]. The goal of video fingerprinting is to judge whether two video have the same contents even under quality-preserving distortions like resizing, frame rate change, and lossy compression [27]. Sequence matching technique detects whether a query sequence is copied version of referenced one or not based on their fingerprints.
Fingerprint Fingerprint of a video sequence satisfy the properties outlined in [27]. It must be *robust* so that the fingerprint of a copied video (degraded to whatever extent possible) and the original one should be similar. On the other hand, the perceptually different video sequences should have different fingerprints. Thus, the selected fingerprint should meet two diverging requirements.

Fingerprint or descriptor of a video sequence can be broadly categorized as *global* or *local* one. Global ones are derived from the whole video sequence or from a subset of sequence, and local descriptors are computed for each individual frame in the video. Local descriptors are transformed into concise form to generate the global fingerprint.

A wide variety of frame level features have been tried by the researchers to generate the fingerprints. Color histogram [8,10] is very widely used one. But, it lacks in terms of discriminability as the spatial distribution of color is not retained in the histogram. Binary signature based on color histogram has been deployed in [28]. Luminance-based descriptors [16,33,34], dominant color [14], gradient-based features [27], and fractal dimension-based texture feature [6] are also tried. Jeong et al. [18] have relied on the singular value decomposition for the signature generation. Joly et al. [19] have considered local descriptors based on Harris corner detector. Wu et al. [45] have suggested trajectory based visual patterns. Scale-invariant feature transform (SIFT) has been tried by Liu et al. [29]. They have also considered locality-sensitive hashing (LSH) and RANdom SAmple Consensus (RANSAC) to maintain the scalability and enhance the robustness. DCT-based hash algorithm has been used by Coskun et al. [9]. Maani et al. [30] have developed local descriptors for identified regions of interest based on angular intensity variation and region geometry. Sarkar et al. [36] have considered Fourier–Mellin transform and scale-invariant feature transform based descriptors to represent the keyframes. Su et al. [41] have relied on visual attention region based fingerprinting. Such regions are defined in a manner so that they can be taken as invariant in a content-preserving distorted video. The success of that scheme relies heavily on the identification of such regions. A graph-based technique has been presented in [46] to define a spatial correlation descriptor where a node in the graph denotes a region in the frame and edges represent the content proximity of the regions in the frame.

In the context of global fingerprint, a lot of work [3,7,23,15,31] have dealt with ordinal measures. In [23], a video frame is partitioned in to 2×2 sub-images and they are ranked based on their average intensity. The 2×2 rank matrix is taken as the fingerprint of the frame. Based on the rank matrix, spatial measure and temporal measures are formulated for matching a query and reference video sequence. In the work of Chen and Stentiford [7], the sub-images are formed in the same way as in case of [23]. But, instead of individual frame, the rank matrix is formed for the whole video. Corresponding sub-images of all the frames in the sequence are ranked, and $n \times m$ matrix is obtained where $n (= 4)$ and m are the number of sub-images in each frame and the number of frames in the sequence, respectively. Finally, a spatio-temporal measure is proposed for matching query and reference video sequence. Similar spatio-temporal measure is proposed in [15], where the rank matrix over the video sequence is formed based on the number of SURF interest points in the sub-images. Two similarity metrics like sequence shape similarity and real sequence similarity have been proposed in [16]. Barrios

Técnicas de matching de videos!

et al. [2] have worked with eight global descriptors based on edge histogram, gray and RGB histogram, reduced image, and ordinal measurement.

Sequence matching The query video sequence and reference sequences in the database are to be matched on the basis of extracted signature. Variety of matching schemes has been tried by the researchers. The schemes can broadly be classified as (i) *dense matching technique* and (ii) *sparse matching technique*. Dense scheme considers all the frames for comparison. But, a sparse technique deals with representative frames (key frames) only. Hence, a sparse technique is faster, but a dense one is more robust.

In dense matching technique, query video sequence (s_q) is matched with the sub-sequence (s_r) of same length taken from the reference video sequence. Different s_r are obtained by shifting the start position of the sub-sequence in reference video. If the distance between s_q and the most similar s_r is less than a pre-defined threshold then s_q is taken as a copy. Selection of threshold is very crucial. Moreover, the query video sequence can not exceed the length of the reference video sequence. Global descriptors like temporal [23] and spatio-temporal measures [7,15] incorporate frame level comparison and reflect the dissimilarity between s_q and s_r. But, in case of local descriptor based system, the distance between s_q and s_r is to be computed by combining the distance between the feature vectors of corresponding frames in s_q and s_r. Depending on the features, measures like Euclidean distance and histogram intersection are widely used. In [46], χ^2 statistics is used to measure the dissimilarity between frame level descriptors of two sequences. For sequence matching, edit distance [1,46] has been tried. Kim and Park [25] have adopted a longest common sub-sequence (LCS) matching technique for measuring the temporal similarity between video clips. Similarity between two sequences also has been measured by calculating the number of frames matched between two shots [6]. Shen et al. [40] proposed to compute similarity based on the volume intersection between two hyper-sphere governed by the video clips. Barrios et al. [2] have considered different distance measures for various descriptors, and finally, a weighted combination of all such distances is taken.

In order to avoid comparison of query video sequence with each of the reference video sequence in the database, indexing scheme has also been proposed in [2,33,37,47]. Hash-based schemes [9] also has been presented. But selecting a suitable hash function is difficult. Moreover, a hash function is very sensitive to the changes in the content and making the system robust against distortions is of a great challenge.

Several keyframe-based schemes (sparse technique) have been reported. In such technique, selection of keyframe is an important task. Jain et al. [17] have proposed sequence matching technique based on a set of keyframes (or sub-sampled frames). Similar approaches have also been reported in [5,8,42]. Modified Hausdorff distance is applied in [24]. Various clustering based schemes [6,12] have also been tried to extract one or more keyframes from each cluster, and finally, the comparison is restricted to keyframes only. Maani et al. [30], in their technique, corresponding to each keyframe in the query sequence, have selected a set of matched keyframes from the database. From the matched set of keyframes, it tried to find out continuous sub-sequence. If the length of such sub-sequence exceeds a threshold then query sequence is considered as a copy. Selection of the threshold here is an important issue.

In order to cope up with the attacks/deformations incorporated in the copied version, researchers have mostly focused on the robustness in designing the signatures of the video sequence and also on the tolerance allowed by the matching strategies. But, unregulated relaxation may lead to false copy detection which is very sensitive. Ordinal measure [7,15,23], radon transform [38], FFT-based features [35], significant point based feature [44], and visual attention region [41] have been used to address various types of attacks. It is quite difficult to handle the post-production attacks. A few work [2,29,46] has taken additional measures in the form of pre-processing to deal with certain post-production attacks. Although sufficient efforts have been put in developing content-based video copy detection system, still an attack invariant robust system is in demand.

