Semi-fragile Watermarking Scheme for H.264/AVC Video Content Authentication Based on Manifold Feature

Chen Ling[12], Obaid Ur-Rehman[3] and Wenjun Zhang[1]
[1] School of Film and TV Arts & Technology, Shanghai University
Shanghai, 200072 – China
[2] School of Communication & Information Engineering, Shanghai University
Shanghai, 200072 – China
[3] Chair for Data Communications Systems, University of Siegen
Siegen, 57076 - Germany
[e-mail: lcrex@shu.edu.cn]
*Corresponding author: Chen Ling

Received December 15, 2013; revised February 15, 2014; revised June 25, 2014; accepted November 4, 2014; published December 31, 2014

Abstract

Authentication of videos and images based on the content is becoming an important problem in information security. Unfortunately, previous studies lack the consideration of Kerckhoffs's principle in order to achieve this (i.e., a cryptosystem should be secure even if everything about the system, except the key, is public knowledge). In this paper, a solution to the problem of finding a relationship between a frame's index and its content is proposed based on the creative utilization of a robust manifold feature. The proposed solution is based on a novel semi-fragile watermarking scheme for H.264/AVC video content authentication. At first, the input I-frame is partitioned for feature extraction and watermark embedding. This is followed by the temporal feature extraction using the Isometric Mapping algorithm. The frame index is included in the feature to produce the temporal watermark. In order to improve security, the spatial watermark will be encrypted together with the temporal watermark. Finally, the resultant watermark is embedded into the Discrete Cosine Transform coefficients in the diagonal positions. At the receiver side, after watermark extraction and decryption, temporal tampering is detected through a mismatch between the frame index extracted from the temporal watermark and the observed frame index. Next, the feature is regenerate through temporal feature regeneration, and compared with the extracted feature. It is judged through the comparison whether the extracted temporal watermark is similar to that of the original watermarked video. Additionally, for spatial authentication, the tampered areas are located via the comparison between extracted and regenerated spatial features. Experimental results show that the proposed method is sensitive to intentional malicious attacks and modifications, whereas it is robust to legitimate manipulations, such as certain level of lossy compression,

This work is supported by the 2012 PPP-Project, funded by the China Scholarship Council (CSC) and the Deutscher Akademischer Austausch Dienst (DAAD).

http://dx.doi.org/10.3837/tiis.2014.12.019

channel noise, Gaussian filtering and brightness adjustment. Through a comparison between the extracted frame index and the current frame index, the temporal tempering is identified. With the proposed scheme, a solution to the Kerckhoffs's principle problem is specified.

Keywords: Semi-fragile watermarking, video authentication, H.264, manifold, Kerchoff’s principle

1. Introduction

Digital multimedia has become an integral part of the modern life. In state-of-the-art networks, together with the development of increasingly powerful signal and image processing techniques, digital multimedia data is susceptible to manipulations and alterations through widely available editing tools. Information integrity authentication for multimedia content protection is becoming an important research problem in the study of information security [1]. There are two main video authentication methods; digital signatures and digital watermarking. Video authentication using digital signature is quite mature and is already used quite widely. One of its shortcomings is that the signature needs extra band-width or a separate secure channel for transmission. Additionally, due to the usage of a hash function, it is susceptible to failed authentication because of the avalanche effect. It might happen that one or a few bits of the multimedia data change during transit, e.g., due to channel noise or source compression. Video authentication using digital watermarking can overcome the aforementioned shortcomings. Since it is difficult to remove or tamper watermarks, the watermarking technology is gaining wide interest in recent research.

Watermarking can be broadly divided into fragile and semi fragile watermarking. Fragile watermarking approach, which fails to be authentic after a slightest modification, is proposed to verify the integrity and authenticity of digital content and location of tampered or modified areas using the embedded data. In practice, a video is always processed by some common image processing operations such as compression and filtering. The fragile watermarking is not suitable for these operations, so the semi-fragile watermarking techniques have been proposed. Semi-fragile watermarking behaves like fragile watermarking against intentional illegitimate modifications and as robust watermarking against casual legitimate manipulations like compression.

When a video is being recorded by a video recording device, it captures the scene frame by frame. Therefore, a video sequence can be viewed as a collection of consecutive frames with temporal dependency in a three dimensional plane. Therefore, videos do not have only two spatial dimensions like images, but three spatiotemporal dimensions. At present, the static image watermarking technology is quite mature. However, video watermarking methods have very different and normally higher requirements than the static image watermarking requirements. In practice, video sequences are stored and transmitted in a compressed form. During compression, the video frame is transformed from the spatial domain into frequency domain. Therefore, video authentication needs to be robust to acceptable operations, such as quantization, compression and filtering. In other words, the watermarking should be semi-fragile.

The state-of-the-art technology for temporal tampering uses the frame index, namely ID, as the temporal feature. It identifies the location of the frame in the time domain. However, according to Kerckhoffs's principle, a cryptosystem should be secure even if everything about the system, except the key, is public knowledge. This means that the security of a video authentication system should closely rely on the secret key. Unfortunately, current video authentication methods using the ID based feature have an assumption that attackers have no knowledge about the watermarking. For a good watermarking system, the embedding and the extraction algorithms should be public. The adversary may get a watermarked video, extract the embedded watermark and then tamper the video. Finally, the extracted watermark might be re-embedded into the tampered data. However, the ID based methods for authentication are

hard to crack because the ID does not relate to the video content. For example, for a video in **Fig. 5 (a)**, a red car and a bus are in the 30th frame. If an attacker knows that the video has been watermarked and also knows the watermark embedding method, he or she can easily extract the original watermark. Due to a secret key, it is hard to forge the watermark. However, the attacker might interchange some frames, e.g., the 30th and the 120th frames of the video. Thus in the 120th frame, the red car disappears as shown in **Fig. 5 (b)**. The attacker might re-embed the extracted watermark into the 120th frame. At the receiver side, the frame ID "30" will be extracted. Thus the system will determine that the video is not tampered in time domain, since the ID did not change. The reason is that the frame ID ("30") has no relationship to the frame content "red car".This paper considers this problem in the context of watermarking for content based authentication. A pioneering robust temporal feature is used to solve this problem. A novel semi-fragile watermarking scheme for H.264/AVC video content authentication is proposed. The proposed method creatively builds a bridge between the contents of a frame and its index. By combining this with spatial watermarking, a temporal-spatial semi-fragile watermark is embedded into the Discrete Cosine Transform (DCT) coefficients. At the receiver, the temporal authentication procedure will authenticate the ID and also whether the re-embedding of watermark has been done by the attacker. The spatial authentication procedure will locate the tampered areas in the tapmered frame.

This paper is organized as follows. In Section 2, we introduce and review the previous related work. The proposed watermark embedding and extraction schemes are given in Section 3 and 4 respectively. In Section 5, experimental results are shown to demonstrate the effectiveness of the proposed method. Finally, conclusion is given in Section 6.

