# Multi-Resolution Based Hybrid Spatiotemporal Compression of Encrypted Videos

*Qiuming Yao, Wenjun Zeng, and Wei Liu*

Computer Science Department, University of Missouri, Columbia, MO, 65211

## ABSTRACT

Compression of encrypted data can be viewed as a special case of distributed source coding and can be achieved by applying Slepian-Wolf Coding. However, how to compress the encrypted video efficiently remains a challenging problem especially for those videos with irregular high motion. This paper proposes a novel multi-resolution based approach which makes it possible not only to effectively derive the temporal side information from previous frames, but also to generate the spatial side information by having partial access to the current frame. The spatial and temporal side information can then be integrated adaptively to facilitate the compression. Simulation results show that the proposed scheme significantly outperforms other schemes, especially for those video clips with irregular high motion.

***Index Terms***— Compression of encrypted data, multi-resolution, hybrid spatiotemporal method, Distributed Source Coding, Context Adaptive Interpolator

## 1. INTRODUCTION

Compression of encrypted data is necessary in some application scenarios. Johnson et al. [1] show that although the encryption masks the source and makes the data appear to be almost totally random, compression of encrypted data (assuming a stream cipher is used) is still possible, as it can be treated as a problem of distributed source coding with side information available at the decoder by performing joint decompression and decryption.

Suppose the plaintext $X$ is independent and identically distributed (i.i.d.), and the encryption function is a stream cipher, represented as $Y = X \oplus K$, where $K$ is the secret key and $Y$ is the ciphertext. The encryption operator "$\oplus$" is bit-wise XOR. According to the Slepian-Wolf theorem for distributed source coding, given the secret key $K$ as the side information at the decoder, the necessary bit rate to reconstruct $Y$ is $R_Y \geq H(Y|K) = H(X \oplus K|K) = H(X)$. Based on this idea, the work on i.i.d. source data [1] shows that given the secret key at the decoder, the encrypted data can still be compressed to the rate of the source entropy.

In order to apply this technique to images, Schonberg et al. [2] introduce a 2-D Markov model to explore the vertical and horizontal correlations of images at the decoder and achieve good results on binary images. However, our previous work [3] shows that 2-D Markov model is somewhat too simple to model grayscale images. Instead, it would help if the decoder could have partial access to the current image such as a lower resolution of it.

Videos can potentially be more compressible than still images because of the high temporal correlations in addition to the spatial correlations. Schonberg et al. extend their 2-D Markov based source model to videos [4]. They utilize three previous decoded frames to estimate the joint distribution between the current frame and its predictor, from which the 2-D Markov model parameters are calculated. This technique works well on those encrypted videos with low motion; however it can hardly compress the encrypted videos with irregular high motion. This is partly due to the limitation of their 2-D Markov based source model as discussed above. Besides, the joint distribution and the prediction from the previous frames are unreliable when dealing with videos with irregular high motion. In this situation, the current spatial information, if available partially, is more reliable than the temporal information.

Here the temporally or spatially predicted information is referred to as the side information. This paper will focus on how to generate more reliable side information and how to utilize it efficiently. As demonstrated in our previous work [3], multi-resolution approach can help to generate the spatial side information of the current frame, which is expected to effectively compensate for the unreliable temporal side information.

The rest of the paper is organized as follows. In Section 2, we briefly review our previous work on compression of encrypted images. Section 3 presents the proposed multi-resolution based hybrid spatiotemporal compression scheme for encrypted videos. Experimental results are shown in Section 4. Section 5 makes some concluding remarks.

## 2. OUR PREVIOUS WORK

Our previous work [3] shows that a global 2-D Markov model is somewhat too simple to characterize the local correlation of images. To address this issue, a multi-resolution based approach is proposed.

Before the encoding, pyramidal downsampling is performed on the original encrypted image (assuming a stream cipher is used so that the spatial location information of pixels is preserved). More specifically, four subbands will first be generated by downsampling with different offsets on the entire encrypted image. After that, the upper-left

subband (with offset (0,0)) is downsampled again, to generate smaller subbands. This process is repeated until the whole image is decomposed into predefined $N$ levels. Then the encoding process is applied on these hierarchically arranged subbands, i.e., the compression is conducted progressively from the highest level (lowest resolution) to the lowest level (highest resolution). The decoding is also performed resolution by resolution. Once a lower resolution image is decoded, it is interpolated to predict a higher resolution image which serves as the side information when a higher resolution is to be decoded. This process iterates until the whole image is decoded. Note that here and in the rest of the paper, the decoding process implies joint decompression and decryption.

