International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 1 Issue 3, November 2013

A Compact Spatio-Temporal Fingerprint for Video Copy Detection System

Divya Devan[1], Gopu Darsan[2]

[1, 2]Department of Computer Science & Engineering, Sree Buddha College of Engineering
Kerala University, Kerala, India

Abstract: ***Video copy detection system is used for video indexing and copy right applications. The fingerprint extraction algorithm extracts unique features of the video based on the content of the video. The key frames are extracted from the video based on the Structural Similarity Measure (SSIM). The fingerprint can be extracted from specially constructed image which is known as Temporally Informative representative Image (TIRI). The TIRI image contains both spatial and temporal information about the video. Then the TIRI-DCT algorithm extracts the fingerprints from the TIRIs. The search algorithm such as header based similarity search method searches close enough match for fingerprint of query video and fingerprint within database. The extracted fingerprint should be robust to frame rate conversion, background and intensity changes and geometric attacks such as rotation, scaling etc. The system focuses on determining pirated version of video from the video database. It consists of three stages. In key frame extraction the SSIM is used for measuring the similarity between the frames. In this way avoid more similar frames and preserve those are distinct or key frames. These key frames are used for generating TIRI images. The weighted average of all the selected frames is taken as TIRI. Then Discrete Cosine Transform (DCT) algorithm is applied to extract the features, because it can precisely capture local variations that exist in the TIRI image. From that the binary feature data or binary fingerprint is generate and stored in the fingerprint database. For fingerprint searching within the database a header based similarity search method is proposed. The system has been tested for around 50 videos. Among these test result gave a true positive rate of 97 percent. This result shows that the proposed copy detection system is more robust and discriminant.***

Keywords: spatiotemporal fingerprint, Structural Similarity Measure, Temporally Informative Representative Images

1. Introduction

Video is the technology of electronically capturing, recording, processing, storing, broadcasting, and reproducing a sequence of still images representing scenes in motion. The Latest advancements in computer technology allowed computers to capture, store, edit and transmit video clips. Frame rate is the number of still pictures per unit of time of video clip. The frame rate ranges from six or eight frames per second (frame/s) for old mechanical cameras but the frame rate for new professional cameras are 120 or more frames per second. The minimum frame rate to achieve the illusion of a moving image is about fifteen frames per second. Aspect ratio describes the proportional relationship between dimensions of video screens and video picture elements. Aspect ratio of videos is commonly described by a ratio between width and height.

Video copy detection is the process of detecting illegally copied or manipulated version of videos by analyzing them and comparing them to original content. The goal of this process is to protect a video creator's intellectual property. Now a day a large number of videos are uploaded to the internet and shared every day. Most of them are illegal copies or manipulated versions of existing media. With the tremendous development of technology, internet has become a fast media to propagate information and a warehouse to access the relevant information. Everyday there are more videos on the internet. This makes Copy right management on the internet is a complicated process. Content-preserving attacks (distortions) are changes that are made to the video unintentionally or intentionally by the users of video-sharing websites. These changes can include signal processing operations, format changes, changes in brightness/contrast, rotation, cropping, added noise, logo insertion, compression, etc. So nowadays video copy detection is widely use and is very essential.

There are two main approaches for identify the distortions of video, one is the well-known Watermarking, and another one is video fingerprinting. Watermarking is done by inserting a distinct pattern or emblem into the video stream, while fingerprinting techniques match content-based signatures to detect copies of video. The bottleneck of watermarking is that the inserted marks are likely to be destroyed or distorted as the format of the video get transformed or by the transmission. But the video signature extraction for the content-based copy detection can be done after the media has been distributed. So video fingerprinting attracts more and more attention recently.

Video Fingerprints are feature vector that uniquely characterize one video clip from another and are used for video identification. A spatial fingerprint characterizes spatial feature of a video frame and is computed independent of other frames. These approaches are based on intensity statistics such as mean, median, variance, centroid and other higher order moments of spatial content of the different frames of the video. The frames are divided into different sub section and within each section features are calculated. It is less robust to geometrical operations such as rotation and scaling. Temporal fingerprints are extracted the characteristics of a video sequence over time. These features work well with long video sequence, but do not perform well with short video clips. Thus, spatio-temporal fingerprints are more efficient. So we propose this method in our proposed method.