3 Proposed methodology

In a video copy detection method, the task is to verify whether or not a test/query sequence is a copied version of a reference sequence present in the database. It has already been discussed that such a system consists of two major modules: *Extraction of fingerprint* (*feature vector*) and *sequence matching*. Fingerprint must fulfill the diverging criteria such as discriminating capability and robustness against various signal distortion. Sequence matching module bears the responsibility of devising the match strategy and verifying the test sequence with likely originals in the database. In this work, we have also considered pre-processing task prior to the computation of fingerprint to combat the attacks. The schematic diagram of the proposed CBVCD system has been shown in Fig. 1. A reference video sequence is first segmented into shots following the technique presented in [32]. Frames in the shot undergo two stages of pre-processing to cope up with geometric and photometric attack. Fingerprint corresponding to the pre-processed frames is computed and stored into the database. Fingerprints are generated for query video sequence following the same steps. Finally, shots of query and reference video sequences are matched to detect whether the query shots are copied version of reference shots or not. Attacks,

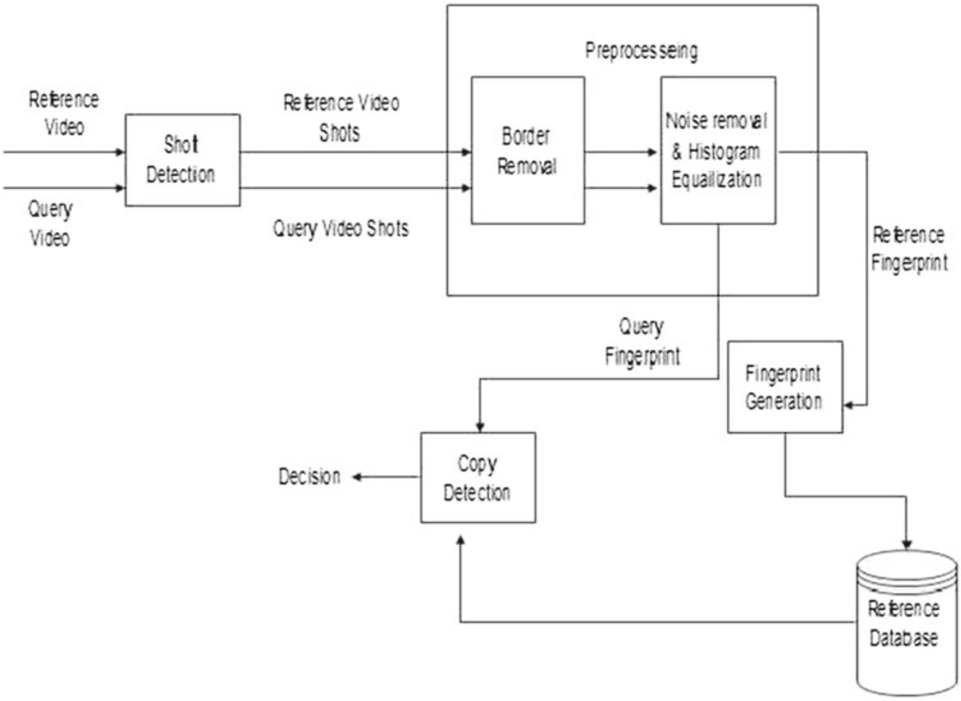


Fig. 1 Schematic diagram of the proposed content-based video copy detection system

pre-processing task, fingerprint generation, and matching technique are elaborated in Sects. 3.1, 3.2, 3.3, and 3.4, respectively.

3.1 Attacks

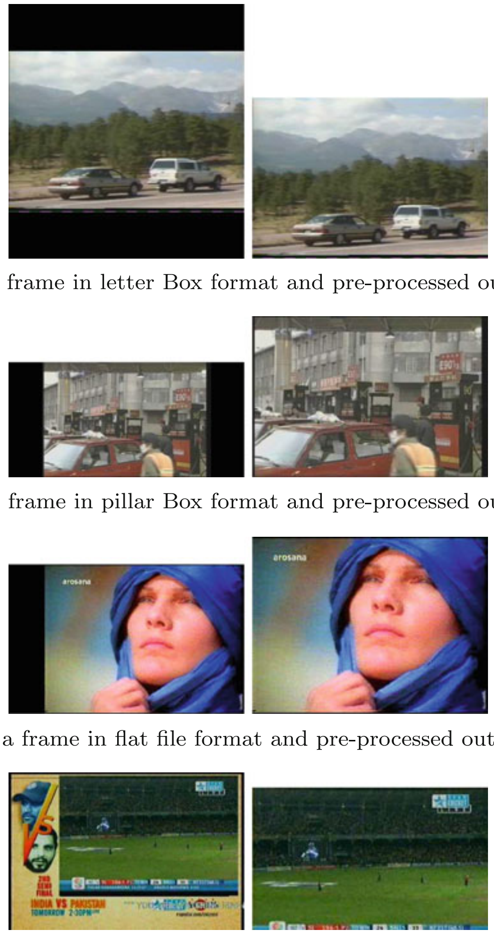
The copied video may undergo various attacks/transformation and that makes it difficult to detect. Such attacks are broadly categorized as *photometric attack* and *geometric attack*. Geometric attacks are mostly post-production attacks. Photometric attacks affect the visual quality, whereas the geometric attacks deal with the frame display format. Various attacks discussed in [9,26,29] can be summarized as follows.

- Photometric attack
  - Increase/decrease in contrast
  - Increase/decrease in brightness
  - Contamination by noise
  - Blurring
- Geometric attack
  - Change in display format
    - Letter box format
    - Pillar box format
    - Flat file format
  - Insertion of logo
  - Picture in picture (PiP)
  - Insertion of L-shaped border (with advertisement)
  - Insertion of horizontal rolling display at bottom
  - Resizing of frames
- Others
  - Change in frame rate
  - Frame drop due to lossy transmission channel
  - Frame clipping
  - Substitution attack

It may be noted that the L-shaped border consists of vertical border on the left and a horizontal one at the bottom. It is very common that such borders contain commercial advertisement with static images or clippings with motion. Examples of few geometric attacks are illustrated in Fig. 2. Apart from geometric transformation, change in frame rate, clipping, and substitution attack are also post-production transformations. Frame rate of a video may be decreased by dropping frames through temporal filtering, and on the other hand, rate can be increased by adding frames through interpolation. In case of clipping frames from both the start and the end of a sequence are dropped. A part of a sequence may be replaced by another sub-sequence with different content in case of substitution attack. Thus, a copied sequence may not be exactly similar or merely a replicate of the reference sequence. It may differ in terms of many features like quality, display pattern, sequence length. Furthermore, combination of various transformations makes the task even more difficult.

