Journal of Signal and Information Processing, 2016, 7, 84-97
Published Online May 2016 in SciRes. http://www.scirp.org/journal/jsip
http://dx.doi.org/10.4236/jsip.2016.72010

Shearlet Based Video Fingerprint for Content-Based Copy Detection

Fang Yuan[1], Lam-Man Po[1], Mengyang Liu[1], Xuyuan Xu[2], Weihua Jian[2], Kaman Wong[2], Keith W. Cheung[3]

[1]Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China
[2]Video Team, Tencent Holdings Limited, Shenzhen, China
[3]School of Communication, Hang Seng Management College, Hong Kong, China
Email: iyuanfang0105@gmail.com, eelmpo@cityu.edu.hk, xuyuanxu2@gmail.com

Received 27 January 2016; accepted 14 May 2016; published 17 May 2016

Abstract

Content-based copy detection (CBCD) is widely used in copyright control for protecting unauthorized use of digital video and its key issue is to extract robust fingerprint against different attacked versions of the same video. In this paper, the "natural parts" (coarse scales) of the Shearlet coefficients are used to generate robust video fingerprints for content-based video copy detection applications. The proposed Shearlet-based video fingerprint (SBVF) is constructed by the Shearlet coefficients in Scale 1 (lowest coarse scale) for revealing the spatial features and Scale 2 (second lowest coarse scale) for revealing the directional features. To achieve spatiotemporal natural, the proposed SBVF is applied to Temporal Informative Representative Image (TIRI) of the video sequences for final fingerprints generation. A TIRI-SBVF based CBCD system is constructed with use of Invert Index File (IIF) hash searching approach for performance evaluation and comparison using TRECVID 2010 dataset. Common attacks are imposed in the queries such as luminance attacks (luminance change, salt and pepper noise, Gaussian noise, text insertion); geometry attacks (letter box and rotation); and temporal attacks (dropping frame, time shifting). The experimental results demonstrate that the proposed TIRI-SBVF fingerprinting algorithm is robust on CBCD applications on most of the attacks. It can achieve an average F1 score of about 0.99, less than 0.01% of false positive rate (FPR) and 97% accuracy of localization.

Keywords

Video Fingerprint, Content-based Copy Detection, Shearlet Transform

1. Introduction

Tens of thousands of videos are being uploaded to the Internet and shared everyday with about 300 hours upload

How to cite this paper: Yuan, F., Po, L.-M., Liu, M.Y., Xu, X.Y., Jian, W.H., Wong, K. and Cheung, K.W. (2016) Shearlet Based Video Fingerprint for Content-Based Copy Detection. *Journal of Signal and Information Processing*, **7**, 84-97.
http://dx.doi.org/10.4236/jsip.2016.72010

per minute [1]. However, a considerable number of these videos are illegal copies or manipulated versions of existing media. This widespread video copyright infringement makes the video copyright management on the Internet a complicated process, at the same time, calls for the development of fast and accurate copy-detection algorithms. Since video is the most complex type of digital media, it has so far received the least attention regarding copyright management. The task of video copy detection determines if a given video (query) has a duplicate in a set of videos. Query videos may be distorted in various ways, such as change of brightness, text insertion, compression, and cropping. If the system finds a matching video segment, it returns the name of the database video and the time stamp at which the query is copied.

Growing broadcasting of digital video content on different media brings the search of copies in large video databases to a new critical issue. Digital videos can be found on TV Channels, Web-TV, video blogs and the public video web servers. The massive capacity of these sources makes the tracing of video content into a very hard problem for video professionals. At the same time, controlling the copyright of the huge number of videos uploaded everyday is a critical challenge for the owner of the popular video web servers. Because videos are available in different formats, it is more efficient to base the copy detection process on the content of the video rather than its name, description, or binary representation. Video fingerprinting [2]-[4] (also known as robust hashing) has been recently proposed for this purpose. A fingerprint is a content-based signature derived from a video (or other form of a multimedia asset) so that it specifically represents the video or asset. To find a copy of a query video in a video database, one can search for a close match of its fingerprint in the corresponding fingerprint database (extracted from the videos in the database). Closeness of two fingerprints represents a similarity between the corresponding videos but two perceptually different videos should have different fingerprints.

Most of the conventional video fingerprint extraction algorithms can be classified into four categories based on the features they extracted, which are 1) Color-space-based, 2) Temporal, 3) Spatial, and 4) Spatiotemporal. For the first category, color-space based fingerprints mostly derived from the histograms of the colors in specific regions in time and/or space within the video, and the RGB image is usually converted into YUV and LAB color-space [2]. However, color features also exists several inherent drawbacks. For example, these features are sensitive to different video formats, and these features also cannot be used for grayscale images. In consideration of these drawbacks, most of the practical video fingerprinting algorithms are based on grayscale images.

For the second category, temporal fingerprints are extracted from the characteristics of a video sequence over time. For example, Chen, L. *et al.* [5] have proposed a video sequence matching method based on temporal ordinal measurements. This method divided each frame into a grid and corresponding grids along a time series are sorted in an ordinal ranking sequence, which gives a global and local description of temporal variation. Temporal features usually work well with long video sequences, but do not perform well for short video clips since they do not contain sufficient discriminant temporal information. Because short video clips occupy a large share of online video databases, temporal fingerprints alone do not suit online applications.

For the third category, spatial fingerprints are usually derived from each frame or from a key frame. They are widely used for both video and image fingerprinting, and there are a lot of researches in this category. For example, Li, T. *et al.* [6] adopted ordinal intensity signature (OIS) as the frame feature descriptor, which divided each frame into a grid and sorted it into an ordinal intensity signature. Radhakrishnan, R. *et al.* [7] have proposed a video signature extraction method based on projections of difference images between consecutive video frames. In this method, the difference images are projected onto random basis vectors to create a low dimensional bit-stream representation of the active content (moving regions) between two video frames. In addition, De Roover, C. *et al.* in [8] also proposed a robust video hashing which is based on radial projections of key frames.

For the fourth category, spatiotemporal fingerprints consider both spatial and temporal information when designing the algorithms. In [9], Kim, C. *et al.* have proposed a video fingerprint method, which is based on spatiotemporal transform. In this method, a segment of video is considered as a 3-D matrix of luminance values. After the preprocessing phase, a 3D-DCT is applied to videos to extract spatiotemporal features. However, the computational and memory requirements of applying the 3-D transform to a video are very high especially for real-time applications. To tackle this problem, Esmaeili, M.M. *et al.* proposed to use temporally informative representative images (TIRIs) [10] [11] of short video segments for fingerprints generation such that spatial and temporal information can be represented in the generated TIRI-based fingerprints. They also developed a TIRI-2D-DCT based fingerprinting system that has been demonstrated to be outperformed the 3D-DCT based fingerprinting system. However, these fingerprinting algorithms are all based on the traditional DCT and this paper attempt to use advance Shearlet transform for video fingerprint generation.

