RC23845 (W0601-064) January 19, 2006
Electrical Engineering

IBM Research Report

Side Information Generation for Low Complexity Video Coding Systems Based on Wyner-Ziv Theorem

Ligang Lu, Vadim Sheinin
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Side Information Generation for Low Complexity Video Coding Systems Based on Wyner-Ziv Theorem

Ligang Lu, Vadim Sheinin

IBM T.J. Watson Research Center.
Yorktown Heights, NY, 10598
lul@us.ibm.com

Abstract

The existing Wyner-Ziv video coding systems have large rate-distortion performance gap comparing to traditional H.26x or MPEG based systems. One of the primary reasons for the performance gap is that the side information generation schemes in these systems use motion based linear extrapolation or interpolation methods which result in low quality side information. We present a side information generation scheme that utilizes both the motion information and the statistical similarity derived from signal statistic features. Results have shown that this scheme can generate side information about 1~2 dB better than H.264's motion compensated prediction.

Keywords

Wyner-Ziv Video Coding,Side Information Generation, Statistical Similarity, Extrapolation, and Interpolation

1. Introduction

Recently some research efforts have been made to develop video coding systems for emerging applications such as distributed surveillance systems and mobile visual communications. The challenges for such systems lie in achieving high performance under very limited computation power and transmission bandwidth. Unlike the traditional video coding systems based on MPEG or H.26x standards, video encoders for these applications need to use low complexity yet efficient algorithms while, at the back-end center, the decoders may have sufficient computation power to handle heavy duty signal processing tasks. Based on Slepian and Wolf's work [1] for lossless coding case, Wyner and Ziv [2] have provided theoretical ground for a new lossy coding paradigm wherein a low complexity encoding and high complexity decoding system may potentially achieve similar rate-distortion performance as the traditional coding systems. However, Wyner and Ziv's paper is not constructive. Recently several papers attempting to build practical video communication systems based on Wyner-Ziv's theoretical frame work have been published [3-5]. But these papers have reported large rate-distortion performance gap (4-6 dB) between their video coding schemes and the traditional coding systems, such as H.263+.

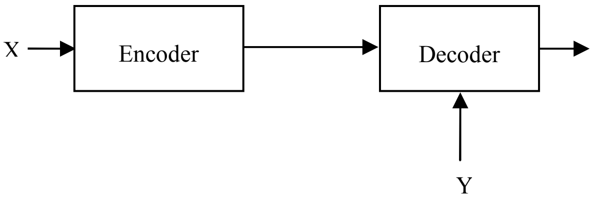

Fig. 1 Wyner-Ziv Coding System

Figure 1 depicts an example of the Wyner-Ziv coding system, wherein the side information Y, which is not available at the encoder, is used in decoding X. A typical Wyner-Ziv video coding system includes a video encoder and a video decoder. The video encoder is a low complexity and low power encoder, so the computation-heavy signal processing tasks, such as the motion estimation, are carried by the decoder instead. In order to decode the received video signals and achieve high efficiency, the Wyner-Ziv decoder needs to exploit the correlation between the source and side information, which is only known to the decoder. The source is the video signal (e.g., a picture) to be encoded at the encoder and transmitted to the decoder for decoding, and the side information is here essentially an estimate or a prediction of the picture to be decoded. The higher the correlation between the source and the side information, the less the bits of information the encoder needs to transmit in the Wyner-Ziv coding system. Since the performance of the Wyner-Ziv system depends heavily on the reliability of the side information, the mechanism used by the decoder to generate the side information plays a very crucial role. Typically, the decoder first performs motion estimation on previously decoded adjacent pictures to generate a set of motion vectors and then uses such motion vectors to generate an estimate of the picture currently being decoded by motion based linear extrapolation or interpolation. This estimate is used as the side information by the decoder for decoding the current picture. Obviously, the higher the correlation exists between the source and the side information, the better the side information is.

In the published papers on Wyner-Ziv video coding, such as [3-5], little research effort was spent on side information generation, only simple motion based linear extrapolation or bi-linear interpolation methods were used to create the side information. In our view, these methods are far from adequate because they have following serious drawbacks:

1. The underlying assumption of linear extrapolation or bi-linear interpolation that the objects moves following a constant linear displacement model from picture to picture often does not hold true for real visual signals; and

2) The extrapolation or interpolation may not result in a one-to-one mapping between the reference picture(s) and the estimate picture. Some pixel positions in the extrapolated or interpolated picture may not get any mapping from the reference picture(s), i.e., leaving empty holes, while other pixel positions in the extrapolated or interpolated picture may have multiple mappings from the reference picture(s), i.e., leaving superimposed spots.