Image feature extraction is a method of capturing visual content of images for indexing and retrieval. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. It can be used in the area of image processing which involves using algorithms to detect and isolate various features of a digitized image. When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant (e.g. the same measurement in both feet and meters) then the input data will be transformed into a reduced representation set of features (features vector). The process of transforming the input data into the set of features or feature vector is called feature extraction.

These features are often corners or edges in the image but should be invariant to scale and stable, meaning that these features should be consistently detected even in a different setting. Because the Harris corner detector is not scale invariant, it was not a good candidate for feature detection. Rather, Scale Invarient Feature Transform (SIFT) and Speeded Up Robust Features (SURF), which are scale-invariant, were decided on as the features of interest. The SIFT and SURF algorithms detect feature in different ways. DCT is a powerful transform to extract proper features. After applying DCT to the entire images, some of the coefficients are selected to construct feature vectors.

SIFT and SURF algorithms detect feature in different ways. DCT is a powerful transform to extract proper features. After applying DCT to the entire images, some of the coefficients are selected to construct feature vectors. So TIRI-DCT algorithm introduced to extract spatio-tempoaral features from the specially constructed images. DCT have the properties of decorrelation, energy compaction, orthogonality and seperability. So it is superior than other methods. The TIRI image is constructed from only three images [1]. Here choose only first, middle and the last frames from the video clip. All other frames or information get lost. So to reduce the necessary information loss a key frame extraction method is proposed. The key frames are extracted based on the similarity between consecutive frames. Based on the similarity the frames are divided into shots, group scenes and so on.

For fingerprint matching inverted file based similarity search method is used in the existing method [1]. This search method is based on the idea that for two fingerprints which are similar enough to be considered as matches. In which divide each fingerprint into small non overlapping blocks of m bits. These are called words. Words are then used to create an inverted file from the fingerprints of database. It is represented as table. The horizontal dimension of the table shows the position of a word inside a fingerprint, and the vertical direction corresponds to possible values of the word. To find a query fingerprint in the database, first the fingerprint is divided into n words (of m bits). The query is then compared to all the fingerprints that start with the same word in the database. The fingerprints with same starting word are found from the corresponding entry in the first column of the inverted file table. The Hamming distance between these fingerprints and the query is then calculated. If a fingerprint has a Hamming distance of less than some predefined threshold, it will be announced as the match or a copy. If no such match is found, the procedure is repeated. It is not practical for word lengths of larger than 16 bits to have a complete inverted file and only the existing words should be indexed.

In order to reduce the complexity here proposed a header based similarity search method for that a centroid is chosen from each fingerprint known as header. Each query fingerprint compare with the dummy fingerprint. Dummy fingerprints are obtained based on the header of the fingerprint stored in the fingerprint database. Then compute the hamming distance and found the fingerprint with minimum hamming distance. And actual comparison is done with that fingerprint. This method reduces the search time.

