Robust Information Hiding in Low-resolution Videos with Quantization Index Modulation in DCT-CS Domain

Huimin Zhao[a,b], **Qingyun Dai**[b,c], **JC Ren**[b,d], **Wenguo Wei**[b,c], **Yinyin Xiao**[a,b] and **Chunying Li**[a,b]

[a] School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China

[b] The Guangzhou Key Laboratory of Digital Content Processing and Security Technologies, Guangzhou, China

[c] School of Electronic and Information, Guangdong Polytechnic Normal University, Guangzhou, China

[d] Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, U.K.

Abstract—Video information hiding and transmission over noisy channels leads to errors on video and degradation of the visual quality notably. In this paper, a video signal fusion scheme is proposed to combine sensed host signal and the hidden signal with quantization index modulation (QIM) technology in the compressive sensing (CS) and discrete cosine transform (DCT) domain. With quantization based signal fusion, a realistic solution is provided to the receiver, which can improve the reconstruction video quality without requiring significant extra channel resource. The extensive experiments have shown that the proposed scheme can effectively achieve the better trade-off between robustness and statistical invisibility for video information hiding communication. This will be extremely important for low-resolution video analytics and protection in big data era.

Index Terms—Video information hiding; Statistical transparency; Compressive sensing; Quantization index modulation; Image Fusion.

1 INTRODUCTION

Video information hiding, especially for low-resolution ones, is an important secret communication technology, where significant amount of secure data is invisibly embedded inside a video carrier signal via digital watermarking and can only be retrieved by those authorized [1]. This process can be applied in several applications including the integrity of the carrier signal, copyright protection and multimedia security. To characterize the performances of information hiding systems, three major criteria are often utilized, which include capacity, robustness and imperceptibility (or transparency) [4]. These respectively refer to the amount of hidden information, the effectiveness of the technique, and any degradation to the visual quality of the original signal. For specific applications, the overall performance usually needs be tuned as a trade-off among these three criteria [2]. With increasing amount of hidden data, the robustness of the hidden signal is improved, yet the hidden signal may become visible and results in degraded imperceptibility. To maintain a certain level of imperceptibility, the capacity is limited, thus the robustness can also be affected.

To enable as much as possible data hidden without noticeably decreasing the imperceptibility, QIM-based techniques and the Scalar Costa Scheme (SCS) are widely used [3][5]. Without using a secret key, the SCS scheme shows lack of security due to the non-statistical transparency while the hidden signal is embedded. As a result, potential attackers will be informed on the presence of the hidden data, where the host videos can be deliberately targeted for illegal usages. To tackle this problem, video information hiding with good statistical transparency is demanded.

Actually, a number of watermarking based approaches with good statistical transparency have been proposed in the last decade. In general, these can be classified into two groups, i.e. quantization based and feature based approaches. In quantization based approaches, a specific quantization scheme is

applied to the signal to be embedded, using techniques such as Fractal based quantization [6] and Trellis Coded Quantization (TCQ) [7]. These quantization techniques help to remove unwanted noise in the probability density function (PDF) of the watermarked signal, caused by the use of a scalar quantizer such as the SCS. As a result, the imperceptibility can be significantly improved at the cost of degraded robustness. Feature based approaches usually work in a transform domain, such as Independent based schemes [8] and the Spread Transform [9]. Due to successfully redundancy removal, the robustness can be improved though the capacity is constrained, especially for high statistical transparency. As can be seen, these two groups of approaches work better in different aspects, and they actually complement to each other. As a result, they can be potentially fused together, and this forms the major motivation of our proposed approach.

Due to the strong capability in significantly reducing the data yet preserving the information of the signal, CS based watermarking has attracted increasing attention in recent years [11, 12, 13]. In Zhao et al [11], distributed CS for secure signal processing in the cloud is discussed. In Zhang et al [12], combining CS and compositive reconstruction is used for watermarking with self-recoverable quality. In Wang et al [13], CS based framework for secure watermarking detection and privacy preserving storage is proposed. Other CS based approaches include Sheik et al. [14] and Zhao et al. [15]. However, statistical transparency is seldom addressed in existing work.

In this paper, we propose a hybrid approach for video information hiding, based on Quantization Index Modulation (QIM) in the DCT-CS Domain. In the proposed approach, DCT and QIM are respectively used for feature extraction and quantization, where compressive sensing (CS) techniques [10] are employed to obtain a sparse representation of the host signal before embedding the watermarking data. With the secure watermarking principle in [16], we evaluate the statistical transparency using **PSNR** (Peak Signal-to-Noise Ratio) [17], the **KLD** (Kullback-Leibler Divergence) [18], and PDF (Probability Density Function). We also measure several evaluation metrics for general watermarking and signal fusion, including robustness and perceptual transparency. Tests with real video sequences are carried out for performance assessment, where it has validated that our proposed DCT-CS methodology can help to maximize the imperceptibility of the watermarking whilst reaching a good tradeoff between robustness and statistical transparency.

The rest of this paper is organized as follows. Section 2 describes the problems of classical data hiding scheme based on Costa's theory. Based on the relation work of CS theory, section 3 introduces the video compressive sensing fusion (VCSF) Scheme proposed in this paper. Section 4 reports the experimental results for the VCSF in video watermarking, and demonstrates the effectiveness of the proposed the scheme. Finally, we give our conclusion in section 5.

2 Related Problems for Information Hiding Communication Scheme

The Costa's quantization based watermarking scheme [4, 5, 20] herein is used as the baseline of our work. In this section, after briefly summarizing the general concepts of this scheme, we will discuss its drawbacks to motivate the proposed approach.

2.1 Costa's Scheme

For a host signal X from a video document with the power σ_X^2, we aim to embed a hiding message $M = [m_1, \cdots, m_n]$ into X. Denote the encoded watermark signal as W (with the power σ_W^2) and the final fused signal as S, we have $S = X+W$. After transferring the fused signal over a

noisy communication channel (modeled as Gaussian noise with power σ_Z^2), the received signal is $R = S + Z$. To extract the embedding information $m_i \in M$ from R, Costa's approach is used to estimate the hidden signal $\bar{M}$ as illustrated in Fig. 1.

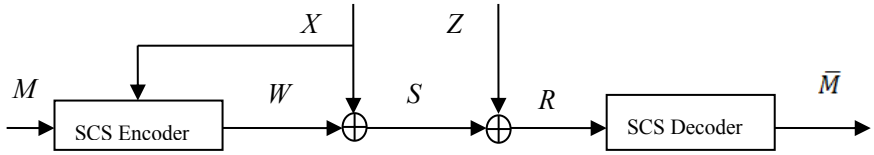


Fig 1. SCS signal fusion scheme

In Fig.1, most of the SCS compression standards quantize the source data for better coding efficiency. This quantization step can be used for embedding data. Depending on the value of the data to be hidden, different quantizers are used in quantizing the source data. In the receiver side, the data hidden can be extracted by determining which quantizer is used. One of the most popular approach which utilizes this idea is Quantization Index Modulation (QIM) [21].

2.2 Analysis of SCS Scheme

The host video signal X contains N components $\{xi\} \in X$, $i \in [1,N]$, which can be in either spatial or transform domain. According to the approach in [21], QIM based data hiding can be processed as follows, where the bit b of $\{xi\}$ is adaptively modulated by using the hidden signal vector $\{mi\} \in M$. The signal $w_i \in W$ is watermarked by:

$$w_i = x_i + \alpha_i m_i \ , \ i = 1, 2, \ldots, n \tag{1}$$

In the SCS signal fusion scheme, the fused signal S has two parts, i.e. $S=X+W$, and the goal of the statistical transparency is to reduce the chance for potential attackers to recognize whether the signal is a fused one or not. This can be represented as the following two hypothesis: H1 (the signal is fused with a PDF of p_S) and H0 (the signal is not fused with a PDF of p_X), where PDF denotes probability density function of the signal.

The two hypotheses mean the false alarm probability P_{fa} can be maximized by considering $P_{fa} = P_r(H_1 \mid H_0)$, i.e. to maximize the probability for the attacker to erroneously believe the video signal is not watermarked. According to the Stein's Lemma [4, 20], the Kullback-Leibler Divergence (KLD) is used for measuring the channel noise Z.

$$D(p_X \parallel p_S) = \int_{-\infty}^{+\infty} p_X(z) \ln \frac{p_X(z)}{P_S(z)} d_z \propto -\ln P_{fa} \tag{2}$$

Herein maximizing P_{fa} is equivalent to minimizing $D(p_X \parallel p_S)$ between the fused signal S and the original signal X. As the perfect statistical transparency with $p_S = p_X$ is hard to achieve, an $\varepsilon_{\text{secure}}$ system with $D(p_X \parallel p_S) < \varepsilon$ is defined for approximation [18]. In our work, we will use the determined KLD for evaluation of the statistical transparency. Smaller the divergence is, higher the P_{fa} will be and more likely the attacker will be confused to decide whether the signal is a fused one or not.