2. Related Work

Tampering of video data can be done for several reasons, for instance to manipulate the integrity of the video content. The availability of a wide range of sophisticated and low cost video editing softwares makes it easier to maliciously manipulate the video content, posing serious challenges for researchers to be solved [2]. Watermarking can protect the video content. Qian Li and Rang-ding Wang [3] proposed a H.264 video integrity authentication algorithm. Through modulate the component of the residual motion vector, this method embeds watermarks into the larger absolute value of the residual component according to the movement of the video. In order to reduce the impact of video quality after watermarking, the selection criteria for the macroblock was restrained.Zhi-yu Hou et al. [4] proposed a scheme for the color video integrity authentication based on fragile watermarking. The RGB color mode is transformed to YST color mode which using for embedding watermarking into cover video. Firstly, the T component is selected to make DCT by 4×4 blocks. After that, according to the relationship of the quantilzed DCTcoefficients of the frames, the authentication code is created, as watermarking is embeded into the image by modifying the last non-zero DCT coefficient. The video anthentication is carried out without refering to the original video, blind detection can be achieved. This method can keep video's quality and also detect the tamper and attack on original video. Yanjiao Shi et al. [5] proposed an object based self-embedding watermarking for video authentication. The principal content of entire video and details of moving objects are protected by a reference sharing mechanism. At the receiver, if the stego-video is authenticated as un-tampered, the details of moving objects can be restored completely. If the stego-video is judged as tampered and the tampered area is not too extensive, the principal content of tampered regions and details of moving objects can be restored.

A wide range of authentication techniques have been proposed in the literature but most of

them primarily focus on spatial tampering (e.g. object removal attacks, object insertion attacks and object modification attacks). However, being three dimensional, when a video sequence is maliciously manipulated, it may affect the content of the video frames as well as the temporal dependency between the frames, which is called as temporal tampering. This kind of manipulations includes frame insertion attacks, frame removal attacks and frame exchange attacks. However, this problem has not been considered much in the literature. Dawen Xu et al. [6] proposed a content-based authentication watermarking scheme for H.264/AVC video, in which the content-based authentication code for spatial tampering is generated using the reliable features extracted from frame blocks. It is then embedded into the DCT coefficients in the diagonal positions. In addition, combining Error Correcting Coding (ECC) and interleaving, the frame index of each video frame is used as watermark information and embedded in the residual coefficients. Temporal tampering can be detected by the mismatch between the extracted and the observed frame index. Weibing Chen et al. [7] proposed a watermarking-based content authentication scheme for Audio Video coding Standard (AVS) in the compression domain. It makes use of the three-dimensional signatures of a video space, where Group of Pictures (GOP) index, GOP structure, frame index, and frame type are used for signature. The information is used to detect temporal manipulations. Furthermore, the mid-frequency DCT coefficients and motion vectors are both utilized for watermark embedding. In additional, authentication is done by comparison between the extracted dual watermarking and restructuring of the signatures.

State-of-the-art on video authentication does not cover the topic of using the frame IDs as part of video authentication, because ID is not related to content. For improved video authentication, a relationship between the temporal features of a video and the video content is necessary and for this a novel Manifold Semi-fragile Watermarking Scheme (MSWS) is proposed in this paper.

A frame, which has many pixels, can be seen as a high-dimensional data in the space domain. However, the content is not a high-dimensional pixel-based presentation, rather a low-dimensional feture-based representation. Feature extraction can be viewed as a method of reduction of dimensions. In this paper, the problem of the reduction of dimensions is creatively solved by a new unsupervised learning method, i.e., the manifold learning, which has become popular in computer vision and pattern recognition. Since 2000, the term “manifold learning” has appeared in quite widely in scientific literature. For nonlinear dimensionality reduction of a data set that lies on or around a low-dimensional manifold, Isometric Mapping (ISOMAP) [8], Locally Linear Embedding (LLE) [9], Laplacian Eigenmap (LE) [10] and Local Tangent Space Alignment (LTSA) [11] have been recently proposed.

In a gray scale video, each frame can be typically represented by the brightness values of the pixels. Thus, if there are $m \times n$ pixels in a frame, each image yields a data point in $\mathbf{R}^{m \times n}$. However, human senses do not feel the higher-dimensional pixels and therefore deals with the $m \times n$ pixels to generate a low-dimensional natural structure. In fact, the natural structure is a low-dimensional manifold embedded in $\mathbf{R}^{m \times n}$. For example, in **Fig. 1**, 16 images of Mona Lisa look the same with the naked eye. In reality, they are all different because they are downsampled from a high resolution image. Neurophysiologists have often found that the firing rate of each neuron in a population can be written as a smooth function of a small number of variables, such as the angular position of the eye or direction of the head. This implies that the population activity of neural firing in brain is constrained to lie on a low-dimensional manifold [12]. Due to the same content, natural structures are the same. This means that they have an embedded low-dimensional manifold.

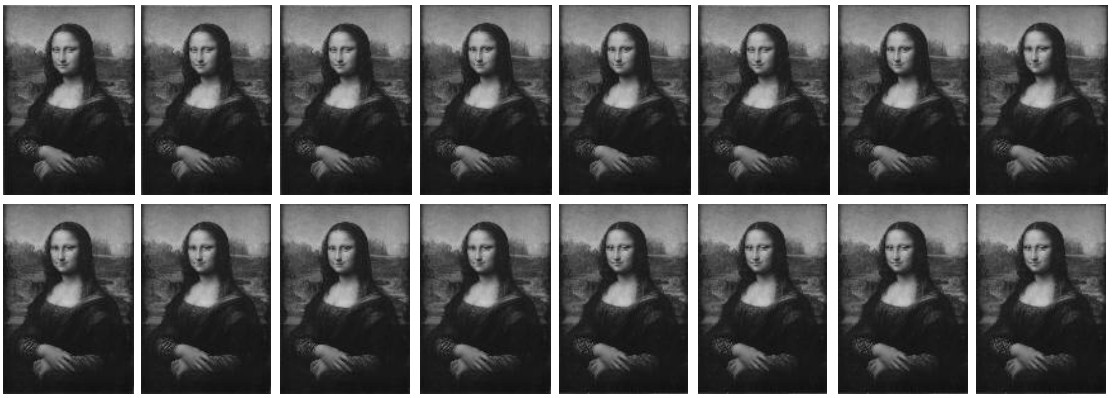


Fig. 1. Downsampling Mona Lisa.

In mathematics, manifold is a topological space that resembles the Euclidean space. More precisely, each point of an *n*-dimensional manifold has a neighbourhood that is homeomorphic to the Euclidean space of dimension *n*. Lines and circles are one-dimensional manifolds. Two-dimensional manifolds are also called surfaces. Assume the data is a low-dimensional manifold, which is sampled at a uniform high-dimensional Euclidean space. Manifold learning restores the low-dimensional manifold structure from high-dimensional sampling data. It looks for the essence of things from the observed phenomena and finds the inherent law generated data. Among them, LLE and ISOMAP are two of the most widely used algorithms.

LLE algorithm [9] computes a different local quantity, that is the coefficients of the best approximation to a data point by a weighted linear combination of its neighbors. Then the algorithm finds a set of low-dimensional points, each of which can be linearly approximated by its neighbors with the same coefficients that were determined from the high-dimensional data points.

In the ISOMAP algorithm [8], the local quantities computed are the distances between neighboring data points. For each pair of the non-neighboring data points, ISOMAP finds the shortest path through the data set connecting them, subject to the constraint that the path must hop from neighbor to neighbor. The length of this path is an approximation to the distance between its end points, as measured within the underlying manifold. Finally, the classical method of Multi-Dimensional Scaling (MDS) is used to find a set of low-dimensional points with similar pairwise distances.