We believe that using the low quality side information is one of the major reasons that result in large rate-distortion performance gap. It is therefore important to develop a better method of side information generation for Wyner-Ziv video coding systems to achieve efficient performance. In this paper we will present a new scheme to generate high quality side information at the Wyner-Ziv decoder by utilizing not only the motion information but also the statistical features. The quality of the side information generated by our scheme is comparable to or better than H.264's motion compensated prediction. We will organize the presentation of the paper as follows. In Section 2 we will describe the key aspects of our scheme. Then in Section3 we will show the methods to fill empty holes and resolve multiple mapping occurrences in our scheme. Finally we will present experimental and comparison results of our scheme and draw conclusions in Section 4.

2. Side Information Generation Based Motion Information and Statistical Features

In the Wyner-Ziv video coding system, the side information is essentially a prediction or an estimate of the picture frame being decoded. Because high temporal correlation usually exists among the adjacent video frames, it is a natural choice to generate the prediction of the current frame, i.e., the side information, base on the previously decoded adjacent frames. Furthermore, motion is of essential important characteristic information of the video signal. There is no doubt that extrapolation and interpolation should make use of motion information to generate the prediction. However as pointed early, motion based linear extrapolation or interpolation alone is far from enough to generate good quality side information.

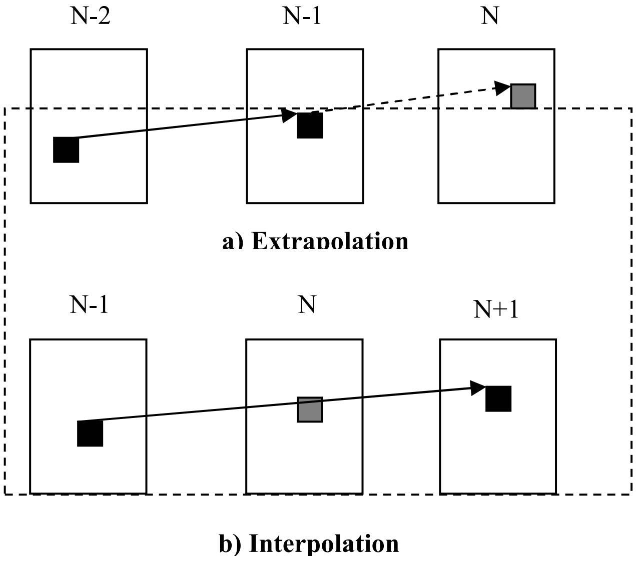

Fig. 2 Motion Based Extrapolation and Interpolation

Since the motion estimation is performed on the previous decoded frames at the decoder and the estimate of the current frame (the side information) is either extrapolated or interpolated from the adjacent decoded frames using the motion vectors as shown in Figure 2, the implicit assumption here is that the motion follows an underlying linear model from frame to frame. However this assumption of linear motion model often does not hold in real video. Furthermore the motion based extrapolation or interpolation is not a one-to-one mapping. Some pixel positions in the side information may not get any mapping while others may have multiple mappings from the reference frames. Actually there are usually a non-trivial number of pixel positions left empty or having multiple mappings using the motion based extrapolation or interpolation. Thus new schemes need to be devised to overcome the short comings and generate better side information by utilizing additional relevant information in the reference frames.

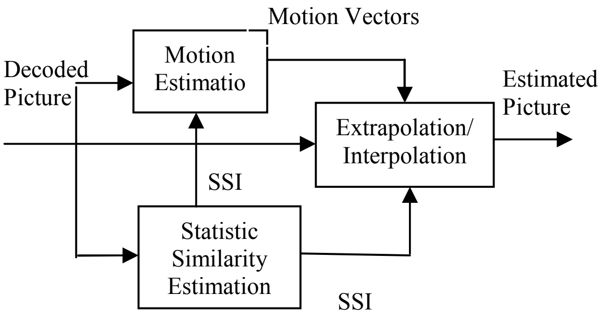

Fig.3 Side Information Generation

2.1 The Frame Work of Our Scheme

In our scheme to generate the side information, we not only make use of the motion information but also utilize the statistical features. As illustrated by Figure 3, statistic similarity is used both in motion estimation and extrapolation/interpolation processes. More specifically, the minimization in the motion estimation is based on the weighted sum of the absolute difference (SAD) and statistical similarity of two blocks of pixels as well as the homogeneity of adjacent motion vectors. The statistic similarity estimation is conducted by comparing the statistics features of two or more pixel blocks, including, but not limited to: the block sample mean, the block sample variance, the neighboring parameters such as the homogeneity of the neighborhood surrounding the block, and the covariance between the two or more pixel blocks. The statistic features of the two blocks provide a good indication on how similar these two blocks are.