In the SCS signal fusion scheme, the main difficulty is the statistical transparency of the hidden information in the fused signal S. As shown in Fig 2, we consider 300 real images extracted from two videos, where the original signal X and the fused signal S show noticeable difference in terms of their statistics PDFs. It has been proven in [20] that this difference is caused by the discontinuity appearing in S when the regularity of the scalar quantization is used. These discontinuities are the evidence of the presence of the watermarking signal embedded in the original signal, which in turn may inform potential attackers the need to remove the hidden signal before any illegal usage. Herein our proposal is

to reduce or remove this discontinuity and reach a good statistical transparency to be validated using *PSNR* and other evaluation metrics.

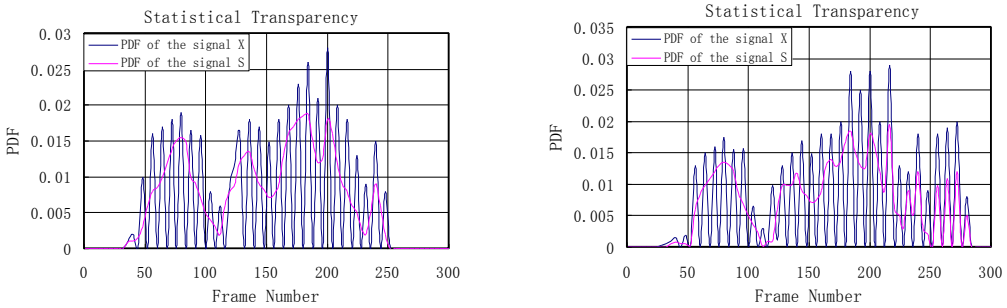


Fig 2. PDFs of the original signal *X* and the fused signal *S* in SCS for Basketball video (left) and Scene video (right).

3 VIDEO COMPRESSIVE SENSING FUSION (VCSF) FRAMEWORK

3.1 Related Works for CS

In [22], [23], and [24], the CS asserts that when a signal can be represented by a small number of non-zero coefficients, it can be perfectly recovered after being transformed by a limited number of incoherent, non-adaptive linear measurements. Suppose a signal $X \in R^N$ is a *K*-sparse vector (only *K* out of the *N* elements of *f* are nonzero), and can be transformed to $Y \in R^M, M < N$, where $Y = A \cdot X = \Phi \cdot \Psi \cdot f$, *Y* is an *M*×1 sampled vector, A is a sensing matrix, and Φ is an $M \times N$ measurement matrix that is incoherent with Ψ, Ψ is called as the sparse matrix, respectively. For images, typical choices of Ψ include the DCT and DWT. If Ψ satisfies RIP (Restricted Isometry Property), [23] shows solving the bellow optimization problem

$$\min \| X \|_1 \qquad s.t. \qquad X = \Psi \cdot f \tag{3}$$

This is equivalent to finding the sparsest solutions to $X = \Psi \cdot x$, provided that $M \geq CK \log(N/K)$, where *C* is a small constant. The CS theory states that such a signal *X* can be reconstructed by taking only *M* linear projection.

Equation (3) presents an *l*1 minimization problem which can be solved by orthogonal matching pursuit algorithm [25]. It has been shown that it is feasible for many signal processing algorithms to be performed in the CS domain [26], and [27]. If the entries of matrix Φ are generated from a Gaussian distribution with zero mean and variance $\sigma \in 1/\sqrt{M}$, Φ is a RIP matrix with overwhelming probability [23]. In our framework, the matrix Φ is generated from such a Gaussian distribution by reference previous work [16], and [19]. The Gaussian CS matrix suits include the seeds and a random function.

In practical application, an image with a size $N = N_1 \times N_2$ is divided into $B \times B$ blocks, and each block is sampled using an appropriately-sized measurement matrix Φ_B in the CS domain. Let x_i be a sparse vector representing *i*th block of the input image $X \in R^M, M < N$, the corresponding measurement sample y_i is determined by:

$$y_i = A_B \cdot x_i = \Phi_B \cdot \Psi f \tag{4}$$

where the length of the signal y_i is *M*, and A_B is a $M \times B^2$ measurement matrix. The size of m_i

is $\lfloor M \cdot B^2 / N \rfloor$ and M is the number of samples needed by the CS measurement for the whole image. In this way, $\boldsymbol{A}$ has a block-diagonal structure as follow:

$$A = \begin{bmatrix} A_B & 0 & \dots & 0 \\ 0 & A_B & \dots & 0 \\ 0 & \dots & A_B & \dots \\ 0 & 0 & 0 & A_B \end{bmatrix} \tag{5}$$

Therefore, the aforementioned approach was called block CS (BCS) [28].

3.2 Proposed VCSF′ Scheme

In our proposed scheme, a measurement signal Y is obtained by compressing the host signal X in the CS domain, where the watermarked signal W is decided using the same procedure as previously described in the ***Costa***'s encoder process yet based on the determined sparse signal M. By applying the QIM method in the DCT-CS domain, the watermarked signal W is generated by $W = \alpha \cdot Q = Y + M$ with $Q = Q_\Delta[\Delta(Y) - m_i] + MOD[\Delta(Y) \oplus m_i]$, where α is the ***Costa***'s robustness optimization parameter. Note that only the non-zero coefficients will be considered in quantizing the signal Y in the CS domain. By adding the determined sparse watermarked signal W to Y, the fused signal $S = Y + W$ is obtained by the VCSF encoder in the DCT-CS domain as shown in Fig 3. This fused signal S is then transmitted along with the additive Gaussian white noise (AGWN) Z over the channel before being sent to the receiver side.

At the receiver side, the received signal with noise is represented as $R = S + Z$, which is compressed by using the same sensing procedure in DCT-CS domain as in the embedding step. From the received signal R, the watermarked signal is extracted to estimate the hidden message $\overline{M}$ in VCSF decoder in Fig 3. In the next several subsections, relevant details are presented for the DCT-CS operations used in our proposed VCSF's scheme, where we first obtain the signal sparsity properties and then the inverse procedure to recover the original signal.

3.2.1 Principle of the VCSF in Encoder

By exploiting the sparsity structure of the signal, CS enables sampling beyond the ***Shannon*** limit [22],[23], and [24]. As a result, signals can be acquired and represented in CS at a significantly lower rate than the ***Nyquist*** rate in conventional solutions. Non-adaptive linear projection is used for fast sampling whilst preserving the signal structure. From these projections, sparsity regularized convex optimization is employed for decoding and signal reconstruction [24]. The CS theory affirms that the original signal can be reconstructed from much fewer measurements than conventional wisdom, though the performance is mainly affected by sparsity and incoherence of the original signal.

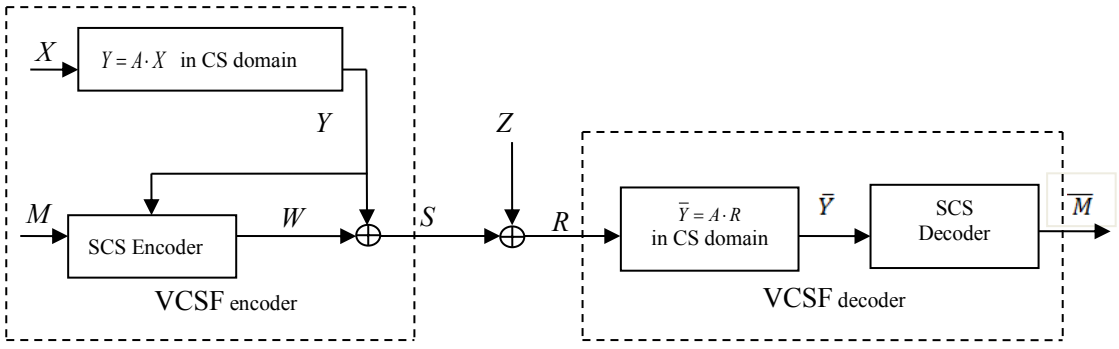


Fig 3. Proposed VCSF scheme in DCT-CS domain

As shown in Fig 3., let $\boldsymbol{X}$ be the input signal from a video stream, a sparse signal in VCSF encoder. Considered as a vector in a finite-dimensional subspace R^n, $X = \{x_1, \dots, x_n\}$ can be proven strictly

or exactly sparse if most of its entries are (very close to) zero, i.e. its support $\Lambda(X) = \{i \in [1,n], x_i \neq 0\}$ is of cardinality $k << n$. A signal $\boldsymbol{X}$ that contains exactly k non-zero valued samples is defined as k-sparse. A non-spare signal $\boldsymbol{X}$ can be made sparse via certain sparse transforms, i.e. in a transform domain. In our application, the original signal $\boldsymbol{X}$ is the frames from a video document f, which is generally not sparse. By using a specific spare base Ψ , we can obtain a sparse signal $\boldsymbol{X}$ as $X = \Psi \cdot f$. In our paper, we choose the sensing matrix $\boldsymbol{A}$ as the Restricted Discrete Cosine Transform (*RDCT*) matrix with *RDCT* $(X)= Y = A \cdot X = \Psi\Phi \cdot X$, where Ψ is a sparse basis and Φ is a DCT measurement matrix which can be determined by multiscale block compressed sensing (BCS) matrix Φ_B in the CS domain [16]. Finally, the fused signal S is generated as shown in Fig. 4 by integrating the sensing signal $Y = [y_1, \cdots, y_M]$ of all sensing matrix A_B and the hidden signal $W = [w_1, \cdots, w_m]$.

