Digital video watermarking algorithm based on asymmetric causal spatio-temporal filter

Yanliang Ge, Hongbo Bi, Di Lu, Xiaoxue Kong, Zheng Zhang and Ning Li

School of Electrical and Communication Engineering, Northeast Petroleum University, Daqing, People’s Republic of China

ABSTRACT

Digital video is vulnerable to accidental or malicious destruction in storage, transmission, processing, which damages the legitimate rights and interests of the product owners. Digital video watermarking technology is a means to protect intellectual property rights. We analyze the basic principle of the 3D-Harris algorithm and Gabor filter method. By considering the temporal causality of the video, the common symmetrical filter is discarded. We design a causal filter which conforms to video characteristics and proposes an asymmetric causal filter in space–time domain. On this basis, an improved video watermarking algorithm is proposed, combining space–time feature points and DCT domain. The experimental results show that proposed algorithm cannot only ensure high invisibility, and can effectively resist various attacks in time domain and space domain.

1. Introduction

Video data integrates content about information in the environment and is widely applied in many visual tasks including surveillance, content authentication, service tracing, etc. However, the feasibility of distributing and acquiring video data causes the issue of protecting intellectual property rights (IPR). Digital video watermarking is one of the most prominent and effective solutions and has been extensively investigated in the recent years.

A video can be considered as a set of sequential static images over time, (i.e. 2 spatial dimensions + 1 temporal dimension). When watermarked video is distributed, it is often attacked by both time and space domains simultaneously, resulting in the loss of synchronization in time or space, which often leads to inaccuracy or even inability to detect/extract watermark information. Therefore, robust video watermarking algorithm research against space–time attacks (including the time-domain operation of video frames and the geometric operation of space domain, signal processing, etc.) is of great significance. (Heidari et al., 2017; Hosseini & Ghofrani, 2014; Tong, Chen, Zhang, & Dong, 2012).

Many interesting events in video can be characterized by strong variations of the data in both the spatial and the temporal dimensions. As a result, video data can be represented by space–time interest points which have been used in video analysis, video interpretation, etc. These space–time interest points reveal the intrinsic structure of video including the spatial and temporal dimension, which can be seen as a potential solution for temporal synchronization in video watermarking while preserving the advantage in spatial domain.

Whereas it seems to be not sufficient enough to applying space–time interest points to the video watermarking. The response function of time–space corner method is insensitive to the changes in the time dimension. The number of detected time–space points of interest is relatively small, and the points of interest may be missed, which is not conducive to follow-up analysis. (Gaj, Rathore, Sur, & Bora, 2017; Gunjan, Mitra, & Gaur, 2012).

To improve the timeline insensitivity, Gabor time filter can be utilized. However, video has a temporal causal relationship, while Gabor filters do not. Some researchers proposed spatio-temporal video watermarking similarly, however, their schemes did not consider the temporal causal relationship either. (Hongbo, Yubo, & Xueming, 2011; Li, Guo, & Pan, 2008). Therefore, we improve the time filter to form an asymmetric causal spatio-temporal filter and combines with the Gaussian kernel filter in space to form a new space–time filter. (Darabkh, 2014; Shabani, Clausi, & Zelek, 2011; Zhou, et al., 2014).

The proposed method of space–time extraction can meet the following needs: (1) The algorithm is simple and can reduce the computational complexity as much as possible due to the huge amount of video data; (2) A temporal filter is introduced that detects enough points...
of interest, not to skip points of slow-moving or corner-insensitive motion; (3) An improved one-dimensional temporal filter that is an asymmetric causal filter, which can meet the causal temporal relationship of video in the real life.

The spatio-temporal feature points are extracted from the objective conditions of video. DCT is used to embed the watermark as the location of video watermarking. Thus a new video watermarking algorithm is formed, and experiments verify that it has good invisibility and robustness.

The outline of the rest of the paper is as follows. In section 2, the related space–time interest points detection algorithm is mentioned. In section 3, the method of spatio-temporal features point detection of asymmetric causal filter is proposed. In section 4, the proposed watermarking scheme is presented, and we give the embedding and extraction of the watermarking. In section 5, some experimental results are mentioned. Finally, conclusions are given in section 6.