3.2 Pre-processing

In this work, we have not dealt with PiP. Pre-processing activity has been carried out to achieve immunity against other



(a) a frame in letter Box format and pre-processed output

(b) a frame in pillar Box format and pre-processed output

(c) a frame in flat file format and pre-processed output

(d) a frame with L-Shaped border and pre-processed output

Fig. 2 Sample frame with different geometric attack and corresponding output after pre-processing. **a** a frame in letter box format and pre-processed output. **b** a frame in pillar box format and pre-processed output. **c** a frame in flat file format and pre-processed output. **d** a frame with L-shaped border and pre-processed output

geometric attack like change in display format, that is, addition of borders and rolling display. Measures have also been taken to minimize the impact of photometric attack like addition of noise, contrast/brightness variation. The major steps for pre-processing are as follows.

será que sem isso o alg funciona ?

- Convert the frame into gray scale image
- Remove the borders (the effect of geometric attacks)
  - Find the edges
  - Detect the lines discriminating content and border in the neighborhood of frame boundary
  - Extract the frame content enclosed by the detected lines
- Minimize the effect of photometric attack
  - Apply mean filter on extracted content to remove noise
  - Apply histogram equalization on filtered image

First of all, we try to minimize the impact of geometric attacks appearing in the form of display format and different types border insertion. As it is indicated, the pre-processing is carried out with the intensity component of the color frame. Intensity values are computed as follows.

$$Y = 0.30 \times R + 0.59 \times G + 0.11B \quad (1)$$

Applying Sobel operator on the intensity image, a gradient image is formed. Considering average gradient as the threshold, pixels with higher gradient are taken as edge points. We try to detect the horizontal (top, bottom) and vertical (left, right) border-content.

In case of letter box (or pillar box), horizontal (resp. vertical) borders appear on top and bottom (resp. left and right) margins of the frame. In a flat file format, a vertical border appears on the left margin. Unlike the plain borders or display formats, rolling display in the horizontal border or activity in the L-shaped borders may be confused with the actual content. We try to develop a common framework based on the observation that there is a sharp transition from the frame borders to the frame content in case of all display formats and all border patterns.

Ideally, the edge image of the frame depicts the vertical and/or horizontal lines demarcating the border and the content. Actual frame content may also give rise to vertical/horizontal lines. To minimize such false detection, we have restricted the detection process only within a range of 30 % (borders are limited within it) of the image width/height from the image boundaries. The border-content discriminating line may have discontinuity. In general, a line may not span over the image width or height and this is particularly true for the L-shaped boundary. To minimize the miss in detecting a possible line, the presence of at least 70 % edge pixels in the vertical (horizontal) direction is considered as a possible line. There is a possibility of having multiple such lines in case of rolling display and L-shaped borders with activity. The contention in such cases is resolved by retaining the line furthest from the nearest image boundary and eliminating the rest. Once the lines discriminating the content and borders are detected, content is segmented out for subsequent use. Sample output of the pre-processing scheme is shown in Fig. 2, and it is observed that the respective borders are successfully removed.

Image obtained after removing the borders is further processed to reduce the effect of various photometric attacks. Noise may be added inherently due to the limitations of production process. To minimize such noises, mean filter of size 5 × 5 is applied as the first step. Manipulation of contrast and/or brightness in a copied sequence may lead to the considerable differences between the intensity histograms of the copied version and the corresponding original one. These changes affect the computed features. To reduce the effects of such errors, histogram equalization technique is applied on

Boa ideia esse pré-processamento!

the filtered image. Finally, the pre-processed image is used for generating the signature.

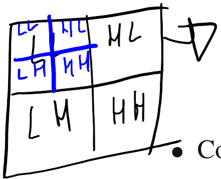
3.3 Fingerprint generation

In order to compute features, we work with the pre-processed, gray scale image where the intensity values are normalized to [0, 1]. We generate the fingerprint/signature for the frames in a sequence based on wavelet decomposed signal and spatial distribution of pixel gradient. The major steps are as follows.

artigo de como implementar encontrado

O Q São?

- Wavelet-based signature
  - F be the image obtained after normalizing the intensity of the pre-processed gray scale frame image
  - Iterate n_w times
    - Apply Haar wavelet transform on F to obtain LL, LH, HL and HH sub-images
    - Compute energy of LH, HL and HH sub-images
    - Consider the LL sub-image as F in next iteration
  - Compute energy of F, that is, energy of the final LL sub-image
  - Consider the $3 \times n_w + 1$-dimensional vector comprising of the energy values as the descriptor



What ?

- Spatial distribution of pixel gradient
  - F be the image obtained after normalizing the intensity of the pre-processed gray scale frame image
  - Apply Haar wavelet transform on F to obtain LL, LH, HL and HH sub-images
  - Compute gradient image corresponding to LL sub-image
  - Divide the gradient image into N_p equisized partitions
  - For each partition, prepare gradient histogram
  - Concatenate the histograms to obtain the descriptor

In order to extract wavelet-based signature, the image is decomposed into four sub-bands (LL, LH, HL and HH) using 2-dimensional Haar wavelet transform. The average intensity or the low frequency component is retained in the LL sub-band, and other three sub-bands show the high-frequency components, that is, the details of the image. Energy in each sub-band is considered as features. Decomposition is continued recursively considering LL sub-band as the image. In successive decomposition, as we deal with the average image (with reduced details) in LL band, the impact of the noise also gets reduced and enables us to cope up with the attacks. In our experiment, we have considered five levels of decompositions to obtain 16-dimensional feature vector.

In general, normalized histogram of gradient is invariant to brightness and so is more invariant to photometric attack in comparison to intensity histogram. It has motivated us to look into gradient-based feature. The frame level gradient histogram is quite global in nature and may not have sufficient discriminating power. Edge-based features can have strong discriminating capability, but spurious edge pixels caused by the noise may reduce its robustness. Thus, a descriptor is required which can cope up with the photometric attack and satisfies the diverging requirement of discriminabilty and robustness. In this context, the histograms of gradient magnitude computed over various partitions in the frame can serve the purpose. So a collection of histograms is used that reflects the spatial distribution of gradient to the extent.