In addition, a general-purpose no-reference image quality assessment (NR-IQA) method is recently proposed in [12] [13] based on the statistical characterization in the Shearlet domain [14]-[19], which is named as SHANIA (SHeArlet based No-reference Image quality Assessment). It is a combination of natural scene statistics (NSS) based and training-based approaches, and can estimate a wide range of image distortions. The main idea of SHANIA is based on the finding that if a natural image is distorted by some common distortions, the linear relationship in coarser scales will retain, but it will be disturbed in fine scales, especially in higher fine scales. Thus, these variations of statistical property in fine scales can be easily detected by Shearlets and applied to describe image quality distortion. Motivated by the NSS model, the sum of Subband Coefficient Amplitudes (SSCA) of coarse scales is viewed as the "natural parts" of a distorted image and the SSCA of fine scales is referred as "distorted parts".

In this paper, we attempt to use the "natural parts" (coarse scales) of the Shearlet coefficients to design a robust transformation-invariant video fingerprint for content-based video copy detection applications. The proposed Shearlet-based video fingerprint (SBVF) is constructed by the Shearlet coefficients in Scale 1 (lowest coarse scale) for revealing the spatial features and Scale 2 (second lowest coarse scale) for revealing the directional features. To achieve spatiotemporal video fingerprint nature, the SBVF is used with TIRIs to build a TIRI-SBVF copy detection system for performance evaluation. With the statistical normalized hamming distance (NHD) and detection and localization performance evaluations using TRECVID 2010 dataset, it is shown that the proposed SBVP is a robust video fingerprint with strong ability of discrimination against different content.

The rest paper is organized as follows. In Section 2, the basic principles of Shearlet transformation are first introduced and then the new SBVF algorithm will be presented. To evaluate the performance of the proposed algorithm, a TIRI-SBVF based CBCD system is presented in Section 3. Experimental results are reported in Section 4 and we conclude this paper in Section 5.

2. Shearlet Based Video Fingerprint (SBVF)

2.1. Shearlet Transform

The major finding in [12] [13] is that if a natural image is distorted by some common distortions, the linear relationship in coarse scales will be retained, but it will be disturbed in fine scales, especially in higher fine scales. Basically, Shearlet transform [12]-[19] is a multi-scale and multi-dimensional wavelet transform. It is capable of addressing anisotropic and directional information at different scales. Take 2-dimensional case as an example, the affine system with composite dilations is defined as:

$$SH_{\phi} f\left(a,s,t\right) = \left\langle f, \phi_{a,s,t} \right\rangle,\ a>0,\ s\in R,\ t\in R^2 \tag{1}$$

where the analyzing factor $\phi_{a,s,t}$ is called Shearlet coefficient, which is defined as:

$$\phi_{a,s,t}\left(x\right) = \left|\det M_{a,s}\right|^{-\frac{1}{2}} \phi\left(M_{a,s}^{-1} x - t\right) \tag{2}$$

in which $M_{a,s}$ is given by

$$M_{a,s} = B_s A_a = \begin{pmatrix} a & \sqrt{as} \\ 0 & \sqrt{a} \end{pmatrix} \text{ with } A_a = \begin{pmatrix} a & 0 \\ 0 & \sqrt{a} \end{pmatrix} \text{ and } B_s = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \tag{3}$$

where A_a is the anisotropic dilation matrix and B_s is the shear matrix. The framework of Shearlet transform is anisotropic and the analyzing functions are defined at different scales, locations and orientations, Thus, Shearlet is more efficient to detect directional information compared with the conventional wavelet transform. From the point of view of optimal approximation, if the signal f (such as an image in the space of C^2) can be reconstructed by partial sums on N largest coefficients and $\hat{f}_N$ is the approximation of f. The approximation characteristic of Shearlet transform, which can achieve $\varepsilon_{\text{Shearlet}} = \left\| f - \hat{f}_N \right\|_2^2 \le CN^{-2}\left(\log N\right)^3$, is better than the wavelet transform $\left(\varepsilon_{\text{Wavelet}} = \left\| f - \hat{f}_N \right\|_2^2 \le CN^{-1}\right)$ and Fourier transform $\left(\varepsilon_{\text{Fourier}} = \left\| f - \hat{f}_N \right\|_2^2 \le N^{-(1/2)}\right)$.

In this paper, we propose to use the coarse scales of the Shearlet coefficients to design a robust fingerprint for content-based video copy detection applications. The proposed SBVF is defined with use of a 4-scale Shearlet transform with 6 directions and this multi-scale Shearlet decomposition is illustrated in **Figure 1**. In this figure, the scale levels are denoted as S1 to S4 and directional filtered coefficients are denoted as D1 to D6.

Basically, Shearlet transform can be considered as a decomposition tool with both scales and directional information into account. Firstly, a two-channel non-subsampled decomposition [20] [21] (called as "a trous") is applied on the input image to recursively decompose the input image into a low-pass image and a high-pass image. This decomposition can be easily achieved with perfect reconstruction condition of $H_0(z)G_0(z) + H_1(z)G_1(z) = 1$, where $H_0(z)$ and $H_1(z)$ are low pass and high pass filter transfer functions, respectively. It is because the non-subsampled filter bank is shift-invariant. Secondly, in each scale of decomposition, the high-passed image is transformed to frequency domain by 2-dimensional Fourier Transform using Fast Fourier Transform (FFT) algorithm and then a Cartesian grid with 6 directions is applied on this frequency domain for generating 6 directional subbands using inverse FFT. However, the low-passed image is further filtered to generate the next scale image of the decomposition. Finally, the multi-scales (S1 to S4) and multi-directions (D1 to D6) information of input image is revealed by Shearlet coefficients.

Figure 2 illustrates a 4-scale Shearlet based on "a trous" decomposition for the well-known Lenna test image, in which S4 image is the high passed image in the first decomposition process while S1 image is the low-passed image in the last decomposition. Basically, the scales of S1 to S4 represent different frequency bands with lower scales for coarse information (low frequency) and higher scales for more detail information (high frequency). In addition, **Figure 3** shows the final Shearlet coefficients in multi-scale and multi-directions representations, in which, the directional information of D1 to D6 are representing the information of singularity points in high frequency images. These singularity points are detected by Shearlet transform in higher frequencies. Therefore the coarse image (S1) is non-directional decomposition. This Shearlet representation consists of more directional information than the conventional transforms such as DCT with representation only on some specific frequency bands. This is one of the motivations that we attempt to use the directional information captured by Shearlet to generate robust video fingerprints.

