ABSTRACT

A novel video watermarking system is presented that uses a spread transform technique to securely embed data into a video sequence, resistant to both spatial and temporal attacks. The system makes use of the 3-D wavelet domain, improving embedding by extending this to include complex as well as discretely sampled wavelet transforms. To ensure imperceptibility an adaptation of a proven video visual model is used. Results are given for both compression and frame dropping attacks.

1. INTRODUCTION

The aim of video data hiding is to securely and imperceptibly embed a watermark message consisting of a string of bits, into a video sequence.

Originally video watermarking schemes viewed the host video as simply a succession of individual frames and embedded the watermark using algorithms mainly designed for still image watermarking. Not only are such schemes computationally intensive but they are also vulnerable to desynchronisation attacks. More recent algorithms [1,2,11] have instead taken advantage of the presence of the temporal dimension and used the 3-D discrete wavelet domain to embed the data.

However the discrete wavelet domain suffers from ‘checker board’ artifacts that can considerably reduce the visual quality of the watermarked video. This problem gets more acute as the number of dimensions to which the transform is applied increases. To combat this it is proposed that 3D complex wavelets should be used instead, as these have the advantage of improved directionality and do not suffer from artifacts like the 3D DWT.

Embedding has been performed by using both spread spectrum [1] and quantization methods [2]. The advantages of both these methods can be combined in what is known as spread transform, Kingsbury et al. [6] have used this algorithm for video watermarking but only in the 2-D domain on individual frames.

2. 3D WAVELETS

Due to the 3D nature of video data it is proposed that 3D wavelets should be used instead of 2D wavelets. The advantages are a reduction in the time required for embedding and more secure embedding by taking advantage of the temporal dimension available in video data.

2.1 3D DWT

The 3D discrete wavelet transform is well documented in literature. A 2D spatio-temporal DWT is performed on each individual frame in the video sequence to give LL, HL, LH and HH subbands. A 1D temporal DWT is then performed across the frame subbands to obtain 3D subbands LLL, LLH, LHL, LLH, LHH, HLL, HHL, HHL and HHH. For multiple levels the process is repeated recursively on the LLL subbands.

2.2 3D DTCWT

The low redundancy version of the dual tree complex wavelet (DTCWT) [9] is used which has a redundancy of 2:1 for 2D signals and produces two diagonal subbands, 2 horizontal subbands and 2 vertical subbands. The 3D DTCWT is performed by taking 4 DWTs acting in parallel upon the same data. It has 4:1 redundancy and so produces 4 times as many subbands as the 3D DWT.

The advantage of the 3D DTCWT is that it separates diagonal features and is free from the checkerboard artifacts of the 3D DWT.

2.3 3D NRCWT

To overcome the redundancy of the DTCWT Fernandes et al. [7] have developed a non-redundant complex wavelet that separates the diagonal features to give 2 diagonal subbands in addition to 1 horizontal and 1 vertical subband.

The filterbank consists of two filters, a real filter NCWTR that is applied to real input to produce a real output and 2 complex outputs. As the two complex outputs are conjugates one can be disregarded. The second filter NCWTC is complex and acts upon complex valued input to produce 3 complex outputs. As a result of the triple output from the filters downsampling at each stage is by 3 rather than 2. The 2D stage form of the NRCWT produces 4 different directional, complex valued subbands. Two diagonal feature subbands, one horizontal and one vertical.

The 2D implementation of [7] is extended to 3D data for use in video watermarking using the filterbank shown in figure 1, creating 13 highpass complex subbands at each decomposition level.
3. VISUAL MODEL

The visual model used is an adaption of that used in [1], a JND (Just Noticeable Distortion) weight is calculated for each coefficient in the wavelet decomposition where $n_x,n_y$ and $n_z$ are the horizontal, vertical and temporal locations of the coefficient respectively. The JND value gives a measure of the maximum distortion that can be applied to the corresponding coefficient without creating a visible distortion.

The jnd value is obtained through a combination of 3 factors, the brightness, the subband noise and the contrast sensitivity function (CSF).

3.1 Luminance Masking

It is well documented in the literature that distortions in bright and dark areas of an image are less perceptible than those in areas of middling brightness. To take this into account the JND profile incorporates a bidimensional luminance mask developed by Chou [4]. The suitability of this model for still image wavelet based watermarking has already been demonstrated by Ghoui [8]. The mask is calculated on the normalized lowpass subband of the lowest level of decomposition.