At this stage, we work with the LL band obtained after applying wavelet transform on the intensity image, and it further enhances the noise immunity. Gradient image is computed by applying Sobel operator on the LL band. The gradient image is then divided into N_p equisized partitions, and normalized histogram is formed for each partition. These histograms are concatenated to represent the frame, and it captures the local description well. It may be noted that, the frame size of a copied sequence may be changed to make it different from the original one. Hence, the size of the partition may also vary from sequence to sequence to keep the total number partitions (N_p) same. A high value for N_p makes the partition very small. As a result, the signature is sensitive to object motion, noise, etc. On the other hand, a very low value of N_p affects the discriminating capability. Hence, in our experiment, a moderate value is chosen for N_p. The fingerprint thus obtained is invariant of frame size, and it reflects the distribution of the dominant edge pixels in the image. It may also be noted that, the frames in a sequence with temporal separation have similarity in the histograms unless there is a strong motion, and it enables to handle the change in the frame rate also.

3.4 Sequence matching

A video sequence given as a query may be a collection of shots, and a shot is a continuous sequence of frames generated by a single non-stop camera operation. A copied sequence may be formed by concatenating the shots from different sequences. To cope up with such situation, in our framework, we have considered shot level matching. There are shots in the query sequence; even if they are copied from various reference sequences then also those can be detected.

The reference video sequences are first partitioned into shots following the technique presented in [32]. Signatures are extracted from each frame in the shots and stored in the database. The query sequence is also segmented into shots in the same manner, and the signature is computed from each frame of the shot. Copy detection proceeds at shot level, and the shot matching strategy forms the core part of the system. In the proposed methodology, a hypothesis test based strategy is utilized that relies on multivariate Wald–Wolfowitz test.

Wald–Wolfowitz run test [43] is used to solve the similarity problem of non-parametric distribution of two samples. Suppose, there are two samples X and Y of size m and n, respectively, and the corresponding distributions are F_x and F_y. H_0, the null hypothesis to be tested and H_1, the alternative hypothesis are as follows:

$$H_0 : X \text{and } Y \text{ are from same population, i.e. } F_x = F_y$$
$$H_1 : \text{They are from different population, i.e. } F_x \neq F_y$$

In our problem, a frame is represented by its signature (feature vector). Thus, a shot may be thought of as $\{v_i\}$, the set of feature vectors where v_i is the feature vector corresponding to ith frame in the shot. Let, s_q and s_r are the query shot and a reference shot which are to be compared. Signatures are extracted for each frame in s_q and s_r to obtain the set of feature vectors $\{v_{q_i}\}$ and $\{v_{r_i}\}$, respectively. $\{v_{q_i}\}$ and $\{v_{r_i}\}$ may be thought of as two samples X and Y, respectively, and accordingly, v_{q_i} and v_{r_i} are the sample points. Cardinality of $\{v_{q_i}\}$ and $\{v_{r_i}\}$ is m and n, respectively.

In classical Wald–Wolfowitz test, it is assumed that the sample points are univariate. $N = n + m$ observations are sorted in ascending order, and the labels X or Y are assigned to them depending on the sample to which they belong. But, in our problem, the sample points are multivariate. Friedman and Rafsky [11] have suggested a multivariate generalization by using the minimal spanning tree (MST) of the sample points. In this approach, each sample point is considered as a node and every node is connected to the closest node (based on the similarity between their feature vectors) to form the MST of the sample points. The steps for forming the MST are as follows.

- remaining_set = set of all nodes (feature vectors) of two samples
- Consider, a node n_i as root
- included_set = $\{n_i\}$
- remaining_set = remaining_set − included_set
    - Repeat the following steps until remaining_set becomes empty
    - For each node, $n_r \in$ remaining_set
        - distance with the tree, $d_k = \min\{\text{dist}(n_r, n_{in})\}$ where, $n_{in} \in$ included_set
    - Let $d_t = \min\{d_k\}$ and it corresponds to the node $n_t \in$ remaining_set
    - Put n_t in included_set
    - Remove n_t from remaining_set

In the proposed scheme, distance between two nodes (feature vectors) is computed as follows.

$$\text{dist} = \text{dist}_w + K \times \text{dist}_g \tag{2}$$

K is a scalar constant that brings dist_w and dist_g to same scale. dist_w is Euclidean distance between the wavelet-based feature vectors. It may be noted that the wavelet-based features are computed with normalized intensity values and dist_w lies within [0, 1]. dist_g, the distance between gradient histogram based features, is computed as

$$\text{dist}_g = 1 - \text{avg}\{\text{sim}_i\} \tag{3}$$

sim_i is the similarity measured by Bhattacharya distance between the gradient histograms of the two frames.

Once the MST is prepared, if we remove all the edges connecting pair of points coming from two different samples, each sub-tree formed would consist of samples from one and only one population and is equivalent to a run of a univariate case. Thus, the number of nodes in each sub-tree is equivalent to the run-length, and R, the number of sub-trees, is equivalent to number of runs. Example of MST has been shown in Fig. 3 where filled nodes and empty nodes denote the sample points of two samples (video sequences). Along a path, sequence of same type of nodes (either all empty nodes or filled nodes) represent a run, and a number of nodes in such sequence are the length of the run. In Fig. 3, nodes forming a run have been encircled. Physically, it may be interpreted that more similar the samples, they will tend to interleave more frequently resulting into large number of runs with smaller length.

Finally, the test statistic W is defined as

$$W = \frac{R - E[R]}{\sqrt{\text{Var}[R]}} \tag{4}$$

where $\text{Var}[R] = \frac{2mn}{N(N-1)} \times (\frac{2mn-N}{N} + \frac{C-N+2}{(N-2)(N-3)} \times (N(N-1)-4mn+2))$, C is the number of edge pairs in MST sharing a common node and $E[R] = \frac{2mn}{N} + 1$. Thus, the steps for sequence matching are as follows.

- Prepare MST considering the frames in query shot (s_q) and reference shot (s_r) as the samples
- Compute the test statistic W

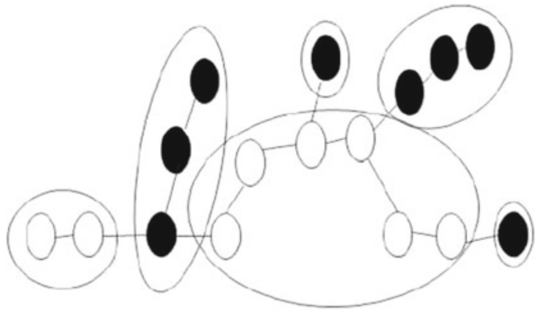


Fig. 3 Minimal spanning tree with filled and empty nodes representing frames of two video sequences and encircled nodes represent a run

- If W falls in critical region then reject H_0, the null hypothesis
- If W is rejected then s_q is not a copy of s_r else s_q is a copy of s_r

As W follows standard normal distribution, a critical region may be chosen for a given level of significance (α) which signifies the maximum probability of rejecting a true H_0. If W falls within the critical region, H_0 is rejected. Physically, low value of R expresses that two samples are less interleaved in the ordered list, and it leads to the interpretation that they are from different populations. In our experiment, α is taken as 0.1.