First, at least two previously decoded and reconstructed pictures, which are referred to hereinafter as the reference pictures, are obtained and stored by the decoder. These two reference pictures are referred to as N-1 and N-2 for extrapolation-based estimation (or as N-1 and N+1 for interpolation-based estimation) as shown in Figure 2.
For each block of pixels in the reference picture N-1, a search process is performed to find its best match in the other reference picture N-2 (or N+1). In order to find the best matching block B* in the reference picture N-2 (or N+1) for a specific block Bi in the reference picture N-1, the search process picks a same size block of pixels, Bp, from the reference picture N-2 (or N+1) and computes a statistic similarity index SSI, which is indicative of the statistic similarities between Bi and Bp, and optionally a prediction error E, which is the differences in pixel values between Bi and Bp. The statistic similarity index SSI and the prediction error E can be combined to determine the best matching block B* in the reference picture N-2 (or N+1).

Once the best matching block B* in the reference picture N-2 (or N+1) is determined, a set of motion vectors can be generated for the block Bi in the reference picture N-1, which are indicative of the movement of block Bi in relation to B*. The motion vectors can be generated by taking the spatial differences (i.e., the horizontal and vertical coordinates) of blocks Bi and B*. The motion vectors are then processed as necessary (e.g., reversed, scaled, shifted, or otherwise altered) for extrapolating or interpolating a location in the picture to be estimated. The pixel values of the estimate block are derived from the pixel values of blocks determined by the motion vector in the reference frames by temporal filtering.

The above-described processing steps are repeated for each block of pixels in the reference picture N-1, so that the estimate of each block of pixels in the reference picture N-1 is mapped, thereby forming a complete estimate picture N.

To determine the statistical similarity between two blocks of pixels P and Q, information directly relating to blocks P and Q are obtained, which include, but are not limited to: pixel values, luminance, contrast, structure, etc. Information relating to the neighborhoods where blocks P and Q reside (i.e., the neighboring information) is also obtained. Various statistic features for blocks P and Q are then calculated based on the information directly relating to blocks P and Q and optionally the neighboring information, which are then compared to determine the statistic similarity between blocks P and Q. For example, statistic features such as block sample mean, block sample variance, neighboring parameters, as well as the covariance between blocks P and Q can be used for determining the statistic similarity. Other statistic features can also be incorporated into the statistical similarity measure.

More specifically, assuming that the blocks P and Q are both characterized by a block size n × m, the pixel values in block P can be referred to as P_{ij}, and the pixel values in block Q can be referred to as Q_{ij}, wherein i = 1, 2, …, n, and j = 1, 2, …, m. The block sample mean for P is defined as $\mu_P = \frac{1}{nm}\sum_{j=1}^{n}\sum_{i=1}^{m} P_{ij}$, and the block sample mean for Q is defined as $\mu_Q = \frac{1}{nm}\sum_{j=1}^{n}\sum_{i=1}^{m} Q_{ij}$. The block sample variance for P is defined as $\sigma_P^2 = \frac{1}{mn-1}\sum_{j=1}^{n}\sum_{i=1}^{m}(P_{ij}-\mu_P)^2$, and the block sample variance for P is defined as $\sigma_Q^2 = \frac{1}{mn-1}\sum_{j=1}^{n}\sum_{i=1}^{m}(Q_{ij} - \mu_Q)^2$. The covariance of blocks P and Q is estimated as $\sigma_{PQ} = \frac{1}{mn-1}\sum_{j=1}^{n}\sum_{i=1}^{m}(P_{ij} - \mu_P)(Q_{ij} - \mu_Q)$.

Moreover, neighboring parameters of blocks P and Q, such as the homogeneity of the neighborhoods surrounding blocks P and Q, can also be used for determining the statistic similarity between blocks P and Q. Besides the above mentioned statistic features, the neighborhood homogeneity can also include additional features, for example, the differences between the motion vectors of the block P or Q and the motion vectors of one or more existing neighboring blocks surrounding the block P or Q.