The lowpass subband is normalised so every value lies in the range [0,256]. The further away from 127 the lowpass coefficient corresponding to the area under consideration is the higher the corresponding luminance factor will be. Distortions in darker areas are even more imperceptible than those in bright areas and so low brightness values will produce the highest luminance factor.

\[
LM(n_1,n_2,n_3) = \begin{cases} 
\left(1 - \frac{LLL(n_1,n_2,n_3)}{127}\right)^{1.7} & \text{if } LLL(n_1,n_2,n_3) < 127 \\
\frac{3}{128} [LLL(n_1,n_2,n_3) - 127] + 3 & \text{if } LLL(n_1,n_2,n_3) \geq 127 
\end{cases}
\]

3.2 Noise Masking

It is also well documented that distortions are less visible in textured regions, edges and areas of high temporal activity. In [1] Campisi et al. calculate the mask using the lowpass subband each time to save computation time. However as our proposed algorithm operates at a lower level the mask can be calculated using the highpass subband under consideration each time. In addition this allows for greater exploitation of the improved directionality provided by complex wavelets. It also allows for the factoring in of spatial or temporal noise only dependent on subband.

For temporal lowpass subbands regions of high noise will correspond to edges and texture regions in the video frames. For temporal highpass subbands areas of high noise will correspond regions in the video where large spatial shifts are occurring.

The mask is calculated using a 3x3x3 gaussian window of variance 0.5 applied to the subband under consideration to obtain \( N_\theta(n_1,n_2,n_3) \).

3.3 Contrast Sensitivity Function

The spatio-temporal CSF defined in [5] attempts to measure the amount of activity betweenand is restated here for convenience:

![Figure 1 - 3D NRCWT filterbank](image-url)
\[ \text{CSF}(\rho, v_R) = k x_0 c_2 v_R \left( \frac{c_1 \rho}{2\pi} \right)^2 \exp \left( -\frac{c_1 \rho}{\pi \rho_{\text{max}}} \right) \] (2)

Where

\[ k = s_1 + s_2 \log(c_2 v_R / 3) \] (3)

\[ \rho_{\text{max}} = p_1 / (c_2 v_R + 2) \] (4)

\[ v_R = v_j - \min(g_{\text{op}} v_j + v_{\text{MIN}}, v_{\text{MAX}}) \] (5)

Where \( \rho \) is the horizontal spatial frequency in c/deg, \( v_R \) is the retinal velocity in deg/sec, \( v_{\text{MIN}} \) the drift velocity of the eye (0.15 deg/sec), \( v_{\text{MAX}} \) the maximum velocity beyond which the eye can not track movements effectively (80 deg/sec) and \( g_{\text{op}} \) is the gain of the smooth pursuit eye movements set equal to 0.82. The constants are set as follows: \( s_1 = 6.1 \), \( s_2 = 7.3 \), \( p_1 = 45.9 \), \( c_0 = 1.14 \), \( c_1 = 0.67 \) and \( c_2 = 1.7 \).

The CSF can then be calculated using equation 2. Finally, the wavelet subband under consideration is discrete Fourier transformed to obtain \( \text{DFT}_\theta(f_1, f_2, f_3) \) with \( f_1, f_2, f_3 \) representing the horizontal, vertical and temporal frequencies respectively. The CSF is then used to weigh each of the frequencies as follows:

\[ \chi_\theta = \sum_{k_2} \left( \sum_{k_1} \left| \text{DFT}_\theta(f_1, f_2, f_3) \right|^2 \cdot \text{CSF}(f_1, f_3) \right) \] (6)

The obtained value is then used to globally scale the JND of the subband under consideration.

\[ jnd_\rho(n_1, n_2, n_3) = \chi_\theta^{-1} \cdot \text{LM}(n_1, n_2, n_3) \cdot N_\rho(n_1, n_2, n_3) \] (7)

Where LM and N are the luminance and noise factors obtained of the individual coefficients of the decomposition.

### 4. PROPOSED ALGORITHM

The proposed algorithm makes use of spread transform watermarking to embed data into the host media. Spread transform is an extension of basic quantization methods where quantization takes place in a reduced dimensional space [3]. Spread transform techniques aim to combine the high capacity provided by quantization with the high robustness provided by spread spectrum. In basic quantization methods quantization is applied to individual host samples individually, in spread transform it is a vector projection of the host samples that is quantized to carry individual watermark bits.

The vector to be quantized is calculated as shown in (8) using the host samples x and a key dependent vector v where r is the length of the vector.

\[ x^{\text{ST}} = \sum_{n=1}^r x_n v_n \] (8)

The vector projection is then quantized using one of two quantizers \( Q_i() \) where \( i \in [0,1] \) and the step size \( \Delta \) is adapted according to the size of the JND values corresponding to the embedding vector. The quantization operates as in (9)

\[ x^{\text{ST}} = Q(x^{\text{ST}}; \Delta / \alpha) + (1 - \alpha)(x^{\text{ST}} - Q(x^{\text{ST}}; \Delta / \alpha)) \] (9)

The value \( \alpha \) is a distortion compensation factor that substitutes accuracy of embedding for larger quantisation bins. In this paper \( \alpha \) is set to 1. To obtain the host samples x the video sequence is divided into scenes using histogram comparison as in [11]

\[ D(k, k + 1) = \sum_{j=1}^n \frac{\left| H_k(j) - H_{k+1}(j) \right|}{H_k(j)} \] (10)

Where H is the histogram of the frame j and k the position of the frame. The value D measures the amount of change in subsequent frames and when it rises above a certain threshold value a cut is declared. This ensures that neighbouring vector components are similar to each other and helps combat the effects of temporal desynchronisation. Each scene is then further partitioned into shots of 32 frames each and 36 for the NRCWT due to the trand structure of its filterbank.