Some of the systems [7,15,23] assumes that query sequence is of shorter duration in comparison with the reference sequence. But, it cannot be ensured as the query sequence may be a slow motion version [31] or may be transformed one with increased frame rate. It may be noted that the proposed sequence matching does not make any assumption regarding the relative length of query and reference sequences. Essentially, it looks for considerable interleaving of the two sequences in the feature space. As a result, the technique has strong immunity to the attacks like frame drop, frame clipping, change in frame rate, frame substitution. Moreover, it is also robust against the photometric attack as long as the feature values of the transformed sequence lies within the same range in the feature space as the original sequence.

4 Experimental results

The experiment has been carried out with a reference video database of approximately 2 h 45 min duration. It contains different sequences of various categories, like news, sports, documentary, movie, and television serials. The sequences are collected from TRECVID 2001, 2002, and 2005 databases. Sequences recorded from television broadcast are also included in the database. The description of the database is shown in Table 1. Please note that, the database contains number of shots belonging to same category like news, soccer, landscape. But, in the context of a CBVCD system, all the shots are considered distinct. So the number of classes is equal to the number of shots or sequences.

Table 1 Description of data and parameters used in the experiments

Reference database		Parameters	
Size of frames	352 × 288	α	0.1
Number of frames	269,628	n_w	5
Number of shots (N_r)	642	N_p	16
Number of class	642	K	1

The value of the different parameters used in the proposed system are summarized in Table 1. n_w denotes the level of decomposition in computing wavelet-based features. N_p stands for number of partitions in which the frame image is divided for obtaining the spatial distribution of pixel gradient. K is the scaling factor used in combining dist_w and dist_g in Eq. 2, and α is required to determine the critical region for test statistic in sequence matching.

Queries are generated by randomly selecting shots from the database. Different types of photometric attack are incorporated in these query shots. Thus, the different versions of query shots include the shots where brightness/contrast is changed and 25–30 % of the pixels are corrupted by random noise. Another set of queries are formed by incorporating different display formats, where the query shots are made to be in pillar box, letter box, flat file format, or even without any such border. So queries are formed with the shots with/without border. Thus, the query shots and the corresponding original ones in the reference database are in different forms. The borders appeared in different display formats, except the L-shaped border, are removed automatically in the pre-processing step, while the L-shaped borders are removed manually. Queries with changed frame rate are formed by randomly dropping the frames from either the query shot or the reference shot. In some cases, query and the reference shots are prepared by randomly choosing the mutually exclusive set of frames from the original sequence. Clipping effect is generated by dropping a set of frames from the leading or trailing or any intermediate part of the reference shots. For certain cases, leading/trailing set of frames of a query are substituted by other frames of a different shot. Number of queries considered for different types of attack is presented in Table 2.

Suppose S_R denotes the set of whole video database at our disposal. For our experiment, for a particular type of attack, we arbitrarily choose a subset $S_r \subset S_R$ as the reference database and another subset $S_Q \subset S_R$ as an initial query database. Here, S_r and S_Q are chosen in such a way that $S_r \neq S_Q$ and $S_c = S_r \cap S_Q \neq \Phi$. So S_c denotes the set of queries that are copies of some shots in S_r, and appropriate labeling is used with the groundtruth for copied sequences. Then, we apply certain attack on the elements of S_Q to generate a set S_q which is finally used as the query database for the experiment. In this experiment, S_r is fixed, but S_q changes for different attacks (see second column of Table 2). For simplicity, we assume the elements of S_Q and S_q can be referred to by the same label S_{q_j}. Further suppose that for a query shot S_{q_j} the set $S_{d_j} \subseteq S_r$ denotes the set of all S_{r_i} of which S_{q_j} is detected as a copy. Let us define $N_r = \#(S_r)$ and $N_{d_j} = \#(S_{d_j})$, where #() stands for cardinality.

Now, given a query shot s_{q_j} from S_q, it is exhaustively verified with each reference shots s_{r_i} in the reference database S_r to infer if the former is a copied version of the

Table 2 Performance comparison of different systems (CR and FR in %)

Attack	No. of queries	Kim's system [23]		Chen's system [7]		Harvey's system [15]		Proposed system	
		CR	FR	CR	FR	CR	FR	CR	FR
No attack	250	100.00	3.85	100.00	3.84	100.00	3.74	100.00	0.13
Brightness changed	200	99.50	3.85	98.50	3.86	99.00	4.75	100.00	0.14
Contrast changed	200	99.50	3.85	99.00	3.86	98.50	3.74	100.00	0.13
Noise added	200	98.50	3.86	98.50	3.86	100.00	3.75	99.50	0.13
Pillar box	125	100.00	3.85	100.00	3.85	99.20	3.75	100.00	0.13
Letter box	125	100.00	3.85	100.00	3.85	99.20	3.75	99.20	0.13
Flat file	125	90.40	3.77	92.80	3.76	99.20	3.75	100.00	0.13
L-shaped border	60	56.67	3.21	58.33	3.27	70.00	2.72	91.67	0.16
Logo inserted	80	100.00	3.85	100.00	3.84	97.50	3.74	97.50	0.13
Changed frame rate	40	100.00	3.82	100.00	3.85	80.00	3.37	100.00	0.20
Clipping	40	75.00	3.82	67.50	3.81	80.00	3.34	100.00	0.16
Substitution	40	92.50	3.82	95.00	3.85	100.00	3.29	100.00	0.16
Overall	1,485	96.22	3.82	96.16	3.82	97.58	3.67	99.39	0.14

latter. The outcome consists of a sequence of N_r elements $\langle O_1, O_2, \ldots, O_{N_r}\rangle$. O_i is true if the system declares s_{q_j} as a copy of s_{r_i}; otherwise, it is false. So $S_{d_j} = \{s_{r_i} \mid O_i = \text{true}\}$. If s_{q_j} is a copy of s_{r_i} then "$O_i = \text{true}$" indicates correct copy detection and "$O_i = \text{false}$" indicates a miss. On the other hand, if s_{q_j} is not a copy (with or without attack) of s_{r_i} then "$O_i = \text{true}$" indicates false alarm. For a CBVCD system, copy detection rate (CR) must be high, while false alarm rate (FR) and miss-detection rate (MR) should be low. For the query set S_q, the CR and FR are computed as follows.

$$\begin{aligned} CR &= \frac{\text{number of correctly detected copy}}{\text{number of actual copy}} \times 100\,\% \\ &= \frac{\sum_j \#(S_{d_j} \cap S_c)}{\#(S_c)} \times 100\,\% \qquad (5) \end{aligned}$$

$$\begin{aligned} FR &= \frac{\text{number of false alarm}}{\text{Total number of trials}} \times 100\,\% \\ &= \frac{\sum_j \#(S_{d_j}) - \sum_j \#(S_{d_j} \cap S_c)}{\#(S_q) \times \#(S_r)} \times 100\,\% \qquad (6) \end{aligned}$$

MR $= (100 - \text{CR})\,\%$ represents percentage of miss-detection.