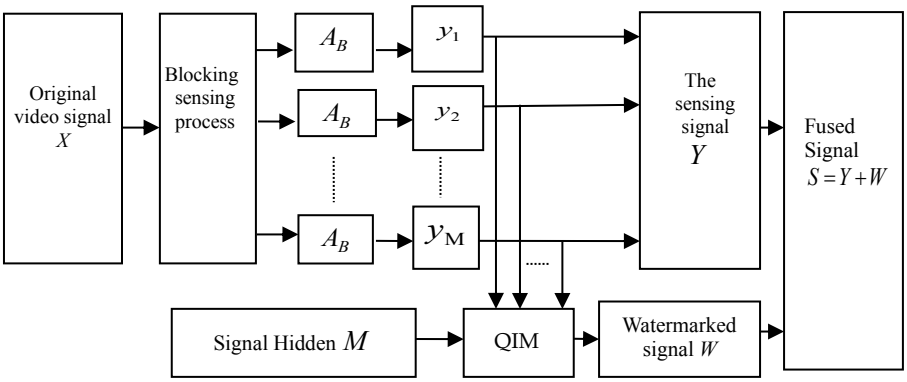


Fig 4. Process of the multi- dimensional signal fusion in VCSF encoder

3.2.2 Fusion Processing of the Signal in DCT-CS Domain

In video coding, the video sequence is first divided into groups of pictures (GOP) of I-frame, B-frame, and P-frame [29, 30]. According to the video sequence characteristics, the B-frame and P-frame are dependent on the I-frame and the raw video data can also be considered as a sequence of still images. As a result, the watermarking signal $m_i \in M$ and the host CS signal $y_i \in Y$ are mainly fused into the luminance component of each I-frame. We extract I frame image by syntactic elements of the video stream [29] before sensing its DCT coefficients using the matrix $\boldsymbol{A}$ to select a suitable area for hiding the fusing signal in the CS domain.

For I-frames extracted from videos in Fig. 4, we take several successive frames as a group, where each image within the group is divided into a number of blocks. Each block is transformed into the DCT domain and separately sensed by using the measurement matrix in CS domain. This process is illustrated in Fig. 5, where four consecutive I-frames are treated as a group. With the CS values obtained in the DCT-CS domain, the sum of the CS values and the watermarking weights is fused as follows.

$$\begin{aligned} y(i,j,k) &= A_B(i,j,k) \cdot x(i,j,k) \\ w(i,j,k) &= \mathrm{y}(i,j,k) + \alpha \cdot m_{i,j,k} \\ s(i,j,k) &= \textstyle\sum_{i,j,k}[y(i,j,k) + w(i,j,k)] \end{aligned} \tag{6}$$

where $\alpha \epsilon [0,1)$, $i \epsilon [1, \mathrm{n}]$, and $s(i,j,k) \epsilon \boldsymbol{S}$ is the sum of all sensing CS values $\mathrm{y}(i,j,k)$, and $w(i,j,k)$ is the watermarking weights. Herein $\mathrm{y}(i,j,k) \epsilon Y$ and $w(i,j,k) = \alpha Q \epsilon \mathrm{W}$ respectively denote the kth CS sensing value and the kth modulated message corresponding to the jth block of successive frames within the ith frame group. Here, the initial weighting value α can be decided by the owner of the source signal. Repeating the above process for all blocks over each frame within the same group, a fused sequence of sums of every block will be obtained. Fig.6 shows an example where each of the

four frames in a group is separated into a few 8×8 sized blocks.

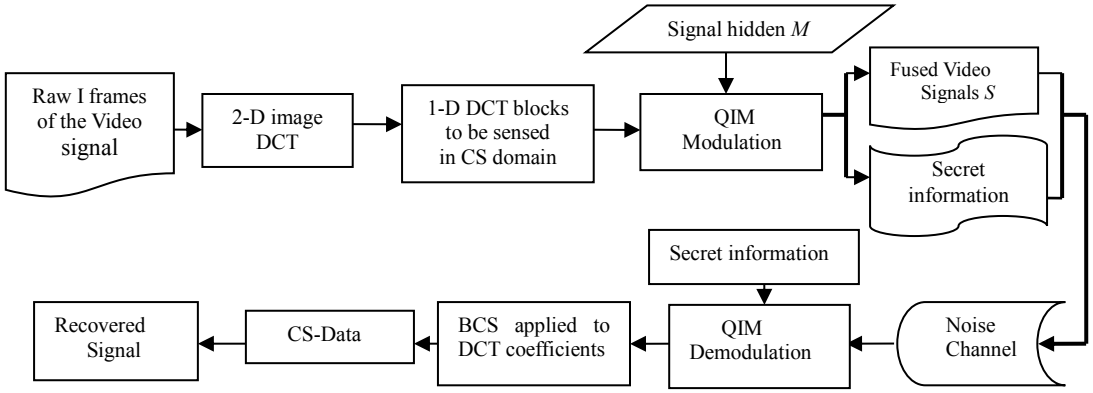


Fig.5. Sketch of the video fusion procedure in DCT-CS domain

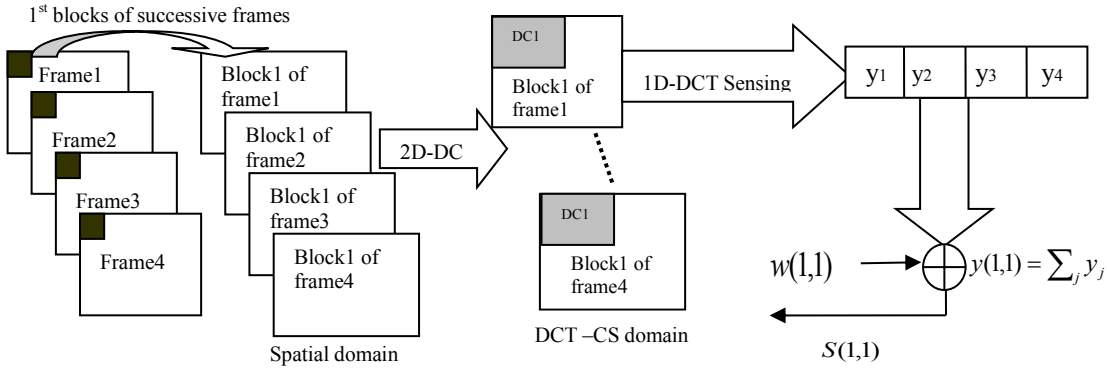


Fig.6 The fused process worked on the signal $S(1,1)$ in DCT-CS domain

After computing $s(i,j,k)$ for all blocks, we use a determined threshold $T(i)$ to balance the robustness and transparency of the hiding messages. $T(i)$ is associated with the characteristic of the input video and the number of bits to be embedded, and it can be adjusted by the weighting value $0 \le \alpha$ <1 depending on the demand and the trade-off between robustness and transparency. Based on the approach in [18], $T(i)$ is determined as follows.

$$T(i) = R + \frac{512}{R \times \log(\frac{512}{R}) \times \log(Max_{sum}(i))} \tag{7}$$

$$R = B \cdot \log(1 + \frac{W}{Y}) \tag{8}$$

when the bandwidth B is decided, the reception variable bit rate R can be estimated to further derive $T(i)$, where $Max_{sum}(i) = \sum_j y_j$. The threshold $T(i)$ can be regarded as the tolerance range for the quantizing process in QIM.