The statistic features of blocks P and Q provide a good indication on how similar these two blocks are. Preferably, a statistic similarity index is computed based on the statistic features of blocks P and Q to provide a quantitative measurement of the statistic similarity between blocks P and Q. The statistical features can be weighted and combined in various fashions for computing the statistic similarity index.
For example, the statistic similarity index SSI can be computed for blocks P and Q by using the following formula:

$$\text{SSI(P,Q)}=\alpha\sqrt{\left[\sigma_{PQ}{}^2 - \sigma_P^2\sigma_Q^2\right]^2} + \beta\left[\mu_P - \mu_Q\right]^2,$$

where α and β are weighting factors. When statistic similarities of multiple pixel blocks are determined to generate multiple statistic similarity indexes, these indexes are normalized, so that each index value falls between 0 and 1. The smaller the value of the statistic similarity index, the more similar the two blocks.

3. Methods for Filling Empty Pixel Positions and Resolving Multiple Mappings

As mentioned above since the extrapolation and interpolation do not generate one-to-one mapping to the estimate picture, there may be pixel positions in the estimate position that do not get any mapping, i.e., leaving empty holes. On the other hand, there may also be pixel positions in the estimate position that get multiple mappings, i.e., leaving superimposed spots. The quality of the estimate picture is adversely affected by existence of the empty holes or superimposed spots. Our scheme therefore provides solutions to these problems, by using statistical similarity estimation to refine the estimate picture, i.e., filling in the empty pixel positions and/or resolving the multiple mappings.

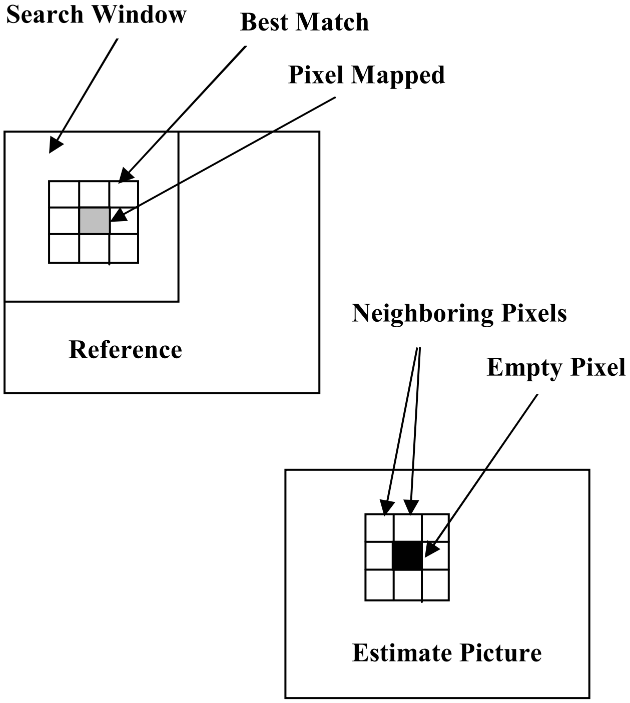

Fig. 4 Fill an Empty Position

Figure 4 shows how statistical similarity estimation can be used to fill in an empty pixel position on an estimate picture N. First, the statistical features of a neighboring block of pixels that surround the empty pixel position on the estimate picture N are calculated. The motion vectors of the pixels in the neighboring block can be used to determine an initial point on the reference picture N-1, from which the estimate picture is generated by extrapolation or interpolation. An appropriate search window surrounding the initial point is then identified. Within this search window, a searching process is performed to find the block that best matches the neighboring block on the estimate picture N. The best matching block is characterized by the highest statistical similarity, and optionally the lowest pixel value difference, with respect to the neighboring block on the estimate picture N. The specific pixel surrounded by this best matching block on the reference picture N-1, as shown in Figure 4, is then used to fill the empty pixel position in the estimate picture N.

Figure 5 shows how statistical similarity estimation can be used to select a best matching pixel for a specific pixel position on the estimate picture N, in the event that multiple pixels on the reference picture N-1 all map to (i.e., by extrapolation or interposition) the same pixel position on the estimate picture N. First, the statistical features of a neighboring block of pixels that surround the specific pixel position on the estimate picture N are calculated. Next, the statistic features for multiple blocks of pixels that each surrounds one of the multiple pixels on the reference picture N-1 are calculated. Among these multiple blocks on the reference picture N-1, the one that best matches the neighboring block on the estimate picture N is identified. The best matching block, as mentioned above, is characterized by the highest statistical similarity, and optionally the lowest pixel value difference, with respect to the neighboring block on the estimate picture N. The specific pixel surrounded by this best matching block on the reference picture N-1 is then selected as the best matching pixel for the specific pixel position in the estimate picture N.