As the web based video databases contain large number of copies with the explosive growth of online videos, effective and efficient copy identification techniques are required for content management and copyrights protection. In recent years, video hashing for video identification turned up using manifold learning due to the stable low-dimensional feature. Xiushan Nie et al. [13] proposed a novel video hashing for video copy identification based on LLE. It maps the video to a low-dimensional space via LLE, which is invariant to translation, rotation and rescaling. In this way, the points mapped from the video can be used as a robust hashing. Ming Tong et al. [14] presented a video dual watermarking method which could effectively resist geometric attacks. Their method generates zero-watermark or dynamic semantics watermark online according to the low-dimensional manifolds of different video shots. In order to balance the robustness and transparency, the method embeds watermark in the intermediate frequency DCT coefficients of AVS predicted residuals with large energy. This method can effectively resist center cutting, irregular cutting, row cutting, rotation, scaling, translation and other high intensity geometric attacks. This is also true for the combined attacks, such as combined rotation and corner cutting, combined center cutting and

rotation etc. We take the advantages of robust manifold video hashing to propose the MSWS method.

3. Watermark Embedding Procedure

In practical video storage and distribution systems, video sequences are stored and transmitted in a compressed format. Therefore, watermarking for video content authentication should be semi-fragile. It should distinguish intentional modifications and malicious attacks from acceptable operations and locate the tampered areas. The state-of-the-art video content authentication methods, which consider temporal tampering, use the frame ID as a temporal feature. However, it does not completely fulfills the criteria laid down by Kerckhoffs's principle.

In this paper, a new watermarking scheme MSWS is designed. Its goals are as follows:

(1) Semi-fragile: MSWS is fragile to malicious manipulation, but robust to legitimate signal processing;
(2) Temporal attacks: MSWS has resistance to frame addition attacks, frame removal attacks and frame exchange attacks;
(3) Kerckhoffs's principle: Suppose the embedding and extraction algorithms are public. The attacker can extract the original watermark, tamper the video and then re-embed the extracted original watermark into the tampered data. MSWS is able to identify this kind of tampering attacks.
(4) H.264/AVC: The lossy compression is of H.264/AVC is considered as legitimate manipulation by MSWS.

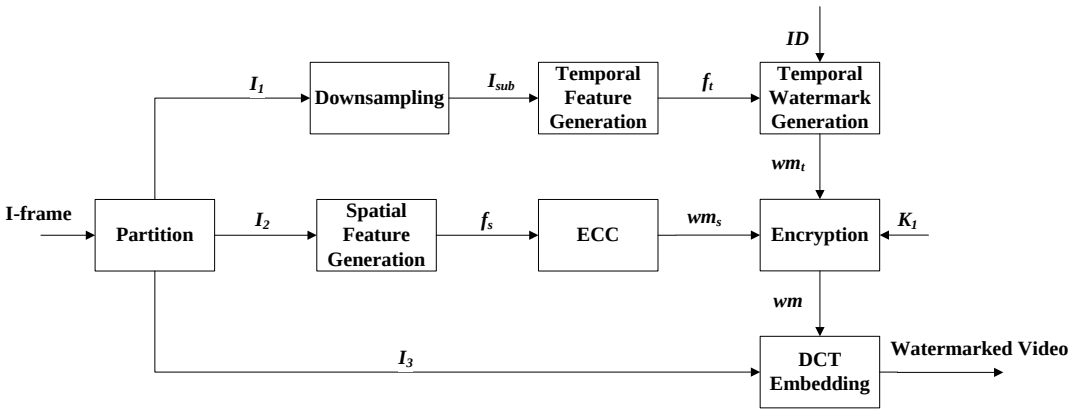


Fig. 2. Block diagram of the watermark embedding procedure.

The framework of the proposed semi-fragile watermarking scheme for H.264/AVC video content authentication is illustrated in **Fig. 2**. The input I-frame is partitioned at first. Then the I-frame is divided into three non-overlapping areas. The first part is downsampled and it gets the temporal feature through ISOMAP algorithm. The feature is combined with the frame index to generate the temporal watermark. The second part of the I-frame is used to extract spatial feature, which is processed by an Error Correction Code (ECC) to get spatial watermark. Both watermarks are scrambled to improve the security using a screct key. Finally, the generated watermark is embedded in the diagonal coefficients of the DCT in the third part of the I-frame. The details are introducted in the following sections.

3.1 Partition

MSWS algorithm for video authentication is based on the H.264/AVC standard [15, 16]. For intra coding, I-frames are independent of other frames. P-frames and B-frames rely on the preceding and following frames using inter frame coding. The P and B frames have low redundancy which is barely sufficient to embed watermarks. Therefore, I-frames are more stable for watermark embedding for video content authentication. Video coding often uses a color representation having three components known as Y, Cb, and Cr. Component Y is called luminance and represents brightness. The two chroma components Cb and Cr represent the extent to which the color deviates from gray towards blue and red, respectively. Because the human visual system is more sensitive to luminance than chroma, often a sampling structure is utilized in which the chroma component arrays have only one-fourth as many samples as the corresponding luminance component array. Therefore, in this paper, only the Y component of an I-frame is used to embed the watermarks.

Assume that the luminance component of the cover I-frame is I, it has N_1 columns and N_2 rows (both N_1 and N_2 are multiples of 16) and the total number of pixels is $N = N_1 \times N_2$. Let $I(x,y)$ denote the pixels, where $1 \le x \le N_1$ and $1 \le y \le N_2$. Each pixel $I(x,y)$ has the range of gray levels [0,255].

The input I-frame I should be devided into three separate parts I_1, I_2 and I_3.

$$\begin{cases} I_1 = I(1:2:N_1, 1:2:N_2) \\ I_2 = I(2:2:N_1, 2:2:N_2) \\ I_3 = I(1:2:N_1, 2:2:N_2) \cup I(2:2:N_1, 1:2:N_2) \end{cases} \quad (1)$$

Here, $I(i:2:N_1, j:2:N_2)$ is the i^{th} and j^{th} $(1 \le i, j \le 2)$ pixel in each 2×2 block of the matrix I. The watermark can also be seen as a kind of noise in the cover image. Once the watermark is embedded, the feature, which was extracted from the cover image or video frame and the one extracted from the watermarked frame are not same. Therefore, the regions from where the feature is extracted and where the watermark is embedded are identifiable through comparison at the receiver.

3.2 Temporal Feature Generation

In MSWS, any robust feature extraction algorithm can be chosen to generate the temporal feature. In this paper, the manifold feature extraction is used. As mentioned above, the brightness component of the cover I-frame I can be down sampled to generate several same content sub-images, as shown in **Fig. 1**. It will resample some pixels (e.g., every 8 pixels) from the original frame I. Sub-images are sampled from a high resolution image. These subimages are points in the high-dimensional space. The principal feature of the content is embedded in the high-dimensional space. A low-dimensional manifold can be found using these points. The embedded manifold is the principal feature of the original frame. We employ this feature as robust temporal feature. In this paper, 8×8 downsampling is used. In other words, these 64 subimages $I_{sub} = \{I_1^i, 1 \le i \le 64\}$ are calculated to get a manifold. Here I_1^i is a $(N_1/8 \times N_2/8) \times 1$ vector.