After determining the $T(i)$ above, an integer quantized quotient is obtained as follows:

$$Q(i,j,k) = \left\lfloor \frac{S(i,j,k)}{T(i)} \right\rfloor \tag{9}$$

where $Q(i,j,k)$ decides the quantization step Δ while utilizing the QIM to perform the embedding operation. Based on the QIM, the embedding domain is divided into several regions. The interval of every region is the same, which equals to $T(i)$, and an index is assigned to every region. Therefore, each region represents a bit (0 or 1) of the watermarking data. For the robustness of the signal hidden, the value of the fusion signal $s(i,j,k)$ is changed to the median value in the corresponding section to

resist for the distortion embedded.

As $s(i,j,k)$ consists of several signals, the modification of it by bit embedding is equal to the change of the CS values. Also note that the low frequency component is more robust and visually more sensitive than the high frequency component. As a result, modulating low frequency component will cause more serious distortion, though it can be more robust to resist attacks than modulating the high frequency components. To this end, under noise $\boldsymbol{Z}$, we apply the CS matrix A to fuse the estimated signal $\bar{s}(i,j,k) \in S$ as follows.

$$\bar{s}(i,j,k) = \sum_{i,j,k} [w(i,j,k) + y(i,j,k) + z(i,j,k)] \tag{10}$$

where $z(i,j,k) \in$ Z denote the k-th noise value corresponding to the j-th block of successive frames within the i-th group, which can be adjusted by the channel noise power σ_Z^2. According to $Q(i,j,k)$ and the k-th bits to be embedded, we can obtain $\bar{s}(i,j)$ by using QIM for $w(i,j)$ modulating $y(i,j)$ as follows.

While $\bar{s}(i,j) = \sum [y(i,j) + w(i,j)]$

$if \quad Q(i,j) = 2p \qquad m_{i,j,k} = 0$

$otherwise \qquad m_{i,j,k} = 1$

end

While $\bar{s}(i,j) = \sum [y(i,j) + w(i,j) + z(i,j)]$

$if \quad Q(i,j) = 2p \qquad m_{i,j,k} = 1$

$otherwise \qquad m_{i,j,k} = 0$

end

where $m_{i,j,k}$ is the embedded values of the watermarked signal W, and $m_{i,j,k} \in \{0,,1\}$ denotes the embedded bit of every block of the hidden signal from the modulated samples in DCT-CS domain. The parameter p is a random nonnegative integer determined the selection of quantizer with a step size Δ.

For a noisy channel, it is easy to know whether $\bar{s}(i,j,k)$ needs be changed or not when $Q(i,j,k)$ is an even or odd number corresponding to the hiding signal $m_{i,j,k}$. After determining the embedding position, the original value of $x(i,j,k) \in X$ in that position can be modulated to the median value of the corresponding section by using $m_{i,j,k}$. This procedure will repeat until all hidden bits are embedded. Finally, the sensing matrix $\boldsymbol{A}$, quantizers $Q(i,j,k)$, the threshold $T(i)$ and all the embedding positions are recorded as the secret information of the embedding key.

Usually, the QIM is independent of the video signal, which therefore may lead to serious degradation to visual quality. However, in our scheme, the conventional QIM algorithm is adjusted with some small amount of the CS signal $y(i,j,k)$ being modulated by using the hidden signal $m_{i,j,k}$ in the video stream. For other CS signal $y(i,j,k)$, they can be changed by using (6) and (10) to maintain a good visual quality and avoid severe distortion.

3.3 Principle of the VCSF in Decoder

3.3.1 Recover of Signal Hidden

In the decoder side of VCSF, the process of signal extraction is actually the inverse operation of embedding. First, the received video sequence is separated into groups of frames and each frame is further divided into blocks. The secret information is then applied to acquire the embedding positions. After determining the embedding blocks, the selected DCT blocks $r(i,j,k) \in$R rather than all blocks are transformed into the DCT-CS domain. By applying sensing matrix $\boldsymbol{A}$ to all selected DCT blocks

$\bar{y}(i,j,k) = A_B(i,j,k) \cdot r(i,j,k)$, we can further obtain an estimated original value $\bar{y}(i,j,k)$ from $\bar{r}(i,j,k)$, which is the sum of *k*th values of *j*th blocks in fused stero-frames within the *i*th group. Then we compute the quotient $\bar{Q}(i,j,k)$ derived from $\bar{r}(i,j,k)$ divided by the threshold $T(i)$, which is recorded in the secret embedding information. After computing $\bar{Q}(i,j,k)$, we can exactly decide which bit in $\bar{y}(i,j,k)$ is embedded as follows.

$$\bar{m}_{i,j,k} = \begin{cases} 0, & if \quad \bar{Q}(i,j,k) = 2p \\ 1, & otherwise \end{cases} \tag{11}$$

By repeating the above steps, the embedded bit $\bar{m}_{i,j,k}$ can be exactly ensured one by one until all bits are extracted. Finally, if $\bar{M}$ denotes estimated signal of the embedded $\bar{m}_{i,j,k}$, $\bar{M} = [\bar{m}_1, \bar{m}_2, ..., \bar{m}_n]^T$ can be recovered using the secret key of video transmission.

3.3.2 Recover of Original Signal with Min-TV Criterion

Assume $\bar{X}$ is an estimated value of original signal *X* from $\bar{Y}$ in the VCSF decoder as shown Fig. 3, *X* is homogenous to a scaled quantization error given by

$$Q = Q_\Delta[\Delta(\bar{Y}) - m_i] + MOD[\Delta(\bar{Y}) \oplus m_i], i = 1, ..., n \tag{12}$$

For the sensing signal $\bar{Y} = RDCT(\bar{X})$ and $\bar{Y} = A_B \cdot \bar{X}$, based on the property that the gradient for natural image generally follows a heavy-tailed distribution [31],[32], we consider the Total Variation (TV) optimization problem such as min- TV(*W*) subject to $\bar{X}$=*IRDCT*($\bar{Y}$) where *IRDCT*($\bar{Y}$) stands for inverse *RDCT*($\bar{Y}$) and

$$TV(\bar{X}) = \sum_{ij} \sqrt{|D_{h;ijk}\bar{Y}|^2 + |D_{v;ijk}\bar{Y}|^2} \tag{13}$$

$$D_{h;ijk}\bar{Y} = \begin{cases} \bar{y}_{i+1,j,k} - \bar{y}_{i,j,k} & if \quad i,j < n \\ 0 & if \quad i,j = n \end{cases} \tag{14}$$

$$D_{v;ijk}\bar{Y} = \begin{cases} \bar{y}_{i,j+1,k} - \bar{y}_{i,j,k} & if \quad i,j < n \\ 0 & if \quad i,j = n \end{cases} \tag{15}$$

For the aforementioned min-TV criterion, it can be solved efficiently with a much reduced computational cost in the CS domain [33], and [34]. Eventually, we can obtain an estimated signal $\bar{X}$ with a reasonably high PSNR over 40dB in the receiver side. In addition, the normalized correlation (*NC*) of structural similarity coefficient (*SSIM*), defined in Section 4, is computed between *M* and $\bar{M}$, and its high value ($NC \geq 0.98$) also confirms the good perceptual transparency performance for the reconstruction of hidden signal [35]. As a stopping criterion, we apply cross validation [36] to predict the *TV* performance.

4 EXPERIMENTS AND RESULTS

To test and verify the performance of the proposed scheme in which the CS procedure was applied to every image of the video document. The experimental results are compared with *SCS*'s method in [5] and Huang's 3D-DCT method in [21] to perform various attacks, including MPEG compression, noise contamination, and collusion attack.

In the experiments, a binary symmetric channel (BSC) is simulated in order to observe the effects of channel bit errors on the bit stream. For the experiments, two test sequences, *Baskedball* and *Scene* are encoded by an MPEG codec in various bit rates and passed through the BSC for two different channel bit error rates (BER). The *Basketball* sequence includes highly textured areas and some various

motions, which provides high frequency also in temporal domain. On the contrary, the *Scene* sequences contain low motion and smooth regions yielding with smaller number of DCT coefficients when compared to the *Basketball* sequence. These variations from the test sequences will ensure a thorough evaluation of the proposed approach. In our experiments, the test sequences are coded in six different bit rates and then transmitted through the BSC with two different BERs at 10^{-4} and 10^{-5}, respectively.

The visual reconstruction quality is determined in terms of Peak Signal-to-Noise ratio (*PSNR*) between the original signal $\boldsymbol{X}$ and the received signal $\boldsymbol{R}$. This process is repeated 100 times with different random seeds for the CS sensing matrix pattern and average reconstructed *PSNR* is calculated for the luminance and chrominance components of each frame. The *PSNR* can clearly judge the every received frame image quality by comparing the degree of diversity between the received signal *R* and the original one *X*. The mean square error (*MSE*) and $PSNR = 20\log(H_{\max} / \sqrt{MSE})$ are used for quantitative evaluation:

$$MSE = \frac{1}{I_h \times I_v} \sum_{i=1}^{I_h} \sum_{j=1}^{I_v} \sum_{k=1}^{n} | r(i,j,k) - x(i,j,k) |^2 \tag{16}$$

where $H_{\max}$ is 255 gray value for a gray-level image; $r(i,j,k) \in R$ and $x(i,j,k) \in X$ denote the received *k*th video data and the original one corresponding to *i* and *j* coordinates in 3D space. I_h and I_v denote the height and width of the video frames, respectively.