2.2. Robustness of the Shearlet Coefficients for Fingerprinting

For robust fingerprint construction, the robustness between the attacked versions of video segments, and

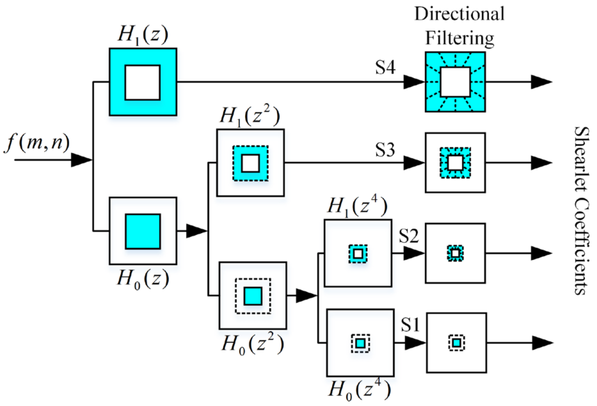


Figure 1. The overview of Shearlet transform. 4 scales (S1 to S4) and 6 directions (D1 to D6).

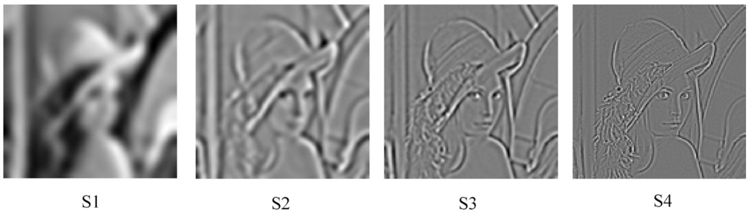


Figure 2. The images of different scales in Shearlet transform.

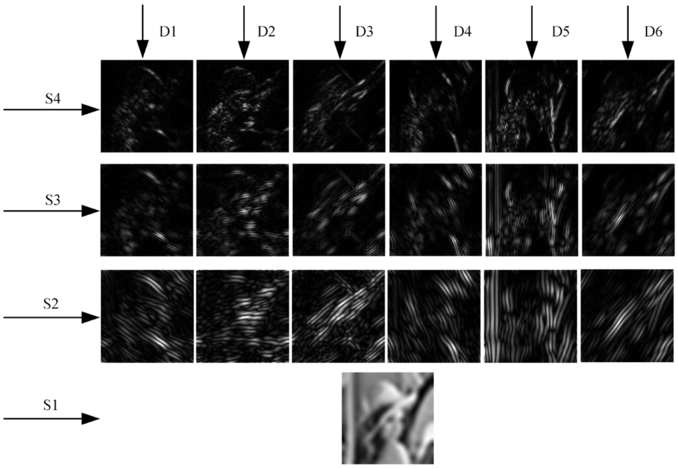


Figure 3. Shearlet coefficients (S1 to S4 for 4-scales, and D1 to D6 for 6-directions).

discrimination between the perceptual different videos are the key issues. The low frequency information of signal has turned to be robust against many distortions like noise corruption. Therefore, the Shearlet coefficients in coarse scales are much preferable as they are robust against different type of distortions and transformations while maintaining high discrimination for perceptual different images. Moreover, the high ordered features are revealed by multi-directional decompositions of Shearlet. To demonstrate the property of coefficients of higher scales in Shearlet domain, the normalized sum of subband coefficient amplitudes (SSCA) [12] [13] is adopted for evaluation, which is defined as:

$$PF(a,s)=\frac{\sum_t \left|SH_\phi f(a,s,t)\right|}{\max\left(\sum_t \left|SH_\phi f(a,s,t)\right|\right)} \tag{4}$$

where the $SH_\phi f(a,s,t)$ are the Shearlet coefficients with a,s,t are the scale, direction and time parameter respectively.

To demonstrate the statistical property of SSCA, a dataset with 366 video frames or images is generated with random selection from the TRECVID 2010 [22] [23] and INRIA Copy Days [24] datasets. Five common types of distortion in video copy detection attacks are applied on these images and they are luminance changing (LC), JPEG, JPEG200 (JP2), Salt and Pepper Noise (PN), Gaussian Noise (GN). **Figure 4** shows the SSCA property of Shearlet coefficients from S2 to S4 for these five types of distortion and **Tables 1-3** illustrate the numerical SSCA in S2 to S4 with 6 directions. In **Figure 4**, the Shearlet transform is 4-scales with 6-directions (subbands) and the horizontal axis represents the number of directions (subbands) in different scales (separated by the dashed line). As shown in this figure, the SSCA in S3 to S4 are seriously affected by distortions, while the S2 is very robust with SSCA are nearly the same on all types of distortion. Thus, it is possible to construct relatively robust fingerprint with directional information from S2.

2.3. Shearlet Based Fingerprint (SBVF) Construction

With the good robustness property of Shearlet S1 for non-directional spatial information and S2 for directional information, we proposed to use S1 and S2 with D1 to D6 images to generate robust video fingerprints. The input image of this fingerprinting algorithm is a pre-processed grayscale image with rescaling to the size of $M \times M$. **Figure 5** shows the block diagram of the proposed SBVF generation process. The S1 image is further down-sampled to an $M_1 \times M_1$ image for generating an S1-Hash, while the six D1 to D6 images (marked as S2-D1 to S2-D6) are down-sampled to six $M_2 \times M_2$ images for generating six directional hashes (D1-Hash to D6-Hash). These hashes are generated by 1-bit differential coding using horizontal snake scanning. It is because it can efficiently signify the variation of the down-sampled Shearlet images. The differential coding rule is very simple, bit "1" is assigned if the current pixel value is greater than or equal to the previous pixel value, otherwise bit "0" is assigned. Thus, the bit length of S1-Hash is an $(M_1 \times M_1 - 1)$ bits and the lengths of the

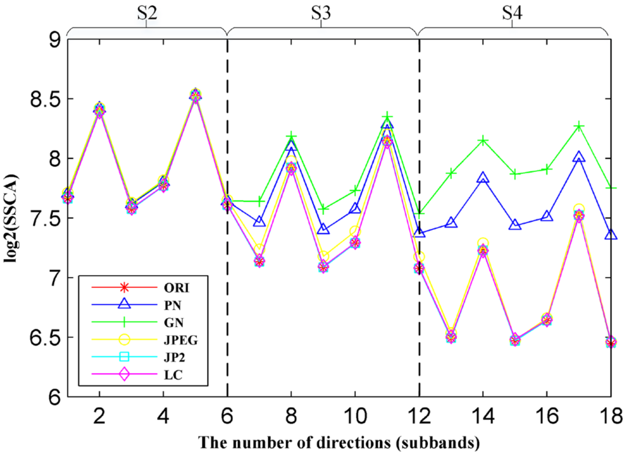


Figure 4. The SSCA property of Shearlet coefficients (S2 to S4) between five attacked versions. ORI: original image; LC: luminance changing; JPEG: JPEG compression; JP2: JPEG-2000 compression; PN: salt and pepper noise; GN: Gaussian noise.

Table 1. SSCA of S2.

	S2-D1	S2-D2	S2-D3	S2-D4	S2-D5	S2-D6
Original	7.66	8.40	7.57	7.77	8.52	7.61
Luminance Change	7.67	8.39	7.58	7.77	8.51	7.61
Salt and Pepper Noise	7.70	8.42	7.61	7.80	8.53	7.64
Gaussian Noise	7.71	8.39	7.63	7.81	8.51	7.64
JPEG	7.71	8.42	7.63	7.82	8.54	7.65
JP2	7.66	8.40	7.57	7.77	8.52	7.61

Table 2. SSCA of S3.