Many state-of-the-art manifold learning algorithms have been proposed in literature. Most of them are local embedding approaches. Although they have low computing complexity and good local properties, their global structure is not good, which will be quite different if few points are different at the receiver. The global nonlinear dimensionality reduction methods,

such as ISOMAP, are iterative. If the point set is large, its computational complexity is very high but it has a good global structure. In this paper, a sub-frame has only 8×8 points, so the computing complexity can be ignored. Due to good global construction, ISOMAP [8] is used for temporal feature generation.

$$f_t^{ori} = \text{ISOMAP}(I_{sub}, k_{NN}, dim) \tag{2}$$

Here ISOMAP(·) denotes the function of ISOMAP algorithm as metioned in **Table 1**. k_{NN} is the nearest neighbor value and *dim* is the manifold dimensionality. For simulations, k_{NN} =8 and *dim*=10 are chosen. The original manifold feature f_t^{ori} is binarized as follows:

$$f_t(i) = \begin{cases} 1 & if\ (dis(i+1) - dis(i)) > med \\ 0 & if\ (dis(i+1) - dis(i)) \le med \end{cases} \tag{3}$$

$$dis(i) = \text{norm}(f_t^{ori}(i), f_t^{ori}(i+1)) \tag{4}$$

Here, norm(·) is Euclidean distance function, $dis(i)$ is its distance, $1 \le i \le 63$, *med* is the median of the distance *dis* and f_t is the result binary temporal feature from f_t^{ori}.

Table 1. ISOMAP Algorithm

INPUT: Data points $X = \{x_1, x_2, ..., x_n \in \mathbf{R}^N\}$, nearest neighbour parameter *k* and low dimensionality *d*.
OUTPUT: Low-dimensional embedded manifold $Y = \{y_1, y_2, ..., y_n \in \mathbf{R}^d\}$.
Init: Compute the *k*-nearest neighbours of each point $x_i, 1 \le i \le n$;
Step 1: Construct the neighborhood graph If x_i and x_j are neighbors, then thier distance is the Euclidean distance $d_x(i,j)$. Otherwise, it is ∞;
Step 2: Compute the shortest path matrix $D_G\{ d_G(i,j)\}$ Calculate the shortest path $d_G(i,j)$ between x_i and x_j.
Step 3: MDS Compute the eigenvector of $\tau(D) = -HSH / 2$. Here, $S = (S_{ij}) = (D_{ij}^2)$, $H = (H_{ij}) = (\delta_{ij} - 1/N)$. Low-dimensional data is the eigenvector corresponding to the least 2nd to (*d*+1)th eigenvalue of $\tau(D)$.

Table 2. The diversity (BER) of different video manifold feature (QP=30), videos are shown in **Fig. 4**.

BER	hall	bus	flower	news
hall	0%	69.8%	50.7%	47.6%
bus	69.8%	0%	47.6%	47.6%
flower	50.7%	47.6%	0%	38.1%
news	47.6%	47.6%	38.1%	0%

The manifold feature has good robustness and Manifold features of different videos have large differences. As presented in **Table 2**, the Bit Error Rate (BER) is employed as the diversity metric in MSWS. According to the four different video sequences, the minimum diversity is more than 35%.

The manifold feature is also tolerant to the common image processing operations, such as H.264 (re)-compression and noise. One of the examples is described in **Table 3**, where the Manifold feature with different variance in Gaussian noise is presented. The given BER is an average of four different videos. It exhibits that the BER is smaller than 35% if the image is not altered too much (VAR≤0.01). It can also be observed that certainly it is hard to identify the content when too much noise (PSNR<20) is added in the video (VAR>0.01).

Table 3. The diversity (BER) of Gaussian noise (different variance)

VAR	0.0001	0.001	0.01	0.1
PSNR	39.9	30.0	20.2	11.6
BER	9.5%	23.0%	25.8%	38.9%

Based on **Table 2** and **Table 3**, a threshold can easily be defined to determine whether the videos have same content or not. Further results are presented in Section 5 on experimental results.

3.3 Temporal Watermark Generation

When the original temporal manifold feature is generated, it should be linked with the frame index *ID*. This increases the robustness against a certain extent of modifications. An additional improvement in the robustness of watermarking is achieved using ECC. The watermark generation algorithm is presented as follows:

- **Step 1: Parity check**

Parity check is employed by the ECC for error detection and correction. The manifold feature f_t has 63 bits. At first, a "0" is appended to the feature data. Now, the result is set to an 8×8 matrix. Each row and column has one parity check bit. Finally, the 16 bit parity check data is concatenated with original 64 bit data to produce the 80 bit watermark wm_t^{ori}.

- **Step 2: Frame index addition**

The current frame index *ID* replaces the 80 bit watermark wm_t^{ori} using binary representation. According to the frame size, the bit length of the *ID* , *len_ID* is self-adapted as:

$$len\_ID = \text{fix}(N / (2 \times 8 \times 8) / (64 + 16)) \tag{5}$$

Here, $\text{fix}(\cdot)$ is the round off function. The frame index *ID* represents binary *ID_one* according to the length *len_ID*. Then the binary inversion is to generate data *ID_zero*.

$$wm_t = \begin{cases} ID\_one & if \quad wm_t^{ori} = 1 \\ ID\_zero & if \quad wm_t^{ori} = 0 \end{cases} \tag{6}$$

Here, wm_t is the resultant temporal watermark. Finally, the *ID_one* and *ID_zero* substitutes for the 80 bit original watermark data wm_t^{ori}. For example, assuming *len_ID*=8, the binary representation *ID_one* of the 5th frame is "0000 0101". In this case, *ID_zero* is "1111 1010". If the bit of the original watermark wm_t^{ori} is "1", the generated temporal watermark is "0000 0101". Whereas, if the bit of the original watermark wm_t^{ori} is "0", the output temporal watermark is *ID_zero*. In this way, when the *ID_one and ID_zero* are extracted at the receiver side, they can be determined by embedded *ID* using the majority voting. Even if some *ID_one and ID_zero* have some error bits, the correct ID will have a maximum probability to obtain.

3.4 Spatial Watermark Generation

Aside from temporal watermark, aspatial watermark is also used. It is utilized to detect and locate tampering areas in the spatial domain. Many state-of-the-art watermarking methods foucses on this problem. The propsosed MSWS can adopt any of these spatial watermarking algorithms. For the convenience of description, this paper uses a familiar spatial watermark, i.e., the gray threshold feature.

The spatial feature, based on the most significant bit (MSB) plane, is extracted from the second part of I-frame I_2. Then the MSB plane is divided into blocks. In this way, each block has only one bit spatial feature wm_s^{ori}.

In order to improve the robustness, the original spatial feature wm_t^{ori} is processed by ECC to obtain the spatial watermark wm_s. Here, two error correcting bits are embedded using (4, 2) repetition code. For example, suppose the watermark is {01}, then the encoded watermark is {01| 01}. Therefore, at the receiver side, the watermark extraction is more robust using majority logic decoding.

3.5 Encryption

It is known that the malicious manipulations have some well designed motives for attackers and normally the tampered regions are always concentrated. For recovery, the data representing the principal feature in a region is embedded into a different region. In this paper, the image scrambling method is based on the logistic map [17]. The logistic map is a classical chaotic system in one-dimension, which is defined by following equation:

$$x_{k+1} = \mu x_k (1 - x_k) \tag{7}$$

Here $0 \le \mu \le 4$ is a parameter of the algorithm and $x_k \in (0,1)$ and x_k is the input coefficient . From a chaotic dynamical system, the logistic map becomes confusion if $3.569945... \le \mu \le 4$, that is to say, chaos sequences x_k are non-periodic, non-convergent and pseudorandom, given two different initial conditions x_0. The temporal and spatial watermarks are scrambled using the logistic map, which depends on a secrct key K_1. Finally, the watermark *wm* is produced.