The performance of statistical transparency of the test sequences are showed by using the PDF of the original video signal *X* and the fused video signal *S*. For secure signal, robustness is also one of important performances in video information hiding communication. In our experiments, a measurement of the normalized correlation (*NC*) used for calculating the difference between the extracted the watermarking signal $\overline{M}$ in VCSF decoder and the original watermark signal *M* in VCSF encoder. The *NC* is defined as

$$NC = \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} m(i,j) \cdot \overline{m}(i,j)}{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} [m(i,j)]^2} \tag{17}$$

where $N = N_1 \times N_2$ denotes the fingerprint size of the watermark image, and N_1 and N_2 are the height and width of one.

4.1 Experiment Dataset

In our experiment, the hidden signal (watermarking message) *M* is generated by measurements values of sensing matrix $\boldsymbol{A}$ for a gray level fingerprint image with the size 160×160 from database FVC 2008. Our ideal of fingerprint image used for the original watermark is to explore a new secure application of e-commerce with the principle of biological recognition in information security field [37]. For example, watermarking data of fingerprint images can be used to secure central databases from which fingerprint images are transmitted on request to intelligence agencies in order to use them for identification and classification purposes.

For the host video signal *X,* experimental frames are extracted from the videos for testing. The size of each frame is 720×480 in the video, and every video sequence consists of 300- 382 frames. We take fifty successive frames as a GOP (group of picture) and each frame is decomposed into several

8×8 non-overlapping blocks and each block is individually and compressively sensed by using a DCT measurement matrix in the CS domain [16]. We consider in this experiment that the sender transmits each frame where each block is compressively sensed with m (m/n=50%) measurements to the transcoder.

4.2 Experiment Results

4.2.1 The Statistical Transparency of the Scheme

The PSNR values in the "***original***" plot belong to the frames of the video encoded by baseline MPEG-2 codec. The bit stream created by the baseline VCSF and SCS as well as 3D-DCT encoder is passed through the BSC and decoded by the corresponding baseline decoder again for 100 times, respectively. The average reconstructed *PSNR* values are labeled as "***VCSF***", and "***SCS***" and "***3D-DCT***", respectively. The reconstructed *PSNR* value versus frame plots for luminance component only are given in the figures from Figs. 7-10 for *Basketball* and *Scene*, respectively.

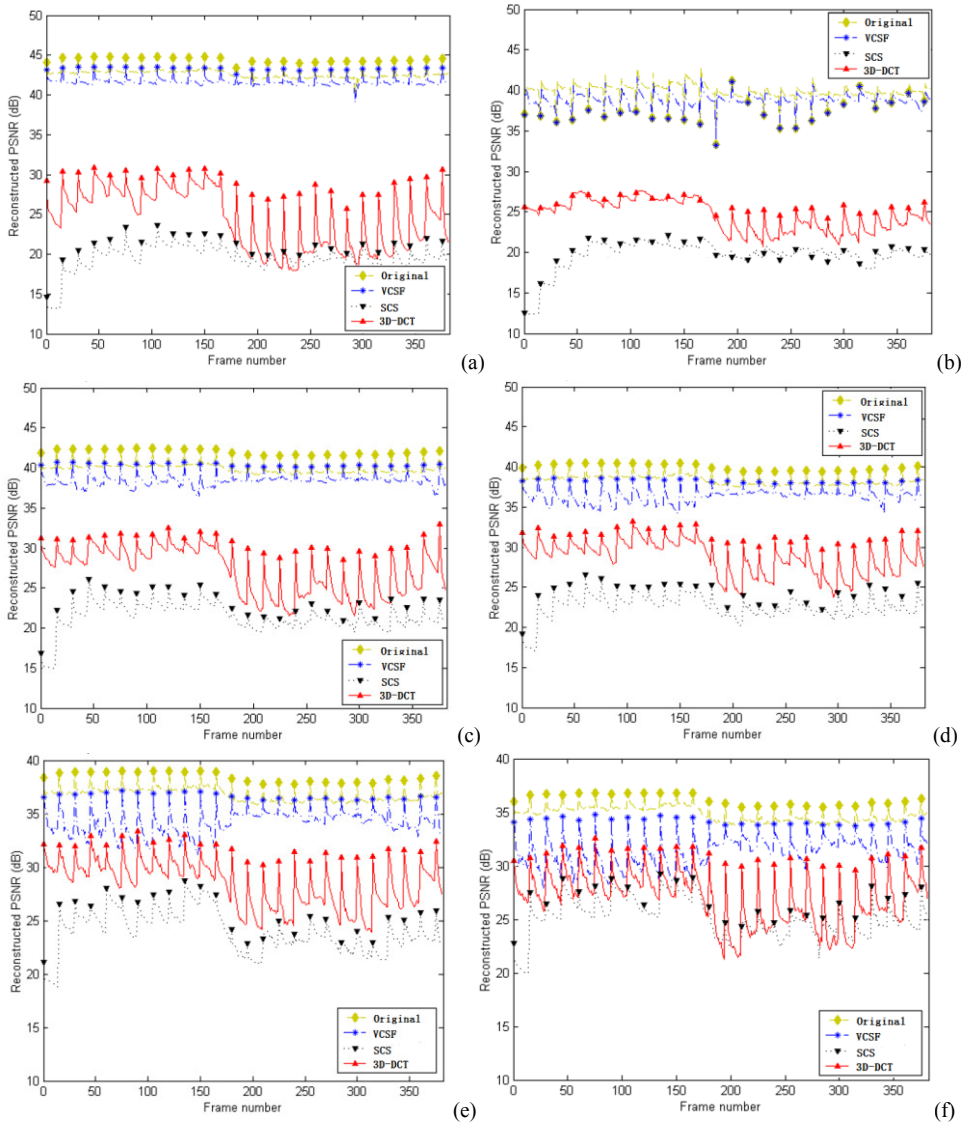


Fig.7 Performance comparison of the proposed ***VCSF*** system with the baseline ***SCS*** and ***3D-DCT*** codec for 382 frames (1intra per 20 frames) *Basketball* sequence at the BER of 10^{-4} and the bit rates of (a) 2 Mbit/sec, (b)1.5M bit/sec, (c) 1.2Mbit/sec,

(d)1.0M kbit/sec, (e) 800 kbit/sec and (f) 600kbit/sec.

For each test video we plot two curves for comparison, one for BER 10^{-4} and one for BER 10^{-5}. In each figure there are four plots, corresponding to six different bit rates including 2M, 1.5M, 1.2M, 1.0M, 800k and 600k bps. In our experiment, the proposed system shows better performance in terms of high bit rates and improved BER due to two reasons. First, more CS measurements will result in higher bit rates of the modulated signal and thus a proportionally small decrease in *PSNR* during data hiding in comparison to the lower bit rates. Second, the small number of errors in low BER decreases the *PSNR* of VCSF slightly, however, the *PSNR* level is still no less than those from "3D-DCT" when the BER is 10^{-5}.