2. Related Work

Changick Kim and Bhaskaran Vasudev [1] proposed a novel sequence matching technique to detect copies of a video.. It effectively deal with format conversion and from letter-box, pillar box and other style. The proposed sequence matching is computationally very efficient and very faster. Another advantage of the proposed method is the low memory requirement since for storing/indexing signatures only 4 bytes/frame is needed. Regunathan Radhakrishnan [2] proposes a novel video signature extraction method based on projection of difference image between consecutive video frames. The difference image can be obtained by computing the absolute values of intensity difference between adjacent video frames. . The signature would not perform as well for frame-rate conversion attacks. They found that the signature is robust to +/-5% change in frame rate but for lower frame rates the difference image between consecutive video frames would not be the same as it was in the original video (especially so for high motion sequences).Alexis Joly, [3] propose a distortion-based probabilistic approximate similarity search technique that is not based on features distribution in the database but rather on the distribution of the feature distortions. One of the major issue in this method is local feature redundancy and the estimation of transformation that occurred between two document. Sunil Lee [4] mentions that the video fingerprints are used to identify a given video query in a database (DB) by measuring the distance between the query fingerprint and the fingerprints in the DB. The main advantage of the proposed fingerprint is robust against minor geometric transformation such as frame rotation up to 1 degree and frame cropping. The major drawback of the video fingerprinting system is not secure against general geometric transformations such as rotation, shift, cropping etc. Geert Willems [5] present a new method for content based video copy detection based on local spatiotemporal features. The use of spatiotemporal features guarantees a robust localization over time. One of the major drawbak is the adaptation of the spatio-temporal detector to streaming video and the application of the features on more and larger datasets are not checked. Matthijs Douze [6], proposes a new video copy detection system which efficiently matches individual frames and then verifies their spatio-temporal consistency. . The proposed system is more

International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 1 Issue 3, November 2013

scalable, still obtains excellent results with a memory footprint and query time reduced 20 times. And it has very accurate image level matching. Here the adaptation of the spatio-temporal detector to streaming video and the application of the features on more and larger datasets are not checked. Zheng Cao, and Ming Zhu **[7]** proposed a novel efficient video copy detection approach. One of the major advantage of this method is the copy detection can be finished in very short time through the volume of database is quite large. It meet the requirement of scalable computing performs well with satisfying recall and precision rate in high efficiency. It improves the search efficiency in large database.

3. Proposed System

Spatial fingerprints are features derived from each frame or from a key frame. The traditional key frame based representation has some drawbacks. One of them is the performance of key frame based shot representation depends on the accuracy of shot segmentation algorithm and the appropriate selection of key frame to characterize the video content. This is not robust to geometric attacks. Temporal fingerprints are extracted from the characteristics of video sequence over time. Which is work well with long video sequence, but do not perform well for short video clip**.** So combining these two such as spatiotemporal fingerprints provide a robust and descriminant features. But the major issue is how these features extract without any loss. And how can reduce the storage space of the database. Which means to make the fingerprint more compact .In this paper a new compact and secure fingerprint extraction algorithm is proposed. The following section consists of proposed system which includes feature extraction, fingerprint generation and similarity search.

3.1 Key Frame Extraction

The reason for not get features from every frame is to make the process fast and fingerprints compact, we'll see that a technique which does complicated computation on a frame prefers to use key frames rather than all the video sequence for building the fingerprints. In boundary detection detecting the boundary of scenes. There are kinds of CBCD methods based on signatures of scenes, the reasons are that frames in a scene are similar or related, and a copy clip may just a scene cut. A scene can be seen as an event in the movie, while another term, shot, means frames coming from the same camera or the same angle, and it's hard for computer to distinguish these two situations. So that to identify the shots the similarity between adjacent frames are computed by using SSIM.SSIM is measure for find the similarity between images.

A great set of key frames should give clear description of a video, and same key frames are expected to be extracted from the original video and its copy. To solve the frame drop situation, we assume even the selected key frame get lost, frames of similar content around it are probably to be selected and maintain the detection accuracy; while techniques without key frame may hardly to solve this problem. So we introduced key frame extraction in our proposed method.

Initially the video is divided into different frames. Then the key frames are extracted from the input video clip by preserving the content semantics. So that compute the structural similarity between consecutive frames and if it less than some predefined threshold then it will be considered as a new shot. These shots are combined to form groups overall procedure is as follows:

- Input: Video shot sequence, S = fshot.
- Output: Video structure in terms of scene, group, and shot.

4. Procedure

1. Initialization: assign shot 0 to group 0 and scene 0; Initialize the group counter num Groups= 1; initialize the scene counter numScenes
2. If S is empty, quit; otherwise get the next shot.
3. Test if shot i can be merged to an existing group:
   (a) Compute the similarities between the current shot and existing groups
   (b) Find the maximum group similarity. Group Similarity is the similarity between (c) Test if this shot can be merged into an existing group:
   i . Merge shot i to group.
   ii .Update the video structure.
   iii. Go to Step 2.

otherwise:

   i. Create a new group containing a single shot i. Let this group be group j.
   ii.Set numGroups = numGroups + 1.