Spatio-temporal measure is widely used in video sequence matching. Hence, for comparison, we have implemented the spatio-temporal measure based systems proposed by Kim et al. [23], Chen et al. [7], and Harvey and Heefeda [15]. All the three methods have assumed that reference sequence is longer than the query sequence, that is, $n \geq m$, where n and m are the number of frames in the reference and the query sequences, respectively. For copy detection, based on a distance measure, the query sequence is compared with the sub-sequences of same length (i.e., m) taken from the reference sequence. Thus, the sub-sequence matching is carried out $(n - m + 1)$ times incurring the cost toward computation of distance measure. The frame rate is, implicitly, assumed to be same for the query and the reference shots. Finally, if the distance of the most similar sub-sequence is less than a predetermined threshold value, the query sequence is declared a copied version of the source sequence. Three systems differ mainly in terms of their features and distance measures.

A spatio-temporal sequence matching technique has been presented in [23], where each frame is partitioned into 2×2, that is, 4 sub-images of equal size and a 2×2 rank matrix is formed for each frame based on the average intensity of the sub-images. Formation of rank matrix involves sorting; otherwise, cost of feature computation varies linearly with the frame size. Based on the rank matrix, an ordinal measure has been presented for spatial matching of two sequences. A temporal distance metric has also been incorporated that relies on the direction of change (increase/decrease/no change) in average intensity of the sub-images over the sequence. A metric combining spatial and temporal measures is used for sequence matching. Cost of computation for both spatial and temporal measure is of $O(N)$ where N is the length of sequence, and it has to be repeated for each partition.

Chen and Stentiford [7] have followed similar philosophy as in [23] and adopted only the temporal metric based on ordinal measure. First of all, the frames are partitioned in the same manner, and each partition is represented by the average intensity of the pixels in the it. Instead of forming the rank matrix for individual frame as in [23], it is formed corresponding to each sub-image (partition) by considering all the frames in a sequence. Thus, the rank matrix is of dimension $k \times N$, where $k(= 4)$ is the number of partitions and N is the number of frames in the sequence. In [23], the rank matrix captures only the spatial information; whereas in [7], the rank matrix reflects the temporal aspect of the sequence. But, rank matrix based ordinal measure used in [7] is similar to that in [23]. Computational cost of rank matrix formation is of $O(N \log N)$ as the N elements in a partition are to be sorted. Furthermore, the task has to be carried out for each of the k partitions. Computation of ordinal measure incurs further cost of $O(N)$. It may be noted that the query sequence is verified with a part of reference sequence. Thus, the computation is to be done for each reference sub-sequence.

Harvey and Hefeeda [15] have considered pre-processing to deal with display formats like pillar and letter box relying on the assumption that the color of introduced borders are of homogeneous in nature. After removing the borders, the content (residual part of the frame) is divided into 3×2 partitions. For each sub-image, thus obtained, the local interest points are generated following speeded up robust features (SURF) methodology. Based on the count of such interest points in the partitions, rank matrix for each partition is formed considering all the frames in a sequence. As in the case of the system in [7], an ordinal measure based on the rank matrix is formulated for sequence matching. It also incurs the similar computational cost as in the case of Chen and Stentiford's system.

In our system, the cost of computation for wavelet-based signature and spatial distribution of gradient varies linearly with the frame size. But, MST creation for sequence matching is computation intensive and cost is of $O(N^2)$ where N is the number of frames participating in the process. In case of a CBVR system, the response time is very crucial, but it is not so for a CBVCD system. In a CBVCD system, a complaint may come with possible target that reduces the search space considerably. More often, copy detection will be an offline activity. Thus, for a CBVCD system, copy detection rate and false alarm rate are more crucial parameters than the response time.

The performance of the proposed system and the other three systems [7,15,23] is presented in Table 2. Number of queries have been executed for each type of attack, and the average values for correct detection rate (CR) and false alarm rate (FR) in percentage are presented. The value of miss-detection rate (MR) is not given here as it is straightaway revealed by CR. The systems presented in [7,23] suffer from a number of limitations. They sustain letter box/pillar box kind of change in display format as such changes affect all the four sub-images in the same manner without leading to any significant change in the rank matrix. But, they cannot cope up with flat file format or L-shaped border. The average intensity of four sub-images form the basic descriptor, and it is almost global. It has made the systems low cost and robust against the photometric attacks. But, the descriptor of too general nature is a compromise in terms of its discriminating power, and it is likely to generate similar rank matrix even for non-similar frames/sequences. As a result, possibility of false alarm is considerably high. However, in case of the system in [7], as the rank matrix is formed for the sequence, its capability to deal with flat file format and L-shaped border is higher than the Kim's system [23]. The system proposed by Harvey and Heefeda [15] have considered pre-processing to remove the borders. As a result, it has performed better than the systems in [7,23] in dealing with flat file format and L-shaped border. Moreover, false alarm rate is also relatively less as the interest point based features having relatively more discriminating power. Table 2 depicts that the proposed method handles wide variety of attacks with high copy detection rate, and the false alarm rate is also very low which is desirable criteria for any CBVCD system. Overall performance indicates that the proposed method outperforms the other systems in terms of both the parameters CR and FR.

Moreover, the assumption made in [7,15,23] that the query sequence is of shorter duration is too restrictive as the situation may be just opposite in case of an attack like increased frame rate, or the copied one is a slow motion version of the reference sequence. Besides, all the systems [7,15,23] rely on a pre-determined threshold for copy detection, and the selection of this threshold is a non-trivial task. In [23], weight adjustment between spatial and temporal measure is also very crucial.

In the proposed system, simple pre-processing technique has made it robust against the geometric attack and has also made immune to the photometric attack. Wavelet-based features are quite general in nature and robust against the photometric attacks. Features based on the spatial distribution of gradient magnitude capture the local structures significantly. So the proposed features are sufficiently representative, and it is well indicated by the fact that the proposed system does not generate any false alarm for the cases shown in Fig. 4a–d. The hypothesis test based sequence matching scheme looks for interleaving of the frame level descriptors of the two sequences under consideration. Thus, it takes care of the temporal similarity of the sequences in a relatively relaxed manner and becomes robust against attacks like changed frame rate, clipping, and substitution. As a whole, the proposed system meets the diverging requirements of robustness and discriminability of a CBVCD system.