The experiments in [3] show that it is feasible to derive, from the lower resolution image, reliable side information of the higher resolution image using a Context Adaptive Interpolator (CAI).

## 3. MULTI-RESOLUTION BASED COMPRESSION OF ENCRYPTED VIDEOS

We extend the multi-resolution based approach to compression of encrypted video in this section.

### 3.1. Spatial and Temporal Side Information

*3.1.1. Estimation of Spatial Side Information*

Suppose the current frame $X(t)$ is decomposed into $N$ levels by downsampling iteratively, which are denoted as $X(t)_1, \ldots X(t)_n, \ldots X(t)_N$. Each level $X(t)_n$ is formed by four subbands denoted as $X(t)_n^{00}$, $X(t)_n^{01}$, $X(t)_n^{10}$, $X(t)_n^{11}$ respectively from the upper-left to the bottom-right. We also have $X(t) = X(t)_1$ and $X(t)_{n+1} = X(t)_n^{00}$ $(1 \le n \le \text{N-1})$.

Let $\widehat{X}(t)$ denote the spatial estimation of $X(t)$, and $x(t)$ be the decoded version of $X(t)$. In the process of multi-resolution based compression, each time before we reconstruct $x(t)_n$, the $x(t)_{n+1}$ ($x(t)_n^{00}$) has already been reconstructed. Thus, $x(t)_n^{00}$ can be used to generate the spatial side information $\widehat{X}(t)_n^{01}$, $\widehat{X}(t)_n^{10}$, $\widehat{X}(t)_n^{11}$ by using the CAI [3]. More specifically, we first generate the spatial side information $\widehat{X}(t)_n^{11}$ using $x(t)_n^{00}$. Having $\widehat{X}(t)_n^{11}$, it is easy to do the decoding and reconstruct $x(t)_n^{11}$. With both $x(t)_n^{00}$ and $x(t)_n^{11}$ further prediction is applied to generate the spatial side information $\widehat{X}(t)_n^{01}$ and $\widehat{X}(t)_n^{10}$.

*3.1.2. Side Information based on Motion Compensated Prediction*

Temporal side information can be derived by Motion Compensated Prediction (MCP). Given the previous two decoded frames $x(t\text{-}1)$ and $x(t\text{-}2)$, motion vectors (MV) can be derived for $x(t\text{-}1)$ and used as an estimate of the MVs for $x(t)$ which can be used to find $\widetilde{X}(t)$. Then by downsampling $\widetilde{X}(t)$, we can get $\widetilde{X}(t)_N^{00}$, which can be used as side information to do the decoding and reconstruct $x(t)_N^{00}$.

Once $x(t)_N^{00}$ is available, we can use it to search for the potentially matched blocks in $x(t\text{-}1)$ and use the resulting MVs to form the temporal side information $\widetilde{X}(t)_N$ ($\widetilde{X}(t)_N^{01}$, $\widetilde{X}(t)_N^{10}$, $\widetilde{X}(t)_N^{11}$). Then through decoding we can reconstruct $x(t)_N^{01}, x(t)_N^{10}, x(t)_N^{11}$, and together with $x(t)_N^{00}$, they form $x(t)_N$ ($x(t)_{N-1}^{00}$), with which we can again do block matching in $x(t\text{-}1)$. This process can be performed iteratively, similar to the one presented in [5], until it reaches the lowest level, i.e., the entire image is reconstructed.

*3.1.3. Side Information Integration*

Except the lowest resolution $X(t)_N^{00}$ which only has the temporal side information $\widetilde{X}(t)_N^{00}$, every subband has both the spatial and temporal side information generated by the above prediction algorithms (CAI or multi-resolution MCP). Therefore, we can generate more reliable side information by integrating the spatial and the temporal side information using an adaptive weighting strategy:

$$S(t)_n^{ij} = \alpha \times \widehat{X}(t)_n^{ij} + (1-\alpha) \times \widetilde{X}(t)_n^{ij} \quad (1)$$

where $(0 \le \alpha \le 1)$. When $\alpha$ =0 or 1, the integration reduces to the scenario where only spatial or temporal side information is used.