4. Test if shot i can be merged to an existing scene:
   (a) Calculate the similarities between the current shot i and existing scenes
   (b) Find the maximum scene similarity:

maxSceneSimi = maxs

where SceneSimi;s is the similarity between shot i and scene s. Let the otherwise:

   i. Create a new scene containing a single shot i and a single group j.
   ii. Set numScenes = numScenes + 1.}

The similarity is computed by using structural similarity between consecutive video frames. Many of the existing algorithms for similarity measures are based on spatial domain features and are fail in the presence of illumination variation or noise.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^{\,2} + \mu_y^{\,2} + C1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (3.1)$$

Where μ_x is the average of x , μy is the average of y σ_x the variance of x , σ_y is the variance of y, σx_y is the covariance of x and y .C1 and C2 are the constants for stabilize the division with weak denominator.

4.1 Pre-processing and TIRI Generation

Preprocessing can be carried out to make the video clip unique both in time and space. Copies of same video with

International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 1 Issue 3, November 2013

different frame sizes and frame rates are taken, so that to make the system robust to changes in frame size and frame rate smoothing and resizing can be performed. To increase the robustness, the video is down sampled in both time and space of width 144, height 176 at 4 frames per second. A Gaussian smoothing filter is applied to both time and space in order to avoid anti- aliasing. Video frames are divided into overlapping segments of fixed length, each containing 'J' frames. It generates a representative image by calculating the weighted average of the frames. Let $l_{m,n,,k}$ be the luminance value of the (m, n) th pixel of kth frame. The pixels of TIRI are then obtained as a weighted sum of the frames

$$l^{|}_{m,n} = \sum_{k=1}^{j} w_k l_{m,n,k} \quad (3.2)$$

Weight can be computed by using exponential weighting function.

$W_{k=1.2}{}^{k}$

where k represents frame number.

Then the TIRI image is divided into overlapping blocks of fixed length and applies 2D- DCT to each block to extract horizontal and vertical features. A 2DDCT is preferred since image is two dimensional and which makes correlation approach convenient. The co-efficients thus obtained per block are arranged in a matrix. The first co-efficient in this matrix is known as the DC component, representing the average intensity of an image, while the rest are the AC co-efficients corresponding to high frequency components of the image. The elimination of the high frequency components also causes the image to be robust to scale variations. So extract first horizontal and vertical frequency co-efficients.

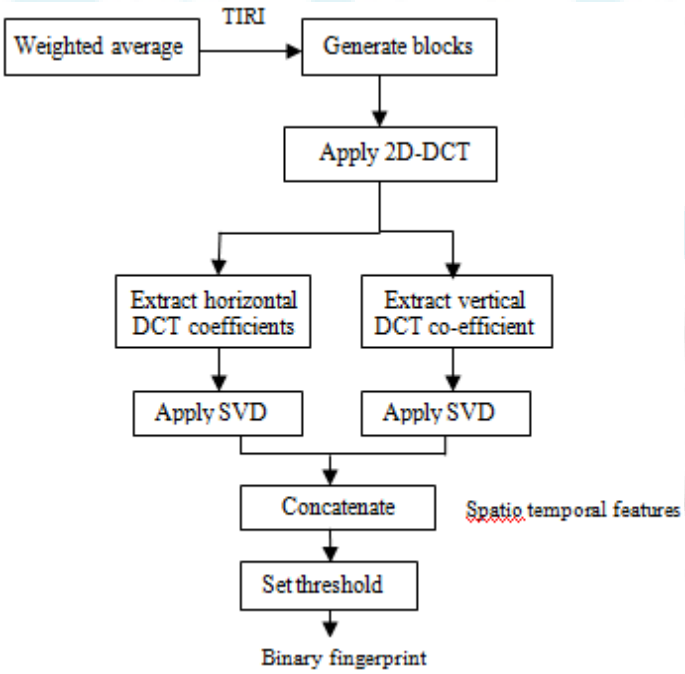

Fig 1: Fingerprint Generation

Algorithm for fingerprint generation

1. Select the key frames
2. Generate TIRI s from each segment of J frames using w_k=0.65^k
3. Segment each TIRI into overlapping blocks of size 2w_2w