2. Related spatio-temporal interest points detection methods

2.1. 3D-Harris algorithm spatio-temporal interest point detection methods

Spatio-Temporal Interest Point was firstly proposed Laptev et al, extending Harris corner detection to 3D space–time domain. (Kalra, Talvar, & Sadawarti, 2015; Laptev, Caputo, Schuldt, & Lindeberg, 2007; Liu, Wang, & Zhu, 2015). The video sequence can be regarded as a 3D space–time function. Because the video sequence has the dual characteristics of time and space and the motion is relatively independent in time and space, the 3D Gaussian convolution function includes two independent variances: time dimension variance and spatial dimension variance, corresponding to the time scale and the spatial scale, which is represented by \( \tau_i \) and \( \sigma_i \) respectively. The linear scale space of space–time three-dimensional video sequence is denoted by \( L \), and the video sequence is expressed by \( f \). The space–time three-dimensional video sequence of linear scale space \( L \) is expressed as Equation (1).

\[
L(\cdot; \sigma^2_i, \tau^2_i) = g(\cdot; \sigma^2_i, \tau^2_i) * f,
\]

where ‘‘ represents 3D coordinates \((x, y, t)\). Where the 3D Gaussian convolution function is defined as

\[
g(\cdot; \sigma^2_i, \tau^2_i) = \exp\left(-\frac{(x^2 + y^2)}{2\sigma^2_i} - \frac{t^2}{2\tau^2_i}\right) \frac{1}{(2\pi)^3 \sigma^4_i \tau^4_i}.
\]

To detect interest points, 3D response function is defined as

\[
H = \det(\mu) - k \cdot \text{trace}^3(\mu),
\]

where \( \mu \) is defined as

\[
\mu = g(\cdot; \sigma^2_i, \tau^2_i) * \begin{bmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_y L_x & L_y^2 & L_y L_t \\ L_t L_x & L_t L_y & L_t^2 \end{bmatrix},
\]

where \( \sigma^2_i \) is an integrated spatial scale and \( \tau^2_i \) is an integrated time scale. \( \sigma^2_i \) and \( \tau^2_i \) is the local scale of space and time respectively, which all has a smooth effect. \( L_x, L_y, L_z \) represents the first-order partial derivatives in the x-axis, y-axis and z-axis respectively, and the expressions of which are defined as \( L_x = \frac{\partial}{\partial x}(g \ast f), L_y = \frac{\partial}{\partial y}(g \ast f), L_z = \frac{\partial}{\partial z}(g \ast f) \).

The local maximum value of the response is a corner point, and the value of the parameter \( k \) will affect the result.

\( \lambda_1, \lambda_2, \lambda_3 \) are the three eigenvalues of \( \mu \) respectively. \( \alpha = \lambda_2/\lambda_1 \) and \( \beta = \lambda_3/\lambda_1 \) are defined. Therefore, the response value can be derived as

\[
H = \lambda_1 \lambda_2 \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3
= \lambda_1^3(\alpha \beta - k(1 + \alpha + \beta)^3).
\]

Since the local maximum value of the response value is selected, the response value is set to be greater than zero first and \( k \leq \alpha \beta / (1 + \alpha + \beta)^3 \) can be derived. The response value is the largest when \( \alpha = \beta = 1 \), and the value of the parameter is 1/27 now.

Corner-based spatio-temporal point of interest detection algorithm has some limit, the reciprocating movement is more sensitive, for example, walking, running, waving and so on. However, for general rotational movement, the corners are not very sensitive and often do not detect the time–space interest points. In addition, the spatio-temporal corner method is equally insensitive to some very slow-moving behaviour. (Nazari, Sharif, & Mollaeefar, 2017; Nyeem, Boles, & Boyd, 2014). Furthermore, because the response function of the corner-based detection algorithm is not sensitive to the changes made in the time dimension, the number of space–time interest points detected on some videos is less. Considering the problems mentioned above, Dollar proposed a linear filter detection method. (Nyeem et al., 2014; Ridzoň & Levický, 2013; Tewari, Saxena, & Gupta, 2014).