3.6 DCT Embedding

The above generated watermark *wm* is embedded into the nonoverlapping 8×8 blocks of the third part of I-frame I_3. This approach is proven to be stable for acceptable image processing [18]. The MSWS algorithm embeds watermarks into DCT coefficients in the diagonal positions. In other words, after zigzag scan, the temporal watermark is embedded by adjusting the polarity between the 28th and the 30th AC coefficients, and the spatial watermark embedding position is the 32nd and 34th AC coefficients. In this way, a good stability and robustness against most of the attacks is obtained. The watermark embedding method is written as follows:

If *wm*=1, then

$$\begin{cases} C_k(u) = C_k(v) + s, C_k(v) = C_k(u) - s & if \ \ C_k(u) < C_k(v) \\ C_k(u) = C_k(u) + 2s, C_k(v) = C_k(v) - 2s & if \ \ C_k(u) = C_k(v) \end{cases} \tag{8}$$

If *wm*=0, then

$$\begin{cases} C_k(u) = C_k(v) - s, C_k(v) = C_k(u) + s & if \ \ C_k(u) > C_k(v) \\ C_k(u) = C_k(u) - 2s, C_k(v) = C_k(v) + 2s & if \ \ C_k(u) = C_k(v) \end{cases} \tag{9}$$

Here $C_k(u)$ and $C_k(v)$ denote the principal diagonal coefficients in the embedding positions u and v. s is a small constant and u and v are the position indices of the diagonal DCT coefficients. After embedding, the watermarked DCT block is inverse transformed. This watermarked frame is used as a reference frame to produce the following P and B frames, which is the same method as is used in the original H.264 encoder.

Let I be the reconstructed pixels of an original I-frame, and IWM be the reconstructed pixels of the I-frame including the watermark. Let P be a reconstructed P-frame based on I, and P′ be the reconstructed version based on IWM. Let E be the residual errors of the P-frame worked out by motion estimation. When there is no secret data hidden in the I-frame, the P-frame is reconstructed on the decoding side by

$$P = I \oplus E \tag{10}$$

where $\oplus$ means motion compensation. When secret bits are embedded into the I-frame, the reconstructed pixels will be different from the original ones, that is,

$$I \neq I_{WM} \tag{11}$$

The decoder uses E and I_{WM} to reconstruct the P-frame by

$$P' = I_{WM} \oplus (E + e) \tag{12}$$

Here, e is the difference of the residual errors based on I and I_{WM}. Comparing P' with P,

$$P' - P = (I_{WM} \oplus (E + e)) - (I \oplus E) = I_{WM} - I + e = wm + e \tag{13}$$

It can be seen that P' and P are different; that is, compared with P, P' is degraded. The following P-frames and B-frames in the stego-video are degraded in a similar manner. This phenomenon is called distortion drift. If changes in the video frames are not big (e.g., adjacent frame), e is very small and can be ignored. The difference between P' and P is the watermark wm. Therefore, when an I-frame is changed to a P/B frame or vice verse by either transcoding or other ways, the watermark will be successfully extracted to a certain degree.

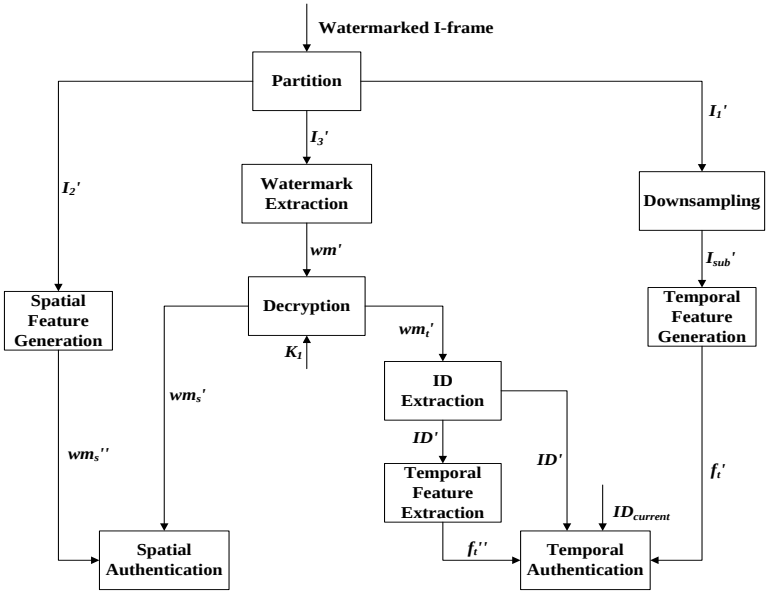


Fig. 3. Sketch of the watermark extraction and authentication procedures.

4. Watermark Extraction and Authentication Procedure

Suppose that an adversary tampers the content of a watermarked video. The proposed MSWS will authenticate the watermarked video using the temporal and spatial tampering. After watermark extraction and decryption, the embedded frame ID is recovered from the temporal watermark for temporal authentication. Next, after the temporal feature regeneration, the regenerated feature is compared with the extracted one. It is judged whether the extracted temporal watermark is the original watermark of the watermarked video. For spatial authentication, the tampered areas are located through a comparison between the extracted and regenerated spatial feature. The extraction and authentication procedure is shown in **Fig.3**.

4.1 Watermark Extraction

Once a watermarked I-frame I' is received, it is divided into three separate parts I_1', I_2' and I_3', as is done in the watermark generation. From I_3', the embedded watermark wm' can be extracted easily as follows:

$$\begin{cases} wm'=1 & if \quad C_k(u) > C_k(v) \\ wm'=0 & if \quad C_k(u) < C_k(v) \\ wm'=-1 & if \quad C_k(u) = C_k(v) \end{cases} \tag{14}$$

Here, wm'=-1 means that the bit has an error. Because the original watermark has some ECC parity, the wm' can be corrected using the ECC. Then, the watermark wm' is inverse permutated using the secret key K_1, and the extracted temporal watermark wm_t' and the spatial watermark wm_s' are obtained.

4.2 Spatial Authentication

The current spatial feature wm_s'' of the watermarked video is regenerated from the frame I_2', whose method is the same as the watermark generation procedure. Then a Tampered Area Location Image **TM** is created to record the tampered blocks.

$$\mathbf{TM} = \begin{cases} 1 & if \quad wm_s' \neq wm_s'' \\ 0 & if \quad wm_s' = wm_s'' \end{cases} \tag{15}$$

In practice, some non-maliciously modified areas in the watermarked frame are marked. For example, if the watermarked video is transmitted through a noisy channel, the noise will result in isolated detection points in **TM**. These points do not result from malicious modifications. Hence, the tampered area location image needs some post-processing. First, calculate the connected domain and remove some small areas. Then use a closed operation to reduce the disconnect area. Finally, high probability region is considered to be the real modified area.

$$\mathbf{TM}_p = \begin{cases} 1 & if \quad \text{Scope}(rs, \mathbf{TM}) / \text{Sum}(rs, \mathbf{TM}) > \tau \\ 0 & if \quad \text{Scope}(rs, \mathbf{TM}) / \text{Sum}(rs, \mathbf{TM}) \le \tau \end{cases} \tag{16}$$

Here, $\mathbf{TM}_p$ is the post-processing tampered area location image. Scope(rs, **TM**) denotes that the number of tampered blocks are within the radius rs. Sum (rs, **TM**) is the total number of

blocks within the radius *rs*. If the tampered proportion is larger than the threshold τ, it can be regarded as a tampered region.