4. Apply DCT Algorithm

For an N_M size image the 2D DCT is

$$F(u,v) = (\frac{2}{N})^{1/2}(\frac{2}{M})^{1/2}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}A(i)A(j)[\cos(\frac{\pi u}{2N}(2i+1))]$$

$$\cos(\frac{\pi v}{2M}(2j+1))\, f(i,j) \quad (3.3)$$

X

5. Concatenate all the co-efficient to form feature vector f.
6. find m which is the median of the element of f.
7. Generate binary hash h from f using the following formula:

$$h_k = \begin{cases} 1, f_k >= m \\ 0, f_k < m \end{cases} \quad (3.4)$$

In this way the fingerprint can be extracted and stored in the database. When a query video is present the same finger print extraction algorithm can be applied and extract the fingerprint. And then find a close enough match for extracted fingerprint within the database.

4.2 Matching of Fingerprint within Database

The fingerprint of the query video and the fingerprint in the database are matched by using the following methods:

4.2.1 Header Based Similarity Search

Clustering is used to reduce the number of queries that are examined within the database. For each fingerprint in the database a header is chosen. The fingerprint is divided into fixed length words. And for each word a header is chosen in such a way that if more number of zeros then it will be set to zero, otherwise it set to one. And the matching is performed with the dummy fingerprint constructed based on the header of the fingerprint. To determine if a query fingerprint matches a fingerprint in the database, the hamming distance between the query and fingerprint in the database are compared and the hamming distance between them is computed by using xor method. Identify the header with minimum hamming distance. And actual comparison is performed with that fingerprint. If no match is found, the query is declared to be out of the database. The cluster heads should be chosen such that a small change in the fingerprint does not result in the fingerprint being assigned to another cluster. In our general setting, we choose cluster heads (centers) as all the binary vectors with length«L. To assign a fingerprint to a cluster the fingerprint is first divided into segments (words) of length m=l/L. Each word is then represented by one bit in the cluster head or header, depending on the majority of word's bit values; e.g., it is represented by 1,if it has more than (or equal to) m/2 1's and it is represented by 0, if it has less than m/2 1's. Equivalently, each bit of the cluster head can be replicated m times and the Hamming distance between the expanded L=m/l bit version of all the cluster heads and the fingerprint is calculated. The cluster head closest to the fingerprint is then assigned to that fingerprint.

International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 1 Issue 3, November 2013

5. Result and Discussion

To evaluate the performance of proposed method the binary fingerprint of the video are extracted and stored in the fingerprint database. And apply many attacks (distortions) for video in the database and create query video. The attacks include changing brightness, frame rate conversion, rotation etc. The entire video copy detection was simulated using MATLAB and calculate true positive (TP) rate and false positive (FP) rate for the algorithm. Consider TP as true positive result correctly labeled as positive (copy).And FP means false positive result incorrectly labeled as positive.

True Positive Rate is computed by
TPR=TP/P

False positive Rate is computed by,
FPR=FP/N

TIRI-DCT demonstrates an acceptable performance for all attacks studied here. As discussed earlier, an important property of a fingerprinting system is its ability to detect and/or reject a query video within a large database in a fast and reliable fashion in the existing method take only three frames for fingerprint generation. The first, middle and the last frames are selected for TIRI generation. And avoid all other informations in the video.