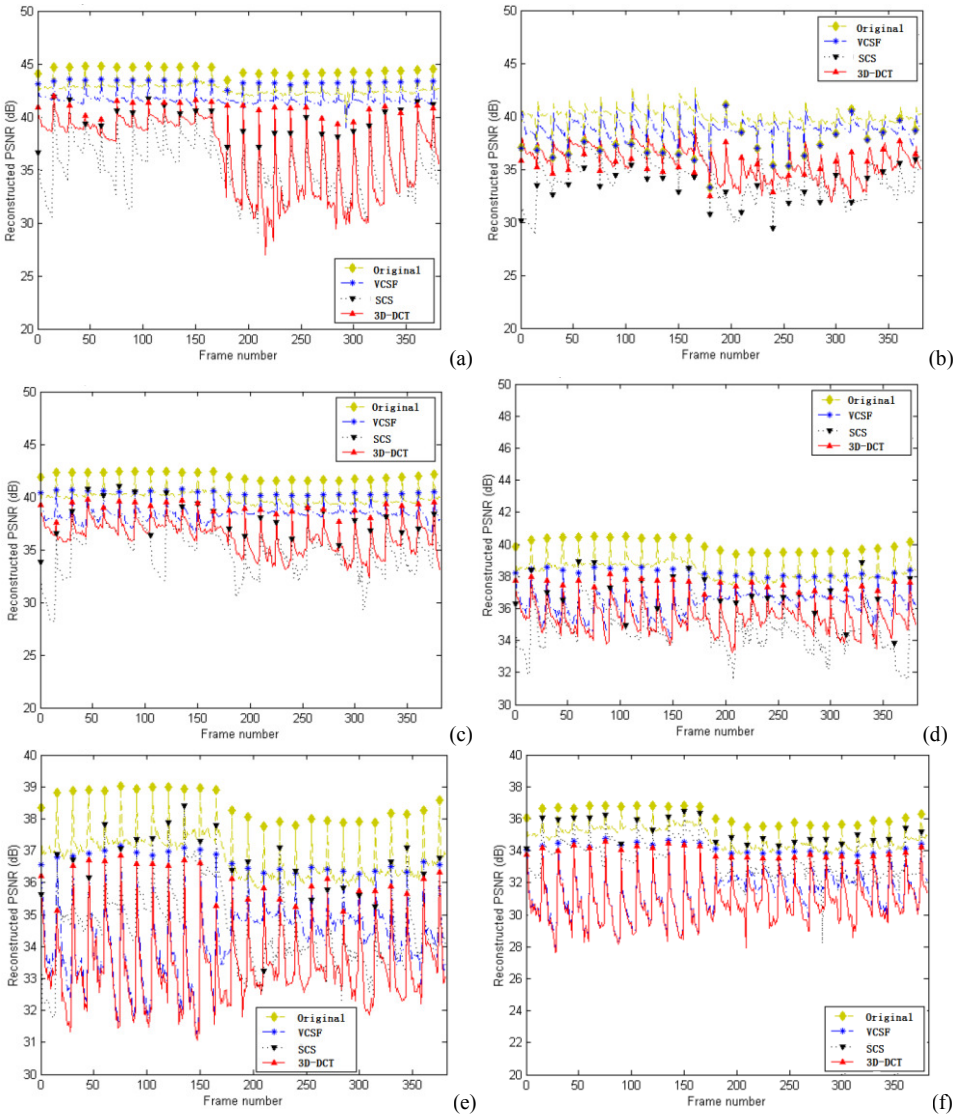


Fig.8 Performance comparison of the proposed ***VCSF*** system with the baseline ***SCS*** and ***3D-DCT*** codec for the 382frames(1intra/20frames) *Basketball* sequence at the BER of 10^{-5} and at the bit rates of (a) 2 Mbit/sec, (b)1.5M bit/sec, (c) 1.2Mbit/sec, (d)1.0M kbit/sec, (e) 800 kbit/sec, (f) 600kbit/sec.

To further analyze the performance in terms of statistical transparency, for both the basketball and scene videos, the average differences in terms of MSE between the PDF of 300 frames of the fused signal *S* and the original signal *X* are compared in Table 1. In comparison to the same technique for the SCS in Fig. 2, the watermarked signal from our approach is widely closer to the non-watermarked one, i.e. less distortion is introduced. By computing the KLD, we are able to find that the document-to-watermark ratio (DWR)

obtained from our proposed ***VCSF***'s approach is generally lower than those from *SCS* and ***3D-DCT*** as shown in Fig 11. This has confirmed again the improved statistical transparency of the proposed method where we have $\varepsilon = 6 \times 10^{-3}$.

Table 1: Comparison of MSE between the PDF of the fused signal S and the original signal X

	SCS	Proposed Approach
Basket ball sequence	1.7%	0.2%
Scene sequence	1.5%	0.14%

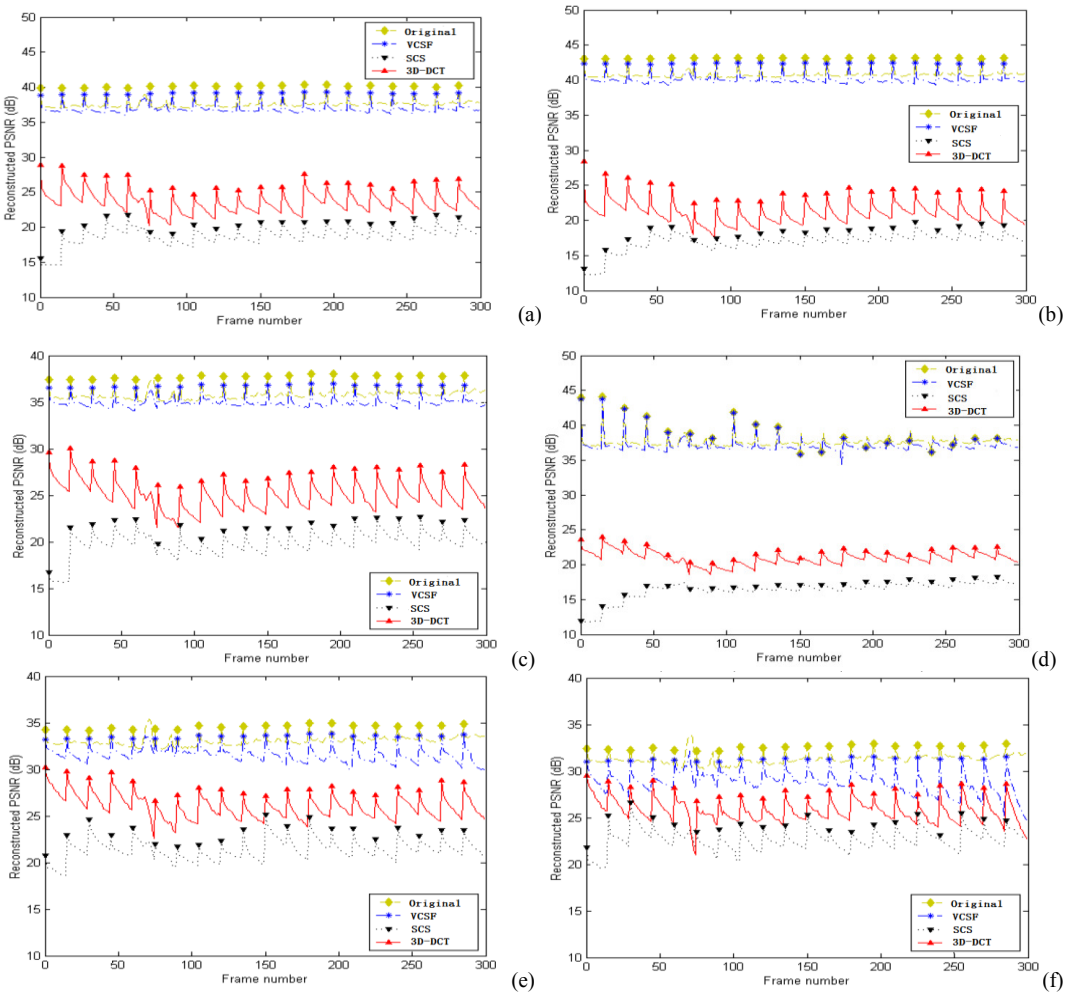
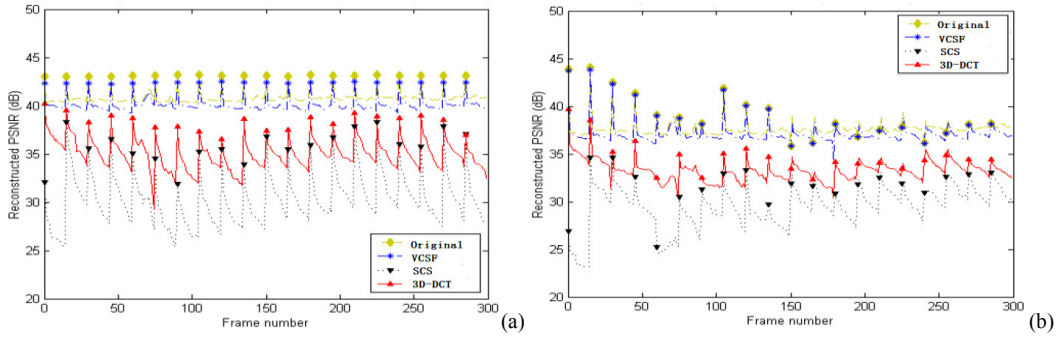


Fig.9 Performance comparison of the proposed ***VCSF*** system with the baseline ***SCS*** and ***3D-DCT*** codec for the 300frames (1intra/20frames) *Scene* sequence at the BER of 10^{-4} and at the bit rates of (a) 2 Mbit/sec, (b)1.5M bit/sec, (c) 1.2Mbit/sec, (d)1.0M kbit/sec, (e) 800 kbit/sec, (f) 600kbit/sec.