### 3.2. Proposed Algorithms

In addition to the side information, another important issue is how to estimate the distribution of the prediction error $P(S(t)_n^{ij} - X(t)_n^{ij})$. We assume that $P$ follows a Laplace distribution, that is $P \sim L(\mu, b)$ where $u$ and $b$ are location and scale parameters. Since the current source data $X(t)_n^{ij}$ is not available, we can estimate the residual distribution $P$ using different strategies as discussed below. In the following schemes ,both $u$ and $b$ are calculated locally for each pixel in a $5 \times 5$ neighbourhood [3].

*3.2.1. Spatial Method*

This is a simple extension of our previous work on images [3]. Each frame of the video can be treated as a single image and thus can be encoded separately, where only spatial side information is used ($\alpha$ =1 in Eq.(1)). The estimation of the residual distribution $P$ is calculated as following:

$$\breve{P}(S(t)_n^{ij} - X(t)_n^{ij}) = P(S(t)_{n+1}^{ij} - x(t)_{n+1}^{ij}) \quad (2)$$

Where $1 \le n \le$ N-1. Note that the lowest resolution $X(t)_N^{00}$ ($X(t)_{N+1}$) is not encoded.

*3.2.2. Temporal Method*

This algorithm only exploits the temporal side information ($\alpha$ =0 in Eq.(1)). The residual distribution $P$ is estimated as the distribution of the corresponding subbands in the previous frame, i.e., :

$$\breve{P}(S(t)_n^{ij} - X(t)_n^{ij}) = P(S(t-1)_n^{ij} - x(t-1)_n^{ij}) \quad (3)$$

Where1 $\le n \le N$, and t≥4. Note that the first three frames are not encoded.

*3.2.3. Hybrid Spatiotemporal Method*

Usually, mean square error (MSE) is used to evaluate how accurate or reliable a prediction is. Let $MSE_p$ and $MSE_t$ be the MSE value of the spatial and the temporal side information respectively. They can be used to derive the weight in Eq.(1):

$$\alpha = \frac{MSE_t}{MSE_p + MSE_t} \quad (4)$$

However, the MSE values of the side information of the current subband being decoded are not available. Similar to the way we estimate $\breve{P}$ , we can find an estimate of $MSE_p$ and $MSE_t$ to derive $\alpha$ . We explore three modes below.

- Mode I: Determination based on co-located block

The weight is estimated based on the co-located blocks in the previous frame. Note that the co-located block in the previous frame has its own spatial and temporal side information, and their respective $MSE_p$ and $MSE_t$ can be calculated and used in Eq.(4) to estimate the weight $\alpha$ for the current block. $\breve{P}$ can be easily derived by Eq. (3).

- Mode II: Determination based on the lower resolution

The weight is estimated based on the blocks in the lower resolution of the current frame. The block in the lower resolution has its own $MSE_p$ and $MSE_t$ , which are used to estimate $\alpha$ in Eq.(4). $\breve{P}$ can be easily derived by Eq(2).

- Mode III: Determination based on the motion compensated previous blocks

The motion compensated block is the block in the previous frame which best matches the block in the current frame. Note that the motion compensated block has its own $MSE_p$ and $MSE_t$ which can be used to estimate $\alpha$ for the current block. $\breve{P}$ is the residual distribution of the motion compensated blocks.

TABLE I
DATA RATE (OUTPUT BITS PER SOURCE BIT) USING DIFFERENT MODES

| Modes | Foreman | Football | Garden |
|---|---|---|---|
| Mode I | 0.5161 | 0.6738 | 0.6337 |
| Mode II | 0.5436 | 0.6856 | 0.6627 |
| Mode III | 0.5190 | 0.6690 | 0.6363 |

From the test results in Table I, we can see that Mode I and III have similar performance and are both better than Mode II, but mode III is more suitable for some videos with high motion (e.g., football) because it explores the similarity based on the moving object rather than the pixel location. So Mode III is chosen as our hybrid spatiotemporal method, which is used to generate the results in Section 4.