2.2. The method of spatio-temporal interest points detection of linear filter based on Dollar

Dollar’s linear filter response function employs a Gaussian filter in the spatial domain and a one-dimensional Gabor
filter in the time domain. Compared with the time–space corner method, the algorithm can detect enough points of interest and show the human motion region better. There are false positives in these detected spatio-temporal points of interest. The response function of a one-dimensional Gabor filter in the time domain is:

\[ R = (f * g_0 * h_e)^2 + (f * g_0 * h_o)^2, \]  

where \( f \) is the value of a pixel in two-dimensional space; \( R \) is the response value of the point; \( g_0 \) is a two-dimensional Gaussian smoothing kernel in the space domain, and is expressed as \( g_0 = \frac{e^{-(x^2+y^2)/2\sigma^2}}{2\pi\sigma^2} \); \( h_e \) and \( h_o \) represent a pair of functions that are orthogonal to each other, establishing a one-dimensional Gabor filter, which plays a role in the time domain, and the expression is: \( h_e(t, \tau, w) = -\cos(2\pi tw)e^{-t^2/\tau^2} \), \( h_o(t, \tau, w) = -\sin(2\pi tw)e^{-t^2/\tau^2} \), \( w = 4/\tau \). The parameters \( \sigma \) and \( \tau \), respectively, represents the detection scale of the response function of the spatial and temporal domains. The spatio-temporal point of interest is the local maximum of the response value is \( R \).

3. The method of spatio-temporal features point detection of asymmetric causal filter

A digital image can be viewed as a graph network in which each pixel \( x \) is a node connected to the neighbour pixel by a resistance \( R \). The brightness \( u \) at each pixel is the charge on a capacitor \( C \). The flux of the brightness between each two neighbouring pixels depends on their relative brightness. (Nazari et al., 2017). We extend this model to a 3D space–time video signal by considering the effect of the potential of the corresponding pixel of the immediate past frame on the potential evolution at the current frame. This consideration satisfies the time causality and is modeled as an external conductance \( R_{ext} \). For simplicity, we show this modeling for a 1D + t signal in Figure 1 and derive the equations for this circuit. From Kirchhoff's Laws, we can derive the diffusion equation as the change of the brightness at a given (time-like) diffusion scale \( s \). The potential at this pixel is denoted by \( u(x; t; s) \) and the current by \( i(x; t; s) \).

\[ i(x, t; s) = (u(x, t; s) - u(x - dx, t; s))/R_1, \]  

\[ i(x + dx, t; s) = (u(x, t; s) - u(x + dx, t; s))/R_2, \]  

\[ i_{ext} = (u(x, t; s) - u(x, t + dt; s))/R_{ext}, \]  

\[ -C\frac{\partial u(x, t; s)}{\partial s} = i(x, t; s) + i(x + dx, t; s) + i_{ext}. \]  

Consider \( R_1 = R_2 = R \) and the per unit quantity, \( R = dx \), \( R_{ext} = \text{ext} dt \), \( C = dx \), and \( i_{ext} = i_{ext}(x, t; s)dx \). Substitute Equations 7–9 in the KCL Equation 10. At the limit \( dt \to 0 \) and \( dx \to 0 \), the KCL equation results in the spatio-temporal diffusion Equation 11.

\[ \frac{\partial u(x, t; s)}{\partial s} = \frac{1}{RC} \frac{\partial^2 u(x, t; s)}{\partial x^2} - \frac{1}{R_{ext}C} \frac{\partial u(x, t; s)}{\partial t}. \]  

The filter in this paper is composed of an asymmetric causal temporal filter and a Gaussian kernel spatial filter. The spatial filter adopts the Gaussian filter with the spatial Gaussian kernel \( G \) with standard deviation \( \sigma = \sqrt{2\sigma s} \). The temporal filter is composed of \( c \) and \( S(t) \), and the time scale of the temporal filter is \( \tau = \beta s \).

The \( S(t) \) denotes the Heaviside step function.