	S3-D1	S3-D2	S3-D3	S3-D4	S3-D5	S3-D6
Original	7.13	7.92	7.08	7.29	8.15	7.08
Luminance Change	7.14	7.91	7.09	7.29	8.14	7.08
Salt and Pepper Noise	7.46	8.10	7.40	7.57	8.28	7.37
Gaussian Noise	7.64	8.18	7.57	7.73	8.35	7.53
JPEG	7.24	7.98	7.18	7.39	8.20	7.18
JP2	7.13	7.92	7.08	7.29	8.15	7.08

Table 3. SSCA of S4.

	S4_D1	S4_D2	S4_D3	S4_D4	S4_D5	S4_D6
Original	6.49	7.23	6.47	6.64	7.52	6.44
Luminance Change	6.50	7.22	6.48	6.64	7.51	6.46
Salt and Pepper Noise	7.45	7.83	7.43	7.51	8.00	7.35
Gaussian Noise	7.87	8.15	7.87	7.91	8.27	7.75
JPEG	6.54	7.29	6.47	6.66	7.57	6.46
JP2	6.49	7.23	6.47	6.64	7.52	6.45

directional hashes (D1-hash to D6-Hash) are ($M_2 \times M_2 - 1$) bits.

In general, a binary hash can uniquely represent 2^L items, where L is the length of the hash. If the length of the hash is too short, the False Positive Rate (FPR) will be high. In order to select appropriate parameters for the proposed SBVF, we performed experiments on FPR with different hash lengths for S1-Hash and D1-Hash to D6-Hash as shown in **Figure 5**. In a parameter selection experiment, we found that the hash length of S1 should be longer than 31 bit for achieving relatively low FPR, while the minimum hash length for directional S2 images is 7 bit. Based on this finding, M = 128 as the input image block size, M_1 = 7 as the S1 down-sampled block size and M_2 = 3 as the down-sampled directional S2 image block size are chosen for generating the proposed SBVF. Thus, the S1-hash is 48 bits and the six directional hashes are 8 bits long, with total bit length of 96 bits.

3. TIRI-SBVF Based Content-Based Copy Detection System

To evaluate the performance of the proposed SBVF, a CBCD system using TIRI [10] [11] based video fingerprints is constructed with a generic structure as shown in **Figure 6**. The system is composed of two processes for fingerprint database generation and query video searching. Basically, the fingerprint database of the system is created off-line from the reference videos, while the query video's fingerprint is extracted on-line and used to search for the closest fingerprint in the database. In addition, the well-known Invert Index File (IIF) [25]-[28] based Hash searching strategy is adopted in this system to identify the best-matched video.

In practice, the input videos are usually in different frame sizes and frame rate. Before fingerprint generation, therefore, the input video has to be pre-processed such that copies of the same video with different frame sizes and frame rates are converted to a pre-defined format. The pre-processing steps of the experimental CBCD system are shown in **Figure 7**, in which each video frame is first converted to grayscale frame and then filtered by Gaussian smoothing filter in both time and space domains for prevent aliasing. After that the input video is down-sampled to a predefined frame size of $W \times H$ pixels and frame rate (FR). Based on the experimental settings in [10] [11] and Shearlet transform property, we selected W = 128, H = 128, and FR = 4 frames per second for the proposed TIRI-SBVF based CBCD system.

In addition, TIRI is also adopted in the pre-processing for achieving spatiotemporal nature of the finally generated fingerprints. Therefore, the pre-processed video frames are also divided into segments with J frames per segment. The TIRI are generated by calculating a weighted average of these J frames. Basically, the TIRI is a

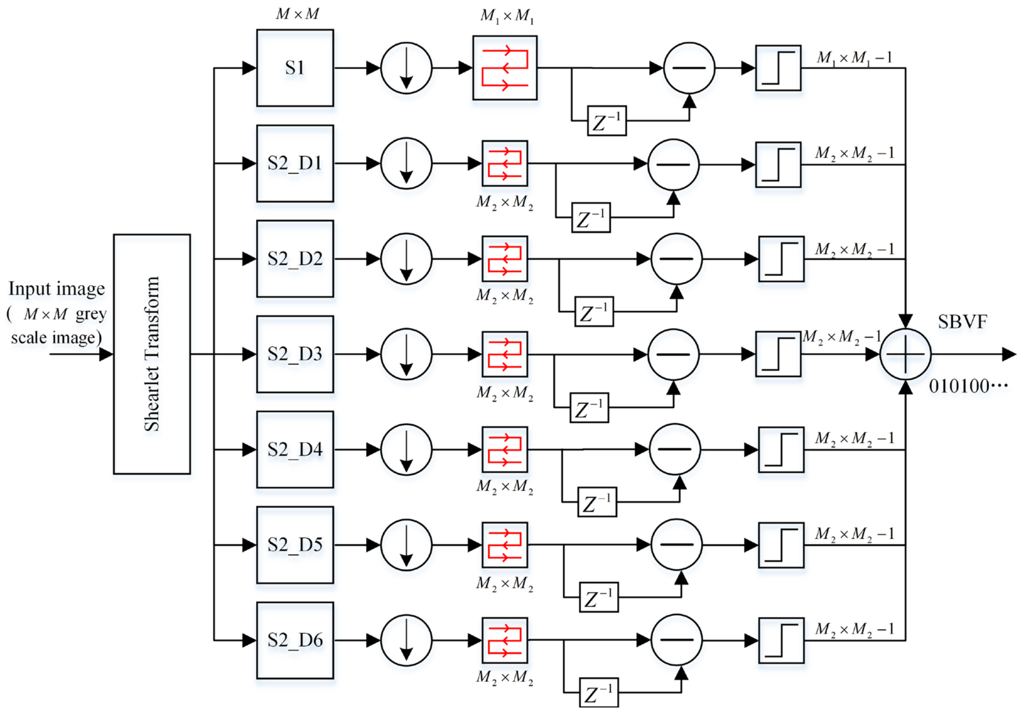


Figure 5. The proposed video fingerprint algorithm (SBVF).

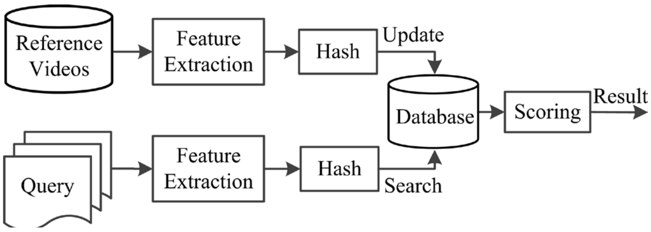


Figure 6. Overview of the general CBCD system.