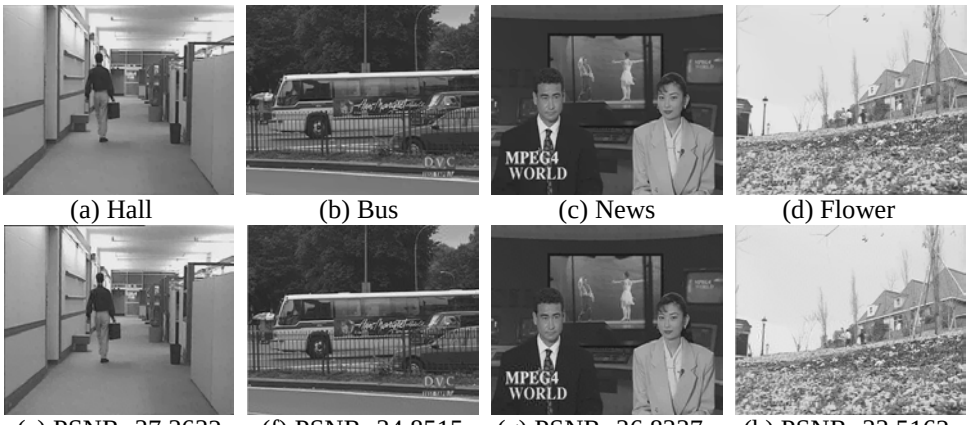
4.3 Temporal Authentication

After decryption, the embedded temporal watermark wm_t' is extracted. Due to the replacement of *ID_one* and *ID_zero*, the *ID_one*, namely the original ID, is obtained using the majority voting. This means most probable occurrence of the extracted ID is considered the original ID. It eliminates some errors, and gets the correct *ID* with a high probability. Then the extracted ID *ID'* is compared with the current frame index *ID_current* to verify whether the video is temporaly attacked and which kind of temporal tampering has been done.

Once the *ID_one* is extracted, the 80 bit original temporal watermark is gained. Even if there are some errors, the 63 bit original manifold feature f_t'' can be corrected using the 16 bit parity check data and the repetition embedded manifold feature f_t''. Similar to the temporal watermarking, the reproduced temporal feature f_t' is obtained from I_1'. Through BER computation between f_t' and f_t'', the difference between the received watermarked video and the original cover video is acquired. If the diversity is bigger than the chosen threshold value, the received video might have been attacked. Its temporal watermark may not be the watermark of the received video because the attacker may have extracted the original watermark first tampered the video and may have re-embedded it into the tampered data. If the diversity is smaller than the threshold, it will pass the authentication.

5. Experimental Results and Analysis

For the performance evaluation of the proposed MSWS algorithm, a comprehensive set of simulations have been performed. Many 4:2:0 CIF YUV format video test sequences, Hall, Bus, News, flower etc. are used in these experiments. These test sequences represent the most common scenarios of video content authentication. For example, the video Hall and the Bus represent video surveillance (VS) system sequences. The Hall is based on a still camera and moving human. The Bus is a video of cars on the street and contains more elements than the former. The News video represents news on the television which can also be seen as a video conference. These video sequences are common as a proof in front of a court of law. The video compressed format employed in this paper is H.264/AVC. The software codec adopted is H.264 Baseline Codec in [16].



(a) Hall (b) Bus (c) News (d) Flower

(e) PSNR=37.3632 (f) PSNR=34.8515 (g) PSNR=36.8337 (h) PSNR=33.5162

Fig. 4. Test videos and corresponding watermarked videos.

5.1 Fidelity

The watermarked videos are given in **Fig. 4** and are perceived identical to the original cover videos due to the human visual perception system. The above row is the four test cover videos. The lower row shows their corresponding watermarked videos. Peak-signal-to-noise ratio (PSNR) is employed as the performance metric. The value of PSNR is higher than 33 dB due to watermark embedding. It can be noticed that the proposed scheme has good imperceptibility.

5.2 Temporal Malicious Attacks

A video can be seen as a series of continuous static images in time domain. However, a video is essentially much different from static images. It has not only spatial but also temporal characteristics. So videos have their own nature of malicious attacks. Temporal tampering includes frame addition attacks, frame removal attacks and frame exchange attacks.

In this paper, the watermarking method is assumed to be public. The adversary can extract the original watermark first, tamper the video and then re-embedded the original watermark into the tampered data. This is compared with the current watermarking methods, like Xu [6] and Chen [7]. They all just straightforwardly use the frame ID. The ID has no relationship with the content of the cover frame, which it is embedded into. Amongst all of the temporal attacks, frame exchange attacks will always face the Kerckhoffs's principle problem, because the adversary needs to manipulate only some shuffled frames. However, other two attacks will tamper all following original frames.

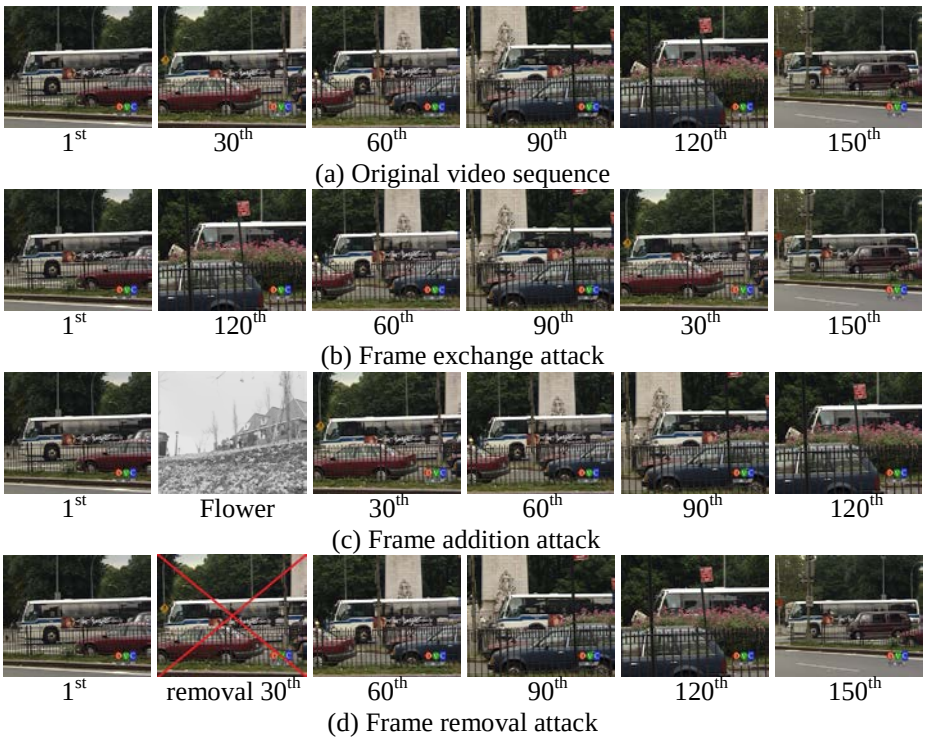


Fig. 5. Bus test video sequence temporal attacks.