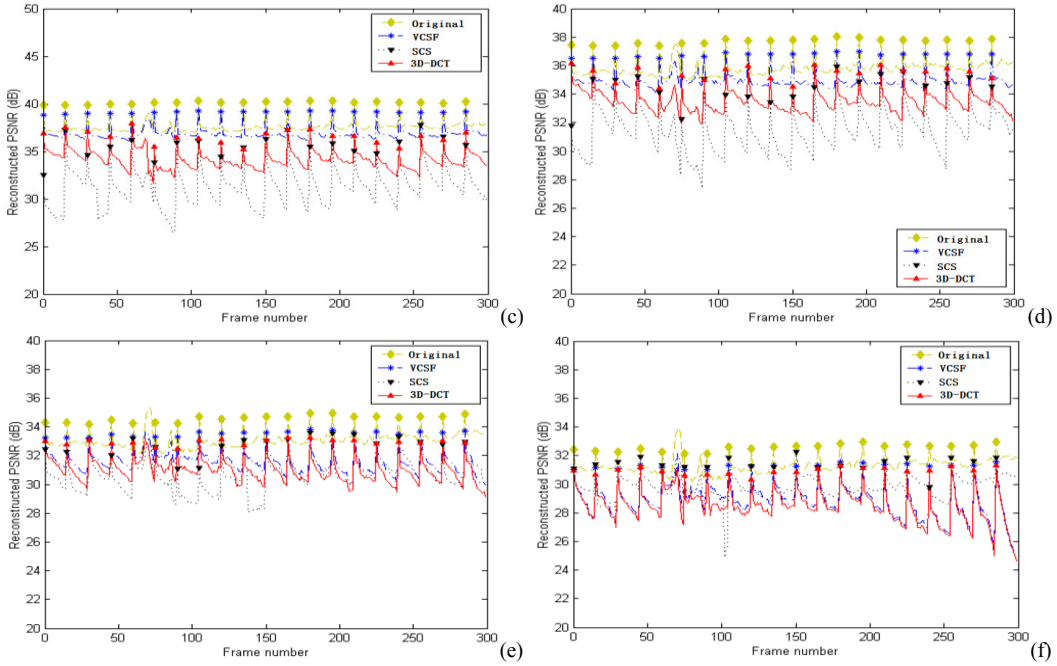


Fig.10 Performance comparison of the proposed ***VCSF*** system with the baseline ***SCS*** and ***3D-DCT*** codec for the 300 frames(1intra/20frames) *Scene* sequence at the BER of 10^{-5} and at the bit rates of (a) 2 Mbit/sec, (b)1.5M bit/sec, (c) 2Mbit/sec, (d)1.0M kbit/sec, (e) 800 kbit/sec, (f) 600kbit/sec.

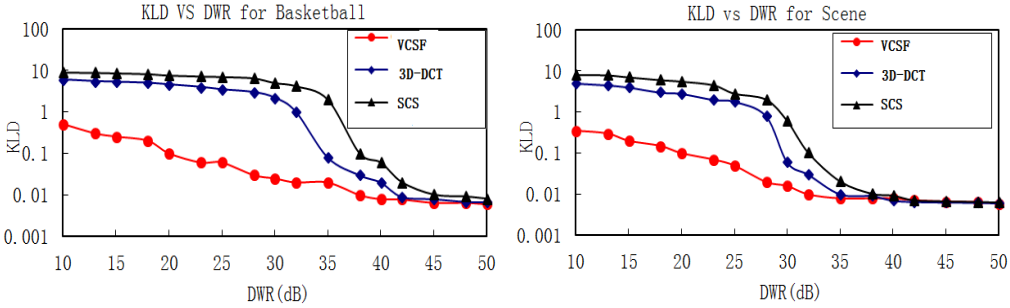


Fig.11 KLD comparison with the proposed VCSF's scheme and SCS's as well as 3D-DCT's ones at the BER of 10^{-5} and at the bit rates of 2 Mbit/sec for Basketball video (left) and Scene video (right).

4.2.2 The Robustness of the Signal Hidden

In order to study the robustness of the signal hidden in the information hiding communication, we consider various intentional or unintentional attacks, such as Gaussian noise (with zero mean and variance 0.05) and wiener filter and median filter attacks to demonstrate the performances of the proposed approach. For the Scene video, Fig.12 present the experimental results of the signal hidden recovered by using the Min-TV Criterion for VCSF and SCS as well as 3D-DCT after various attacks. From these results, we can find that no matter what the attacks are, the *NC* values of the fingerprint image of signal hidden recovered from our proposed ***VCSF*** scheme can still exceed 0.98, and the fingerprint image can reconstruct with higher quality than the methods of SCS and 3D-DCT. In other words, the ***VCSF*** proposed in this paper have a better ability to resist various attacks. In this case, we take fully advantage of CS values in which a signal can be retrieved with a high probability by using a relatively small number of measurements.

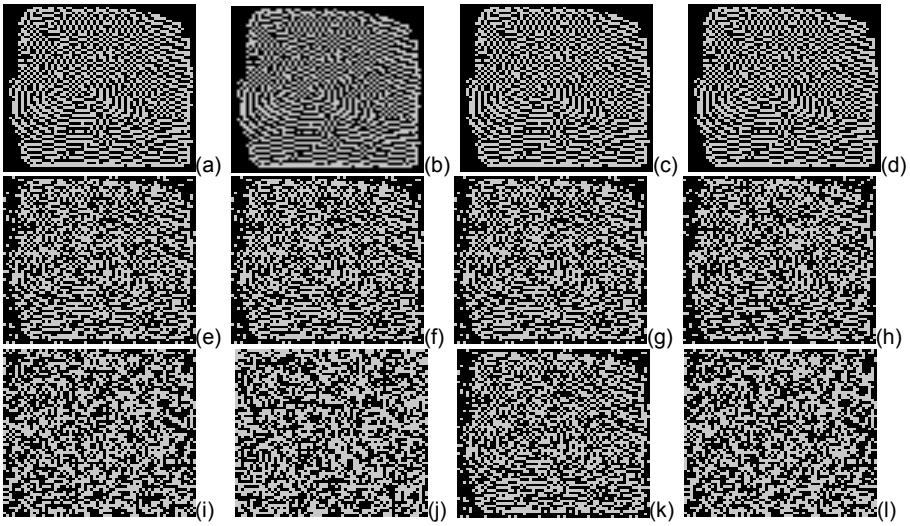


Fig.12 Comparison result against different attack for 2 Mbit/sec Scene video at the BER of 10^{-5}. From top to down, the three rows are results from VCSF, 3D-DCT and SCS approaches. From left to right, the four columns respectively are for results under attacks of winner filtering, median filtering, Gaussian noise and Pepper&salt noise. The NC values for (a-l) are 0.998, 0.990, 0.999, 0.980, 0.769, 0.778,0.878, 0.782, 0.615, 0.665, 0.777, and 0.651, respectively.

5. CONCLUSIONS

In this paper, a hybrid approach for robust video information hiding and communication is proposed. With Quantization Index Modulation (QIM) and Discrete Cosine Transform based feature extraction and quantization, the compressive sensing (CS) techniques have effectively obtained a sparse representation of the host signal before embedding the watermarking data. Using real video test sequences, we have demonstrated the efficacy of the proposed approach in terms of statistical transparency, robustness and perceptual transparency. The proposed VCSF approach can effectively resist attacks such as compressions, noise, filtering attacks, and can maintain a good tradeoff between statistical transparency and robustness. Also it is found that the proposed approach can help to maximize the imperceptibility of the watermarking in VCSF. Future work will involve further study on CS related information forensics and secure communications, including visual perception [38], deep learning [39], frequency domain processing [40] even for embedded implementation [41] and intrusion detection [42]. Another direction is to address hashing based techniques for data hiding, such as dictionary-based hashing [43], probability based hashing [44], machine learning based hashing [45] and even for hardware implementation of fingerprint generation [46].

Acknowledgment

This work was supported by the National Natural Science Foundation of China (61672008), Guangdong Provincial Application-oriented Technical Research and Development Special fund project (2016B010127006, 2015B010131017), the Natural Science Foundation of Guangdong Province (2016A030311013,2015A030313672), and International Scientific and Technological Cooperation Projects of Education Department of Guangdong Province (2015KGJHZ021, 2017A050501039).

REFERENCES

[1] I. J. Cox, M. L. Miller, and J. A. Bloom, Digital Watermarking. San Mateo, CA: Morgan Kaufmann, 2001.