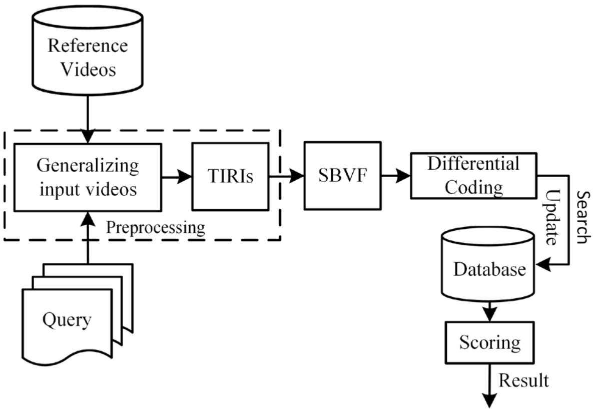


Figure 7. Overview of the CBCD system based on TIRI-SBVF.

blurred image that contains information about possible existing motions in a video sequence. The generation process of a TIRI can be defined as

$$p'_{(x,y)} = \sum_{i=1}^{J} \omega_i p_{(x,y,i)} \tag{5}$$

where $p_{(x,y,i)}$ is the luminance value of (x, y) th pixel of i-th frame in a set of J frames. The ω_i is the weight factor that could be constant, linear, or exponential. Experimental results have shown that the exponential weights has good ability in capturing the motion information of videos. Therefore, the exponential weights $\omega_i = \beta^i, \beta = 0.65$ are adopted in the experimental CBCD system. After preprocessing, the video fingerprints are extracted by applying the proposed SBVF to TIRIs during the query video searching process and fingerprint database generation.

The main challenge of CBCD system is to determine if a query video is a pirated version of a video in the database. However, the pirated videos are always not exactly same as the original video in the database. Then, fingerprints of two different copies of the same video content are similar but not necessarily identical. Thus, it is preferable to use searching methods that are looking for a close match of the query instead of an exact match. This problem is a nearest neighbor problem in the binary space [29], which can be simply handled by an exhaustive search but the computational requirement is extremely high. For a practical CBCD system, the size of the video database is huge with tens of millions of videos; fast approximate search algorithms are commonly used. In addition, the inverted index file (IIF) [25]-[28] has been proved as an efficient and fast searching strategy, especially using Hash data structure. The main idea of IIF is using the consumption of space to get the efficiency of time. Thus, the IIF searching strategy is adopted to build the TIRI-SBVF system and the IIF-based SBVF database structure is illustrated in **Figure 8**.

As shown in **Figure 8**, the reference database has been pre-organized using a look up table (LUT) [28] based on hash table, in which an exact partial fingerprint is a hash key (such as 0X0000). While the matched reference video ID and the position of this exact partial fingerprint is a hash value of <IDi, position>. The hash key (an exact partial fingerprint) is pointed to the reference video and the position of this fingerprint. Moreover, the hash value is stored in a linked list since this fingerprint may occur at multiple positions in different videos. The details of hash value are shown as [<ID1, position>, <IDm, position>] in **Figure 8**. The efficiency and searching

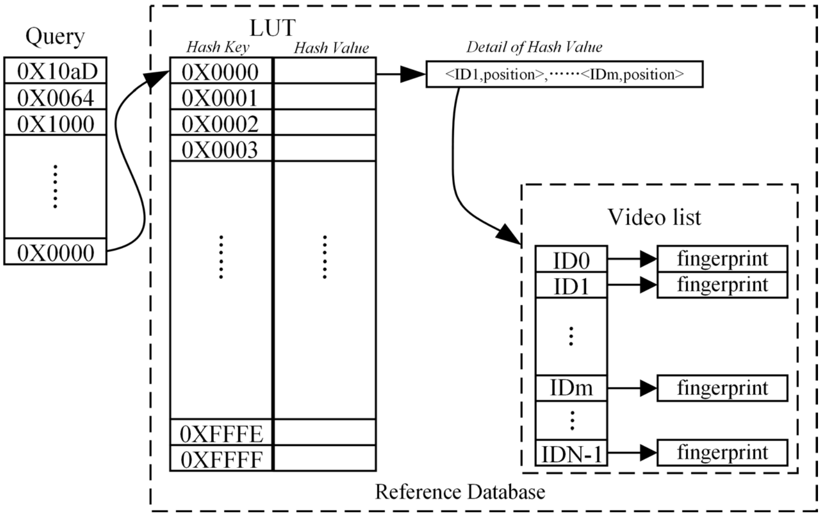


Figure 8. Inverted index file based searching strategy.

speed of IIF actually are guaranteed by the LUT as all the fingerprints with different possibility are pre-ordered in the LUT with the reference video ID and position. Therefore, this method can reduce the searching complexity from $O(n)$ of exhaustive searching to $O(1)$ of IIF. The restoring ability (R) of reference database is defined by the length of one fingerprint (L) with $R = 2^L$. In the experimental CBCD system, L is set to 16 bits with practical considerations in the tradeoff of the searching speed and memory requirement.

4. Experimental Results

4.1. Statistical Evaluation of the Proposed Shearlet Based Video Fingerprint (SBVF)

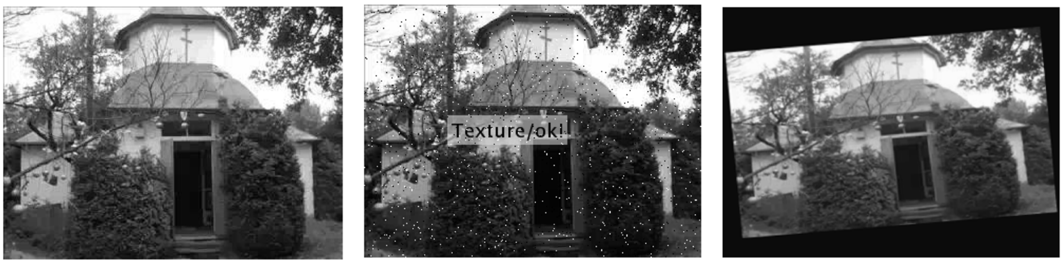
In general, a good video fingerprint should be robust for perceptual similar video segments under different types of distortions but discriminative for the different videos. The normalized Hamming distance (NHD) is a well-known metric to measure the similarity between different fingerprints, which is equal to the different bit counts between two fingerprints with normalization of length. Thus, NHD is adopted to evaluate the robustness of the proposed SBVF on individual image or video frame from TRECVID 2010 [22] [23] and INRIA Copy Day [24] datasets. The evaluation dataset is created by randomly select 3 frames from 122 videos with total of 366 frames from TRECVID 2010 datasets and 143 images from INRIA Copy Day dataset.

To test the robustness, common types of distortion are applied to these selected frames such as geometrical distortions of letter box and rotation. For luminance distortions, luminance change, salt and pepper noise, Gaussian noise, text insertion, and JPEG compression are used. The details of these types of distortion are listed in **Table 4**. To achieve a comprehensive evaluation, some distortions are combined to create more challenging attacks. The combined-1 distortion emphases on luminance attacks, which combine the distortions of luminance change, salt and pepper noise, Gaussian noise, JPEG compression and text insertion. While the combined-2 distortion emphases on geometrical attacks, which combine the distortions of letter box and rotation. **Figure 9** shows an example of a video frame with these two types of combined distortions. With these 9 types of distortion, there are total 509 original images and 4581 distorted images as testing images.