The expressions of the spatial filter \( G \) and the temporal filter \( k \) are as follows:

\[ G(x, y; \sigma) = \frac{e^{-(x^2+y^2)/2\sigma^2}}{2\pi\sigma^2}, \]  

\[ k(t; \tau) = \sin \beta (t - \tau)S(t). \]

To get a phase insensitive animation, we need a set of orthogonal kernels to capture all the phase information. Therefore, we can build a function \( k_h \) orthogonal to \( k \), which can be obtained by convolution of \( k \) with \( h(t) \), \( k_h(t; \tau) = k(t; \tau) * h(t) \), \( h(t) = \frac{1}{\pi t} \). To get spatio-temporal points of interest, \( R \) filter can be used to represent the space–time feature response value, and the local maximum of which is the space–time feature point sought. Express the filter as follows:

\[ R = (G * k * u_0)^2 + (G * k_h * u_0)^2, \]  

where \( u_0 \) represents the original video sequence value.

4. Video watermarking algorithm based on space–time feature points

Space–time feature point detection method based on asymmetric filter is used to extract feature points, which
shields the shortcoming of Harris corner detection being insensitive to motion. It is also possible to collect enough interest points by using linear filter space–time interest detection method, which is more in line with video characteristics because of the use of asymmetric time filters. (Heidari et al., 2017).

The execution of the proposed scheme are described in section 4.1, 4.2 and 4.3, respectively.

### 4.1. Extraction of feature points and selection of feature regions

According to the space–time feature point detection method of asymmetric filter to extract the feature points, enough points of interest can be obtained, and the space–time coordinates and the corresponding response values of the related points of interest can be obtained. (Laptev et al., 2007; Liu et al., 2015). Because using this method can get most point of interest and too many points of interest increase the difficulty of computation, a few interests with large response value are selected from all feature points of the video. As showed in the 10th frame in Figure 2, the 16th frame in Figures 3–5 show the feature frame points of large response values in Figures 3 and 4 respectively. Since the range of points of interest for which a large response value is selected is not in frames, the number of points of interest in each frame is uncertain, and there may be more than one or none.

For the selected point of interest, we need to discover the feature region to embed the watermark. To create the feature region, we need a certain area, but the edge of the point does not have enough area to embed the watermark. If the feature region created by the feature point of a frame coincides, select the region with larger response value.

![Figure 2](image1.png)  
*Figure 2.* Feature points in the 10th frame.

![Figure 3](image2.png)  
*Figure 3.* Feature points in the 10th frame with larger response values.

![Figure 4](image3.png)  
*Figure 4.* Feature points in the 16th frame.

![Figure 5](image4.png)  
*Figure 5.* Feature points in the 16th frame with larger response values.

### 4.2. Embedding the watermark

The repeated watermark is embedded in the generated feature area.
Transform the transformation of the domain

Suppose the watermark is a binary sequence \( m, m_j \in \{0, 1\} \). For each region, it is first transformed into the DCT domain. Then the DCT coefficients are reordered according to the ZigZag scanning order to form a sequence ranging from low frequency to high frequency. Then the mid-frequency coefficients are selected using a private key. The selected mid-frequency coefficients are segmented into non-overlapping blocks denoted as \( F \) to form a new matrix.

2. SVD

The new matrix \( F \) is decomposed using Singular Value Decomposition (SVD).

\[
F = UBV^H = \sum_{i=1}^{r} \lambda_i u_i v_i^H, \quad (15)
\]

where \( U = (u_1, u_2, \cdots, u_M) \in R \times R \), \( V = (v_1, v_2, \cdots, v_N) \in R \times R \) is a unit matrix, and \( UU^H = E \), \( VV^H = E \), \( E \) is an identity matrix, and \( H \) represents the conjugate transpose. \( B \) is a non-negative diagonal matrix of diagonal elements.

\[
B = \begin{pmatrix}
\lambda_1 \\
\lambda_2 \\
\vdots \\
\lambda_r \\
\vdots \\
0
\end{pmatrix}, \quad (16)
\]

where \( r \) is the rank of \( F \), \( \lambda_i = \sqrt{\sigma_i} \) is named singular value (SV), while \( \sigma_i \) is named eigenvalue.

The SV of the image represents the energy distribution, which is not easily changed, and the SV of the matrix has rotation and proportional invariance. The SV respectively reflects the internal characteristics of the image. The watermark embedding position can be determined according to the internal features, which are beneficial to the visual effect and make the watermark more concealed. The singular value of \( B \) decreases with the increase of the number of rows, so the SV of the first row can be chosen as the watermark embedding scheme. The robustness of the watermark is beneficial because the singular values of the matrix have rotation and proportional invariance.