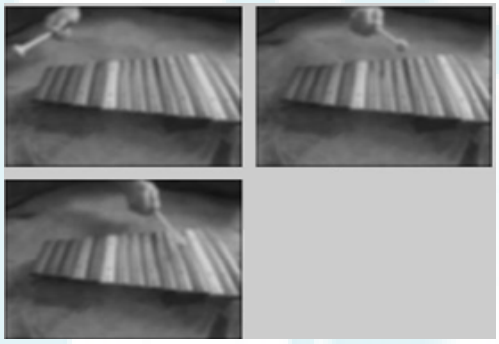

Figure 3.1: Select first, middle and the last frames for TIRI generation

In the key frame extraction method we proposed SSIM. In which take the threshold as 0.5 for shot detection. Then compute the similarity between the shots. For that the threshold is set as 0.45.The frames have structural similarity more than 0.45 are selected and merge into similar groups. Then TIRI image is computed from the selected frames. The gaussian blur is a type of image blurring filter that uses a Gaussian function for calculating the transformation to apply to each pixel in the image. Each pixel new value is set to a weighted average of that pixel's neighborhood the original pixel value receives the heaviest weight (having the highest Gaussian value) .And the neighboring pixels receive smaller weights as their distance to the original pixel increases. This results in a blur that preserves boundaries and edges. This can also be regarded as low pass results where the low frequencies are preserved. It also removes detail and noise from images. The gaussian filter has created a smoothing result. So the Gaussian smoothing filter makes the edges and orientation feature stable and also it reduce the influence of noise. We have examined different weight factors (constant, linear, and exponential) and observed that exponential weighting generates images that best capture the motion. In exponential method the weight factor is obtained by $w_k = \text{Ш}^k$. A very large, or one close to 0, generates a TIRI with detailed spatial information and low temporal information, resulting in a more discriminate representative. On the other hand, a Ш close to 1(giving the same weight to all the frames along the time line) results in a blurred image that is a more robust representative (containing averaged temporal information). By changing Ш from 0 to 1, we can move from a single frame selection (high spatial information) to selecting all frames with equal weights (high temporal information).

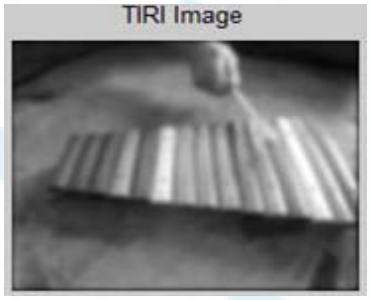

Figure 3.2: TIRI constructed from the selected frames

Figure 3.3: 25 frames are selected for TIRI generation

The hamming distance between the extracted fingerprints are computed by xor operation. If the difference or hamming distance is more than 300 then it will be announced as a match or a copy. When the threshold is set to too low, then the system will not detect small variations in the video. And if it is too high the false positive rate is more.

Figure 3.4: TIRI constructed in our proposed method

Contrast adjustment operation modifies the range of pixel values of an image without changing their mutual dynamic relationship. The proposed video signatures are robust up to

90 percentage contrast increase. Brightness adjustment is performed by either adding or removing a specific percent of the frame mean luminance value to or from each pixel in the frame. In brightness adjustment the detection fails only when brightness increase is at an extreme level. For other instances, the video signatures are observed to be robust and discriminant; so that we were able to detect all the pirated version of videos without any false positives identification.

Attack	*TPR*	*FPR*
Noise	96.42	3.52
Brightness change	97.65	2.35
Rotation	97.71	2.29
frame rate conversion	96.68	3.32
color to gray	98	2
average	97.26	2.708

Blurring is performed by filtering each frame using a standard Gaussian filter function with parameter (i.e., standard deviation). Since blurring will remove much of the medium- to high-frequency component from the image. But the preprocessing step preserves the low frequency co-efficient. So the average frequency preserved and detects the copy video.