5.2.1 Frame exchange attacks

In frame exchange attack, frames of a given video are shuffled or reordered in such a way that the correct frame sequence is mingled and wrong information is produced by the video as compared to the original recorded video. **Fig. 5(b)** shows a typical example of frame exchange attacks. Thirty frames, from position 30th till 59th, are interchanged with another thirty frames from position 120th till 149th.

ID is a side information, which helps to locate the frame positions in time domain. If the attacker does not know that the video has been watermarked, the type of temporal tampering can be confirmed via comparison between the current ID and the extracted ID. As shown in **Fig. 6**, the 30 frames have been shuffled.

If the adversary knows the watermarking scheme, as assumed in this paper, the forged information can be embedded into the tampered video. As shown in **Table 4**, current frame index is 30. The attacker has extracted the watermark of the 30th frame, and then re-embedded it into the exchanged 120th frame. **Table 4** shows that the extracted ID is 30 through our proposed method, Xu method [6] and Chen method [7]. Our proposed scheme has a second chance to authenticate whether the ID is correct. The diversity between regenerated temporal feature and extracted temporal feature is 44.4%, which is bigger than the defined threshold value (in this paper, we set it 35%). ID trust worthiness shows that the extracted ID is incorrect. The video has been attacked and the MSWS algorithm has identified the frame shuffle.

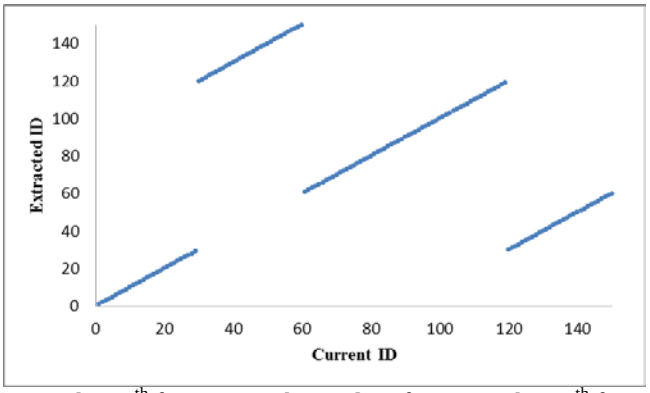


Fig. 6. The 30th frame is exchanged 30 frames with 120th frame.

Table 4. The ID extraction result of the frame exchange attack (current ID=30)

	MSWS	**Xu** [6]	**Chen** [7]
Extracted ID	30	30	30
BER	44.4%	Null	Null
ID trust worthiness	No	Yes	Yes

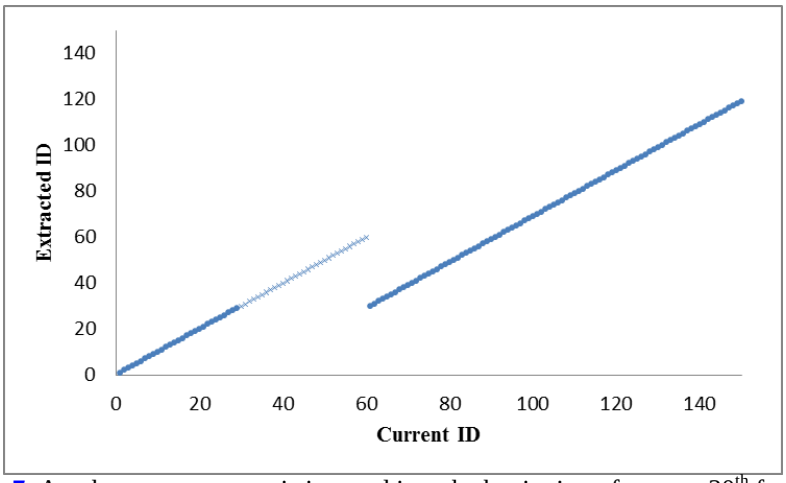


Fig. 7. Another test sequence is inserted into the beginning of current 30th frame.

5.2.2 Frame insertion attacks

In the frame insertion attack, additional frames are intentionally inserted at some locations in a given video. This attack is intended to camouflage the actual content and provide incorrect information. **Fig. 5(c)** shows a typical example of frame insertion attack. A Flower test sequence is inserted into the beginning of 30th Bus frame.

If the attacker does not know the video has been watermarked, the result of the frame insertion attack is shown in **Fig. 7**. It is obvious that some frames are inserted into the original video. Here, points "×" means the incorrect ID points.

If the adversary knows the watermarking scheme, he can embed extracted correct temporal watermark into forged frames. The proposed MSWS scheme can authenticate whether the extracted watermark is correct. As presented in **Table 2**, different videos have a huge diversity. Through a threshold, the temporal feature can be distinguished. However Xu method [6] and Chen method [7] will not be able to detect this.

5.2.3 Frame removal attacks

In frame removal attack, the frames of a given video are intentionally eliminated. In this kind of attack, the frames or set of frames can be removed from a specific location to a fixed location or can be removed from different locations. Commonly this kind of tampering attack is performed on surveillance video where an intruder wants to remove his presence in the video. **Fig. 5 (d)** shows a typical example of frame removal attacks. 30th to 59th original frames have been removed. The authentication result can been seen in **Fig. 8**.

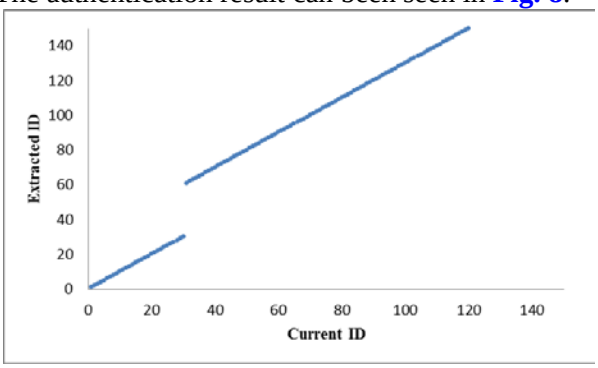


Fig. 8. 30 frames have been eliminated.

Table 5. Temporal watermark agaist different common image processing operations

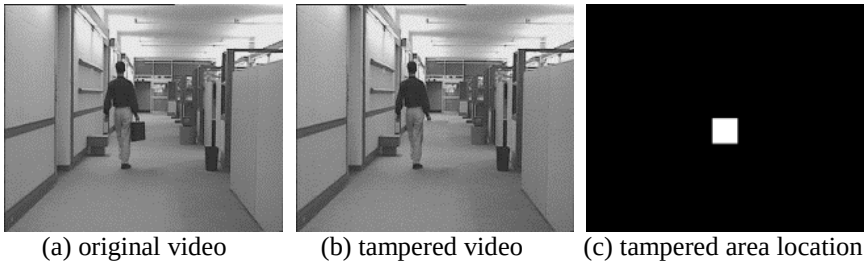
Image processing operations		BER (QP=30)
Re-compression (QP)	0	0
	10	4.0%
	20	10.3%
	40	20.6%
Salt & pepper noise	0.0001	1.6%
	0.001	19.0%
	0.01	29.4%
Gaussian noise	0.0001	9.5%
	0.001	23.0%
Gaussian filtering		12.7%
Brightness enhancement		8.7%

5.3 Common image processing operations

Our watermarking scheme is semi-fragile. In other words, the proposed method tolerates a certain extent of lossy compression, noise, Gaussian filtering and brightness enhancement. As illustrated in **Table 5**, different common image processing operations are simulated. The result is an average of different videos. In this paper, the threshold is chosen as 35%. It can be observed from **Table 5** that the BER for most of the operations is smaller than the defined threshold value. This shows that the MSWS algorithm is robust to such common image processing operations.