The complete algorithm of decoding the current frame $X(t)$ is summarized here:

1. Get MCP of $X(t)$ from $x(t\text{-}1)$ and $x(t\text{-}2)$, derive $\widetilde{X}(t)_N^{00}$ and decode the lowest resolution. Set $n = N$.
2. Using the decoded data $x(t)_n^{00}$ , derive spatial and temporal side information $\widehat{X}(t)_n$ and $\widetilde{X}(t)_n$ .
3. Find motion compensated blocks of $x(t)_n^{00}$ , determine the estimated weight $\alpha$ , calculate $\breve{P}$ and derive the side information $S(t)_n$.
4. Perform the joint decompression and decryption.
5. $n = n$ -1 and if $n > 0$, go to step 2.

This is done when t≥3, until the whole video clip is decoded, while the first two frames are not encoded.

## 4. SIMULATIONS

In the following experiments, the channel coding tool we use is the Low Density Parity Code (LDPC) [6]. The test videos are 'foreman.cif', 'football.cif' and 'garden.cif'. Table II compares the performance of different approaches. The experiment is done on the first twelve frames of each video clip to be comparable to the experimental settings in [7]. The data rates in the table (including the directly quoted numerical results from [7]) show that among our three methods, the hybrid spatiotemporal method performs the best, and significantly outperforms the scheme in [7].

TABLE II
PERFORMANCE OF DIFFERENT ALGORITHMS

| Output bit per source bit | Foreman | Football | Garden |
|---|---|---|---|
| Spatial Method Only | 0.6003 | 0.7167 | 0.7646 |
| Temporal Method Only | 0.5579 | 0.7585 | 0.6493 |
| Hybrid Spatiotemporal Method | 0.5190 | 0.6690 | 0.6363 |
| Scheme in [7] | 0.6700 | 0.9283 | 0.8236 |

Figures 1, 2, 3 show the data rate frame by frame. The spatial method performs more stable than the other two methods. Obviously, the hybrid spatiotemporal method always has the best performance.

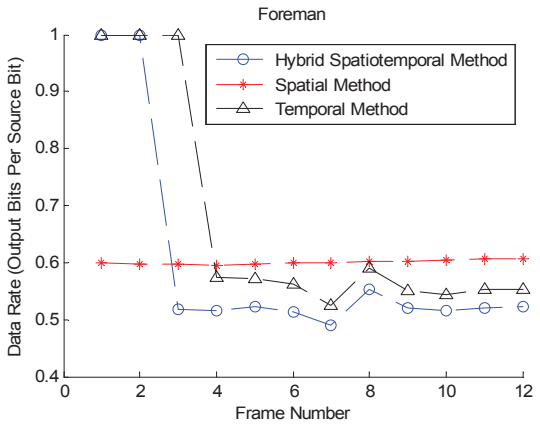


Fig.1. Data rates by frames for 'Foreman'

Comparing Figure 2 to Figure 3, we can see that for 'Foreman', the temporal method performs better than the

spatial method since it is a video with low and stable motion and thus the motion prediction is more reliable. But for 'Football', the spatial method outperforms the temporal method since it is a video with irregular high motion and thus the spatial side information is relatively more reliable.

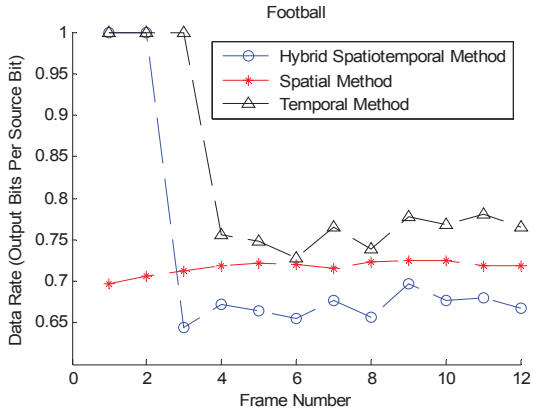


Fig.2. Data rates by frames for 'Football'