[2] S. Voloshynovskiy, F. Deguillaume, S. Pereira, and T. Pun, “Optimal adaptive diversity watermarking with channel state estimation,” in Proc. SPIE Security and Watermarking of Multimedia Contents III, Bellingham, WA, vol. 4134, pp. 23–27, SPIE Press, 2011.
[3] B. Chen, G.W. Wornell, Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. IEEE Trans. Inform. Theory, vol. 47, no. 4, pp. 1423-1443, 2001.
[4] C. Delpha, S. Hijazi and R. Boyer, “A compressive sensing based quantized watermarking scheme with statistical transparency constraint,” Lecture Notes in Computer Science, vol. 8389, pp. 409-422, 2014
[5] J. J. Eggers, R. Bauml, R. Tzchoppe, and B. Girod, “Scalar costa scheme for information embedding,” *IEEE Trans. on Signal Processing*, vol.51, no.4, pp.1003-1019, April 2003.
[6] A. Komaty, C. Delpha, and A. Fraysse. Floating costa scheme with fractal structure for information embedding. In IEEE International Conference on Telecommunications (ICT 2012), Jounieh, Lebanon, April 2012.
[7] G. Le Guelvouit. Trellis-coded quantization for public-key steganography. IEEE Conf. on Acoustics, Speech and Signal Proc., March 2005.
[8] I. Benkara Mostefa, S. Braci, C. Delpha, R. Boyer, and M. Khamadja. Quantized based image watermarking in an independent domain. Signal Processing: Image Communication, vol.26, no.3, pp:194-204, March 2011.
[9] S. Braci, C. Delpha, and R. Boyer. How quantization based schemes can be used in image steganographic context. Signal Processing: Image Communication, 26(8-9):567-576, October 2011.
[10] R. Calderbank, S. Jafarpour and R. Schapire, “Compressed learning: universal sparse dimensionality deduction and learning in the measurement domain”, *oai:CiteSeerX.psu:10.1.1.154.7564*, 2009.
[11] H. Zhao, W. Wei, J. Cai, F. Lei and J. Luo, “Distributed compressed sensing for multi-sourced fusion and secure signal processing in private cloud,” Multidimensional Systems and Signal Processing, 27(4): 891-908, Oct. 2016.
[12] X. Zhang, Z. Qian, Y. Ren, and G. Feng, “Watermarking with flexible self-recovery quality based on compressive sensing and compositive reconstruction,” IEEE Transaction on Information Forensics and Security, vol.6, no.4, 2011, pp: 1223-1232.
[13] Q. Wang, W. J. Zeng, and J. Tian, “A Compressive Sensing based Secure Watermark Detection and Privacy Preserving Storage Framework,” IEEE Transactions on Image Processing, vol. 23, no. 3, pp:1317-1328, 2014.
[14] M. A. Sheikh and R. G. Baraniuk. Blind error-free detection of transform-domain watermarks. In IEEE International Conference on Image Processing (ICIP), San Antonio, Texas, USA, September 2007.
[15] C.-H. Zhao, W. Liu, “Block Compressive Sensing Based Image Semi-fragile Zero-watermarking Algorithm”, Acta Automatica Sinica , vol. 38, no. 4, April, 2012, pp.609-617.
[16] W. M. Chen, C. J. Lai, H. C. Wang, et al, “H.264 Video Watermarking with Secret Image Sharing,” IEEE Trans. Image Processing, vol.5, no.4, pp: 349-354, 2011.
[17] H. M. Zhao, J. H. Lai , J. Cai, X. L. Chen, “A Video Watermarking Algorithm for Intraframe Tampering Detection Based Compressed Sensing”, Acta Electronica Sinica, 41(6), pp.1153-1158, 2013.
[18] C. Cachin. An information-theoretic model for steganography. Lecture Notes in Computer Science, 1525:306-318, 1998.
[19]. J. Jiang et al, “Live: an integrated production and feedback system for intelligent and interactive broadcasting,” IEEE Trans. Broadcasting, 57(3), pp. 646-661, 2011
[20] S. Braci, C. Delpha, R. Boyer, and G. Le Guelvouit. Informed stego-systems in active warden context: Statistical undetectability and capacity. IEEE Proc. MMSP, October 2008.
[21] H.Y. Huang, C.H. Yang, W.H. Hsu, “A Video Watermarking Technique Based on Pseudo-3-D DCT and Quantization Index Modulation,” IEEE Transactions On Information Forensics and Security, vol.5, no.4, pp:625-627, 2010.
[22] D. Donoho, “Compressed sensing”, IEEE Transaction on Information Theory, vol. 52, No. 4, pp.1289-1306, 2006.
[23] D.L. Donoho, Y. Tsaig, “Extensions of compressed sensing”, *Signal Processing*, vol. 86, no.3, pp.533-548. 2006.
[24] E. Candes and M.Wakin, “An introduction to compressive sampling”, IEEE Signal Processing Magazine, Volume 25, Issue. 2, pp.21-30, Mar. 2008.
[25] H. Rauhut, K. Schnass, P. Vandergheynst, “Compressed sensing and redundant dictionaries,” IEEE Transaction on Information Theory, vol.54, no.5, pp:2210-2219, 2008.
[26] A. Masoum, N. Merana, P. Havinga. A distributed compressive sensing technique for data gathering in wireless sensor network. *Procedia Computer Science*, *21*, 207–216, 2013.
[27] X. Hou, L. Zhang, C. Gong, L. Xiao, J. Sun, X. Qian. SAR image Bayesian compressive sensing exploiting the interscale and intrascale dependencies in directional lifting wavelet transform domain. *Neurocomputing*, *133*, 358–368, 2014.
[28] E. J. Fowler, S. Mun, and W.E. Tramel, “Multiscale Block Compressed Sensing with Smoothed Projected Landweber Reconstruction”, 19th European Signal Processing Conference (EUSIPCO 2011), Barcelona, Aug 29-Sep 2, 2011,pp: 564-568.
[29]S. Biswas, R. Das, and M. Petriu, “An adaptive compressed MPEG-2 video watermarking scheme”, IEEE Transactions on Instrumentation and Measurement, vol. 54, no. 5, pp. 1853-1861, Oct. 2005.
[30] M. Barni, F. Bartolini and N. Checcacci, “Watermarking of MPEG-4 Video Objects”, IEEE Transactions on Mutimedia, vol. 7, no. 1, pp.23-32, Feb. 2005.
[31] E. Candes and J. Romberg. L1-magic: Recovery of sparse signals via convex programming, October 2005.
[32] J. Portilla, V. Strela, M.J. Wainwright, and E.P. Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. IEEE Transactions on Image Processing, vol.12, no.11, pp. 1338 -1351,

November 2003.
[33] D. Hsu, S. M. Kakade, J. Langford and T. Zhang. "Multi-label prediction via compressed sensing", In *Neural Information Processing Systems* (NIPS), 2009.
[34]M. Davenport, P. Boufounos, M. Wakin, and R. Baraniuk, "Signal processing with compressive measurements," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, No.2, pp. 445-460, 2010.
[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, vol.13, no.4, pp:1-14, April 2004.
[36] R.Ward, "Compressed sensing with cross validation," IEEE Transactions on Information Theory, vol.55, no.11 , pp: 5773-5782, December 2009.
[37] B. Zhang, Q. Lei, W. Wang, and J.S. Mu, "Distributed video coding of secure compressed sensing," Security and Communication Networks, Special Issue Paper 8:2416–2419, 2015.
[38] Y. Zhou, et al, "Hierarchical visual perception and two-dimensional compressive sensing for effective content-based color image retrieval," Cognitive Computation, 8(5): 877-889, Oct. 2016
[39] J Han et al, "Background prior-based salient object detection via deep reconstruction residual," IEEE Trans. Circuits and Systems for Video Technology, 25(8): 1309-1321, 2015
[40] J. Ren et al, "Gradient-based subspace phase correlation for fast and effective image alignment," J. Visual Communication and Image Representation, 25 (7): 1558-1565, 2014
[41] J. Zabalza, et al, "Robust PCA micro-Doppler classification using SVM on embedded systems," IEEE Trans. Aerospace and Electronic Systems, 50 (3): 2304-2310, 2014
[42] P. Gao and J. Ren, "Analysis and Realization of Snort-based intrusion detection system," Computer Applications and Software, 23(8): 134-135, 2006.
[43] C. Qin, C.-C. Chang and P.-L. Tsou, "Dictionary-based data hiding using image hashing strategy," Int. J. Innovative Computing, Information & Control: IJICIC, 9(2):599-610, Jan 2013.
[44] Z. Lin, G. Ding, J. Han and J. Wang, "Cross-view retrieval via probability-based semantics-preserving hashing," IEEE Trans. Cybernetics, In Press
[45] Y. Guo, G. Ding, L. Liu, J. Han and L. Shao, "Learning to hash with optimized anchor embedding for scalable retrieval," IEEE Trans. Image Processing, 26(3): 1344-1354, 2017.
[46] J. Han, G. C. Langelaar, "Method and device for generating fingerprints of information signals," EP Patent App. EP20150719863, 2017