5 Conclusion

In this work, we have proposed a content-based video copy detection system which is robust against various post-production attacks (both geometric and photometric). The fingerprints or features of video data are generated based on the wavelet decomposed sub-bands and localized intensity gradient histograms in the low sub-band. The proposed fingerprint is robust with considerable discriminating power.

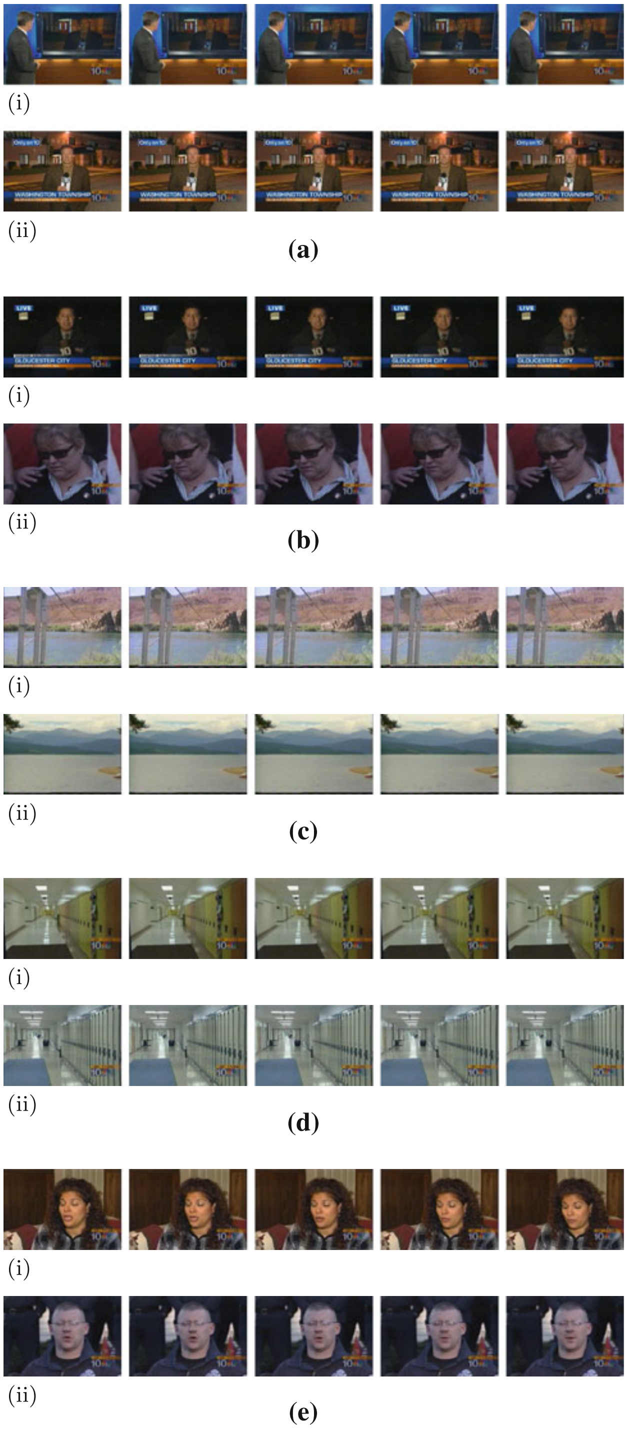


Fig. 4 Sample cases of false detection by different systems **a** A sample case of false detection by Kim's System [23]. (*i*) Sample frames of query sequence. (*ii*) Sample frames of matched reference sequence. **b** A sample case of false detection by Kim's System [23] and Harvey's System [15]. (*i*) Sample frames of query sequence. (*ii*) Sample frames of matched reference sequence. **c** A sample case of false detection by Chen's System [7] and Harvey's System [15]. (*i*) Sample frames of query sequence. (*ii*) Sample frames of matched reference sequence. **d** A sample case of false detection by Chen's System [7] and Kim's system [23] and Harvey's System [15]. (*i*) Sample frames of query sequence. (*ii*) Sample frames of matched reference sequence. **e** A sample case of false detection by Chen's System [7], Kim's system [23], Harvey's System [15] and the proposed system. (*i*) Sample frames of matched reference sequence. (*ii*) Sample frames of matched reference sequence

Simple pre-processing is adopted to enhance the attack handling capability. For sequence matching, a hypothesis test based scheme is employed which also adds to the strength of the system. As the computation cost of the proposed system is comparatively higher, in future, an indexing scheme may be adopted to carry out the matching with a reduced set of reference sequences. An experiment is carried out with a reference database consisting of 642 distinct shot. 1,485 query sequences have been formulated to represent different types of attacks. Proposed system achieves high copy detection rate (99.39 %) and low false alarm rate (0.14 %). Thus, it fulfills the diverging criteria of robustness and discriminability. Performance of our system has also been compared with three other well-known systems that rely on widely used concept of spatio-temporal matching. Experimental result indicates that the proposed system is a more effective content-based video copy detection system than others.