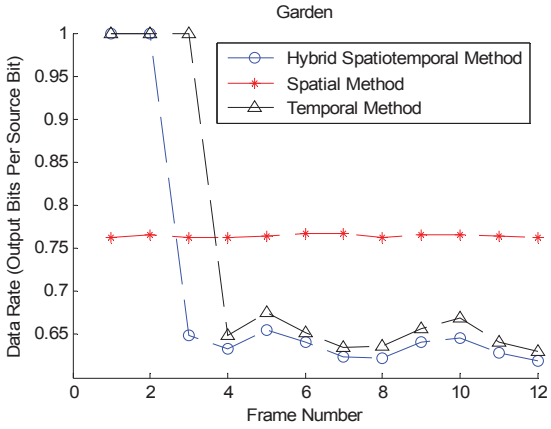


Fig.3. Data rates by frames for 'Garden'

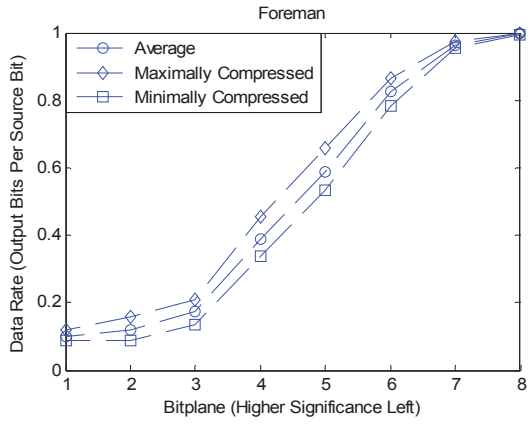


Fig.4. Data rates by bitplanes for 'Foreman'

For 'Garden', the performance gap between the spatial method and the other two methods is quite large. That's probably because there are too many irregular edges and textures within the frame which spoil the spatial prediction results. It is noticed that the hybrid spatiotemporal method performs similarly to the purely temporal method, which indicates that the integration process is inclined to the temporal side information.

Figure 4 shows the data rate of the spatiotemporal method for 'Foreman' by different bitplanes. Circles represent the average data rate over the frame 3-12 on different bitplanes, and diamonds represent the maximal value across these frames while squares show the minimum. The most significant bitplanes (MSBs) can be compressed efficiently. The highest two bitplanes can hardly be compressed due to the randomness. This result is also much better than the results shown in [2]. We also observe that if the two least significant bitplanes are left unencoded and sent first to help the decoding of more significant bitplanes, the overall compression will be slightly more efficient.

## 5. CONCLUSIONS

A novel lossless compression method for encrypted video based on a multi-resolution approach is proposed in this paper. The multi-resolution approach makes it possible to have access to part of the spatial source data to generate more reliable spatial and temporal side information. The proposed hybrid spatiotemporal method exploits both the spatial and the temporal side information so that it is more adaptive than other methods. Our experiments show that it significantly outperforms other techniques.

## 6. ACKNOWLEDGMENT

This research is supported by the Center for Cyber Security Research, Computer Science Department, University of Missour, and by NSF grant CNS-0423386.

## 7. REFERENCES

[1] M. Johnson, P. Ishwar, V. M. Prabhakaran, D. Schonberg, and K. Ramchandran, "On Compressing Encrypted Data," *IEEE Transactions on Signal Processing, Supplement on Secure Media I*, Volume: 52, Issue: 10, October 2004, Pages 2992 - 3006.

[2] D. Schonberg, S. C. Draper, and K. Ramchandran, "On Compression of Encrypted Images," *International Conference on Image Processing*, Atlanta, GA, October 2006.

[3] W. Liu, W. Zeng, L. Dong and Q. Yao, "Resolution-progressive Compression of Encrypted Grayscale Images", *IEEE International Conference on Image Processing* (ICIP), 2008.

[4] D. Schonberg, C. Yeo, S. C. Draper, and K. Ramchandran, "On Compression of Encrypted Video," *Data Compression Conference*, 2007. *Proceedings DCC* 2007, 27-29 March 2007, pp. 173 - 182.

[5] W. Liu, L. Dong and W. Zeng, "Wyner-Ziv Video Coding with Multi-resolution Motion Refinement: Theoretical Analysis and Practical Significance," *Visual Communications and Image Processing* (VCIP), San Jose, Jan. 2008.

[6] [Online].Available: http://www.stanford.edu/~divad/ldpca.html

[7] D. Schonberg, *Practical Distributed Source Coding and Its Application to the Compression of Encrypted Data*, Ph.D. Dissertation, Univ. of California, Berkeley, CA 2007.