3. Watermark embedding

We compute the Euclidean norm of SVs.

\[
N = \sqrt{\sum_{i=1}^{r} \lambda_i^2}, \quad (17)
\]

And the modification of the SV is (Hongbo et al., 2011):

\[
N = \begin{cases}
N - Q - \varepsilon/4 & m = 1, 0 \leq Q \leq \varepsilon/4 \\
N - Q + 3 \varepsilon/4 & m = 1, 3 \varepsilon/4 \leq Q \leq \varepsilon \\
N - Q + \varepsilon/4 & m = 0, 0 \leq Q \leq \varepsilon/4 \\
N - Q + 5 \varepsilon/4 & m = 0, 3 \varepsilon/4 \leq Q \leq \varepsilon 
\end{cases}, \quad (18)
\]

Suppose \( \varepsilon \) is suitable quantization value which is selected combined with HVS experimentally, \( Q = \text{rem}(N, \varepsilon) \) and \( m \) represent the watermark bit, where \( \text{rem} \) represents remaining operation, we change the norm according to (18). Let the biggest SV \( B(1,1) = N \), and we get new \( B^m \). Applying inverse SVD to \( B^m \), \( U \), \( V \), and we obtain \( F = UBV^H \). At this point, the watermark embedding within a feature area is completed.

For each non-overlapping block, repeat the step 2, 3 until the whole watermark is embedded into the host frame. Then perform the inverse ZigZag scanning and inverse DCT combined with low frequency and high frequency coefficients, and we get the watermarked frame. Repeat this for each frame containing STIP then the watermarked video will be obtained.

4.3. Watermark extraction and detection

Extracting the watermark is simple and it is a blind procedure. Watermark transforms the transformation of the domain, then the singular value decomposition. Specific steps are:

1. Find the frame containing the feature points, segment blocks and compute the Euclidean norm of SVs and remainder

\[
N_0 = \text{norm}(B_0^m), \quad (19)
\]

\[
Q_0 = \text{rem}(N_0, \varepsilon), \quad (20)
\]

2. Recover the watermark bit

\[
m_0 = \begin{cases}
1 & Q_0 > \varepsilon/2 \\
0 & Q_0 \leq \varepsilon/2
\end{cases}. \quad (21)
\]

5. Experimental results and analysis

For assessing the performance of the proposed algorithm, a variety of experiments were carried out. The block size for embedding is 80*80, and the length of the watermark is 128 bits, the threshold of the watermark is 86, as a result, the false alarm probability is \( 3 \times 10^{-6} \).

5.1. Invisibility test

According to the algorithm, we compare the host 10th frame and the watermarked 10th frame. The naked eye
is difficult to distinguish watermark traces. Objective evaluation index PSNR is nearly 31.4db, so the visual invisibility is significant (Figures 6 and 7).

### 5.2. **Robustness test**

Table 1 lists the results after various attacks on the watermarked frames.

| Attack            | Our scheme | Li’s scheme | Bi’s scheme |
|-------------------|------------|-------------|-------------|
| No attack         | 3/3        | 17/17       | 9/9         |
| Frame insert      | 2/4        | ×           | 5/11        |
| Frame dropping    | 3/3        | 7/14        | 7/10        |
| Gaussian noise    | 3/3        | 11/13       | 9/9         |
| Median filtering  | 3/5        | ×           | 8/11        |
| Gaussian filtering| 3/5        | 14/17       | 8/11        |

The 3D-Harris algorithm is fast, but it is insensitive to the time axis, and it may miss the detection of some slow-moving corners points, resulting in fewer detected feature points, which is not conducive to analysis. Gabor algorithm introduces time filter, and it will detect enough points, but its time filter does not match the actual video. Aiming at the shortage of time filter, an asymmetric time cause filter is introduced, and a novel video watermarking technique is proposed by combining the spatio-temporal point extraction method with the watermark in DCT domain. Experiments on video watermarking algorithm based on improved Gabor show that it has good invisibility and robustness.