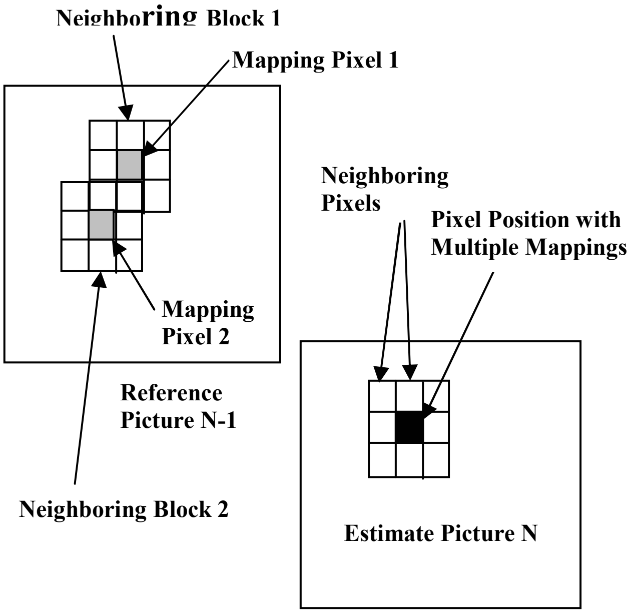

Fig. 5 Multiple Mappings

4. Results and Conclusions

The side information generated by our scheme is implemented in software for simulation and experimented on several standard test sequences. In the test, we used ¼-pel precision motion estimation with both full search and diamond fast search motion estimations. Its performance is compared with H.264 motion compensated prediction results. Note that H.264 uses the current frame and reference frame to perform motion estimation and motion compensated prediction while our scheme only uses the reference frames to do motion estimation and estimate the side information. Table 1 shows the testing results on two CIF (352x288) sequences Coast Guard and Mobile Calendar. The listed average PSNR results are obtained on 18 extrapolated or interpolated frames generated from the first 20 reconstructed frames. The tests and PSNR computation are only performed on the luminance signal of the video sequences. The motion compensated prediction results of H.264 does not include residues.

The results show that, in extrapolation case, our scheme can generate side information that is comparable to H.264's motion compensated prediction while in interpolation case, the quality of our side information is significantly better. When using fast search Diamond algorithm in motion estimation, the results of our scheme only has a less than 0.5dB degradation from the results obtained using full search motion estimation algorithm.

In this paper, we have presented a new side information generation scheme for Wyner-Ziv video coding systems. Our scheme utilizes both motion information and statistic features in generating an estimate of the current frame to be decoded. The utilization of the statistical similarity not only helps in estimating better motion vectors but also provides methods to solve the empty positions and multiple mapping problems manifested in the existing side information generation schemes. Consequently, as the experiment results have shown, our scheme can generate very good quality side information, especially in the interpolation case.

Table 1. Quality Comparison of Side Information Generated by Our Scheme and H.264 Motion Compensated Prediction

	Extrapolation	Interpolation	H.264
Coast Guard	30.79dB	32.19dB	31.00dB
Mobile Calendar	27.77dB	29.60dB	27.74dB

REFERENCES

[1] D. Slepian and J.K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. on Information Theory*, Vol. 19, pp.471-480, July 1973.

[2] A. Wyner and J. Ziv, "The Rate-Distortion Function for Source Coding with Side Information at the Decoder," *IEEE Transactions on Information Theory*, Vol. 22, No. 1, pp.1-10, January 1976.

[3] A. Aaron, E. Setton, and B. Girod, "Toward Practical Wyner-Ziv Coding of Video," in *Proc. IEEE Int. Conf on Image Processing*, 2003.

[4] R. Puri and K. Ramchandran, "PRISM: A NewRobust Video Coding Architecture based on Distributed Compression Principles, " in *Allerton Conference on Communication, Control and Computing*, 2002.

[5] A. Liveris, Z. Xiong, and C. Georghiades, A distributed source coding technique for correlated images using turbo codes," *IEEE Communications Letters*, Vol.6, no.9, pp.379-381, Sept.2002.