For comparison, the well-known 2D-DCT [11] and 2D-DCT-2AC [10] fingerprinting algorithms are used in the experiments. The 2D-DCT [11] fingerprint is widely used as a perceptual hash for image searching, which is based on applying 2-dimensional DCT transform to down-sampled grayscale image and then the 8×8 low frequency DCT coefficients with median value as threshold are used to generate 64 bits fingerprint. To tackle the problem of the various dynamic ranges in different DCT coefficients, only two lowest frequency AC DCT coefficients (first horizontal and vertical AC coefficients) with similar dynamic ranges are used in [10] to achieve more robust fingerprint. In which overlapping 32×32 blocks with 50% overlapping are used to generate 96-bit 2D-DCT-2AC fingerprints. In addition, the OIS [6] is a traditional video fingerprint for CBCD applications, which is derived by dividing each frame into a grid and sorted it into an ordinal intensity signature. As these three fingerprinting algorithms are widely used in the implementation of CBCD system, they are used to

Table 4. The detail on different types of distortion.

Distortions	Description
Luminance Change	Illumination = −25%
Salt and Pepper Noise	Density = 0.02
Gaussian Noise	Mean = 0.01; Variance = 0.001
Text Insertion	Font size = 8
JPEG	Quality = −15%
Combined-1	Illumination = −25%; Density = 0.001; Font size = 8; Quality = 5%
Letter Box	Height = 80%
Rotation	Degree = 5
Combined-2	Height = 80%; Degree = 5



(a) (b) (c)

Figure 9. (a) Original video frame, (b) video frame with combined-1 distortion, and (c) video frame with combined-2 distortion.

compare with the proposed SBVF in the robustness evaluation.

Based on NHD as similarity measure with use of different threshold values in matching, two commonly used metrics of True Positive Rate (TPR) and False Positive Rate (FPR) are adopted in the evaluation. They are defined as:

$$\text{TPR} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{6}$$

$$\text{FPR} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}} \tag{7}$$

In this experiment, 300 images are used for evaluation, which are selected from Copy Days dataset [24]. The similarity between each frame and its distorted version is calculated with use of NHD. In order to illustrate the statistical distribution of similarity of the testing frames in the dataset, the TPR and FPR are calculated to demonstrate the performance. As shown in **Table 5**, two threshold values of NHD are used to define the matching, which means that images are considered perceptual similar if the NHD is smaller than the NHD threshold (thr1 = 0.1 and thr 2 = 0.2). These two threshold values of 0.1 and 0.2 are commonly used in the multimedia copy detection system [4] [11] [12].

As shown in **Table 5** with threshold = 0.1, the proposed SBVF is robust against most of luminance distortions, such as luminance change, salt and pepper noise and Gaussian noise. However, all four evaluated algorithms cannot perform well for letter box and rotation types of distortion using 0.1 threshold. The main reason is that the threshold of 0.1 is too strict for similarity comparison using NHD. For a practical system, the threshold of 0.2 can achieve higher TPR performance as shown in **Table 6**. It is because all the four algorithms are improved especially on geometric types of distortion. Moreover the proposed SBVF is outstanding in terms of TPR performance on most types of distortion such as luminance change, salt and pepper noise, JPEG, combined-1, letter box, rotation, and combined-2. On the other hand, the FPR performance (discrimination) is also a key property of fingerprinting algorithm, a good fingerprint should also make sure of the low FPR characteristics. From the

Table 5. The TPR and FPR performance using threshold value of 0.1 for different fingerprinting algorithms.

(Thr1 = 0.1)	TPR				FPR			
	2D-DCT [11]	2D-DCT-2AC [10]	OIS [6]	SBVF	2D-DCT [11]	2D-DCT-2AC [10]	OIS [6]	SBVF
Luminance Change	1.00	0.98	1.00	1.00	0.00	0.00	0.00	0.00
Salt and Pepper Noise	0.99	0.98	0.92	0.97	0.00	0.00	0.00	0.00
Gaussian Noise	0.98	0.96	0.88	0.86	0.00	0.00	0.00	0.00
Text Insertion	0.59	0.47	0.68	0.76	0.00	0.00	0.00	0.00
JPEG	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00
Combined-1	0.55	0.32	0.49	0.72	0.00	0.00	0.00	0.00
Letter Box	0.08	0.06	0.11	0.53	0.00	0.00	0.00	0.00
Rotation	0.01	0.03	0.02	0.03	0.00	0.00	0.00	0.00
Combined-2	0.01	0.01	0.01	0.03	0.00	0.00	0.00	0.00

Table 6. The TPR and FPR performance using threshold value of 0.2 for different fingerprinting algorithms.

(Thr = 0.2)	TPR				FPR			
	2D-DCT [11]	2D-DCT-2AC [10]	OIS [6]	SBVF	2D-DCT [11]	2D-DCT-2AC [10]	OIS [6]	SBVF
Luminance Change	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00
Salt and Pepper Noise	1.00	0.99	0.98	1.00	0.00	0.00	0.00	0.00
Gaussian Noise	1.00	0.97	0.97	0.99	0.00	0.00	0.00	0.00
Text Insertion	0.97	1.00	1.00	0.99	0.01	0.00	0.01	0.00
JPEG	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00
Combined-1	0.96	0.94	0.95	0.97	0.00	0.00	0.01	0.00
Letter Box	0.85	0.67	0.65	1.00	0.00	0.00	0.00	0.00
Rotation	0.47	0.35	0.68	0.70	0.01	0.00	0.02	0.00
Combined-2	0.11	0.20	0.26	0.54	0.00	0.01	0.02	0.00

FPR results as shown in **Table 5** and **Table 6**, most of the testing algorithms including the proposed SBVF can achieve low FPR with good discrimination property. These experimental results demonstrate that the proposed SBVF can achieve high robustness with better performance than the three compared fingerprint algorithms.

4.2. TIRI Based CBCD Systems Evaluation

In this section, the performance of the proposed TIRI-SBVF video copy detection algorithm as described in section 3 is evaluated using the TRECVID 2010 dataset [22] [23]. For performance comparison, we also implemented the three well-known video fingerprints as used in section 4.1 for the TIRI based system evaluation. These systems are named as TIRI-2D-DCT [11], TIRI-2D-DCT-2AC [10] and TIRI-OSI [6]. Moreover, the commonly used preprocessing procedure as shown in **Figure 7** is adopted in these systems implementation, which includes generalizing down-sampled grayscale input video and TIRIs generation. In order to have a fair evaluation, the parameters selection is according to systems in [10] [11]. In which 128×128 down-sampled fame size and frame rate (FR) of 4 frames per second are used to generate the pre-processed input video, and $\beta = 0.65$ and $J = 4$ are used to generate the TIRIs.