### 6. Conclusion

The 3D-Harris algorithm is fast, but it is insensitive to the time axis, and it may miss the detection of some slow-moving corners points, resulting in fewer detected feature points, which is not conducive to analysis. Gabor algorithm introduces time filter, and it will detect enough points, but its time filter does not match the actual video. Aiming at the shortage of time filter, an asymmetric time cause filter is introduced, and a novel video watermarking technique is proposed by combining the spatio-temporal point extraction method with the watermark in DCT domain. Experiments on video watermarking algorithm based on improved Gabor show that it has good invisibility and robustness.

### Disclosure statement

No potential conflict of interest was reported by the authors.

### ORCID

Hongbo Bi [http://orcid.org/0000-0003-2442-330X](http://orcid.org/0000-0003-2442-330X)

### References

Darabkh, K. A. (2014). Imperceptible and robust DWT-SVD-based digital audio watermarking algorithm. *Journal of Software Engineering and Applications*, 07(10), 859–871.

Gaj, S., Rathore, A. K., Sur, A., & Bora, P. K. (2017). A robust watermarking scheme against frame blending and projection attacks. *Multimedia Tools and Applications*, 76(20), 20755–20779.

Gunjan, R., Mitra, P., & Gaur, M. S. (2012). Contourlet based image watermarking scheme using Schur factorization and SVD. *Advances in Communication, Network, and Computing*, 108, 337–340.

Heidari, M., Samavi, S., Soroushmehr, S. M. R., Shirani, S., Karimi, N., & Najarian, K. (2017). Framework for robust blind image watermarking based on classification of attacks. *Multimedia Tools and Applications*, 76(22), 23459–23479.

Hongbo, B., Yubo, Z., & Xueming, L. (2011). Video watermarking robust against spatio-temporal attacks. *Journal of Networks*, 6, 932–936.

Hosseini, S. A., & Ghofrani, S. (2014). Digital watermarking based on contourlet-SVD. *Majlesi Journal of Electrical Engineering*, 8(1), 63–68.

Kalra, G. S., Talwar, R., & Sadawarti, H. (2015). Adaptive digital image watermarking for color images in frequency domain. *Multimedia Tools and Applications*, 74(17), 6849–6869.

Laptev, I., Caputo, B., Schult, C., & Lindeberg, T. (2007). Local velocity-adapted motion events for spatio-temporal
recognition. *Computer Vision and Image Understanding, 108*(3), 207–229. doi:10.1016/j.cviu.2006.11.023

Li, L.-D., Guo, B.-L., & Pan, J.-S. (2008). Video watermarking by space-time interest points. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, E91-A*(8), 2252–2256. doi:10.1093/ietfec/e91-a.8.2252

Liu, Y., Wang, Y., & Zhu, X. (2015). Novel robust multiple watermarking against regional attacks of digital images. *Multimedia Tools and Applications, 74*(13), 4765–4787.

Nazari, M., Sharif, A., & Mollaeefar, M. (2017). An improved method for digital image fragile watermarking based on chaotic maps. *Multimedia Tools and Applications, 76*(15), 16107–16123.

Nyeem, H., Boles, W., & Boyd, C. (2014). Digital image watermarking: Its formal model, fundamental properties and possible attacks. *EURASIP Journal on Advances in Signal Processing, 2014*(1), 1–22.

Ridzoň, R., & Levický, D. (2013). Content protection in grayscale and color images based on robust digital watermarking. *Telecommunication Systems, 52*(3), 1617–1631.

Shabani, A.-H., Clausi, D., & Zelek, J. S. (2011). Improved spatio-temporal salient feature detection for action recognition. 100.101–100.100. doi:10.5244/c.25.100

Tewari, T. K., Saxena, V., & Gupta, J. P. (2014). A digital audio watermarking scheme using selective mid band DCT coefficients and energy threshold. *International Journal of Speech Technology, 17*(4), 365–371.

Tong, M., Chen, T., Zhang, W., & Dong, L. (2012). New video watermark scheme resistant to super strong cropping attacks. *Journal of Information Security, 03*(2), 138–148.

Zhou, W., Jiang, G., Yu, M., Wang, Z., Peng, Z., & Shao, F. (2014). Reduced reference stereoscopic image quality assessment using digital watermarking. *Computers & Electrical Engineering, 40*(8), 104–116.