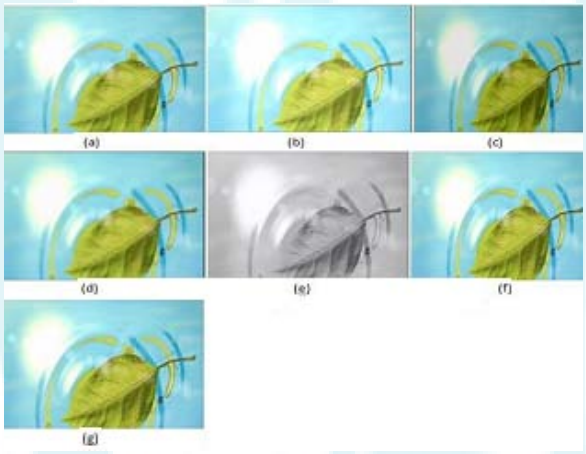

Figure 6.5: (a) original video clip (b) Brightness increased (c) Brightness decreased (d) Blurred (e) Gray scale conversion (f) Frame resized conversion (g) Frame rate conversion attack

Frame rate is the number of still pictures per unit of time of video. We try to change the frame rate by making them very fast and very slow. In both conditions detects the original video from the database. Frame size is referred to as the dimensions of the array of pixels. We try to change the size of the frames. Initially the video frames are resized to fixed size. So it detects the original video from the database. Convert the video to gray scale such as black and white. The proposed system detects the grayscale conversion also. The system has been tested for many videos. The experimental results shows that the proposed method detects major distortion such as frame rate conversion, frame resizing, brightness changes, contrast adjustment, blurring, cropping, rotation etc. From which average true positive rate and false negative rate are computed. Among these result gave an average true positive rate of 97 percent. The processing time is less due to key frame extraction. Also it makes the fingerprint more compact.

6. Conclusion and Future Work

The proposed video copy detection system is applicable for copyright management and indexing applications. The system consists of a fingerprint extraction algorithm followed by an approximate search method. The proposed fingerprinting algorithm (TIRI-DCT) extracts robust, discriminant, and compact fingerprints from videos in a fast and reliable fashion. These fingerprints are extracted from TIRIs containing both spatial and temporal information about a video segment. We demonstrate that TIRI-DCT generally outperforms the well-established (3D-DCT) algorithm and maintains a good performance for different attacks on video signals, including noise addition, changes in brightness/contrast, rotation, spatial/temporal shift and frame loss. To improve the security a random code word stored in the database instead of fingerprint and decode by using the hamming distance.

7. Acknowledgment

Our thanks to all the faculty members and colleagues who have contributed towards development of this paper

References

[1] Mani Malek Esmaeili, Mehrdad Fatourechi, and Rabab Kreidieh Ward, Fellow, IEEE" A Robust and Fast Video Copy Detection System Using Content-Based Fingerprinting", IEEE Transactions On Information Forensics And Security, Vol. 6, No. 1, March 2011

[2] Kim and B. Vasudev, "Spatiotemporal sequence matching for efficient video copy detection," IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 1, pp. 127–132, Jan. 2005.

[3] R. Radhakrishnan and C. Bauer, "Content-based video signatures based on projections of difference images," in *Proc. MMSP*, Oct. 2007, pp. 341–344.

[4] Joly, O. Buisson, and C. Frelicot, "Content-based copy retrieval using distortion-based probabilistic similarity search," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 293–306, Feb. 2007.

[5] S. Lee and C. Yoo, "Robust video fingerprinting for content-based video identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 7, pp. 983–988, Jul. 2008.

[6] G.Willems, T. Tuytelaars, and L. Van Gool, "Spatio-temporal features for robust content-based video copy detection," in *Proc. ACM Int. Conf. Multimedia Information Retrieval*, New York, NY, 2008, pp. 283–290, ACM.

[7] Matthijs Douze, Herv´e J´egou and Cordelia Schmid "An image-based approach to video copy detection with spatio-temporal post-filtering" IEEE 2008

[8] Zheng Cao, and Ming Zhu "An Efficient Video Copy Detection Method Based on Video Signature" IEEE International Conference on Automation and Logistics August 2009.

[9] Xing Su, Tiejun Huang, Wen Gao," Robust Video Fingerprinting Based On Visual Attention Regions", IEEE 2009.