5.4 Spatial Malicious Attacks

The most important security issue for watermarking based authentication system is the spatial malicious attacks also known as block replacement attacks. The attacker would replace a watermarked block with another block to edit something important in the video. **Fig. 9** shows a simulation result for the replacement attack. In the original watermarked test sequences for Hall, a man with a briefcase is walking in the office, as shown in **Fig. 9(a)**. The tampered version is shown in **Fig. 9(b)**, where the briefcase is removed. One can see clearly that the man carries nothing. MSWS algorithms can locate tampered areas as shown in **Fig.9(c)**. As mentioned above, MSWS can employ any spatial feature **making it a flexible approach.**



(a) original video (b) tampered video (c) tampered area location

Fig. 9. Simulation results of the replacement attack.

6. Conclusion and Future Work

In this paper, a novel semi-fragile watermarking scheme for H.264/AVC video content authentication is proposed. In order to circumvent the difficulty of the relationship between the frame index and the frame content, manifold feature is utilized to produce a robust temporal watermark. By combining the temporal watermark and the spatial watermark, a temporal-spatial semi-fragile watermark is embedded into the DCT coefficients. At the receiver, the extracted watermark is divides into temporal and spatial watermarks again. They are independently verified using temporal and spatial authentication procedures. Experiments implementing the proposed scheme can authenticate whether the extracted ID should be the ID of the the current frame.

However, some problems still exist. First, if the spatial tampered areas are too big, the temporal authentication will fail. On the contrary, if the spatial tampered areas are too small, the temporal authentication may believe it is correct, although the frame might have been attacked in the temporal domain. Second, the rich texture video will weaken the performance of the robust feature, because the details can be seen as Gaussian noise, which is hard to distinguish.Third, the proposed method is based on I-frame. When an I frame is changed to a P/B slice or vice verse by either transcoding or other ways, the watermark would not have reliable performance. These problems are identified here and left for future work.

References

[1] X. Li, Y. Shoshan, A. Fish, et al, "Hardware implementations of video watermarking," *Information Technologies & Knowledge*, vol. 3, pp. 103-120, 2009. Article (CrossRef Link)

[2] S. Upadhyay, S. K. Singh, "Video Authentication: Issues and Challenges," *International Journal of Computer Science*, vol. 9, no. 2012. Article (CrossRef Link)

[3] Q. Li, R. Wang, "Watermarking Algorithm for the Integrity Authentication of H.264 Video," *Microelectronics & Computer*, vol. 29, no. 9, pp. 189-192, 2012. Article (CrossRef Link)

[4] Z. Hou, X. Tang, H. Chen, "Integrity Authentication Scheme of Color Video Based on the Fragile Watermarking," *Journal of Hangzhou Dianzi University*, vol. 31, no. 5, pp. 135-138, Article (CrossRef Link)

[5] Y. Shi, M. Qi, Y. Lu, et al, "Object based self-embedding watermarking for video authentication," in *Proc. of 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)*, pp.519-522, 2011. Article (CrossRef Link)

[6] D. Xu, R. Wang, J. Wang, "A novel watermarking scheme for H. 264/AVC video authentication," *Signal processing: Image Communication*, vol. 26, no. 6, pp. 267-279, 2011. Article (CrossRef Link)

[7] W. Chen, G. Zhang, G. Liu, "Algorithm study on video watermarking authentication based on AVS in compressed domain," in *Proc. of World Automation Congress (WAC)*, pp.283-287, 2012. Article (CrossRef Link)

[8] J. B. Tenenbaum, V. De Silva, J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319-2323, 2000. Article (CrossRef Link)

[9] S. T. Roweis, L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323-2326, 2000. Article (CrossRef Link)

[10] M. Belkin, P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373-1396, 2003. Article (CrossRef Link)

[11] Z. Zhang, H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *Journal of Shanghai University (English Edition)*, vol. 8, no. 4, pp. 406-424, 2004. Article (CrossRef Link)

[12] H. S. Seung, D. D. Lee, "The manifold ways of perception," *Science*, vol. 290, no. 5500, pp. 2268-2269, 2000. Article (CrossRef Link)
[13] X. Nie, J. Qiao, J. Liu, et al, "LLE-based video hashing for video identification," in *Proc. of 2010 IEEE 10th International Conference on Signal Processing (ICSP)*, pp.1837-1840, 2010. Article (CrossRef Link)
[14] M. Tong, T. Xu, J. Zhang, "Video dual watermarking method for resisting geometric attacks based on low-dimensional manifold," *Journal of Xidian University*, vol. 3, no. 2, 2011. Article (CrossRef Link)
[15] G. J. Sullivan, T. Wiegand, "Video compression-from concepts to the H. 264/AVC standard," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 18-31, 2005. Article (CrossRef Link)
[16] A. A. Muhit, M. R. Pickering, M. R. Frater, et al, "Video coding using elastic motion model and larger blocks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 5, pp. 661-672, 2010. Article (CrossRef Link)
[17] G. Ye, "Image scrambling encryption algorithm of pixel bit based on chaos map," *Pattern Recognition Letters*, vol. 31, no. 5, pp. 347-354, 2010. Article (CrossRef Link)
[18] T. Kuo, Y. Lo, "A hybrid scheme of robust and fragile watermarking for H. 264/AVC video," in *Proc. of 2010 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp.1-6, 2010. Article (CrossRef Link)

Chen Ling received his bachelor of engineering in Film and TV Arts & Technology, Shanghai University, Shanghai, China in 2010. For the continuous academic program that involves postgraduate and doctoral study, he is currently working towards his PhD in Shanghai University. He has published more than 13 scientific research publications in international journals and conferences. His research interests include multimedia, video authentication, watermarking, augmented reality, human-computer interaction and somatic science.

Obaid Ur-Rehman received his PhD in Electrical Engineering from the University of Siegen, Germany in 2012 and his MS in Computer Engineering from the University of Engineering and Technology Taxila in 2004. He has more than 8 years of industrial and academic experience. He serves as the reviewer of various international journals and has authored more than 25 scientific research publications in international journals and conferences. Currently, Dr. Ur-Rehman is a post doctoral fellow at the University of Siegen, Germany. His areas of interest include error correcting codes, data authentication and network security.

Wenjun Zhang received his bachelor and Master degrees in electronics from Faculty of Electrical Engineering at University of Belgrade in 1984 and 1986, respectively, and his PhD in Telecommunications from University of Belgrade in 1989. He is currently a Professor in School of Film and TV Art & Technology and in School of Communication and Information Engineering of Shanghai University. He has published 5 books and more than 100 papers. His research interests include digital media technology and applications, especially in the fields of digital image processing, digital image communication, digital right management, digital content design and productions.