The 3D wavelet decomposition is then performed upon the luminance band of each shot and the watermark bits embedded. For the DWT and DTCWT the decomposition is done to 3 levels and 2 levels for the NRCWT due the trand structure of its filterbank which downsamples by 3 at each level rather than by 2. This coarse level embedding is done not only to save on computation time but also to make the embedding more robust against common forms of video compression.

Embedding is performed on all 3rd level subbands for the DWT and DTCWT apart from the low pass subband which is neglected to maintain imperceptibility of the watermark. The embedding in both temporal low pass and high pass subbands is consistent with the notion stated in [11] that watermarks should be embedded in both the spatial and temporal domains for resistance against both temporal and spatial attacks. For the NRCWT the same process is performed upon the 2nd level coefficients. The coefficients that make up the vectors for embedding are grouped together based on their spatial and temporal proximity to each other within each subband.

At the receiver the watermark is decoded by again calculating the vector projection as in (8), the received bit b is calculated as follows:
Figure 2 - (a) Original Frame (b) DWT distortion (c) DTCWT distortion (d) NRCWT distortion

\[ \text{err}_i = \left\| Q_i(x^{ST}) - x^{ST} \right\| \]  

\[ b = \begin{cases} 0 & \text{if } \text{err}_0 < \text{err}_i \\ 1 & \text{otherwise} \end{cases} \]  

Watermarks from the luminance colour band of a frame of the ‘foreman’ sequence are shown in figure 2, multiplied by 20 for clarity and added to a grey level image.

5. RESULTS

A variety of common test video sequences were used of frame size 352x288 and sequence length 384 frames (351x288 frames and sequence length 396 frames in case of 3D NRCWT). The host video is split up into individual shots and the luminance colour band is watermarked. The sequence is then subjected to compression and frame dropping attacks. The video is then re-divided into shots. Bit error rate (BER) results from across the test sequences are then obtained. The watermark payload is 600 bits in all cases. In order to ensure that the watermark distortion is imperceptible the Video Quality Metric (VQM) model is used. The VQM attempts to measure the subjective quality of the video rather than using more simplistic objective measures that do not take into account the ability of the HVS to detect distortion in video sequences. A VQM value of 0.2 which is accepted as being imperceptible to the naked eye is used.

To illustrate the improved HVS adaptability of the complex wavelet transforms table 1 lists the average per pixel distortion from individual frames when embedding is limited by a VQM value of 0.2. This objective measure gives an indication of how much more effective embedding power is provided by the complex wavelets.

<table>
<thead>
<tr>
<th>Transform</th>
<th>Average Distortion</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D DWT</td>
<td>0.1399</td>
</tr>
<tr>
<td>3D DTCWT</td>
<td>0.1653</td>
</tr>
<tr>
<td>3D NRCWT</td>
<td>0.1549</td>
</tr>
</tbody>
</table>

Results are compared with the 3-D DCT QIM algorithm of [12] in the case of all considered attacks.

Figure 3 shows BER results after MPEG2 compression has been applied to the video sequence for values from 4500 kbit/s to 500 kbit/s.

Figure 4 shows the BER results after MPEG4 compression has been applied to the video sequence for values from 2500 kbit/s to 500 kbit/s.
6. CONCLUSION

As expected embedding the payload in both the spatial and temporal features at a coarse level renders the embedded data highly robust. It is able to withstand both compression attacks in the form of MPEG2 and MPEG4 compression as well as showing improved levels of performance against temporal de-synchronisation attacks such as frame dropping.

In addition the improved directionality of the complex wavelets allows for much better adaptation of the watermark to the host signal. However despite this the 3D NRCWT performs worse than both the 3D DWT and 3D DTCWT. This inferior performance for the frame dropping attack in particular can be attributed to the larger proportion of coefficients in the 3D NRCWT decomposition that represent temporal features.

In conclusion the results demonstrate the improved video watermarking robustness and visual quality offered by the 3D dual tree complex wavelet transform combined with spread transform embedding. In addition, by embedding in both coarse spatial and temporal subbands high degrees of robustness to both compression and temporal de-synchronisation can be obtained.

REFERENCES