In this experiment, we selected 122 videos from TRECVID 2010 dataset as reference videos, which are used to generate query with copy issue. And then, another 122 videos as non-reference are used to form the query with no copy issue. In the evaluation, the queries are randomly extracted from each reference and non-reference video sets with 15 seconds length. In addition, eight types of attacks are imposed to these queries to generate a query pool with 1952, in which 976 copied queries and 976 no copied queries. The eight types of distortion are geometrical attacks including letter box and rotation, luminance attacks including luminance change, salt and pepper noise, Gaussian noise, text insertion, and temporal attacks including dropping frames, time shifting. Before searching database, the fingerprint reference database has pre-generated using the same fingerprint algorithm. The threshold value of 0.2 is used for NHD based similarity matching, which is commonly used in most of the CBCD system implementation.

For a robust CBCD system, it should be achieve the balance between precision (discrimination) and recall (robustness). To evaluate the performance of the proposed TIRI-SBVF, TIRI-2D-DCT [11], TIRI-2D-DCT-2AC [10] and TIRI-OIS [6], we adopt the F-score (F_λ) as a combined metric. The F_λ is defined as:

$$F_\lambda = \left(1+\lambda^2\right)\frac{\text{precision} * \text{recall}}{\lambda^2 \text{precision} + \text{recall}} \tag{8}$$

where λ is a weight of combination between precision and recall. In this paper, the balanced F-score ($\lambda = 1$) is used as F_λ score can capture the precision and recall property more generally.

Detection and localization of the copied video segment are two main tasks of CBCD system. The purpose of detection is to detect any copied segments in the reference video, while the purpose of localization is to locate the copied segment in the matched video. The performance of detection is shown in **Table 7**, which shows that the proposed TIRI-SBVF can achieve outstanding average F1 performance of about 0.99 in luminance and temporal types of attacks. Moreover, the propose TIRI-SBVF is always better than compared methods. The proposed TIRI-SBVF can achieve specially a good performance in the geometry distortions such as letter box. Additionally, the FPR of proposed SBVF is much lower than 0.01%. It is worth to mention that the rotate and letter box attack is the common challenge to all compared video fingerprints, in the practical CBCD system, some preprocessing method like the algorithm of letter box removing which can be used to overcome this challenge. However, the proposed SBVF has best performance in these two challenging distortions especially in letter box distortion with 0.95 of F1-Score.

The accuracy of localization of copied video segment is usually correlated with the performance of detection, and it is defined as how many queries are correctly localized in the queries, which are detected with copy issue. The accuracy of localization is shown in **Table 8**. It is turned out that the task of localization could be well handled by most of algorithms, the proposed TIRI-SBVF also has very good property of localization shown as **Table 8**, it can achieve the average accuracy of about 97%.

5. Conclusions

In this paper, a novel Shearlet based video fingerprint (SBVF) with use of Temporal Informative Reprehensive Images as a global spatiotemporal video fingerprint is proposed for content-based video copy detection applications. The SBVF design is motivated by the multi-scale and multi-directional decomposition characteristics of Shearlet transform. With high robustness on different types of distortion for S1 and S2 with six directional subbands of 4-Scale Shearlet transform, the SBVF is constructed by 1-bit differential coding on down-sampled images of these Shearlet images. In statistical evaluation based on normalized hamming distance, the proposed

Table 7. TPR, FPR and F1 score of detection performance.

Detection Performance	TPR				FPR				F_1-Score			
	TIRI-2D-DCT [11]	TIRI-DCT-2AC [10]	TIRI-OIS [6]	TIRI-SBVF	TIRI-2D-DCT [11]	TIRI-2D-DCT-2AC [10]	TIRI-OIS [6]	TIRI-SBVF	TIRI-2D-DCT [11]	TIRI-2D-DCT-2AC [10]	TIRI-OIS [6]	TIRI-SBVF
Luminance Change	0.97	0.98	1.00	0.99	0.00	0.00	0.09	0.00	0.98	0.98	1.00	1.00
Text Insertion	0.98	0.83	0.99	0.93	0.10	0.01	0.07	0.00	0.94	0.91	1.00	0.96
Gaussian Noise	0.93	0.96	0.93	0.98	0.00	0.00	0.05	0.00	0.96	0.98	0.96	0.99
Salt and Pepper Noise	0.93	0.96	0.93	0.98	0.00	0.00	0.07	0.00	0.97	0.98	0.97	0.99
LetterBox	0.38	0.48	0.56	0.92	0.02	0.00	0.10	0.00	0.54	0.65	0.65	0.95
Rotation	0.16	0.12	0.39	0.42	0.02	0.01	0.30	0.01	0.27	0.22	0.55	0.59
TimeShift	0.93	0.96	1.00	0.98	0.00	0.00	0.07	0.00	0.96	0.98	1.00	0.99
DropFrame	0.93	0.97	0.99	1.00	0.02	0.00	0.09	0.00	0.97	0.98	0.99	1.00

Table 8. The accuracy of localization.

Localization (Accuracy)	TIRI-2D-DCT [11]	TIRI-2D-DCT-2AC [10]	TIRI-OIS [6]	TIRI-SBVF
LuminanceChange	0.97	0.98	0.96	0.98
Text Insertion	0.95	0.97	0.97	0.98
GaussianNoise	0.98	0.98	0.98	0.99
Salt and PepperNoise	0.98	0.99	0.98	0.98
LetterBox	0.93	0.97	0.94	0.97
Rotation	0.87	0.90	0.87	0.96
TimeShift	0.95	0.96	0.95	0.95
DropFrame	0.98	0.98	0.97	0.98
Average	0.95	0.97	0.95	0.97

SBVF can achieve relatively high robustness as compared three well-known fingerprinting algorithms of 2D-DCT, 2D-DCT-2AC and OIS. In addition, TIRI-SBVF based CBCD system is constructed to evaluate the video detection and localization performance with comparison with TIRI-2D-DCT, TIRI-2D-DCT-2AC and TIRI-OSI based copy detection systems.

With use of the well-known TRECVID 2010 dataset, experimental results show that proposed TIRI-SBVF is a robust video fingerprint with strong ability of discrimination and robustness for variety of video copy attack. The proposed TIRI-SBVF can achieve average F1 performance of about 0.99, and very low false positive rate (FPR) of less than 0.01%, and the average localization accuracy of about 97%.