References

1. Adjeroh, D.A., Lee, M.C., King, I.: A distance measure for video sequences. Comput. Vis. Imag. Underst. **75**, 25–45 (1999)
2. Barrios, J.M., Bustos, B.: Competitive content-based video copy detection using global descriptors. Multimed. Tools Appl. **62**(1), 75–110 (2011)
3. Bhat, D.N., Nayar, S.K.: Ordinal measures for image correspondence. IEEE Trans. PAMI **20**(4), 415–423 (1998)
4. Chang, E.Y., Wang, J.Z., Li, C., Wiederhold, G.: Rime: a replicated image detector for the world-wide-web. In: Proceedings of the SPIE multimedia storage and archiving systems III, pp. 68–71 (1998)
5. Chang, S.F.S., Chen, W., Meng, H.J., Sundaram, H., Zhong, D.: Videoq: an automated content based video search system using visual cues. In: Proceedings of 5th ACM international conference on Multimedia, ACM, pp. 313–324 (1997)
6. Chen, L., Chua, T.S.: A match and tiling approach to content-based video retrieval. In: Proceedings of the international conference on multimedia and expo (2001)
7. Chen, L., Stentiford, F.W.M.: Video sequence matching based on temporal ordinal measurement. Pattern Recognit. Lett. **29**, 1824–1831 (2008)
8. Cheung, S.C.S., Zakhor, A.: Efficient video similarity measurement with video signature. IEEE Trans. CSVT **13**(1), 59–74 (2003)
9. Coskun, B., Sankur, B., Memon, N.: Spatio-temporal transform based video hashing. IEEE Trans. Multimed. **8**(6), 1190–1208 (2006)
10. Ferman, A.M., Tekalp, A.M., Mehrotra, R.: Robust color histogram descriptors for video segment retrieval and identification. IEEE Trans. IP **11**(5), 497–508 (2002)
11. Friedman, J.H., Rafsky, L.C.: Multivariate generalizations of the Wald–Wolfowitz and smirnov two-sample tests. Ann. Stat. **7**(4), 697–717 (1979)
12. Guil, N., Gonzalez-Linares, J.M., Cozar, J.R., Zapata, E.L.: A clustering technique for video copy detection. In: Proceedings of the Iberian conference on pattern recognition and image, analysis, pp. 451–458 (2007)
13. Hampapur, A., Bolle, R.: Feature based indexing for media tracking. In: Proceedings of the International conference on multimedia and expo, pp. 67–70 (2000)
14. Hampapur, A., Bolle, R.: Comparison of sequence matching techniques for video copy detection. In: Proceedings of the international conference on multimedia and expo, pp. 188–192 (2001)
15. Harvey, R.C., Heefeda, M.: Spatio-temporal video copy detection. In: Proceedings of the multimedia system conference, pp. 35–46 (2012)
16. Hua, X.S., Chen, X., Zhang, H.J.: Robust video signature based on ordinal measure. In: Proceedings of the ICIP, pp. 685–688 (2004)
17. Jain, A.K., Vailaya, A., Xiong, W.: Query by clip. Multimed. Syst. J. **7**(5), 369–384 (1999)
18. Jeong, K.M., Lee, J.J., Ha, Y.H.: Video sequence matching using singular value decomposition. In: Proceedings of the ICIAR, pp. 426–435 (2006)
19. Joly, A., Buisson, O., Frelicot, C.: Content-based copy retrieval using distortion-based probabilistici similarity search. IEEE Trans. Multimed. **9**(2), 293–306 (2007)
20. Ke, Y., Sukthankar, R., Houston, L.: Efficient near duplicate detection and sub-image retrieval. In: Proceedings of the MM (2004)
21. Kim, C.: Content-based image copy detection. Signal Process. Image Commun. **18**(3), 169–184 (2003a)
22. Kim, C.: Ordinal measure of DCT coefficients for image correspondence and its application to copy detection. In: Proceedings of the for SPIE storage and retrieval for media databases, pp. 199–210 (2003b)
23. Kim, C., Vasudev, B.: Spatiotemporal sequence matching for efficient video copy detection. IEEE Trans. CSVT **15**(1), 127–132 (2005)
24. Kim, S.H., Park, R.H.: An efficient algorithm for video sequence matching using the modified hausdroff distance and the directed divergence. IEEE Trans. CSVT **12**(7), 592–596 (2002)
25. Kim, Y.T., Chua, T.S.: Retrieval of news video using video sequence matching. In: Proceedings of the CIVR (2007)
26. Law-To, J., Chen, L., Joly, A., Laptev, I., Buisson, O., Gouet-Brunet, V., Boujemaa, N., Stentiford, F.: Video copy detection: a comparative study. In: Proceedings of the CIVR (2007)
27. Lee, S., Yoo, C.D.: Video fingerprinting based on centroids of gradient orientations. In: Proceedings of the ICASSP, pp. 401–404 (2006)
28. Li, Y., Jin, J.S., Zhou, X.: Video matching using binary signature. In: Proceedings of the international symposium on intelligent signal processing and communication Systems, pp. 317–320 (2005)
29. Liu, Z., Liu, T., Gibbon, D., Shaararay, B.: Effective and scalable video copy detection. In: Proceedings of the MIR'10 (2010)
30. Maani, E., Tsaftaris, S.A., Katsaggelos, A.K.: Local feature extraction for video copy detection. In: Proceedings of the ICIP, pp. 1716–1719 (2008)
31. Mohan, R.: Video sequence matching. In: Proceedings of the ICASSP, pp. 3697–3700 (1998)
32. Mohanta, P.P., Saha, S.K., Chanda, B.: A model-based shot boundary detection technique using frame transition parameters. IEEE Trans. Multimed. **14**(1), 223–233 (2012)
33. Oostveen, J., Kalker, T., Haitsma, J.: Feature extraction and a database strategy for video fingerprinting. In: Proceedings of the VISUAL, pp. 117–128 (2002)
34. Radhakrishnan, R., Bauer, C.: Robust video fingerprints based on subspace embedding. In: Proceedings of the ICASSP, pp. 2245–2248 (2008)
35. Ravandadi, S., Aarabi, P.: Rotation invariance in imaging. In: Proceedings of the ICASSP(2007)
36. Sarkar, A., Ghosh, P., Moxley, E., Manjunath, B.S.: Video fingerprinting: features for duplicate and similar video detection and query-based video retrieval. SPIE—multimedia content access: algorithms and systems II (2008)
37. Schmid, C., Mohr, R.: Local grayvalue invariants for image retrieval. IEEE Trans. PAMI **19**(5), 530–535 (1997)

38. Seo, J.S., Haitsma, T.J., Yoo, C.D.: A robust image fingerprinting system using the radon transform. In: Proceedings of hte ACM MM, vol. **19**, pp. 325–339 (2004)
39. Seo, J.S., Jin, M., Lee, S., Jang, D., Lee, S.J., d Yoo, C.: Audio fingerprinting based on normalized spectral subband centroids. In: Proceedings of the ICASSP, pp. 213–216 (2005)
40. Shen, H., Ooi, B.C., Zhou, X.: Towards effective indexing for very large video sequence database. In: Proceedings of the SIGMOD, pp. 730–741 (2005)
41. Su, X., Huang, T., Gao, W.: Robust video fingerprinting based on visual attention regions. In: Proceedings of the ICASSP (2009)
42. Sze, K.W., Lam, K.M., Qiu, G.: A new keyframe representation for video segment retrieval. IEEE Trans. CSVT **15**(9), 1148–1155 (2005)
43. Wald, A., Wolfowitz, J.: On a test whether two samples are from the same population. Ann. Math. Stat. **11**, 147–162 (1940)
44. Willems, G., Tuytelaars, L.T.: Spatio-temporal features for robust content-based video copy detection. In: Proceedings of the ACM MM (2008)
45. Wu, X., Zhang, Y., Wu, Y., Guo, J., Li, J.: Invariant visual patterns for video copy detection. In: Proceedings of the ICPR, pp. 1–4 (2008)
46. Yeh, M.C., Cheng, K.Y.: A compact, efective descriptor for video copy detection. In: Proceedings of the ACM Multimedia (2009)
47. Zhao, H.V., Wu, M., Wang, Z.J., Liu, K.J.R.: Forensic analysis of nonlinear collusion attacks for multimedia fingerprinting. IEEE Trans. IP **14**(5), 646–661 (2005)