References

[1] YouTube Statistics, YouTube. https://www.youtube.com/yt/press/en-GB/statistics.html

[2] Lu, J. (2009)Video Fingerprinting for Copy Identification: From Research to Industry Applications. *Proceedings of SPIE*, *Media Forensics and Security*. http://dx.doi.org/10.1117/12.805709

[3] Hampapur, A. and Bolle, R.M. (2001) Comparison of Distance Measures for Video Copy Detection. *ICME* 2001, *IEEE International Conference on Multimedia and Expo* 2001, Tokyo, 22-25 August 2001, 737-740. http://dx.doi.org/10.1109/ICME.2001.1237827

[4] Hampapur, A., Hyun, K. and Bolle, R.M. (2002) Comparison of Sequence Matching Techniques for Video Copy Detection, *Proceeding of SPIE* 4676, *Storage and Retrieval for Media Databased*, **4676**, 194-201. http://dx.doi.org/10.1117/12.451091

[5] Chen, L. and Stentiford, F. (2008) Video Sequence Matching Based on Temporal Ordinal Measurement. *Pattern Recognition Letters*, **29**, 1824-1831. http://dx.doi.org/10.1016/j.patrec.2008.05.015

[6] Li, T., Nian, F., Wu, X., Gao, Q. and Lu, Y. (2014) Efficient Video Copy Detection Using Multi-modality and Dynamic Path Search. *Multimedia System*, **22**, 29-39. http://dx.doi.org/10.1007/s00530-014-0387-8

[7] Radhakrishnan, R. and Bauer, C. (2007) Content-Based Video Signatures Based on Projections of Difference Images, *IEEE* 9*th Workshop on Multimedia Signal Processing*, Crete, 1-3 October 2007, 341-344. http://dx.doi.org/10.1109/MMSP.2007.4412886

[8] De Roover, C., De Vleeschouwer, C. and Macq, B. (2005) Robust Video Hashing Based on Radial Projections of Key Frames. *IEEE Transactions on Signal Processing*, **53**, 4020-4037. http://dx.doi.org/10.1109/TSP.2005.855414

[9] Kim, C. and Vasudev, B. (2005) Spatiotemporal Sequence Matching for Efficient Video Copy Detection, *IEEE Transactions on Circuits and Systems for Video Technology*, **15**, 127-132. http://dx.doi.org/10.1109/TCSVT.2004.836751

[10] Esmaeili, M.M., Fatourechi, M., and Ward, R.K. (2011) A Robust and Fast Video Copy Detection System Using Content-Based Fingerprinting, *IEEE Transactions on Information Forensics and Security*, **6**, 213-226. http://dx.doi.org/10.1109/TIFS.2010.2097593

[11] Esmaeili, M.M., Fatourechi, M. and Ward, R.K. (2009) Video Copy Detection Using Temporally Informative Representative Images. *International Conference on Machine Learning and Applications* (*ICMLA*), Miami Beach, December 2009, 69-74. http://dx.doi.org/10.1109/ICMLA.2009.32

[12] Li, Y., Po, L.M., Xu, X., Feng, L. and Yuan, F. (2015) No-Reference Image Quality Assessment with Shearlet Transform and Deep Neural Networks. *Neurocomputing*, **154**, 94-109. http://dx.doi.org/10.1016/j.neucom.2014.12.015

[13] Li, Y., Po, L.M., Cheung, C.H., Xu, X., Feng, L. and Yuan, F. (2015) No-Reference Video Quality Assessment with 3D Shearlet Transform and Convolutional Neural Networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 1. http://dx.doi.org/10.1109/TCSVT.2015.2430711

[14] Yi, S., Labate, D., Easley, G.R. and Krim, H. (2009) A Shearlet Approach to Edge Analysis and Detection. *IEEE Transactions on Image Processing*, **18**, 929-941. http://dx.doi.org/10.1109/TIP.2009.2013082

[15] Kutyniok, G. and Lim, W. (2010) Image Separation Using Wavelets and Shearlets. *Curves and Surfaces*, **6920**, 416-430. http://dx.doi.org/10.1007/978-3-642-27413-8_26

[16] Kutyniok, G., Shahram, M. and Zhuang, X. (2011) ShearLab: A Rational Design of A Digital Parabolic Scaling Algorithm. arXiv: 1106.1319v1. http://arxiv.org/abs/1106.1319v1

[17] Kutyniok, G., Lim, W. and Zhuang, X. (2012) Digital Shearlet Transforms. In: Kutyniok, G. and Labate, D., Eds., *Shearlets*, Birkhäuser, Boston, 239-282.

[18] Lim, W.-Q. (2010) The Discrete Shearlet Transform: A New Directional Transform and Compactly Supported Shearlet Frames. *IEEE Transactions on Image Processing*, **19**, 1166-1180. http://dx.doi.org/10.1109/TIP.2010.2041410

[19] ShearLab. http://www.shearlet.org/

[20] Zhou, J.P., Cunha, A.L. and Do, M.N. (2005) Nonsubsampled Contourlet Transform Construction and Application Enhancement. *IEEE International Conference on Image Processing*, **1**, 469-472.

[21] Easley, G., Labate, D. and Lim, W.-Q. (2008) Sparse Directional Image Representations Using the Discrete Shearlet Transform. *Applied and Computational Harmonic Analysis*, **25**, 25-46. http://dx.doi.org/10.1016/j.acha.2007.09.003

[22] Gupta, V., Boulianne, G. and Cardinal, P. (2012) CRIM's Content-Based Audio Copy Detection System for TRECVID 2009. *Multimedia Tools and Applications*, **60**, 371-387. http://dx.doi.org/10.1007/s11042-010-0608-x

[23] Smeaton, A.F., Kraaij, W. and Over, P. (2004) The TREC Video Retrieval Evaluation (TRECVID): A Case Study and Status Report. *RIAO* 2004, *International Conference of Computer-Assisted Information Retrieval*, Avignon, 26-28 April 2004, 25-37.

[24] Jegou, H., Douze, M. and Schmid, C. (2008) Hamming Embedding and Weak Geometry Consistency for Large Scale Image Search. *The* 10*th European Conference on Computer Vision*, Marseille, 12-18 October 2008, 304-317.

[25] Sivic, J. and Zisserman, A. (2003) Video Google: A Text Retrieval Approach to Object Matching in Videos. *The* 9*th IEEE International Conference on Computer Vision*, Nice, 13-16 October 2003, 1470-1477.

[26] Douze, M., Jegou, H. and Schmid, C. (2010) An Image-Based Approach to Video Copy Detection with Spatio-Temporal Post-Filtering. *IEEE Transactions on Multimedia*, **12**, 257-266. http://dx.doi.org/10.1109/TMM.2010.2046265

[27] Oostveen, J., Kalker, T. and Haitsma, J. (2002) Feature Extraction and a Database Strategy for Video Fingerprinting. 5*th International Conference*, *VISUAL* 2002, Hsin Chu, 11-13 March 2002, 117-128.

[28] Haitsma, J. and Kalker, T. (2003) A Highly Robust Audio Fingerprinting System with an Efficient Search Strategy. *Journal of New Music Research*, **32**, 211-222. http://dx.doi.org/10.1076/jnmr.32.2.211.16746

[29] Law-To, J., Chen, L., Alexis, J., Ivan, L., Olivier, B., Valerie, G.B., Nozha, B. and Fred, S. (2007) Video Copy Detection: A Comparative Study. *Proceedings of the* 6*th ACM International Conference on Image and Video Retrieval*, Amsterdam, 9-11 July 2007, 371-378. http://dx.doi.org/10.1145/1282280.1282336